DATA ANALYSIS CERTIFICATE PROGRAM



Alex Mata
Calbright College

LibreTexts™



Data Analysis Certificate Program

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (https://LibreTexts.org) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden borne from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by NICE CXone Expert and were supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptions contact info@LibreTexts.org or visit our main website at https://LibreTexts.org.

The LibreTexts name, logo, and book covers are trademarked by LibreTexts, Inc. (a 501(c)(3) not-for-profit organization) and protected against unauthorized use.

This text was compiled on 09/22/2026

TABLE OF CONTENTS

Licensing

1: Welcome to Data Analysis

2: Introduction to Working in Microsoft Excel

- 2.1: Using Excel Cells
- 2.2: Using Excel Data
- 2.3: Using Excel Functions
- 2.4: Using Excel Charts
- 2.5: Using Conditional Formatting in Excel

3: Working in Google Sheets

- 3.1: Creating a New Spreadsheet
- 3.2: Using Google Sheets
- 3.3: Add-Ons

4: SQL Structured Query Language

- 4.1: Create Database
- 4.2: Constraints
- 4.3: User Defined Types and Altering Tables
- 4.4: Key Terms and Exercises

5: SQL Data Manipulation Language

- 5.1: Using SELECT Statement
- 5.2: INSERT Statement
- 5.3: UPDATE and DELETE Statements
- 5.4: Built-In Functions
- 5.5: Joining Tables
- 5.6: Key Terms and Exercises

6: The Relational Data Model

- 6.1: Fundamental Concepts in the Relational Data Model
- 6.2: Key Terms

7: The Entity Relationship Data Model

- 7.1: Entity, Entity Set and Entity Type
- 7.2: Types of Attributes
- 7.3: Keys
- 7.4: Nulls
- 7.5: Relationships
- 7.6: Key Terms and Exercises

8: Visual Media

8.1: Tables, Charts, and Infographics
8.2: Matching Graphics and Objectives
8.3: Placement, Style, and Coloring
8.4: Captions and Titles
8.5: Introduction to Contemporary Visual Aids
8.6: Video as a Visual Aid
8.7: Designing a Video for Your Needs
8.8: Products and Resources for Creating Videos
8.9: Introduction to Accessible Visual Aids
8.10: Accessibility Matters
8.11: Why It Matters- Visual Media
8.12: Introduction to Using Visuals
8.13: Visuals in a Report
8.14: Increasing Impact with Media
8.15: Revising and Enhancing Visual Media for Impact
8.16: Evaluating the Effectiveness of your Message
8.17: Putting It Together- Visual Media
8.18: Discussion- Visual Media
8.19: Assignment- Visual Media
8.20: Introduction to Media and Your Message
8.21: Using Visual Media Resources
8.22: Visual Design Principles
8.23: Introduction to Images
8.24: Images Overview
8.25: Using Images for Impact
8.26: Using Images Legally
8.27: Introduction to Charts, Diagrams, and Graphic Organizers

9: Research and Statistics

9.1: Introduction
9.2: Measures of Center
9.3: Measures of Variation
9.4: There's More to Research Methods than Statistics
9.5: A brief Introduction to Research Design
9.6: Introduction to Psychological Measurement
9.7: Scales of measurement
9.8: Assessing the Reliability of a Measurement
9.9: The role of variables — predictors and outcomes
9.10: Experimental and non-experimental research
9.11: Assessing the validity of a study

9.12: Confounds, Artifacts and other Threats to Validity
9.13: Videos
9.14: References

10: Data Validation

10.1: Importance of Statistics and Experimental Measurements

11: Describing Data

11.1: This is what too many numbers looks like
11.2: Look at the data
11.3: Important Ideas - Distribution, Central Tendency, and Variance
11.4: Measures of Central Tendency (Sameness)
11.5: Measures of Variation (Differentness)
11.6: Using Descriptive Statistics with data
11.7: Rolling your own descriptive statistics
11.8: Remember to look at your data
11.9: Videos
11.10: References

12: Correlation

12.1: If something caused something else to change, what would that look like?
12.2: Pearson's r
12.3: Turning the numbers into a measure of co-variance
12.4: Examples with Data
12.5: Regression — A mini intro
12.6: Interpreting Correlations
12.7: Summary
12.8: References

13: Probability, Sampling, and Estimation

13.1: Prelude to Probability, Sampling, and Estimation
13.2: How are Probability and Statistics Different?
13.3: What Does Probability Mean?
13.4: Basic Probability Theory
13.5: The Binomial Distribution
13.6: The normal distribution
13.7: Other useful distributions
13.8: Summary of Probability
13.9: Samples, populations and sampling
13.10: The Law of Large Numbers
13.11: Sampling distributions and the central limit theorem
13.12: The Central Limit Theorem
13.13: z-scores

13.14: Estimating population parameters
13.15: Estimating a confidence interval
13.16: Summary
13.17: Videos
13.18: References

14: Public Speaking

14.1: How You Speak
14.2: Introduction to Audience Engagement
14.3: Audience Attention and Rapport
14.4: Audience Participation
14.5: Introduction to Speech Tips and Techniques
14.6: Overcoming Fears and Anxiety
14.7: Pro Tips
14.8: Putting It Together- Public Speaking
14.9: Discussion- Public Speaking
14.10: Why It Matters- Public Speaking
14.11: Introduction to Effective Public Speaking
14.12: What is Public Speaking?
14.13: Benefits of Public Speaking
14.14: Audience Expectations
14.15: Developing an Effective Speech
14.16: Introduction to Delivery Techniques
14.17: Starting Your Speech
14.18: Body Language and Gestures

15: Developing and Delivering Business Presentations

15.1: Text Boxes and Themes
15.2: Images, Videos, and Arranging Objects
15.3: Introduction to Making a Presentation for a Meeting
15.4: Parts of a Good Presentation
15.5: What's my Presentation About
15.6: Bad Presentations
15.7: Making a Presentation for a Meeting
15.8: Putting It Together- Developing and Delivering Business Presentations
15.9: Assignment- Developing and Delivering Business Presentations
15.10: Why It Matters- Developing and Delivering Business Presentations
15.11: Introduction to Visual Aids
15.12: Presentation Tools and Visual Aids
15.13: Introduction to Using Microsoft PowerPoint
15.14: New Presentations
15.15: Editing and Enhancing a PowerPoint Presentation

15.16: Visuals- Images, Videos, and Objects
15.17: Introduction to Using Google Slides
15.18: Creating and Organizing Presentations

Index

Glossary

Detailed Licensing

Detailed Licensing

This page titled Data Analysis Certificate Program is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Alex Mata.

Licensing

A detailed breakdown of this resource's licensing can be found in ***Back Matter/Detailed Licensing***.

This page titled Licensing is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

1: Welcome to Data Analysis

Thank you for exploring the Data Analysis certificate program textbook published by Calbright College, a proud member of the California Community Colleges system.

This textbook was created through a compilation of Open Educational Resources (OER) that align to the learning objectives included in the certificate program. Calbright prepares data analysts to be versatile and capable of contributing to any industry. Therefore, readers will find a broad range of topics presented in this text, including spreadsheets, Structured Query Language (SQL), databases, data visualization, public speaking, research, and statistics. This resource is written for anyone seeking to become a well-rounded data professional and wants an introduction to the specializations that exist within the varied data analyst roles of today's workforce. We hope you find this text to be an excellent resource, first as students and later on as a reference early in your career.

This page titled 1: Welcome to Data Analysis is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

2: INTRODUCTION TO WORKING IN MICROSOFT EXCEL

Learning Objectives

Create workbooks and format data in Microsoft Excel

Imagine that you have a lot of business data. Perhaps you have names and addresses for a mailing list. Maybe you have inventory data or quarterly sales values. All this information could be kept in a Word document, but Microsoft Office actually has an extremely useful program for organizing, storing, and even manipulating data: Microsoft Excel.



Learning to use Microsoft Excel is one of the most helpful and versatile workplace skills you can acquire, and creating a worksheet in a workbook is the first step. Many of the skills you learned for Microsoft Word can also be applied to Microsoft Excel, such as basic text formatting and file extensions. The file extension for a Microsoft Excel workbook is **.xlsx**, although pre-2003 versions of Excel might use **.xls**.

In this page, we'll focus on the manipulation of data, rather than the appearance of the worksheet. Additionally, while this page only provides one method of completing each task, there can be multiple ways to accomplish a single goal. For more in-depth instruction, check out this online course covering the basics of Microsoft Excel.

2.1: Using Excel Cells

2.2: Using Excel Data

2.3: Using Excel Functions

2.4: Using Excel Charts

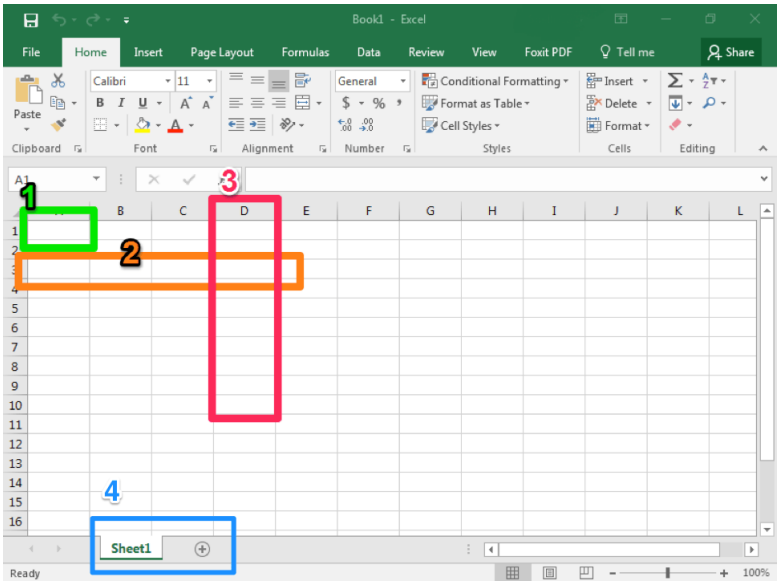
2.5: Using Conditional Formatting in Excel

This page titled 2: Introduction to Working in Microsoft Excel is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Freedom Learning Group via source content that was edited to the style and standards of the LibreTexts platform.

- 4.25: Working in Microsoft Excel by Freedom Learning Group is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

2.1: Using Excel Cells

Before using a workbook, it is helpful to know a few key terms.

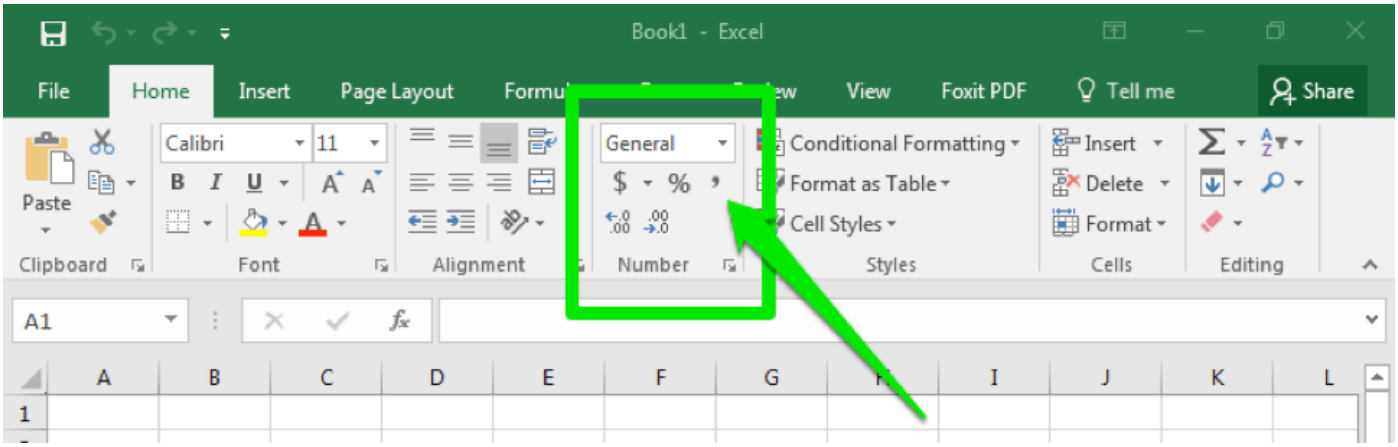


1. **Cell.** This is the area where you will enter data.
2. **Row.** Rows are cells aligned horizontally.
3. **Column.** Columns are cells aligned vertically.
4. **Worksheet.** A worksheet is a single page within a workbook. Like the tabs in an internet browser, the tabs in an Excel workbook show different pages, or worksheets. A workbook may have many worksheets included in it. In this screenshot, the workbook only has one worksheet and one tab, which is labeled Sheet1. The selected tab shows the selected worksheet. Clicking the + button will add another worksheet. When you save a workbook in Excel, all of the worksheets in that workbook are saved.

Comma Styles

At times, you may also wish to use a specific comma style with numbers entered into an Excel worksheet. For example, you may wish “1234” to display as typed or with a comma like “1,234.”

Comma styles are easy to change in Excel using a quick select option in the Number group in the ribbon. Simply to click on the Comma Style button in the Number group.



When clicking the comma style button, the comma style default is to display numbers with a comma in the thousands place and include two decimal places (Ex: "1200" becomes "1,200.00). This will also change the visible cell styles in the Style" area of the ribbon so you can easily select different options for comma and display format.

Listed below are the three most common options for comma and display format.

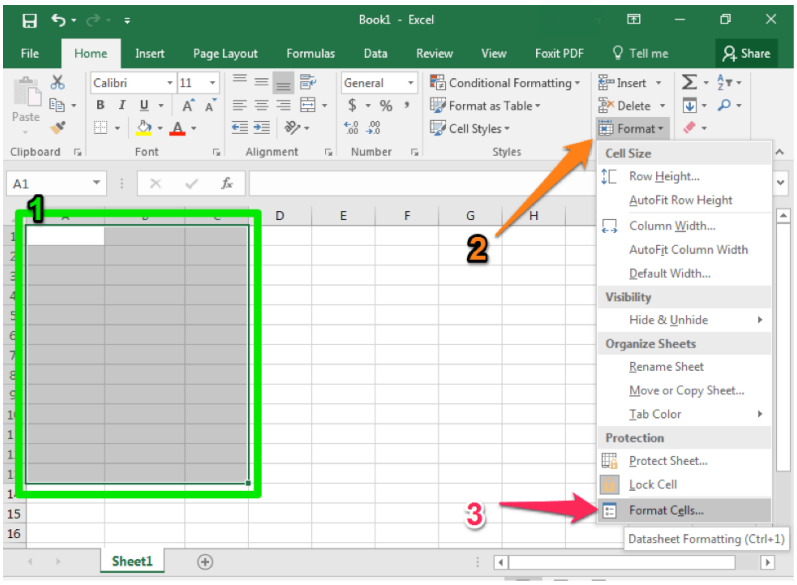
1. **Comma**: Comma with two decimal points (e.g., 1,234.00)
2. **Comma [0]**: Comma with no decimal points (e.g., 1,234)
3. **Currency**: Comma with two decimal points and a dollar sign (e.g., $1,234.00)

Cell Format

As mentioned previously, Excel will default to certain styles when you create a new worksheet. In particular, this includes the way that numbers are displayed and whether or not commas are automatically included. In this section, we will take a look at changing these defaults.

When you type numbers into an Excel workbook, it will often default to a specific format. For example, if you type "12/15/17," Excel will convert this to read "12/15/2017," assuming you were entering month, day, and abbreviated year. Similarly, "3/4" will display at "4-Mar," the fourth day of March. However, it is possible that you may have been entering fractions, so "3/4" was meant to indicate three-quarters instead.

If this is the case, you will need to format your cells to properly display the information you are entering. When possible, consider formatting your cells before you enter the data. Otherwise, Excel may convert some of the entries and you will need to re-enter that information.

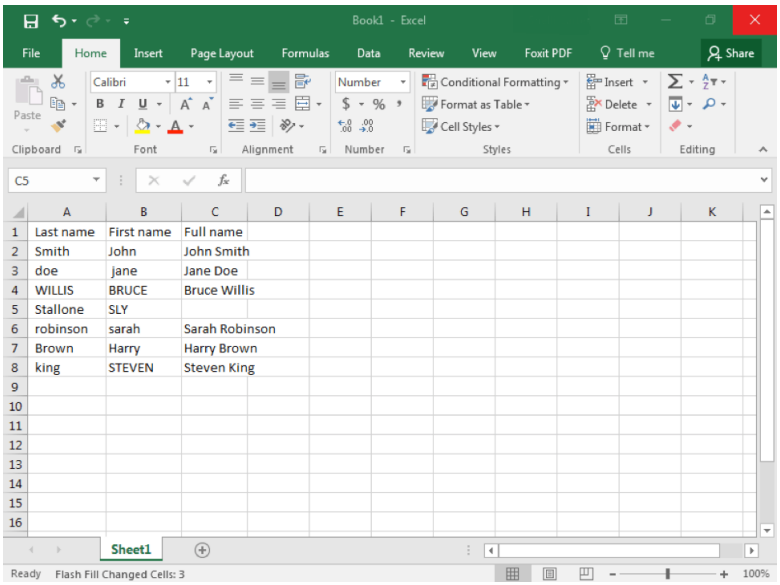


1. Begin by highlighting the cells you plan to use.
2. Select the Format dropdown from the Cells group of the ribbon.
3. Select the Format cells option at the bottom of the dropdown menu.

Flash Fill

Like many modern software programs, Excel is designed to recognize certain patterns. For example, perhaps you are creating a table that lists the last and first names of attendees at a company training session. After all the names have been entered into two separate columns, you realize you would like a single column to correctly display the full name. An easy way to achieve this without having to manually retype the entire list is to use Flash Fill.

1. Create a new column for the combined information you wish to display.
2. In the first cell, type the name as you wish it to display. In our screenshots, this would be “John Smith.”
3. Begin typing the next piece of data in the next cell. Excel should automatically suggest a Flash Fill option.
4. If the Flash Fill suggestion matches how you would like the information displayed, simply hit the Enter key and the rest of your column should fill in automatically.



Flash Fill is especially helpful if your data is initially in different forms but you want the final information to display in the same

fashion. For example, in our attendee list, some of the names were capitalized, in all caps, or had no capitalization. Sometimes you may need to manually update more than one option but Excel will detect your pattern.

Flash Fill should automatically be turned on in Excel but if it is not, you can turn it on using the File > Options > Advanced menus. You can also turn Flash Fill on or off using the shortcut Ctrl+E. Be aware that the Mac version of Excel does not have Flash Fill.

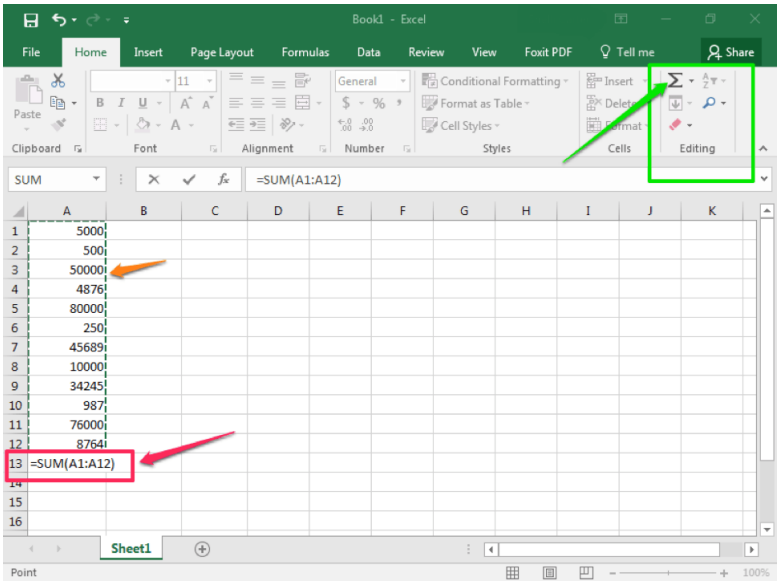
This page titled 2.1: Using Excel Cells is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

2.2: Using Excel Data

SUM Data

One of the main uses for Excel is to organize and manipulate numerical data. Often you may wish to add up all the numbers in a column or row. Excel has formulas and commands to automatically add your data, and the easiest way to use this feature is the AutoSum button.

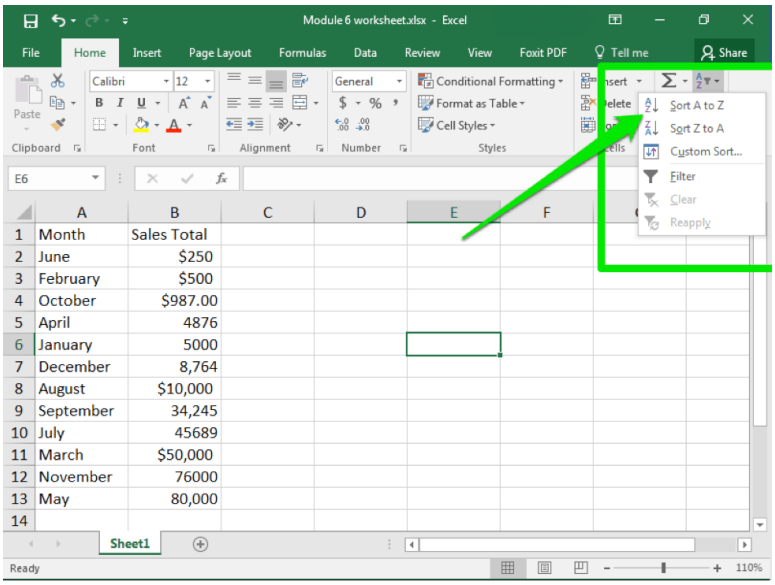
1. Once your numbers are organized in either a row or column, click on the cell where you would like the total sum to display. In the screenshot below this was A13.
2. Click on the AutoSum button from the Editing group of the ribbon.
3. Excel will highlight the cells that it is adding up and will apply the SUM formula.
4. Hit Enter to accept the highlighted cells and see the total value of your data.



Note that it is possible to SUM several columns (or rows) at once. Select all the cells you wish to display a SUM and click AutoSum. Excel will individually add up the columns.

Sorting Data

1. Select the column or row you wish to sort.
2. From the Sort & Filter button in the Editing group in the ribbon, click the Sort button.
3. From the menu, choose how you would like to sort the data. For example, A to Z or Z to A. Note that A to Z is equivalent to Smallest to Largest and Z to A is equivalent to Largest to Smallest.

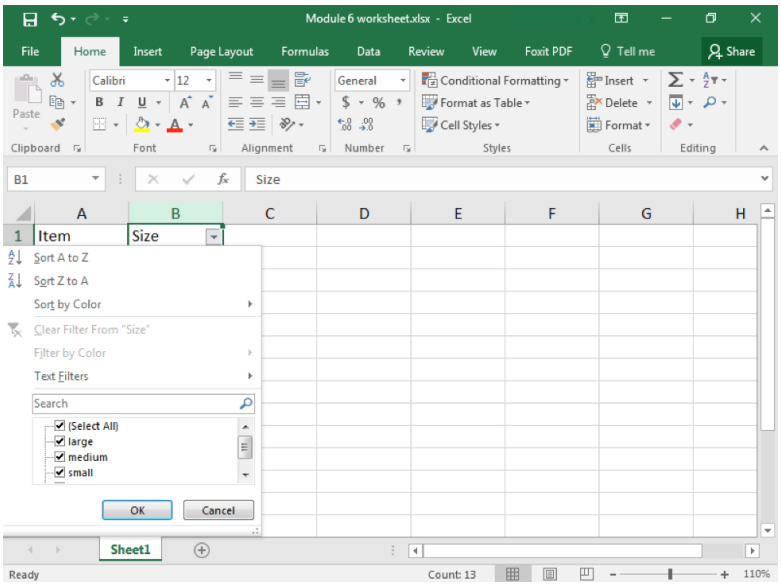


	A	B
1	Month	Sales Total
2	June	$250
3	February	$500
4	October	$987.00
5	April	4876
6	January	5000
7	December	8,764
8	August	$10,000
9	September	34,245
10	July	45689
11	March	$50,000
12	November	76000
13	May	80,000
14		

Filtering Data

After entering data in Excel, it is also possible to filter, or hide some parts of the data, based on user-indicated categories. When using the Filter option, no data is lost; it is just hidden from view.

1. Select the column or row you wish to sort.
2. From the Sort & Filter button in the Editing group in the ribbon, click the Filter button.
3. When the Filter menu appears, you can choose which categories of data to hide and deselect the appropriate buttons. For example, you can deselect the button next to large and you will no longer see the large cells in your table.



This page titled 2.2: Using Excel Data is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

2.3: Using Excel Functions

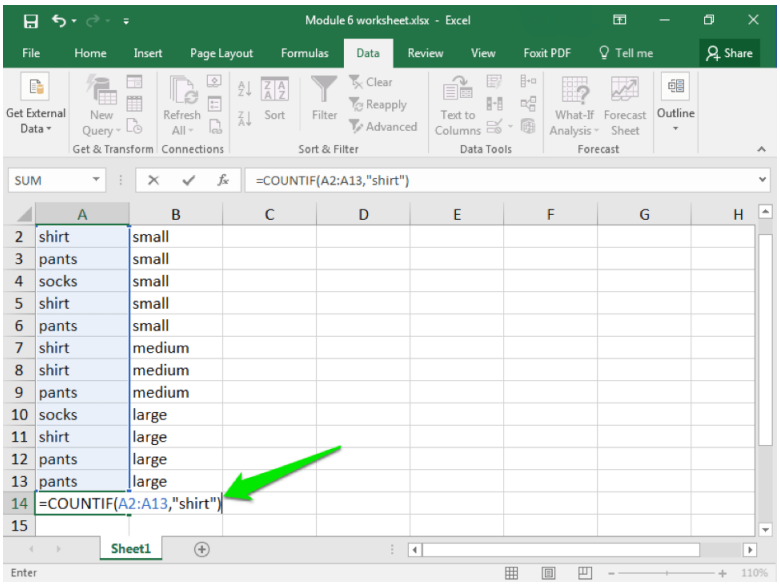
Functions

Excel can perform a variety of really nice data analysis features for you. We've already touched upon how you can filter data. But you can also look for other connections, or screen large numbers of cells to determine how often something occurs.

COUNTIF

COUNTIF is a way for you to ask Excel to count how many times a certain piece of information appears in your worksheet. For example, perhaps you want to know how often "shirt" appear in an inventory list. All you need to do is ask Excel to count the number of cells that contain the word "shirt."

1. Determine which cells you want Excel to look at. In our example, we will look at A2 though A13.
2. Click on the cell you wish your count to be displayed in.
3. Type the formula for a count
   =COUNTIF(A2:A13, "shirt")



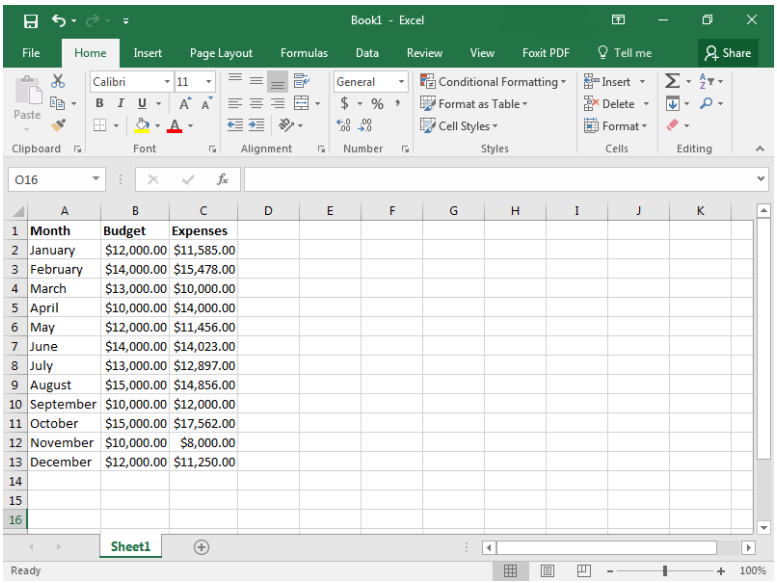
Here you are telling Excel which cells to examine—A2 through A13—and what to look for: "shirt." Note that your text must match exactly what is typed in the cells, and if you are looking for a specific word it needs to be enclosed in quotation marks (so "shirt" instead of shirt).

4. Hit enter and your results will appear.

IF

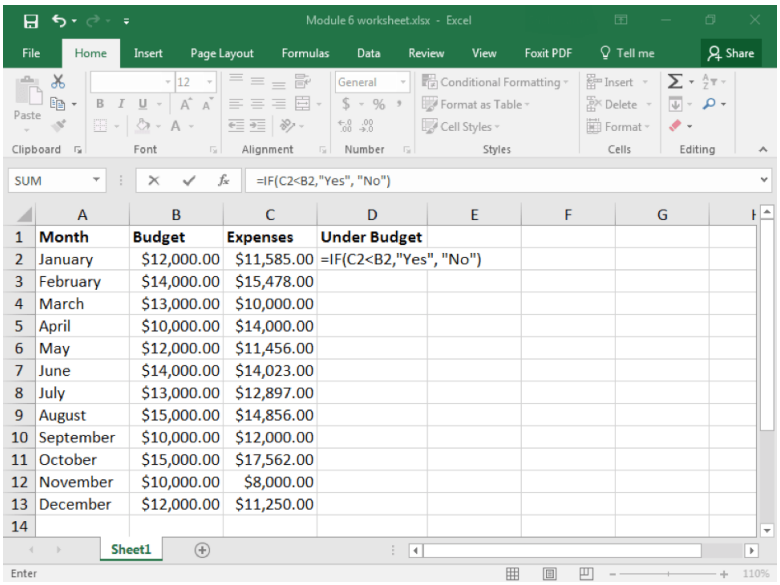
Another commonly used function in Excel is the "IF" function. In this case, you are asking Excel to look for something and then tell you if that something occurred. For example, perhaps you want to compare whether your monthly expenses were under your monthly budget. That is the scenario we will look at in our example.

In this case, let us just ask for a simple "yes" or "no" answer. Looking at the screenshot below, you can see how the worksheet has all the data at hand. We are looking for whether the information in the C column is less than the information in the B column. We would like the D column to display the answer (yes or no).



	A	B	C
1	Month	Budget	Expenses
2	January	$12,000.00	$11,585.00
3	February	$14,000.00	$15,478.00
4	March	$13,000.00	$10,000.00
5	April	$10,000.00	$14,000.00
6	May	$12,000.00	$11,456.00
7	June	$14,000.00	$14,023.00
8	July	$13,000.00	$12,897.00
9	August	$15,000.00	$14,856.00
10	September	$10,000.00	$12,000.00
11	October	$15,000.00	$17,562.00
12	November	$10,000.00	$8,000.00
13	December	$12,000.00	$11,250.00
14			
15			
16			

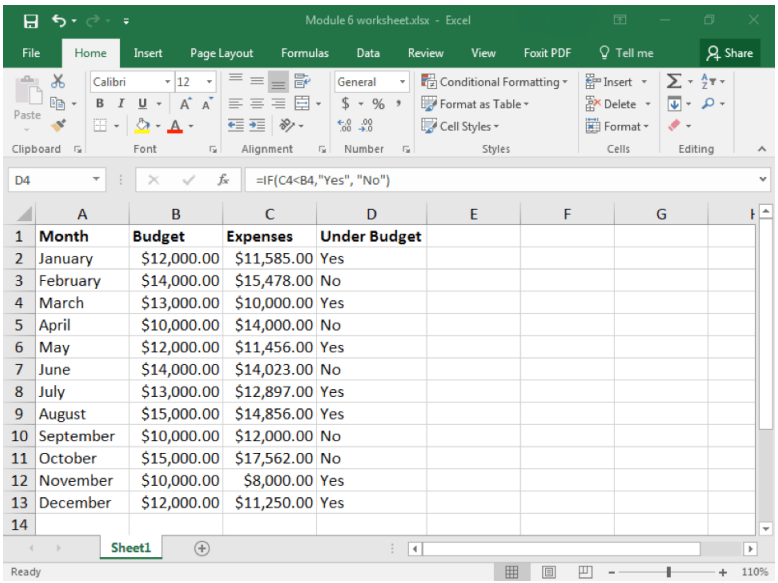
1. Click on D2 and enter the “IF” function for what you want Excel to compare and do.
   =IF(C2



	A	B	C	D	E	F	G
1	Month	Budget	Expenses	Under Budget			
2	January	$12,000.00	$11,585.00	=IF(C2<B2,"Yes", "No")			
3	February	$14,000.00	$15,478.00				
4	March	$13,000.00	$10,000.00				
5	April	$10,000.00	$14,000.00				
6	May	$12,000.00	$11,456.00				
7	June	$14,000.00	$14,023.00				
8	July	$13,000.00	$12,897.00				
9	August	$15,000.00	$14,856.00				
10	September	$10,000.00	$12,000.00				
11	October	$15,000.00	$17,562.00				
12	November	$10,000.00	$8,000.00				
13	December	$12,000.00	$11,250.00				
14							

2. You do not have to manually reenter the formula into the other cells in D. Instead you can copy and paste the formula from D2 into D3, D4, and so on. Each time you do this, the formula should automatically update with the correct cell number to compare.

As you can see, the D cells begin to display “Yes” or “No.” “Yes” means that the expenses in the C column were less than the monthly budget entered into the B column. “No” means that expenses were higher than the budget. Just as in the COUNTIF function, you need to enclose text in quotation marks.



	A	B	C	D
1	**Month**	**Budget**	**Expenses**	**Under Budget**
2	January	$12,000.00	$11,585.00	Yes
3	February	$14,000.00	$15,478.00	No
4	March	$13,000.00	$10,000.00	Yes
5	April	$10,000.00	$14,000.00	No
6	May	$12,000.00	$11,456.00	Yes
7	June	$14,000.00	$14,023.00	No
8	July	$13,000.00	$12,897.00	Yes
9	August	$15,000.00	$14,856.00	Yes
10	September	$10,000.00	$12,000.00	No
11	October	$15,000.00	$17,562.00	No
12	November	$10,000.00	$8,000.00	Yes
13	December	$12,000.00	$11,250.00	Yes
14				

This page titled 2.3: Using Excel Functions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

2.4: Using Excel Charts

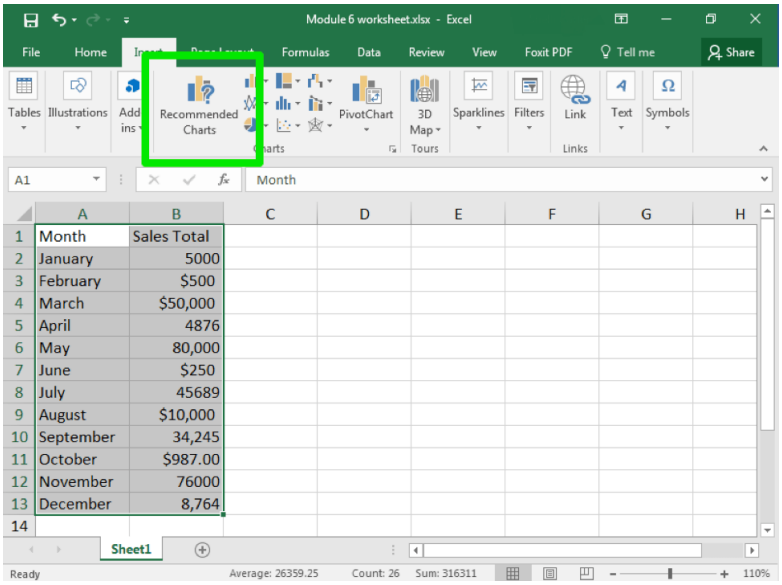
Clustered Column Charts

Excel is not just used for organizing and processing data and formulas. It also can be used to visually represent data in the form of charts and graphs. In this page, we will work on creating a basic chart, the clustered column chart, and then modifying a chart style.

A clustered column chart is sometimes called a bar graph, because it shows data organized in solid shapes like pillars. A clustered column chart organizes these pillars up and down, so they are “columns.” On the other hand, a clustered bar graph organizes these pillars left to right, so they are “bars.” Bar graphs are useful charts when looking at changes from month to month or across employees.

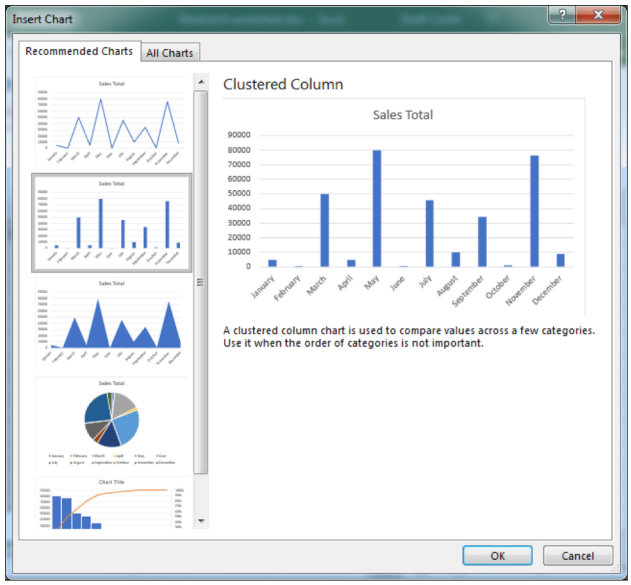
The first step to creating any chart is to organize your data. It is definitely a good idea to include headers in the first cell of each column. By default, a clustered column chart will cluster the data by the columns in your table, so try to keep that in mind when setting up the worksheet.

1. After organizing your data, select the cells you wish to include in the chart. This should be at least two columns.
2. Click on the Insert tab and find the Charts group of the ribbon.

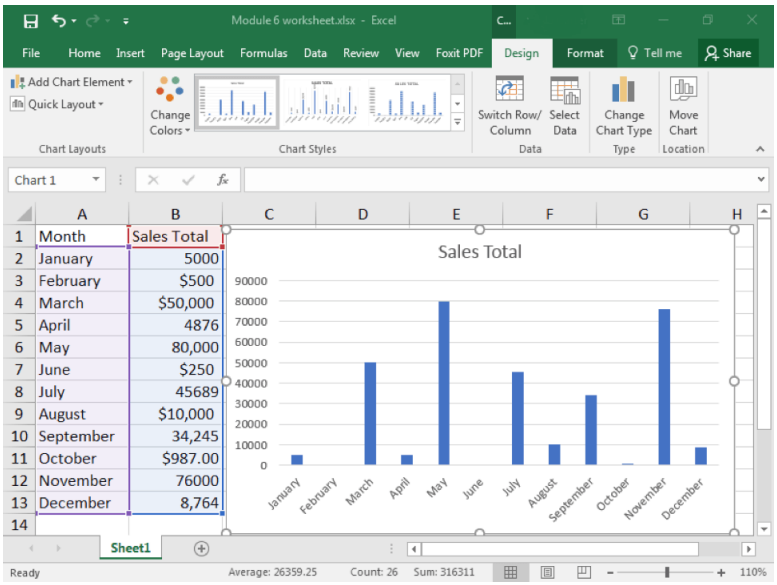


	A	B	C	D	E	F	G	H
1	Month	Sales Total						
2	January	5000						
3	February	$500						
4	March	$50,000						
5	April	4876						
6	May	80,000						
7	June	$250						
8	July	45689						
9	August	$10,000						
10	September	34,245						
11	October	$987.00						
12	November	76000						
13	December	8,764						
14								

3. “Clustered column chart” is actually a recommended chart. Click on that chart.



4. When you select the chart, you will see colored boxes surrounding the data that connect to the different categories of the chart.

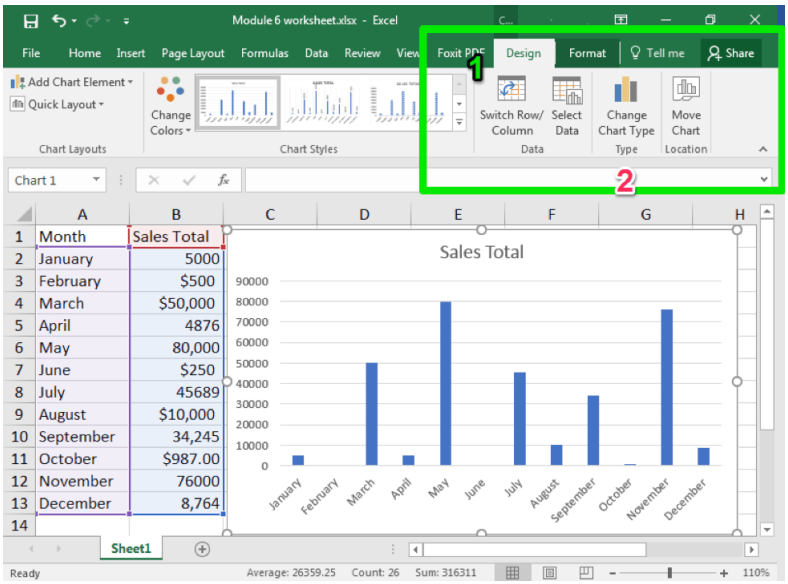


	Month	Sales Total
2	January	5000
3	February	$500
4	March	$50,000
5	April	4876
6	May	80,000
7	June	$250
8	July	45689
9	August	$10,000
10	September	34,245
11	October	$987.00
12	November	76000
13	December	8,764

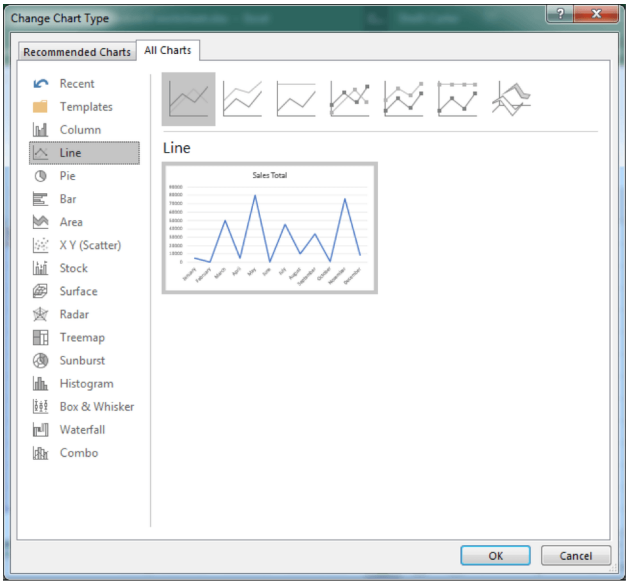
Chart Styles

Once you have created a chart, or if you are given a worksheet that contains a chart, it is very easy to change the chart style.

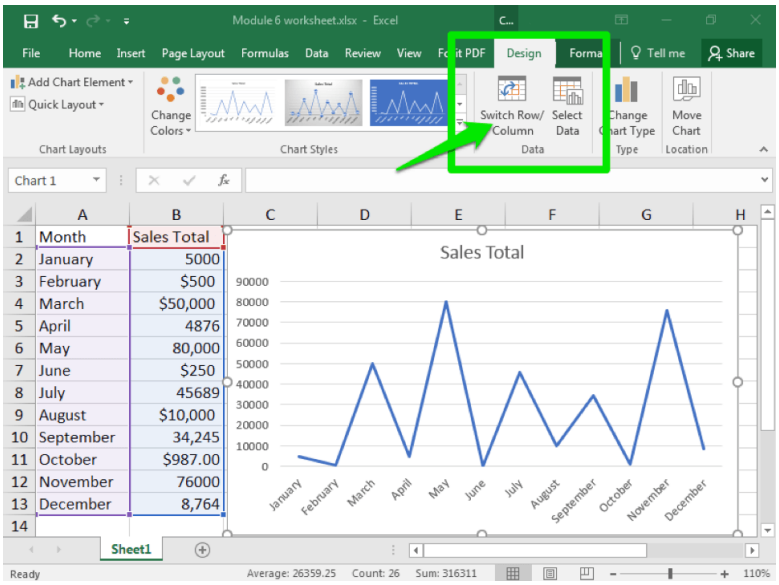
1. Click on the chart you wish to change. The Design tab should appear in the ribbon area.
2. Click on Change Chart Type button



3. Click on the type of chart you would like.



From this same window, you can also switch the data that is being charted. For example, you can switch which data from a row to a column or change which data is arranged on the x- or y-axis.



This page titled 2.4: Using Excel Charts is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

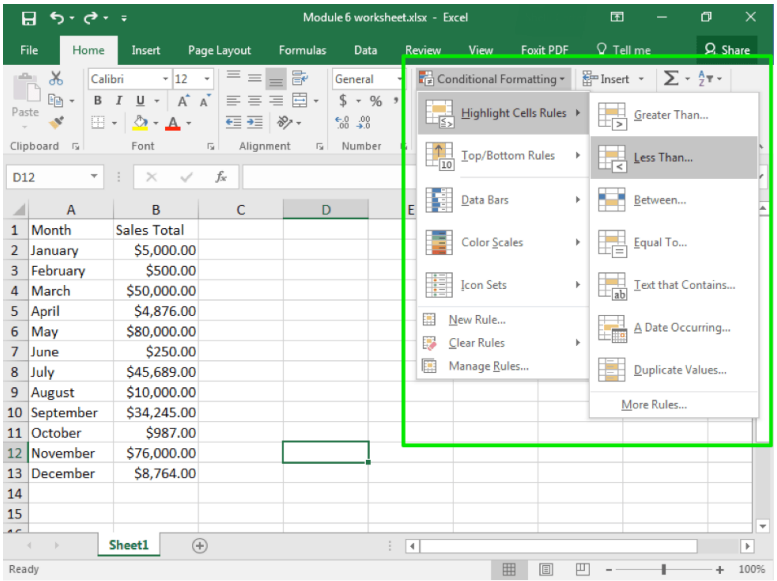
2.5: Using Conditional Formatting in Excel

Conditional Formatting

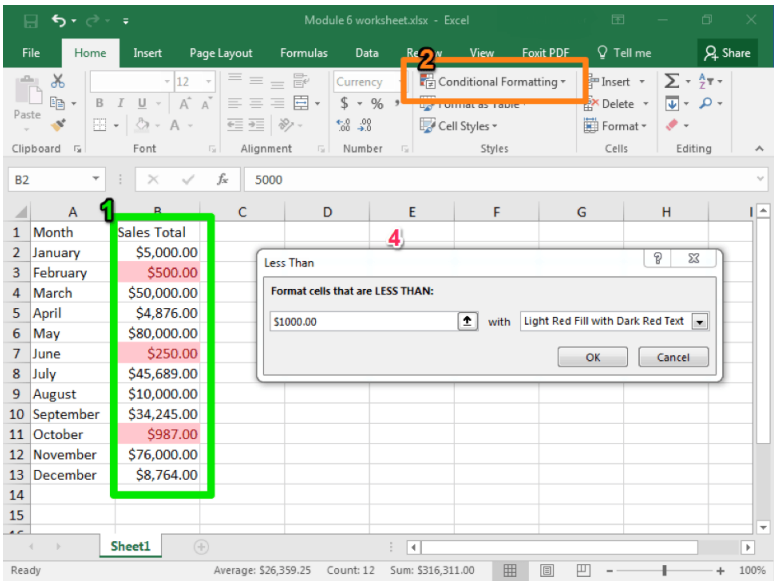
As we have learned so far, Excel has a wide variety of easy to use tools for organizing, sorting, and otherwise marking information. Think back to when we applied styles to a cell to indicate good information or information that needs to be verified. Excel also has the ability to automatically apply such markings through conditional formatting.

With conditional formatting, you provide Excel with a rule, such as “less than 10,” and the program will scan through your data and highlight all the cells that meet that rule. There are several rules already available, but you can also create and apply your own rules and visual clues.

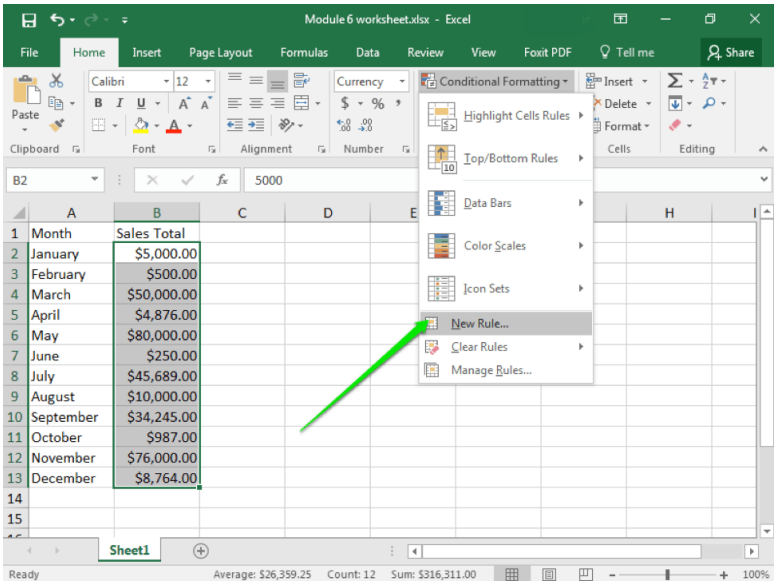
1. Select the cells, rows, or columns you wish to have conditional formatting.
2. From the Styles group, click on the Conditional Formatting button
3. Select the style of formatting you would like. Here we have Highlight Cell Rules.
4. Select the specific type of rule you would like to use and then apply your target value. Here we have selected Less Than.

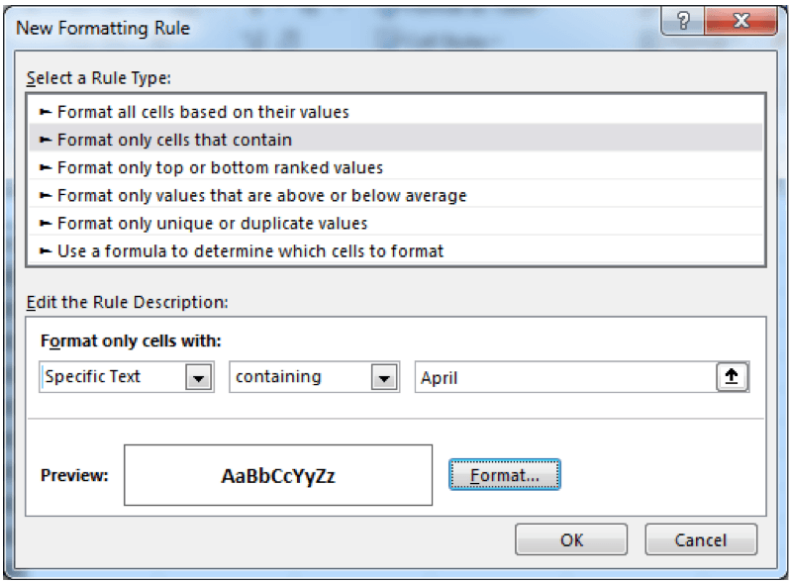


5. The formatting will appear automatically so you can see what it will look like. Note that Excel will automatically provide a value, but you can manually change it.
6. Hit OK if you wish to apply the formatting. Otherwise, when you leave the formatting menu, it will disappear.



One important tool to keep in mind is the ability to enter your own rules. This can include applying formatting to specific date ranges, to specific text (like names), or even cells that are blank. In this case, you also set the format, so instead of highlighting cells you can choose to strikethrough text or change the font, change the size, or bold the text.





This page titled 2.5: Using Conditional Formatting in Excel is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

3: WORKING IN GOOGLE SHEETS

Learning Objectives

Create workbooks and format data in Google Sheets

Google Sheets is very similar to Microsoft Excel. Perhaps the greatest difference between the two is Google Sheets' online functionality, which allows for real-time collaboration. Additionally, as Google Sheets is a free online tool, you can be assured everyone you'll work with has or can get access to Google Sheets, as long as they have the internet.

In Google Sheets, the creation of a spreadsheet, the manipulation of data, and the type of use and results are all very similar to usage in Microsoft Excel. Below is a screenshot of a raw and unused Google Sheet. Note the same columns and rows orientation, various ways to adjust the display of numerical data via the menu bar buttons, and formula buttons along the top bar (Figure 1).

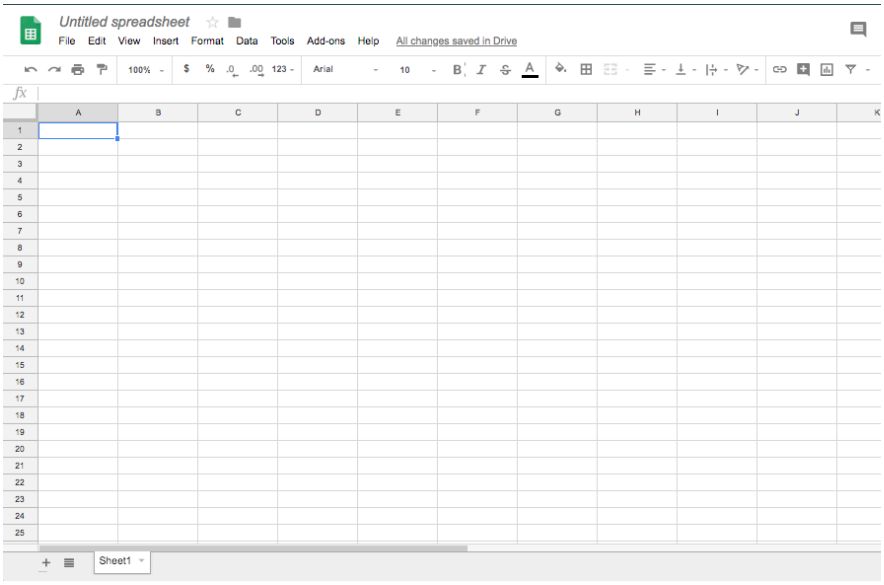


Figure 1. A new spreadsheet in Google Sheets

Learning to use Sheets is similarly helpful to learning Excel; both use the same type of data organization and manipulation, and creating a worksheet is nearly identical. Many of the skills you learned for Excel can also be applied to Google Sheets, such as basic text formatting. File extensions for Google work differently, and arguably a strength of Google Sheets over Microsoft is the ability to collaborate in the cloud real time on a given sheet.

In this page, you will learn the basics of using Google Sheets, focused on rearranging information into tables and changing style elements.

Creating a New Speadsheet

You can open Google Sheets by logging into Google Drive on any web browser or by going directly to sheets.google.com. Once there, you'll be taken to the Google Sheets home screen (Figure 2).

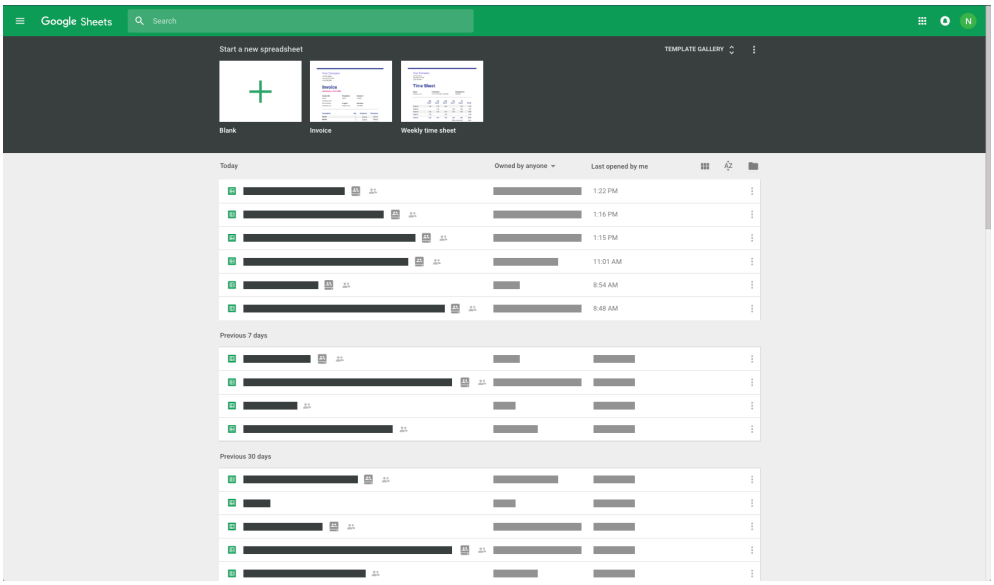


Figure 2. Creating a new spreadsheet

To create a new Google Sheet, click on the blank tile with the large green + symbol (as seen in the upper left of Figure 2). This will open a new sheet identical to the blank sheet seen in Figure 1.

Using Google Sheets

Most of what you see is very similar to Excel. Arguably, Excel has more functionality, and it probably should be used if large data crunching—pages and pages of data—is necessary. For minor data analysis, or for looking at flow charts or other data visually, Google Sheets is probably better than Excel. This is true particularly if you consider the robust sharing features that come with Google Drive. Microsoft Office also has online, cloud-based sharing. Our recommendation would be to experiment with both for your research needs.

Like Excel, Sheets has the following exact terms:

1. Cell. This is the area where you will enter data.
2. Row. Rows are cells aligned horizontally.
3. Column. Columns are cells aligned vertically.
4. Worksheet. A worksheet is a single page within a workbook. Like the tabs in an internet browser, the tabs in a Google Sheets workbook show different pages, or worksheets. A workbook may have many worksheets included in it. In this screenshot, the workbook has only one worksheet and one tab, which is labeled Sheet1. The selected tab shows the selected worksheet. Clicking the + button will add another worksheet. Saving is automatic in Google; when you build more sheets, they will all be saved in that workbook.

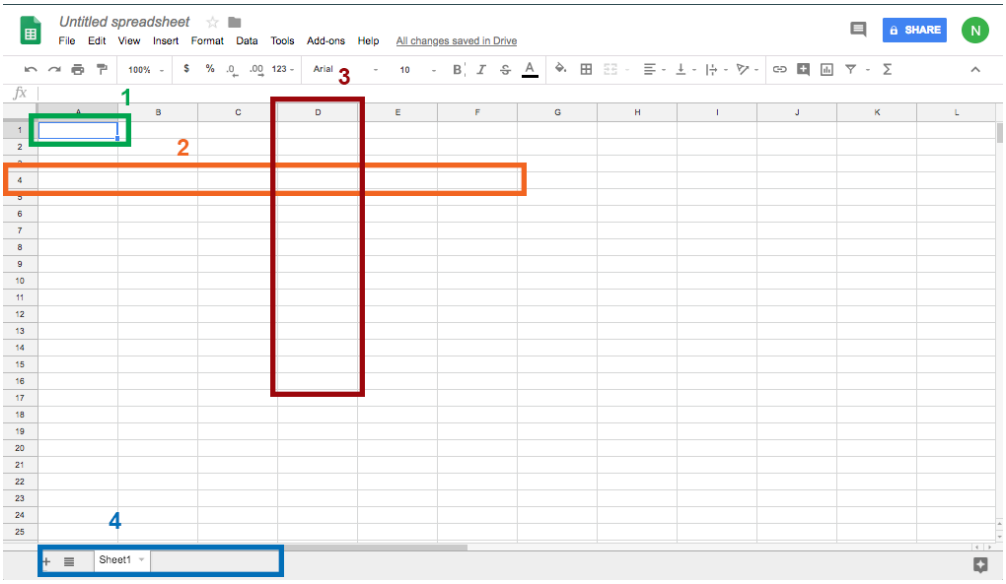


Figure 3. Four primary components of a spreadsheet

Add-ons

While the majority of Google Sheets is almost identical to Microsoft Excel, there are notable differences. The primary difference we'll focus on here are Google Sheets add-ons, which you can access in a drop-down menu (Figure 4).

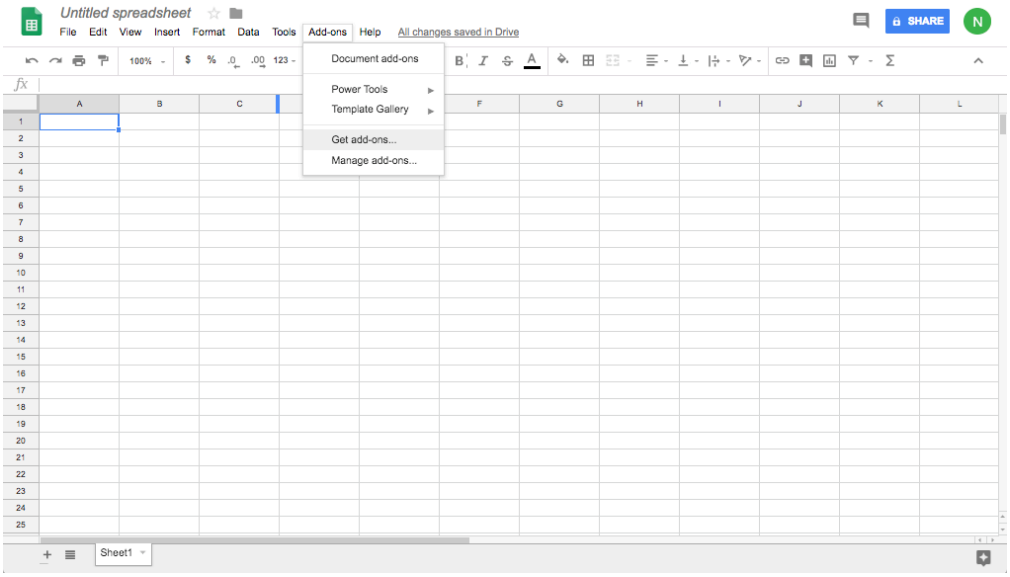


Figure 4. Add-ons are accessible through the menu bar

A strength of Google Drive, and Google's office suite writ large, is the addition of third party add-ons. These range from styles that can change the look and feel of a given Sheet (see Figure 5), to the addition of macros or tools that help with grading or other type of data analysis. Depending on your data and goals for using the data, you can choose from literally thousands of options. If you are looking for a particular functionality, you can use the search bar in the upper-right corner of the add-ons window (Figure 5).

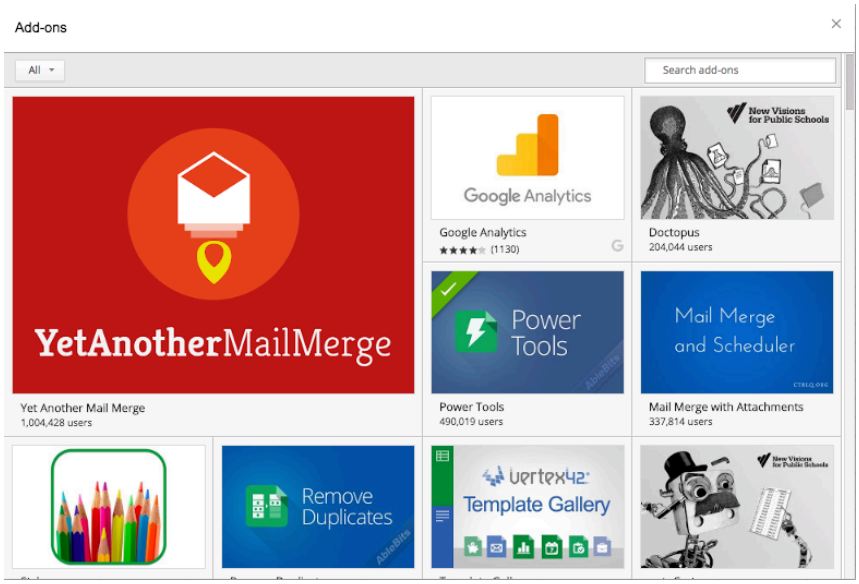


Figure 5. There are a huge variety of add-ons available in Google Sheets.

These add-ons can be used for a variety of purposes, most of which revolve around adding or "popping" functionality into the spreadsheet itself.

Vertex Template Gallery

The Vertex Template Gallery (seen in the bottom row of Figure 5) gives you the option to use several pre-built templates ranging from project management/waterfall planning tools, to budget sheets and even survey forms.

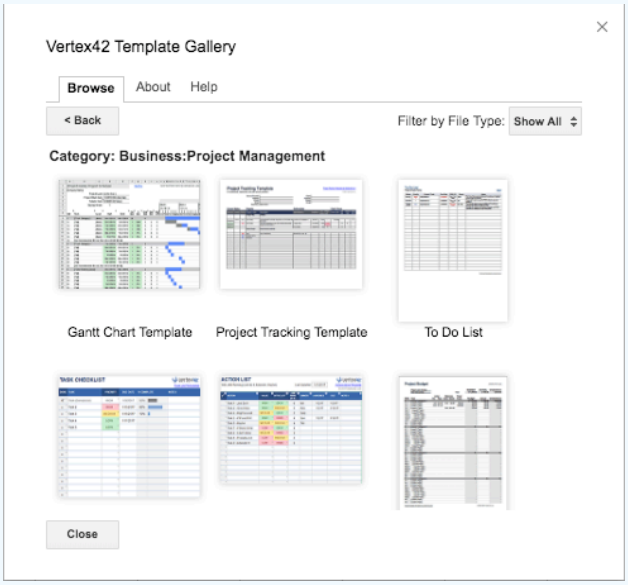


Figure 6. Templates available through Vertex Template Gallery

For your business research needs, there are a few key strengths of Google Sheets to consider. Michael Ansaldo, of *PC World*, wrote that users should consider Sheets when the following are key aspects of your project:[1]

1. **Cost:** Google is free to use, which can be particularly useful for smaller or newer teams where money is often much tighter.
2. **Collaboration:** Google Sheets is better suited to collaboration between multiple parties, particularly if you need to write on the same spreadsheet in real time from various computers and places. Using a shared spreadsheet can cut down on time and confusion.
3. **Google Integration.** Google has created a host of tools that provide unique functions, and these tools can integrate

seamlessly with Google Sheets.
4. **Tracking Changes.** Google Sheets automatically keeps track of the version history of a document, as well as keeping track of who made changes (as long as a user is signed in).

In his article, Anslado expands on the utility of Google integration:[2]

Because of its tight integration with Google, Sheets can import all kinds of data from other Google services and the web at large. You can translate the contents of a cell using the function GOOGLETRANSLATE(), or you can fetch current or historical securities info from Google Finance with the function GOOGLEFINANCE(). And with Sheets IMPORTFEED and IMPORTDATA functions, you can pull information from the internet directly into your spreadsheet.

Learning Objectives

Both Microsoft Excel and Google Sheets are excellent spreadsheet tools. Both have similar functionality, and use formulas to "crunch" quantitative data. Arguably for business report writing, Google Sheets might be better, especially if there is a need to collaborate on the same sheet in real time.

---

1. Ansaldo, Michael. "Microsoft Excel vs. Google Sheets: The 4 key ways Sheets beats Excel." *PCWorld*. Sep 28, 2015. Web. 12 June 2018.
2. Ibid.

3.1: Creating a New Spreadsheet
3.2: Using Google Sheets
3.3: Add-Ons

This page titled 3: Working in Google Sheets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Freedom Learning Group via source content that was edited to the style and standards of the LibreTexts platform.

- 4.26: Working in Google Sheets by Freedom Learning Group is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

3.1: Creating a New Spreadsheet

Creating a New Spreadsheet

You can open Google Sheets by logging into Google Drive on any web browser or by going directly to sheets.google.com. Once there, you'll be taken to the Google Sheets home screen (Figure 2).

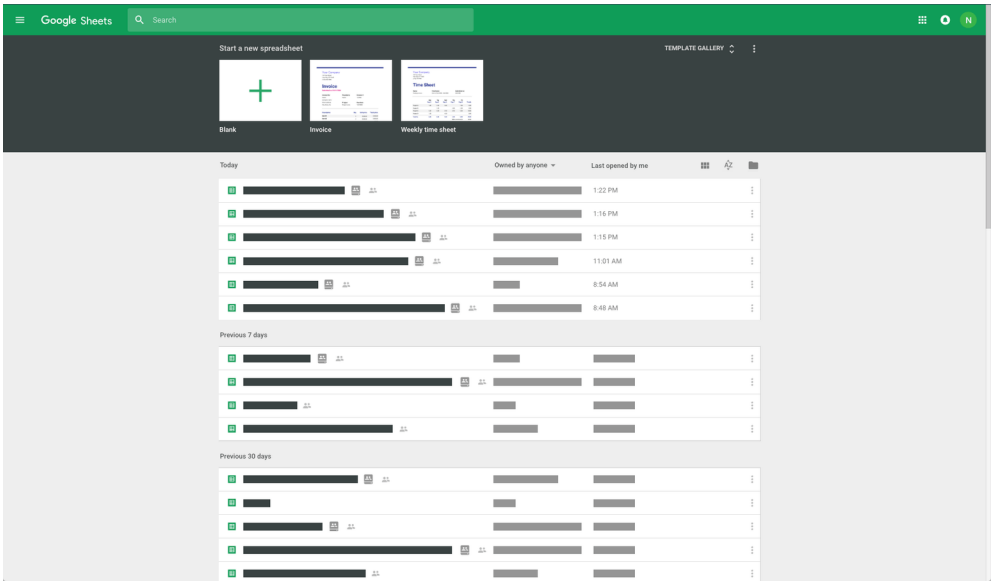


Figure 2. Creating a new spreadsheet

To create a new Google Sheet, click on the blank tile with the large green + symbol (as seen in the upper left of Figure 2). This will open a new sheet identical to the blank sheet seen in Figure 1.

This page titled 3.1: Creating a New Spreadsheet is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

- 4.26: Working in Google Sheets by Freedom Learning Group is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

3.2: Using Google Sheets

Using Google Sheets

Most of what you see is very similar to Excel. Arguably, Excel has more functionality, and it probably should be used if large data crunching—pages and pages of data—is necessary. For minor data analysis, or for looking at flow charts or other data visually, Google Sheets is probably better than Excel. This is true particularly if you consider the robust sharing features that come with Google Drive. Microsoft Office also has online, cloud-based sharing. Our recommendation would be to experiment with both for your research needs.

Like Excel, Sheets has the following exact terms:

1. Cell. This is the area where you will enter data.
2. Row. Rows are cells aligned horizontally.
3. Column. Columns are cells aligned vertically.
4. Worksheet. A worksheet is a single page within a workbook. Like the tabs in an internet browser, the tabs in a Google Sheets workbook show different pages, or worksheets. A workbook may have many worksheets included in it. In this screenshot, the workbook has only one worksheet and one tab, which is labeled Sheet1. The selected tab shows the selected worksheet. Clicking the + button will add another worksheet. Saving is automatic in Google; when you build more sheets, they will all be saved in that workbook.

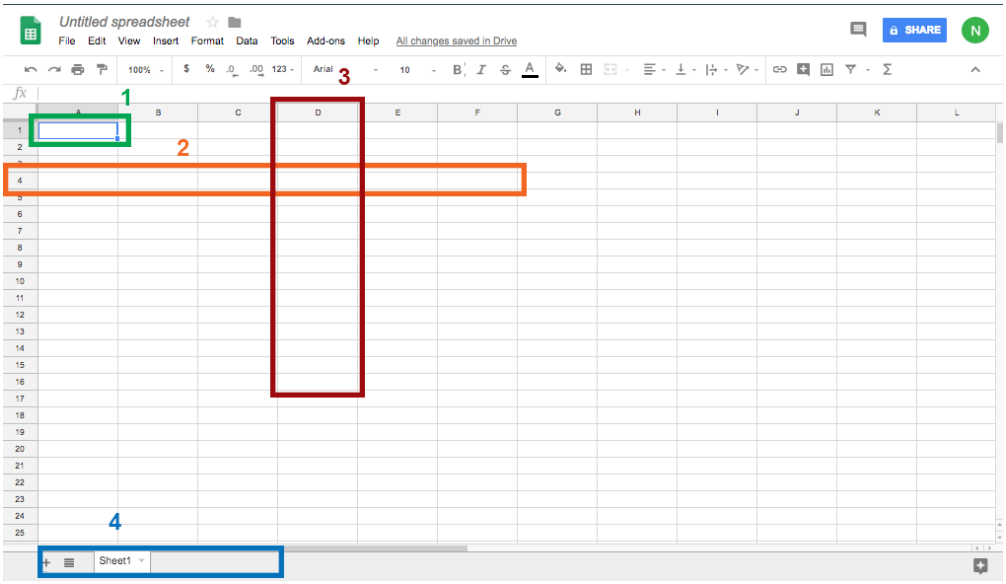


Figure 3. Four primary components of a spreadsheet

This page titled 3.2: Using Google Sheets is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

3.3: Add-Ons

Add-ons

While the majority of Google Sheets is almost identical to Microsoft Excel, there are notable differences. The primary difference we'll focus on here are Google Sheets add-ons, which you can access in a drop-down menu (Figure 4).

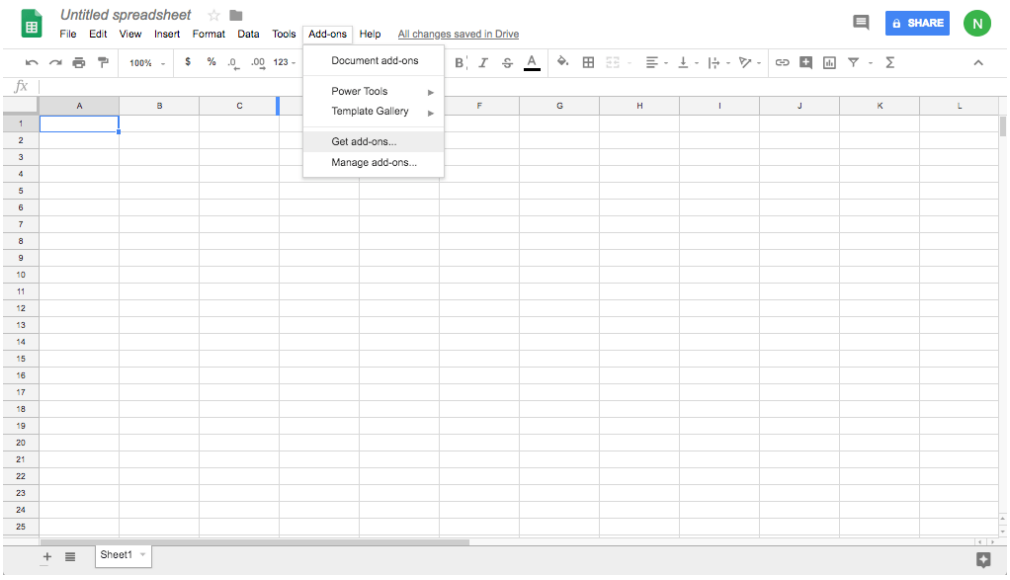


Figure 4. Add-ons are accessible through the menu bar

A strength of Google Drive, and Google's office suite writ large, is the addition of third party add-ons. These range from styles that can change the look and feel of a given Sheet (see Figure 5), to the addition of macros or tools that help with grading or other type of data analysis. Depending on your data and goals for using the data, you can choose from literally thousands of options. If you are looking for a particular functionality, you can use the search bar in the upper-right corner of the add-ons window (Figure 5).

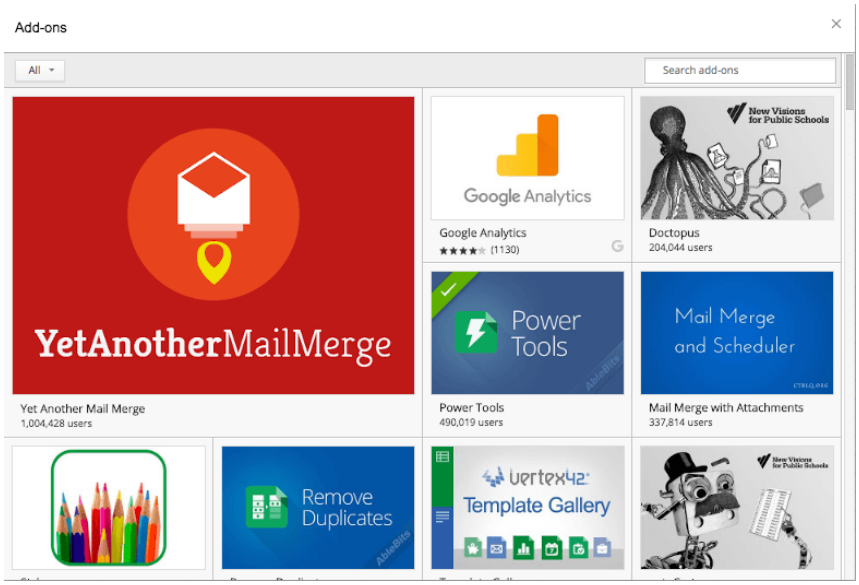


Figure 5. There are a huge variety of add-ons available in Google Sheets.

These add-ons can be used for a variety of purposes, most of which revolve around adding or “popping” functionality into the spreadsheet itself.

Vertex Template Gallery

The Vertex Template Gallery (seen in the bottom row of Figure 5) gives you the option to use several pre-built templates ranging from project management/waterfall planning tools, to budget sheets and even survey forms.

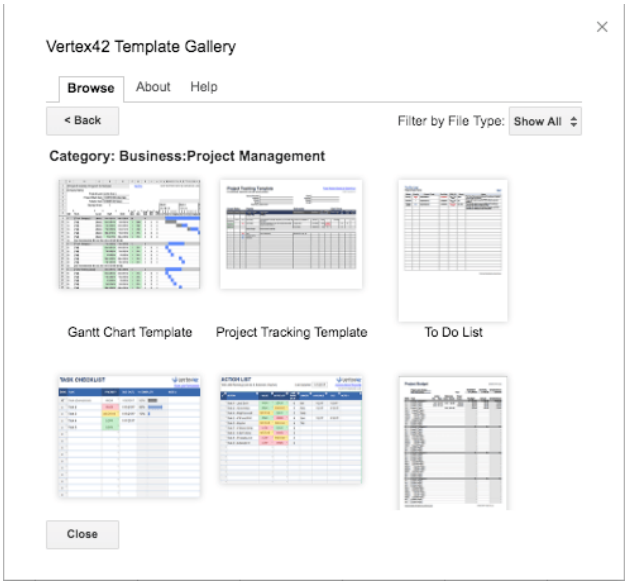


Figure 6. Templates available through Vertex Template Gallery

For your business research needs, there are a few key strengths of Google Sheets to consider. Michael Ansaldo, of *PC World*, wrote that users should consider Sheets when the following are key aspects of your project:[1]

1. **Cost:** Google is free to use, which can be particularly useful for smaller or newer teams where money is often much tighter.
2. **Collaboration:** Google Sheets is better suited to collaboration between multiple parties, particularly if you need to write on the same spreadsheet in real time from various computers and places. Using a shared spreadsheet can cut down on time and confusion.
3. **Google Integration.** Google has created a host of tools that provide unique functions, and these tools can integrate seamlessly with Google Sheets.
4. **Tracking Changes.** Google Sheets automatically keeps track of the version history of a document, as well as keeping track of who made changes (as long as a user is signed in).

In his article, Anslado expands on the utility of Google integration:[2]

Because of its tight integration with Google, Sheets can import all kinds of data from other Google services and the web at large. You can translate the contents of a cell using the function GOOGLETRANSLATE(), or you can fetch current or historical securities info from Google Finance with the function GOOGLEFINANCE(). And with Sheets IMPORTFEED and IMPORTDATA functions, you can pull information from the internet directly into your spreadsheet.

Learning Objectives

Both Microsoft Excel and Google Sheets are excellent spreadsheet tools. Both have similar functionality, and use formulas to “crunch” quantitative data. Arguably for business report writing, Google Sheets might be better, especially if there is a need to collaborate on the same sheet in real time.

---

1. Ansaldo, Michael. "Microsoft Excel vs. Google Sheets: The 4 key ways Sheets beats Excel." *PCWorld*. Sep 28, 2015. Web. 12 June 2018.

2. Ibid.

This page titled 3.3: Add-Ons is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

4: SQL STRUCTURED QUERY LANGUAGE

Structured Query Language (SQL) is a database language designed for managing data held in a relational database management system. SQL was initially developed by IBM in the early 1970s (Date 1986). The initial version, called *SEQUEL* (Structured English Query Language), was designed to manipulate and retrieve data stored in IBM's quasi-relational database management system, System R. Then in the late 1970s, Relational Software Inc., which is now Oracle Corporation, introduced the first commercially available implementation of SQL, Oracle V2 for VAX computers.

Many of the currently available relational DBMSs, such as Oracle Database, Microsoft SQL Server (shown in Figure 15.1), MySQL, IBM DB2, IBM Informix and Microsoft Access, use SQL.

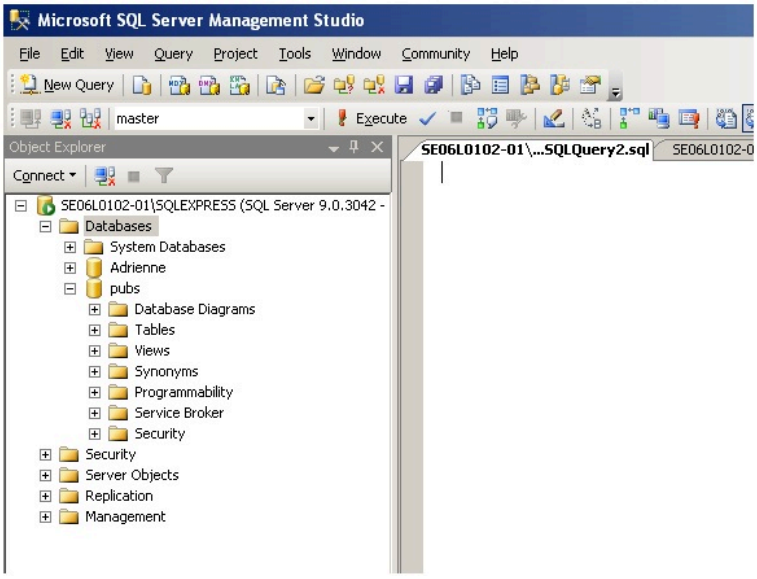


Figure 15.1. Example of Microsoft SQL Server, by A. Watt.

In a DBMS, the SQL database language is used to:

- Create the database and table structures
- Perform basic data management chores (add, delete and modify)
- Perform complex queries to transform raw data into useful information

In this chapter, we will focus on using SQL to create the database and table structures, mainly using SQL as a data definition language (*DDL*). In Chapter 16, we will use SQL as a data manipulation language (*DML*) to insert, delete, select and update data within the database tables.

4.1: Create Database

4.2: Constraints

4.3: User Defined Types and Altering Tables

4.4: Key Terms and Exercises

This page titled 4: SQL Structured Query Language is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

4.1: Create Database

Create Database

The major SQL DDL statements are CREATE DATABASE and CREATE/DROP/ALTER TABLE. The SQL statement CREATE is used to create the database and table structures.

Example: CREATE DATABASE SW

A new database named **SW** is created by the SQL statement CREATE DATABASE SW. Once the database is created, the next step is to create the database tables.

The general format for the CREATE TABLE command is:

CREATE TABLE

(
ColumnName, Datatype, Optional Column Constraint,
ColumnName, Datatype, Optional Column Constraint,
Optional table Constraints
);

Tablename is the name of the database table such as **Employee**. Each field in the CREATE TABLE has three parts (see above):

1. ColumnName
2. Data type
3. Optional Column Constraint

ColumnName

The ColumnName must be unique within the table. Some examples of ColumnNames are FirstName and LastName.

Data Type

The data type, as described below, must be a system data type or a user-defined data type. Many of the data types have a size such as CHAR(35) or Numeric(8,2).

Bit–Integer data with either a 1 or 0 value

Int –Integer (whole number) data from -2^31 (-2,147,483,648) through 2^31 – 1 (2,147,483,647)

Smallint –Integer data from 2^15 (-32,768) through 2^15 – 1 (32,767)

Tinyint–Integer data from 0 through 255

Decimal –Fixed precision and scale numeric data from -10^38 -1 through 10^38

Numeric –A synonym for **decimal**

Timestamp –A database-wide unique number

Uniqueidentifier –A globally unique identifier (GUID)

Money – Monetary data values from -2^63 (-922,337,203,685,477.5808) through 2^63 – 1 (+922,337,203,685,477.5807), with accuracy to one-ten-thousandth of a monetary unit

Smallmoney –Monetary data values from -214,748.3648 through +214,748.3647, with accuracy to one-ten-thousandth of a monetary unit

Float –Floating precision number data from -1.79E + 308 through 1.79E + 308

Real –Floating precision number data from -3.40E + 38 through 3.40E + 38

Datetime –Date and time data from January 1, 1753, to December 31, 9999, with an accuracy of one-three-hundredths of a second, or 3.33 milliseconds

Smalldatetime –Date and time data from January 1, 1900, through June 6, 2079, with an accuracy of one minute

Char –Fixed-length non-Unicode character data with a maximum length of 8,000 characters

Varchar –Variable-length non-Unicode data with a maximum of 8,000 characters

Text –Variable-length non-Unicode data with a maximum length of 2^31 – 1 (2,147,483,647) characters

Binary –Fixed-length binary data with a maximum length of 8,000 bytes

Varbinary –Variable-length binary data with a maximum length of 8,000 bytes

Image –Variable-length binary data with a maximum length of 2^31 – 1 (2,147,483,647) bytes

This page titled 4.1: Create Database is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

4.2: Constraints

Optional Column Constraints

The Optional ColumnConstraints are NULL, NOT NULL, UNIQUE, PRIMARY KEY and DEFAULT, used to initialize a value for a new record. The column constraint NULL indicates that null values are allowed, which means that a row can be created without a value for this column. The column constraint NOT NULL indicates that a value must be supplied when a new row is created.

To illustrate, we will use the SQL statement CREATE TABLE EMPLOYEES to create the employees table with 16 attributes or fields.

```
USE SW
CREATE TABLE EMPLOYEES
(
EmployeeNo CHAR(10) NOT NULL UNIQUE,
DepartmentName CHAR(30) NOT NULL DEFAULT “Human Resources”,
FirstName CHAR(25) NOT NULL,
LastName CHAR(25) NOT NULL,
Category CHAR(20) NOT NULL,
HourlyRate CURRENCY NOT NULL,
TimeCard LOGICAL NOT NULL,
HourlySalaried CHAR(1) NOT NULL,
EmpType CHAR(1) NOT NULL,
Terminated LOGICAL NOT NULL,
ExemptCode CHAR(2) NOT NULL,
Supervisor LOGICAL NOT NULL,
SupervisorName CHAR(50) NOT NULL,
BirthDate DATE NOT NULL,
CollegeDegree CHAR(5) NOT NULL,
CONSTRAINT Employee_PK PRIMARY KEY(EmployeeNo
);
```

The first field is EmployeeNo with a field type of CHAR. For this field, the field length is 10 characters, and the user cannot leave this field empty (NOT NULL).

Similarly, the second field is DepartmentName with a field type CHAR of length 30. After all the table columns are defined, a table constraint, identified by the word CONSTRAINT, is used to create the primary key:

```
CONSTRAINT EmployeePK PRIMARY KEY(EmployeeNo)
```

We will discuss the constraint property further later in this chapter.

Likewise, we can create a Department table, a Project table and an Assignment table using the CREATE TABLE SQL DDL command as shown in the below example.

```
USE SW
CREATE TABLE DEPARTMENT
(
DepartmentName Char(35) NOT NULL,
BudgetCode Char(30) NOT NULL,
OfficeNumber Char(15) NOT NULL,
Phone Char(15) NOT NULL,
CONSTRAINT DEPARTMENT_PK PRIMARY KEY(DepartmentName)
);
```

In this example, a project table is created with seven fields: ProjectID, ProjectName, Department, MaxHours, StartDate, and

EndDate.

```
USE SW
CREATE TABLE PROJECT
(
ProjectID Int NOT NULL IDENTITY (1000,100),
ProjectName Char(50) NOT NULL,
Department Char(35) NOT NULL,
MaxHours Numeric(8,2) NOT NULL DEFAULT 100,
StartDate DateTime NULL,
EndDate DateTime NULL,
CONSTRAINT ASSIGNMENT_PK PRIMARY KEY(ProjectID)
);
```

In this last example, an assignment table is created with three fields: ProjectID, EmployeeNumber, and HoursWorked. The assignment table is used to record who (EmployeeNumber) and how much time(HoursWorked) an employee worked on the particular project(ProjectID).

```
USE SW
CREATE TABLE ASSIGNMENT
(
ProjectID Int NOT NULL,
EmployeeNumber Int NOT NULL,
HoursWorked Numeric(6,2) NULL,
);
```

Table Constraints

Table constraints are identified by the CONSTRAINT keyword and can be used to implement various constraints described below.

IDENTITY constraint

We can use the optional column constraint IDENTITY to provide a unique, incremental value for that column. Identity columns are often used with the PRIMARY KEY constraints to serve as the unique row identifier for the table. The IDENTITY property can be assigned to a column with a tinyint, smallint, int, decimal or numeric data type. This constraint:

- Generates sequential numbers
- Does not enforce entity integrity
- Only one column can have the IDENTITY property
- Must be defined as an integer, numeric or decimal data type
- Cannot update a column with the IDENTITY property
- Cannot contain NULL values
- Cannot bind defaults and default constraints to the column

For IDENTITY[(seed, increment)]

- Seed – the initial value of the identity column
- Increment – the value to add to the last increment column

We will use another database example to further illustrate the SQL DDL statements by creating the table tblHotel in this HOTEL database.

```
CREATE TABLE tblHotel
(
HotelNo Int IDENTITY (1,1),
Name Char(50) NOT NULL,
Address Char(50) NULL,
City Char(25) NULL,
```

)

UNIQUE constraint

The UNIQUE constraint prevents duplicate values from being entered into a column.

- Both PK and UNIQUE constraints are used to enforce entity integrity.
- Multiple UNIQUE constraints can be defined for a table.
- When a UNIQUE constraint is added to an existing table, the existing data is always validated.
- A UNIQUE constraint can be placed on columns that accept nulls. *Only one row can be NULL*.
- A UNIQUE constraint automatically creates a unique index on the selected column.

This is the general syntax for the UNIQUE constraint:

```
[CONSTRAINT constraint_name]
UNIQUE [CLUSTERED | NONCLUSTERED]
(col_name [, col_name2 [..., col_name16]])
[ON segment_name]
```

This is an examle using the UNIQUE constraint.

```
CREATE TABLE EMPLOYEES
(
EmployeeNo CHAR(10) NOT NULL UNIQUE,
)
```

FOREIGN KEY constraint

The FOREIGN KEY (FK) constraint defines a column, or combination of columns, whose values match the PRIMARY KEY (PK) of another table.

- Values in an FK are automatically updated when the PK values in the associated table are updated/changed.
- FK constraints must reference PK or the UNIQUE constraint of another table.
- The number of columns for FK must be same as PK or UNIQUE constraint.
- If the WITH NOCHECK option is used, the FK constraint will not validate existing data in a table.
- No index is created on the columns that participate in an FK constraint.

This is the general syntax for the FOREIGN KEY constraint:

```
[CONSTRAINT constraint_name]
[FOREIGN KEY (col_name [, col_name2 [..., col_name16]])]
REFERENCES [owner.]ref_table [(ref_col [, ref_col2 [..., ref_col16]])]
```

In this example, the field HotelNo in the tblRoom table is a FK to the field HotelNo in the tblHotel table shown previously.

```
USE HOTEL
GO
CREATE TABLE tblRoom
(
HotelNo Int NOT NULL ,
RoomNo Int NOT NULL,
Type Char(50) NULL,
Price Money NULL,
PRIMARY KEY (HotelNo, RoomNo),
FOREIGN KEY (HotelNo) REFERENCES tblHotel
)
```

CHECK constraint

The CHECK constraint restricts values that can be entered into a table.

- It can contain search conditions similar to a WHERE clause.
- It can reference columns in the same table.
- The data validation rule for a CHECK constraint must evaluate to a boolean expression.
- It can be defined for a column that has a rule bound to it.

This is the general syntax for the CHECK constraint:

```
[CONSTRAINT constraint_name]
CHECK [NOT FOR REPLICATION] (expression)
```

In this example, the Type field is restricted to have only the types ‘Single’, ‘Double’, ‘Suite’ or ‘Executive’.

```
USE HOTEL
GO
CREATE TABLE tblRoom
(
HotelNo Int NOT NULL,
RoomNo Int NOT NULL,
Type Char(50) NULL,
Price Money NULL,
PRIMARY KEY (HotelNo, RoomNo),
FOREIGN KEY (HotelNo) REFERENCES tblHotel
CONSTRAINT Valid_Type
CHECK (Type IN (‘Single’, ‘Double’, ‘Suite’, ‘Executive’))
)
```

In this second example, the employee hire date should be before January 1, 2004, or have a salary limit of $300,000.

```
GO
CREATE TABLE SALESREPS
(
Empl_num Int Not Null
CHECK (Empl_num BETWEEN 101 and 199),
Name Char (15),
Age Int CHECK (Age >= 21),
Quota Money CHECK (Quota >= 0.0),
HireDate DateTime,
CONSTRAINT QuotaCap CHECK ((HireDate < “01-01-2004”) OR (Quota <=300000))
)
```

DEFAULT constraint

The DEFAULT constraint is used to supply a value that is automatically added for a column if the user does not supply one.

- A column can have only one DEFAULT.
- The DEFAULT constraint cannot be used on columns with a timestamp data type or identity property.
- DEFAULT constraints are automatically bound to a column when they are created.

The general syntax for the DEFAULT constraint is:

```
[CONSTRAINT constraint_name]
DEFAULT {constant_expression | niladic-function | NULL}
[FOR col_name]
```

This example sets the default for the city field to ‘Vancouver’.

```
USE HOTEL
ALTER TABLE tblHotel
Add CONSTRAINT df_city DEFAULT ‘Vancouver’ FOR City
```

This page titled 4.2: Constraints is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

4.3: User Defined Types and Altering Tables

User Defined Types

User defined types are always based on system-supplied data type. They can enforce data integrity and they allow nulls.

To create a user-defined data type in SQL Server, choose types under “Programmability” in your database. Next, right click and choose ‘New’ –>‘User-defined data type’ or execute the sp_addtype system stored procedure. After this, type:

sp_addtype ssn, ‘varchar(11)’, ‘NOT NULL’

This will add a new user-defined data type called SIN with nine characters.

In this example, the field EmployeeSIN uses the user-defined data type SIN.

```
CREATE TABLE SINTable
(
EmployeeID INT Primary Key,
EmployeeSIN SIN,
CONSTRAINT CheckSIN
CHECK (EmployeeSIN LIKE
‘ [0-9][0-9][0-9] – [0-9][0-9] [0-9] – [0-9][0-9][0-9] ‘)
)
```

ALTER TABLE

You can use ALTER TABLE statements to add and drop constraints.

- ALTER TABLE allows columns to be removed.
- When a constraint is added, all existing data are verified for violations.

In this example, we use the ALTER TABLE statement to the IDENTITY property to a ColumnName field.

```
USE HOTEL
GO
ALTER TABLE tblHotel
ADD CONSTRAINT unqName UNIQUE (Name)
```

Use the ALTER TABLE statement to add a column with the IDENTITY property such as ALTER TABLE TableName.

```
ADD
ColumnName int IDENTITY(seed, increment)
```

DROP TABLE

The DROP TABLE will remove a table from the database. Make sure you have the correct database selected.

```
DROP TABLE tblHotel
```

Executing the above SQL DROP TABLE statement will remove the table tblHotel from the database.

This page titled 4.3: User Defined Types and Altering Tables is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

4.4: Key Terms and Exercises

Key Terms

DDL: abbreviation for *data definition language*

DML: abbreviation for *data manipulation language*

SEQUEL: acronym for *Structured English Query Language;* designed to manipulate and retrieve data stored in IBM's quasi-relational database management system, System R

Structured Query Language (SQL): a database language designed for managing data held in a relational database management system

Exercises

1. Using the information for the Chapter 9 exercise, implement the schema using Transact SQL (show SQL statements for each table). Implement the constraints as well.
2. Create the table shown here in SQL Server and show the statements you used.
   Table: Employee

ATTRIBUTE (FIELD) NAME	DATA DECLARATION
EMP_NUM	CHAR(3)
EMP_LNAME	VARCHAR(15)
EMP_FNAME	VARCHAR(15)
EMP_INITIAL	CHAR(1)
EMP_HIREDATE	DATE
JOB_CODE	CHAR(3)

3. Figure 15.2. Employee table with data for questions 4-10, by A. Watt.

   Use Figure 15.2 to answer questions 4 to 10.

4. Write the SQL code to change the job code to 501 for the person whose personnel number is 107. After you have completed the task, examine the results, and then reset the job code to its original value.
5. Assuming that the data shown in the Employee table have been entered, write the SQL code that lists all attributes for a job code of 502.
6. Write the SQL code to delete the row for the person named William Smithfield, who was hired on June 22, 2004, and whose job code classification is 500. (*Hint*: Use logical operators to include all the information given in this problem.)
7. Add the attributes EMP_PCT and PROJ_NUM to the Employee table. The EMP_PCT is the bonus percentage to be paid to each employee.
8. Using a single command, write the SQL code that will enter the project number (PROJ_NUM) = 18 for all employees whose job classification (JOB_CODE) is 500.
9. Using a single command, write the SQL code that will enter the project number (PROJ_NUM) = 25 for all employees whose job classification (JOB_CODE) is 502 or higher.
10. Write the SQL code that will change the PROJ_NUM to 14 for those employees who were hired before January 1, 1994, and whose job code is at least 501. (You may assume that the table will be restored to its original condition preceding this question.)

Also see *Appendix C: SQL Lab with Solution*

References

Date, C.J. *Relational Database Selected Writings*. Reading: Mass: Addison-Wesley Publishing Company Inc., 1986, p. 269-311.

This page titled 4.4: Key Terms and Exercises is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

5: SQL DATA MANIPULATION LANGUAGE

The SQL data manipulation language (DML) is used to query and modify database data. In this chapter, we will describe how to use the SELECT, INSERT, UPDATE, and DELETE SQL DML command statements, defined below.

- *SELECT* – to query data in the database
- *INSERT* – to insert data into a table
- *UPDATE* – to update data in a table
- *DELETE* – to delete data from a table

In the SQL DML statement:

- Each clause in a statement should begin on a new line.
- The beginning of each clause should line up with the beginning of other clauses.
- If a clause has several parts, they should appear on separate lines and be indented under the start of the clause to show the relationship.
- Upper case letters are used to represent reserved words.
- Lower case letters are used to represent user-defined words.

5.1: Using SELECT Statement

5.2: INSERT Statement

5.3: UPDATE and DELETE Statements

5.4: Built-In Functions

5.5: Joining Tables

5.6: Key Terms and Exercises

This page titled 5: SQL Data Manipulation Language is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

5.1: Using SELECT Statement

SELECT Statement

The SELECT statement, or command, allows the user to extract data from tables, based on specific criteria. It is processed according to the following sequence:

SELECT DISTINCT item(s)
FROM table(s)
WHERE predicate
GROUP BY field(s)
ORDER BY fields

We can use the SELECT statement to generate an employee phone list from the Employees table as follows:

SELECT FirstName, LastName, phone
FROM Employees
ORDER BY LastName

This action will display employee's last name, first name, and phone number from the Employees table, seen in Table 16.1.

Last Name	First Name	Phone Number
Hagans	Jim	604-232-3232
Wong	Bruce	604-244-2322

Table 16.1. Employees table.

In this next example, we will use a Publishers table (Table 16.2). (You will notice that Canada is mispelled in the *Publisher Country* field for Example Publishing and ABC Publishing. To correct mispelling, use the UPDATE statement to standardize the country field to Canada – see UPDATE statement later in this chapter.)

Publisher Name	Publisher City	Publisher Province	Publisher Country
Acme Publishing	Vancouver	BC	Canada
Example Publishing	Edmonton	AB	Cnada
ABC Publishing	Toronto	ON	Canda

Table 16.2. Publishers table.

If you add the publisher's name and city, you would use the SELECT statement followed by the fields name separated by a comma:

SELECT PubName, city
FROM Publishers

This action will display the publisher's name and city from the Publishers table.

If you just want the publisher's name under the display name city, you would use the SELECT statement with *no comma* separating pub_name and city:

SELECT PubName city
FROM Publishers

Performing this action will display only the pub_name from the Publishers table with a "city" heading. If you do not include the

comma, SQL Server assumes you want a new column name for pub_name.

SELECT statement with WHERE criteria

Sometimes you might want to focus on a portion of the Publishers table, such as only publishers that are in Vancouver. In this situation, you would use the SELECT statement with the WHERE criterion, i.e., WHERE city = 'Vancouver'.

These first two examples illustrate how to limit record selection with the WHERE criterion using BETWEEN. Each of these examples give the same results for store items with between 20 and 50 items in stock.

Example #1 uses the quantity, *qty BETWEEN 20 and 50*.

```
SELECT StorID, qty, TitleID
FROM Sales
WHERE qty BETWEEN 20 and 50 (includes the 20 and 50)
```

Example #2, on the other hand, uses *qty >=20 and qty <=50* .

```
SELECT StorID, qty, TitleID
FROM Sales
WHERE qty >= 20 and qty <= 50
```

Example #3 illustrates how to limit record selection with the WHERE criterion using NOT BETWEEN.

```
SELECT StorID, qty, TitleID
FROM Sales
WHERE qty NOT BETWEEN 20 and 50
```

The next two examples show two different ways to limit record selection with the WHERE criterion using IN, with each yielding the same results.

Example #4 shows how to select records using *province*= as part of the WHERE statement.

```
SELECT *
FROM Publishers
WHERE province = 'BC' OR province = 'AB' OR province = 'ON'
```

Example #5 select records using *province IN* as part of the WHERE statement.

```
SELECT *
FROM Publishers
WHERE province IN ('BC', 'AB', 'ON')
```

The final two examples illustrate how NULL and NOT NULL can be used to select records. For these examples, a Books table (not shown) would be used that contains fields called Title, Quantity, and Price (of book). Each publisher has a Books table that lists all of its books.

Example #6 uses NULL.

```
SELECT price, title
FROM Books
WHERE price IS NULL
```

Example #7 uses NOT NULL.

```
SELECT price, title
FROM Books
WHERE price IS NOT NULL
```

Using wildcards in the LIKE clause

The LIKE keyword selects rows containing fields that match specified portions of character strings. LIKE is used with char, varchar, text, datetime and smalldatetime data. A *wildcard* allows the user to match fields that contain certain letters. For example, the wildcard province = 'N%' would give all provinces that start with the letter 'N'. Table 16.3 shows four ways to specify wildcards in the SELECT statement in regular express format.

%	Any string of zero or more characters
_	Any single character
[ ]	Any single character within the specified range (e.g., [a-f]) or set (e.g., [abcdef])
[^]	Any single character not within the specified range (e.g., [^a – f]) or set (e.g., [^abcdef])

Table 16.3. How to specify wildcards in the SELECT statement.

In example #1, LIKE 'Mc%' searches for all last names that begin with the letters "Mc" (e.g., McBadden).

```
SELECT LastName
FROM Employees
WHERE LastName LIKE 'Mc%'
```

For example #2: LIKE '%inger' searches for all last names that end with the letters "inger" (e.g., Ringer, Stringer).

```
SELECT LastName
FROM Employees
WHERE LastName LIKE '%inger'
```

In, example #3: LIKE '%en%' searches for all last names that have the letters "en" (e.g., Bennett, Green, McBadden).

```
SELECT LastName
FROM Employees
WHERE LastName LIKE '%en%'
```

SELECT statement with ORDER BY clause

You use the ORDER BY clause to sort the records in the resulting list. Use *ASC* to sort the results in ascending order and *DESC* to sort the results in descending order.

For example, with ASC:

```
SELECT *
FROM Employees
ORDER BY HireDate ASC
```

And with DESC:

```
SELECT *
FROM Books
ORDER BY type, price DESC
```

SELECT statement with GROUP BY clause

The *GROUP BY* clause is used to create one output row per each group and produces summary values for the selected columns, as shown below.

```
SELECT type
FROM Books
GROUP BY type
```

Here is an example using the above statement.

```
SELECT type AS 'Type', MIN(price) AS 'Minimum Price'
FROM Books
WHERE royalty > 10
GROUP BY type
```

If the SELECT statement includes a WHERE criterion where *price is not null*,

```
SELECT type, price
FROM Books
WHERE price is not null
```

then a statement with the GROUP BY clause would look like this:

```
SELECT type AS 'Type', MIN(price) AS 'Minimum Price'
FROM Books
WHERE price is not null
GROUP BY type
```

Using COUNT with GROUP BY

We can use COUNT to tally how many items are in a container. However, if we want to count different items into separate groups, such as marbles of varying colours, then we would use the COUNT function with the GROUP BY command.

The below SELECT statement illustrates how to count groups of data using the COUNT function with the GROUP BY clause.

```
SELECT COUNT(*)
FROM Books
GROUP BY type
```

Using AVG and SUM with GROUP BY

We can use the AVG function to give us the average of any group, and SUM to give the total.

Example #1 uses the AVG FUNCTION with the GROUP BY type.

```
SELECT AVG(qty)
FROM Books
GROUP BY type
```

Example #2 uses the SUM function with the GROUP BY type.

```
SELECT SUM(qty)
FROM Books
GROUP BY type
```

Example #3 uses both the AVG and SUM functions with the GROUP BY type in the SELECT statement.

```
SELECT 'Total Sales' = SUM(qty), 'Average Sales' = AVG(qty), stor_id
FROM Sales
GROUP BY StorID ORDER BY 'Total Sales'
```

Restricting rows with HAVING

The HAVING clause can be used to restrict rows. It is similar to the WHERE condition except HAVING can include the aggregate function; the WHERE cannot do this.

The HAVING clause behaves like the WHERE clause, but is applicable to groups. In this example, we use the HAVING clause to exclude the groups with the province 'BC'.

```
SELECT au_fname AS 'Author"s First Name', province as 'Province'
FROM Authors
GROUP BY au_fname, province
HAVING province <> 'BC'
```

This page titled 5.1: Using SELECT Statement is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

5.2: INSERT Statement

INSERT statement

The *INSERT statement* adds rows to a table. In addition,

- INSERT specifies the table or view that data will be inserted into.
- Column_list lists columns that will be affected by the INSERT.
- If a column is omitted, each value must be provided.
- If you are including columns, they can be listed in any order.
- VALUES specifies the data that you want to insert into the table. VALUES is required.
- Columns with the IDENTITY property should not be explicitly listed in the column_list or values_clause.

The syntax for the INSERT statement is:

INSERT [INTO] Table_name | view name [column_list]
DEFAULT VALUES | values_list | select statement

When inserting rows with the INSERT statement, these rules apply:

- Inserting an empty string (‘ ‘) into a varchar or text column inserts a single space.
- All char columns are right-padded to the defined length.
- All trailing spaces are removed from data inserted into varchar columns, except in strings that contain only spaces. These strings are truncated to a single space.
- If an INSERT statement violates a constraint, default or rule, or if it is the wrong data type, the statement fails and SQL Server displays an error message.

When you specify values for only some of the columns in the column_list, one of three things can happen to the columns that have no values:

1. A default value is entered if the column has a DEFAULT constraint, if a default is bound to the column, or if a default is bound to the underlying user-defined data type.
2. NULL is entered if the column allows NULLs and no default value exists for the column.
3. An error message is displayed and the row is rejected if the column is defined as NOT NULL and no default exists.

This example uses INSERT to add a record to the publisher’s Authors table.

INSERT INTO Authors
VALUES(‘555-0993-467’, ‘Martin’, ‘April’, ‘281 555-5673’, ‘816 Market St.,’ , ‘Vancouver’, ‘BC’, ‘V7G3P4’, 0)

This following example illustrates how to insert a partial row into the Publishers table with a column list. The country column had a default value of Canada so it does not require that you include it in your values.

INSERT INTO Publishers (PubID, PubName, city, province)
VALUES (‘9900’, ‘Acme Publishing’, ‘Vancouver’, ‘BC’)

To insert rows into a table with an IDENTITY column, follow the below example. Do not supply the value for the IDENTITY nor the name of the column in the column list.

INSERT INTO jobs
VALUES (‘DBA’, 100, 175)

Inserting specific values into an IDENTITY column

By default, data cannot be inserted directly into an IDENTITY column; however, if a row is accidentally deleted, or there are gaps

in the IDENTITY column values, you can insert a row and specify the IDENTITY column value.

IDENTITY_INSERT option

To allow an insert with a specific identity value, the IDENTITY_INSERT option can be used as follows.

```
SET IDENTITY_INSERT jobs ON
INSERT INTO jobs (job_id, job_desc, min_lvl, max_lvl)
VALUES (19, 'DBA2', 100, 175)
SET IDENTITY_INSERT jobs OFF
```

Inserting rows with a SELECT statement

We can sometimes create a small temporary table from a large table. For this, we can insert rows with a SELECT statement. When using this command, there is no validation for uniqueness. Consequently, there may be many rows with the same pub_id in the example below.

This example creates a smaller temporary Publishers table using the CREATE TABLE statement. Then the INSERT with a SELECT statement is used to add records to this temporary Publishers table from the publis table.

```
CREATE TABLE dbo.tmpPublishers (
PubID char (4) NOT NULL ,
PubName varchar (40) NULL ,
city varchar (20) NULL ,
province char (2) NULL ,
country varchar (30) NULL DEFAULT ('Canada')
)
INSERT tmpPublishers
SELECT * FROM Publishers
```

In this example, we're copying a subset of data.

```
INSERT tmpPublishers (pub_id, pub_name)
SELECT PubID, PubName
FROM Publishers
```

In this example, the publishers' data are copied to the tmpPublishers table and the country column is set to Canada.

```
INSERT tmpPublishers (PubID, PubName, city, province, country)
SELECT PubID, PubName, city, province, 'Canada'
FROM Publishers
```

This page titled 5.2: INSERT Statement is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

5.3: UPDATE and DELETE Statements

UPDATE statement

The *UPDATE statement* changes data in existing rows either by adding new data or modifying existing data.

This example uses the UPDATE statement to standardize the country field to be Canada for all records in the Publishers table.

```
UPDATE Publishers
SET country = 'Canada'
```

This example increases the royalty amount by 10% for those royalty amounts between 10 and 20.

```
UPDATE roysched
SET royalty = royalty + (royalty * .10)
WHERE royalty BETWEEN 10 and 20
```

Including subqueries in an UPDATE statement

The employees from the Employees table who were hired by the publisher in 2010 are given a promotion to the highest job level for their job type. This is what the UPDATE statement would look like.

```
UPDATE Employees
SET job_lvl =
(SELECT max_lvl FROM jobs
WHERE employee.job_id = jobs.job_id)
WHERE DATEPART(year, employee.hire_date) = 2010
```

DELETE statement

The *DELETE statement* removes rows from a record set. DELETE names the table or view that holds the rows that will be deleted and only one table or row may be listed at a time. WHERE is a standard WHERE clause that limits the deletion to select records.

The DELETE syntax looks like this.

```
DELETE [FROM] {table_name | view_name }
[WHERE clause]
```

The rules for the DELETE statement are:

1. If you omit a WHERE clause, all rows in the table are removed (except for indexes, the table, constraints).
2. DELETE cannot be used with a view that has a FROM clause naming more than one table. (Delete can affect only one base table at a time.)

What follows are three different DELETE statements that can be used.

1. Deleting all rows from a table.

```
DELETE
FROM Discounts
```

2. Deleting selected rows:

```
DELETE
FROM Sales
```

WHERE stor_id = '6380'

3. Deleting rows based on a value in a subquery:

DELETE FROM Sales
WHERE title_id IN
(SELECT title_id FROM Books WHERE type = 'mod_cook')

This page titled 5.3: UPDATE and DELETE Statements is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

5.4: Built-In Functions

Built-in Functions

There are many built-in functions in SQL Server such as:

1. *Aggregate:* returns summary values
2. *Conversion:* transforms one data type to another
3. *Date:* displays information about dates and times
4. *Mathematical:* performs operations on numeric data
5. *String:* performs operations on character strings, binary data or expressions
6. *System:* returns a special piece of information from the database
7. *Text and image:* performs operations on text and image data

Below you will find detailed descriptions and examples for the first four functions.

Aggregate functions

Aggregate functions perform a calculation on a set of values and return a single, or summary, value. Table 16.4 lists these functions.

FUNCTION	DESCRIPTION
AVG	Returns the average of all the values, or only the DISTINCT values, in the expression.
COUNT	Returns the number of non-null values in the expression. When DISTINCT is specified, COUNT finds the number of unique non-null values.
COUNT(*)	Returns the number of rows. COUNT(*) takes no parameters and cannot be used with DISTINCT.
MAX	Returns the maximum value in the expression. MAX can be used with numeric, character and datetime columns, but not with bit columns. With character columns, MAX finds the highest value in the collating sequence. MAX ignores any null values.
MIN	Returns the minimum value in the expression. MIN can be used with numeric, character and datetime columns, but not with bit columns. With character columns, MIN finds the value that is lowest in the sort sequence. MIN ignores any null values.
SUM	Returns the sum of all the values, or only the DISTINCT values, in the expression. SUM can be used with numeric columns only.

Table 16.4 A list of aggregate functions and descriptions.

Below are examples of each of the aggregate functions listed in Table 16.4.

Example #1: AVG

SELECT AVG (price) AS ‘Average Title Price’
FROM Books

Example #2: COUNT

SELECT COUNT(PubID) AS ‘Number of Publishers’
FROM Publishers

Example #3: COUNT

```
SELECT COUNT(province) AS 'Number of Publishers'
FROM Publishers
```

Example #3: COUNT (*)

```
SELECT COUNT(*)
FROM Employees
WHERE job_lvl = 35
```

Example #4: MAX

```
SELECT MAX (HireDate)
FROM Employees
```

Example #5: MIN

```
SELECT MIN (price)
FROM Books
```

Example #6: SUM

```
SELECT SUM(discount) AS 'Total Discounts'
FROM Discounts
```

Conversion function

The conversion function transforms one data type to another.

In the example below, a price that contains two 9s is converted into five characters. The syntax for this statement is SELECT 'The date is ' + CONVERT(varchar(12), getdate()).

```
SELECT CONVERT(int, 10.6496)
SELECT title_id, price
FROM Books
WHERE CONVERT(char(5), price) LIKE '%99%'
```

In this second example, the conversion function changes data to a data type with a different size.

```
SELECT title_id, CONVERT(char(4), ytd_sales) as 'Sales'
FROM Books
WHERE type LIKE '%cook'
```

Date function

The date function produces a date by adding an interval to a specified date. The result is a datetime value equal to the date plus the number of date parts. If the date parameter is a smalldatetime value, the result is also a smalldatetime value.

The DATEADD function is used to add and increment date values. The syntax for this function is DATEADD(datepart, number, date).

```
SELECT DATEADD(day, 3, hire_date)
FROM Employees
```

In this example, the function DATEDIFF(datepart, date1, date2) is used.

This command returns the number of datepart "boundaries" crossed between two specified dates. The method of counting crossed boundaries makes the result given by DATEDIFF consistent across all data types such as minutes, seconds, and

milliseconds.

SELECT DATEDIFF(day, HireDate, ‘Nov 30 1995’)
FROM Employees

For any particular date, we can examine any part of that date from the year to the millisecond.

The date parts (DATEPART) and abbreviations recognized by SQL Server, and the acceptable values are listed in Table 16.5.

DATE PART	ABBREVIATION	VALUES
Year	yy	1753-9999
Quarter	qq	1-4
Month	mm	1-12
Day of year	dy	1-366
Day	dd	1-31
Week	wk	1-53
Weekday	dw	1-7 (Sun.-Sat.)
Hour	hh	0-23
Minute	mi	0-59
Second	ss	0-59
Millisecond	ms	0-999

Table 16.5. Date part abbreviations and values.

Mathematical functions

Mathematical functions perform operations on numeric data. The following example lists the current price for each book sold by the publisher and what they would be if all prices increased by 10%.

SELECT Price, (price * 1.1) AS ‘New Price’, title
FROM Books
SELECT ‘Square Root’ = SQRT(81)
SELECT ‘Rounded‘ = ROUND(4567.9876,2)
SELECT FLOOR (123.45)

This page titled 5.4: Built-In Functions is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

5.5: Joining Tables

Joining Tables

Joining two or more tables is the process of comparing the data in specified columns and using the comparison results to form a new table from the rows that qualify. A join statement:

- Specifies a column from each table
- Compares the values in those columns row by row
- Combines rows with qualifying values into a new row

Although the comparison is usually for equality – values that match exactly – other types of joins can also be specified. All the different joins such as inner, left (outer), right (outer), and cross join will be described below.

Inner join

An *inner join* connects two tables on a column with the same data type. Only the rows where the column values match are returned; unmatched rows are discarded.

Example #1

```
SELECT jobs.job_id, job_desc
FROM jobs
INNER JOIN Employees ON employee.job_id = jobs.job_id
WHERE jobs.job_id < 7
```

Example #2

```
SELECT authors.au_fname, authors.au_lname, books.royalty, title
FROM authorsINNER JOIN titleauthor ON authors.au_id=titleauthor.au_id
INNER JOIN books ON titleauthor.title_id=books.title_id
GROUP BY authors.au_lname, authors.au_fname, title, title.royalty
ORDER BY authors.au_lname
```

Left outer join

A *left outer join* specifies that all left outer rows be returned. All rows from the left table that did not meet the condition specified are included in the results set, and output columns from the other table are set to NULL.

This first example uses the new syntax for a left outer join.

```
SELECT publishers.pub_name, books.title
FROM Publishers
LEFT OUTER JOIN Books On publishers.pub_id = books.pub_id
```

This is an example of a left outer join using the old syntax.

```
SELECT publishers.pub_name, books.title
FROM Publishers, Books
WHERE publishers.pub_id *= books.pub_id
```

Right outer join

A *right outer join* includes, in its result set, all rows from the right table that did not meet the condition specified. Output columns

that correspond to the other table are set to NULL.

Below is an example using the new syntax for a right outer join.

```
SELECT titleauthor.title_id, authors.au_lname, authors.au_fname
FROM titleauthor
RIGHT OUTER JOIN authors ON titleauthor.au_id = authors.au_id
ORDERY BY au_lname
```

This second example show the old syntax used for a right outer join.

```
SELECT titleauthor.title_id, authors.au_lname, authors.au_fname
FROM titleauthor, authors
WHERE titleauthor.au_id =* authors.au_id
ORDERY BY au_lname
```

Full outer join

A *full outer join* specifies that if a row from either table does not match the selection criteria, the row is included in the result set, and its output columns that correspond to the other table are set to NULL.

Here is an example of a full outer join.

```
SELECT books.title, publishers.pub_name, publishers.province
FROM Publishers
FULL OUTER JOIN Books ON books.pub_id = publishers.pub_id
WHERE (publishers.province <> “BC” and publishers.province <> “ON”)
ORDER BY books.title_id
```

Cross join

A *cross join* is a product combining two tables. This join returns the same rows as if no WHERE clause were specified. For example:

```
SELECT au_lname, pub_name,
FROM Authors CROSS JOIN Publishers
```

This page titled 5.5: Joining Tables is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

5.6: Key Terms and Exercises

Key Terms

aggregate function: returns summary values

ASC: ascending order

conversion function: transforms one data type to another

cross join: a product combining two tables

date function: displays information about dates and times

DELETE statement: removes rows from a record set

DESC: descending order

full outer join: specifies that if a row from either table does not match the selection criteria

GROUP BY: used to create one output row per each group and produces summary values for the selected columns

inner join: connects two tables on a column with the same data type

INSERT statement: adds rows to a table

left outer join: specifies that all left outer rows be returned

mathematical function: performs operations on numeric data

right outer join: includes all rows from the right table that did not meet the condition specified

SELECT statement: used to query data in the database

string function: performs operations on character strings, binary data or expressions

system function: returns a special piece of information from the database

text and image functions: performs operations on text and image data

UPDATE statement: changes data in existing rows either by adding new data or modifying existing data

wildcard: allows the user to match fields that contain certain letters.

Exercises

For questions 1 to 18 use the PUBS sample database created by Microsoft. To download the script to generate this database please go to the following site: www.microsoft.com/en-ca/downl....aspx?id=23654.

1. Display a list of publication dates and titles (books) that were published in 2011.
2. Display a list of titles that have been categorized as either traditional or modern cooking. Use the Books table.
3. Display all authors whose first names are five letters long.
4. Display from the Books table: type, price, pub_id, title about the books put out by each publisher. Rename the column type with "Book Category." Sort by type (descending) and then price (ascending).
5. Display title_id, pubdate and pubdate plus three days, using the Books table.
6. Using the datediff and getdate function determine how much time has elapsed in months since the books in the Books table

were published.
7. List the title IDs and quantity of all books that sold more than 30 copies.
8. Display a list of all last names of the authors who live in Ontario (ON) and the cities where they live.
9. Display all rows that contain a 60 in the payterms field. Use the Sales table.
10. Display all authors whose first names are five letters long , end in O or A, and start with M or P.
11. Display all titles that cost more than $30 and either begin with T or have a publisher ID of 0877.
12. Display from the Employees table the first name (fname), last name (lname), employe ID(emp_id) and job level (job_lvl) columns for those employees with a job level greater than 200; and rename the column headings to: “First Name,” “Last Name,” “IDENTIFICATION#” and “Job Level.”
13. Display the royalty, royalty plus 50% as “royalty plus 50” and title_id. Use the Roysched table.
14. Using the STUFF function create a string “12xxxx567” from the string “1234567.”
15. Display the first 40 characters of each title, along with the average monthly sales for that title to date (ytd_sales/12). Use the Title table.
16. Show how many books have assigned prices.
17. Display a list of cookbooks with the average cost for all of the books of each type. Use the GROUP BY.

Advanced Questions (Union, Intersect, and Minus)

1. The relational set operators UNION, INTERSECT and MINUS work properly only if the relations are union-compatible. What does union-compatible mean, and how would you check for this condition?
2. What is the difference between UNION and UNION ALL? Write the syntax for each.
3. Suppose that you have two tables, Employees and Employees_1. The Employees table contains the records for three employees: Alice Cordoza, John Cretchakov, and Anne McDonald. The Employees_1 table contains the records for employees: John Cretchakov and Mary Chen. Given that information, what is the query output for the UNION query? List the query output.
4. Given the employee information in question 3, what is the query output for the UNION ALL query? List the query output.
5. Given the employee information in question 3, what is the query output for the INTERSECT query? List the query output.
6. Given the employee information in question 3, what is the query output for the EXCEPT query? List the query output.
7. What is a cross join? Give an example of its syntax.
8. Explain these three join types:
    1. left outer join
    2. right outer join
    3. full outer join
9. What is a subquery, and what are its basic characteristics?
10. What is a correlated subquery? Give an example.
11. Suppose that a Product table contains two attributes, PROD_CODE and VEND_CODE. The values for the PROD_CODE are: ABC, DEF, GHI and JKL. These are matched by the following values for the VEND_CODE: 125, 124, 124 and 123, respectively (e.g., PROD_CODE value ABC corresponds to VEND_CODE value 125). The Vendor table contains a single attribute, VEND_CODE, with values 123, 124, 125 and 126. (The VEND_CODE attribute in the Product table is a foreign key to the VEND_CODE in the Vendor table.)
12. Given the information in question 11, what would be the query output for the following? Show values.
    1. A UNION query based on these two tables
    2. A UNION ALL query based on these two tables
    3. An INTERSECT query based on these two tables
    4. A MINUS query based on these two tables

Advanced Questions (Using Joins)

1. Display a list of all titles and sales numbers in the Books and Sales tables, including titles that have no sales. Use a join.
2. Display a list of authors’ last names and all associated titles that each author has published sorted by the author’s last name. Use a join. Save it as a view named: Published Authors.
3. Using a subquery, display all the authors (show last and first name, postal code) who receive a royalty of 100% and live in Alberta. Save it as a view titled: AuthorsView. When creating the view, rename the author’s last name and first name as ‘Last Name’ and ‘First Name’.
4. Display the stores that did not sell the title *Is Anger the Enemy?*
5. Display a list of store names for sales after 2013 (Order Date is greater than 2013). Display store name and order date.
6. Display a list of titles for books sold in store name “News & Brews.” Display store name, titles and order dates.
7. List total sales (qty) by title. Display total quantity and title columns.

8. List total sales (qty) by type. Display total quantity and type columns.
9. List total sales (qty*price) by type. Display total dollar value and type columns.
10. Calculate the total number of types of books by publisher. Show publisher name and total count of types of books for each publisher.
11. Show publisher names that do not have any type of book. Display publisher name only.

This page titled 5.6: Key Terms and Exercises is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

6: THE RELATIONAL DATA MODEL

author: Adrienne Watt

The relational data model was introduced by C. F. Codd in 1970. Currently, it is the most widely used data model.

The relational model has provided the basis for:

- Research on the theory of data/relationship/constraint
- Numerous database design methodologies
- The standard database access language called s*tructured query language (SQL)*
- Almost all modern commercial database management systems

The relational data model describes the world as “a collection of inter-related relations (or tables).”

Fundamental Concepts in the Relational Data Model

Relation

A *relation,* also known as a *table* or *file*, is a subset of the Cartesian product of a list of domains characterized by a name. And within a table, each row represents a group of related data values. A *row*, or record, is also known as a *tuple*. The columns in a table is a field and is also referred to as an attribute. You can also think of it this way: an attribute is used to define the record and a record contains a set of attributes.

The steps below outline the logic between a relation and its domains.

1. Given *n* domains are denoted by D1, D2, ... Dn
2. And *r* is a relation defined on these domains
3. Then r ⊆ D1×D2×...×Dn

Table

A database is composed of multiple tables and each table holds the data. Figure 7.1 shows a database that contains three tables.

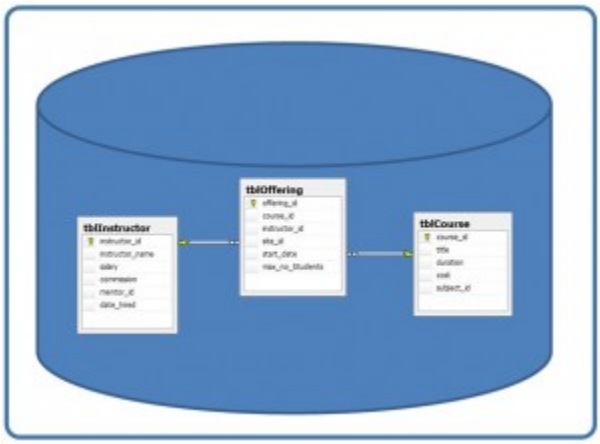


Figure 7.1. Database with three tables.

Column

A database stores pieces of information or facts in an organized way. Understanding how to use and get the most out of databases

requires us to understand that method of organization.

The principal storage units are called *columns* or *fields* or *attributes*. These house the basic components of data into which your content can be broken down. When deciding which fields to create, you need to think generically about your information, for example, drawing out the common components of the information that you will store in the database and avoiding the specifics that distinguish one item from another.

Look at the example of an ID card in Figure 7.2 to see the relationship between fields and their data.

Field Name	Data
First Name	Isabelle
Family Name	Whelan
Nationality	British
Salary	109,900
Date of Birth	15 September 1983
Marital Status	Single
Shift	Mon, Wed
Place of issue	Addis Ababa
Valid until	17 December 2003

Figure 7.2. Example of an ID card by A. Watt.

Domain

A *domain* is the original sets of atomic values used to model data. By *atomic value*, we mean that each value in the domain is indivisible as far as the relational model is concerned. For example:

- The domain of Marital Status has a set of possibilities: Married, Single, Divorced.
- The domain of Shift has the set of all possible days: {Mon, Tue, Wed...}.
- The domain of Salary is the set of all floating-point numbers greater than 0 and less than 200,000.
- The domain of First Name is the set of character strings that represents names of people.

In summary, a domain is a set of acceptable values that a column is allowed to contain. This is based on various properties and the data type for the column. We will discuss data types in another chapter.

Records

Just as the content of any one document or item needs to be broken down into its constituent bits of data for storage in the fields, the link between them also needs to be available so that they can be reconstituted into their whole form. Records allow us to do this. *Records* contain fields that are related, such as a customer or an employee. As noted earlier, a tuple is another term used for record.

Records and fields form the basis of all databases. A simple table gives us the clearest picture of how records and fields work together in a database storage project.

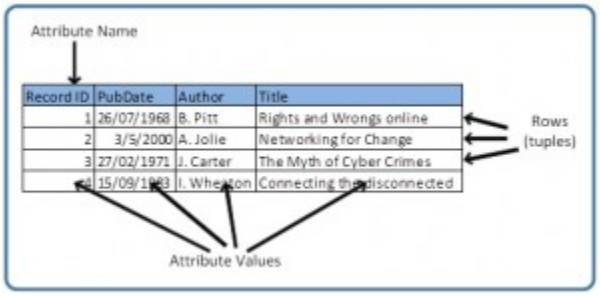


Figure 7.3. Example of a simple table by A. Watt.

The simple table example in Figure 7.3 shows us how fields can hold a range of different sorts of data. This one has:

- A Record ID field: this is an ordinal number; its data type is an integer.
- A PubDate field: this is displayed as day/month/year; its data type is date.
- An Author field: this is displayed as Initial. Surname; its data type is text.
- A Title field text: free text can be entered here.

You can command the database to sift through its data and organize it in a particular way. For example, you can request that a selection of records be limited by date: 1. all before a given date, 2. all after a given date or 3. all between two given dates. Similarly, you can choose to have records sorted by date. Because the field, or record, containing the data is set up as a Date field, the database reads the information in the Date field not just as numbers separated by slashes, but rather, as dates that must be ordered according to a calendar system.

Degree

The *degree* is the number of attributes in a table. In our example in Figure 7.3, the degree is 4.

Properties of a Table

- A table has a name that is distinct from all other tables in the database.
- There are no duplicate rows; each row is distinct.
- Entries in columns are atomic. The table does not contain repeating groups or multivalued attributes.
- Entries from columns are from the same domain based on their data type including:
  - number (numeric, integer, float, smallint,...)
  - character (string)
  - date
  - logical (true or false)
- Operations combining different data types are disallowed.
- Each attribute has a distinct name.
- The sequence of columns is insignificant.
- The sequence of rows is insignificant.

Key Terms

atomic value: each value in the domain is indivisible as far as the relational model is concerned

attribute: principle storage unit in a database

column: see *attribute*

degree: number of attributes in a table

domain: the original sets of atomic values used to model data; a set of acceptable values that a column is allowed to contain

field: see *attribute*

file: see *relation*

record: contains fields that are related; *see tuple*

relation: a subset of the Cartesian product of a list of domains characterized by a name; the technical term for table or file

row: *see tuple*

structured query language (SQL): the standard database access language

table: see *relation*

tuple: a technical term for row or record

Terminology Key

Several of the terms used in this chapter are synonymous. In addition to the Key Terms above, please refer to Table 7.1 below. The terms in the Alternative 1 column are most commonly used.

Formal Terms (Codd)	Alternative 1	Alternative 2
Relation	Table	File
Tuple	Row	Record
Attribute	Column	Field

Table 7.1. Terms and their synonyms by A. Watt.

6.1: Fundamental Concepts in the Relational Data Model

6.2: Key Terms

This page titled 6: The Relational Data Model is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

6.1: Fundamental Concepts in the Relational Data Model

Fundamental Concepts in the Relational Data Model

Relation

A *relation,* also known as a *table* or *file*, is a subset of the Cartesian product of a list of domains characterized by a name. And within a table, each row represents a group of related data values. A *row*, or record, is also known as a *tuple*. The columns in a table is a field and is also referred to as an attribute. You can also think of it this way: an attribute is used to define the record and a record contains a set of attributes.

The steps below outline the logic between a relation and its domains.

1. Given *n* domains are denoted by D1, D2, ... Dn
2. And *r* is a relation defined on these domains
3. Then r $\subseteq$ D1×D2×...×Dn

Table

A database is composed of multiple tables and each table holds the data. Figure 7.1 shows a database that contains three tables.

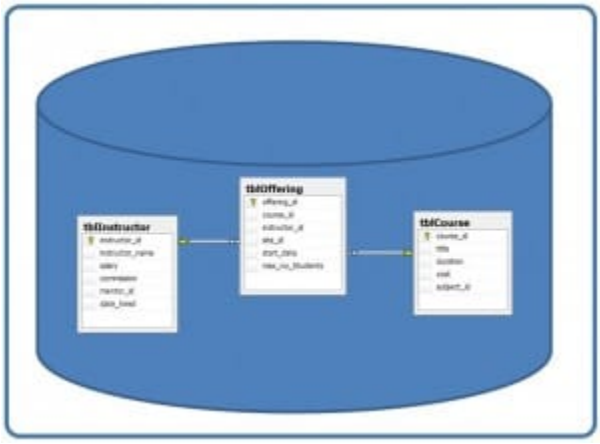


Figure 7.1. Database with three tables.

Column

A database stores pieces of information or facts in an organized way. Understanding how to use and get the most out of databases requires us to understand that method of organization.

The principal storage units are called *columns* or *fields* or *attributes*. These house the basic components of data into which your content can be broken down. When deciding which fields to create, you need to think generically about your information, for example, drawing out the common components of the information that you will store in the database and avoiding the specifics that distinguish one item from another.

Look at the example of an ID card in Figure 7.2 to see the relationship between fields and their data.

Field Name	Data
First Name	Isabelle
Family Name	Whelan
Nationality	British
Salary	109,900
Date of Birth	15 September 1983
Marital Status	Single
Shift	Mon, Wed
Place of issue	Addis Ababa
Valid until	17 December 2003

Figure 7.2. Example of an ID card by A. Watt.

Domain

A *domain* is the original sets of atomic values used to model data. By *atomic value*, we mean that each value in the domain is indivisible as far as the relational model is concerned. For example:

- The domain of Marital Status has a set of possibilities: Married, Single, Divorced.
- The domain of Shift has the set of all possible days: {Mon, Tue, Wed...}.
- The domain of Salary is the set of all floating-point numbers greater than 0 and less than 200,000.
- The domain of First Name is the set of character strings that represents names of people.

In summary, a domain is a set of acceptable values that a column is allowed to contain. This is based on various properties and the data type for the column. We will discuss data types in another chapter.

Records

Just as the content of any one document or item needs to be broken down into its constituent bits of data for storage in the fields, the link between them also needs to be available so that they can be reconstituted into their whole form. Records allow us to do this. *Record*s contain fields that are related, such as a customer or an employee. As noted earlier, a tuple is another term used for record.

Records and fields form the basis of all databases. A simple table gives us the clearest picture of how records and fields work together in a database storage project.

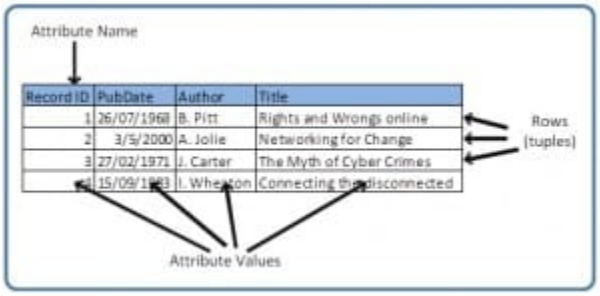


Figure 7.3. Example of a simple table by A. Watt.

The simple table example in Figure 7.3 shows us how fields can hold a range of different sorts of data. This one has:

- A Record ID field: this is an ordinal number; its data type is an integer.
- A PubDate field: this is displayed as day/month/year; its data type is date.
- An Author field: this is displayed as Initial. Surname; its data type is text.
- A Title field text: free text can be entered here.

You can command the database to sift through its data and organize it in a particular way. For example, you can request that a selection of records be limited by date: 1. all before a given date, 2. all after a given date or 3. all between two given dates.

Similarly, you can choose to have records sorted by date. Because the field, or record, containing the data is set up as a Date field, the database reads the information in the Date field not just as numbers separated by slashes, but rather, as dates that must be ordered according to a calendar system.

Degree

The *degree* is the number of attributes in a table. In our example in Figure 7.3, the degree is 4.

Properties of a Table

- A table has a name that is distinct from all other tables in the database.
- There are no duplicate rows; each row is distinct.
- Entries in columns are atomic. The table does not contain repeating groups or multivalued attributes.
- Entries from columns are from the same domain based on their data type including:
  - number (numeric, integer, float, smallint,...)
  - character (string)
  - date
  - logical (true or false)
- Operations combining different data types are disallowed.
- Each attribute has a distinct name.
- The sequence of columns is insignificant.
- The sequence of rows is insignificant.

This page titled 6.1: Fundamental Concepts in the Relational Data Model is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

6.2: Key Terms

Key Terms

atomic value: each value in the domain is indivisible as far as the relational model is concerned

attribute: principle storage unit in a database

column: see *attribute*

degree: number of attributes in a table

domain: the original sets of atomic values used to model data; a set of acceptable values that a column is allowed to contain

field: see *attribute*

file: see *relation*

record: contains fields that are related; *see tuple*

relation: a subset of the Cartesian product of a list of domains characterized by a name; the technical term for table or file

row: *see tuple*

structured query language (SQL): the standard database access language

table: see *relation*

tuple: a technical term for row or record

Terminology Key

Several of the terms used in this chapter are synonymous. In addition to the Key Terms above, please refer to Table 7.1 below. The terms in the Alternative 1 column are most commonly used.

Formal Terms (Codd)	Alternative 1	Alternative 2
Relation	Table	File
Tuple	Row	Record
Attribute	Column	Field

Table 7.1. Terms and their synonyms by A. Watt.

This page titled 6.2: Key Terms is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

7: THE ENTITY RELATIONSHIP DATA MODEL

The *entity relationship (ER) data model* has existed for over 35 years. It is well suited to data modelling for use with databases because it is fairly abstract and is easy to discuss and explain. ER models are readily translated to relations. ER models, also called an ER schema, are represented by ER diagrams.

ER modelling is based on two concepts:

- Entities, defined as tables that hold specific information (data)
- *Relationships,* defined as the associations or interactions between entities

Here is an example of how these two concepts might be combined in an ER data model: Prof. Ba (entity)teaches (relationship)the Database Systems course (entity).

For the rest of this chapter, we will use a sample database called the COMPANY database to illustrate the concepts of the ER model. This database contains information about employees, departments and projects. Important points to note include:

- There are several departments in the company. Each department has a unique identification, a name, location of the office and a particular employee who manages the department.
- A department controls a number of projects, each of which has a unique name, a unique number and a budget.
- Each employee has a name, identification number, address, salary and birthdate. An employee is assigned to one department but can join in several projects. We need to record the start date of the employee in each project. We also need to know the direct supervisor of each employee.
- We want to keep track of the dependents for each employee. Each dependent has a name, birthdate and relationship with the employee.

Entity, Entity Set and Entity Type

An *entity* is an object in the real world with an independent existence that can be differentiated from other objects. An entity might be

- An object with physical existence (e.g., a lecturer, a student, a car)
- An object with conceptual existence (e.g., a course, a job, a position)

Entities can be classified based on their strength. An entity is considered weak if its tables are existence dependent.

- That is, it cannot exist without a relationship with another entity
- Its primary key is derived from the primary key of the parent entity
    - The Spouse table, in the COMPANY database, is a weak entity because its primary key is dependent on the Employee table. Without a corresponding employee record, the spouse record would not exist.

An entity is considered strong if it can exist apart from all of its related entities.

- Kernels are strong entities.
- A table without a foreign key or a table that contains a foreign key that can contain nulls is a strong entity

Another term to know is *entity type* which defines a collection of similar entities.

An *entity set* is a collection of entities of an entity type at a particular point of time. In an entity relationship diagram (ERD), an entity type is represented by a name in a box. For example, in Figure 8.1, the entity type is EMPLOYEE.

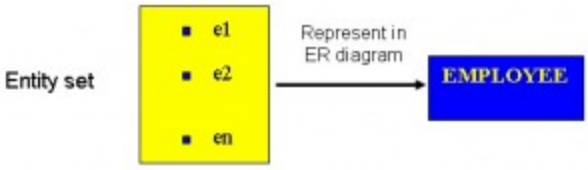


Figure 8.1. ERD with entity type EMPLOYEE.

Existence dependency

An entity's existence is dependent on the existence of the related entity. It is existence-dependent if it has a mandatory foreign key (i.e., a foreign key attribute that cannot be null). For example, in the COMPANY database, a Spouse entity is existence -dependent on the Employee entity.

Kinds of Entities

You should also be familiar with different kinds of entities including independent entities, dependent entities and characteristic entities. These are described below.

Independent entities

Independent entities, also referred to as kernels, are the backbone of the database. They are what other tables are based on. *Kernels* have the following characteristics:

- They are the building blocks of a database.
- The primary key may be simple or composite.
- The primary key is not a foreign key.
- They do not depend on another entity for their existence.

If we refer back to our COMPANY database, examples of an independent entity include the Customer table, Employee table or Product table.

Dependent entities

Dependent entities, also referred to as *derived entities*, depend on other tables for their meaning. These entities have the following characteristics:

- Dependent entities are used to connect two kernels together.
- They are said to be existence dependent on two or more tables.
- Many to many relationships become associative tables with at least two foreign keys.
- They may contain other attributes.
- The foreign key identifies each associated table.
- There are three options for the primary key:
  1. Use a composite of foreign keys of associated tables if unique
  2. Use a composite of foreign keys and a qualifying column
  3. Create a new simple primary key

Characteristic entities

Characteristic entities provide more information about another table. These entities have the following characteristics:

- They represent multivalued attributes.
- They describe other entities.
- They typically have a one to many relationship.
- The foreign key is used to further identify the characterized table.
- Options for primary key are as follows:
  1. Use a composite of foreign key plus a qualifying column

2. Create a new simple primary key. In the COMPANY database, these might include:
   - Employee (EID, Name, Address, Age, Salary) – EID is the simple primary key.
   - EmployeePhone (EID, Phone) – EID is part of a composite primary key. Here, EID is also a foreign key.

Attributes

Each entity is described by a set of attributes (e.g., Employee = (Name, Address, Birthdate (Age), Salary).

Each attribute has a name, and is associated with an entity and a domain of legal values. However, the information about attribute domain is not presented on the ERD.

In the entity relationship diagram, shown in Figure 8.2, each attribute is represented by an oval with a name inside.

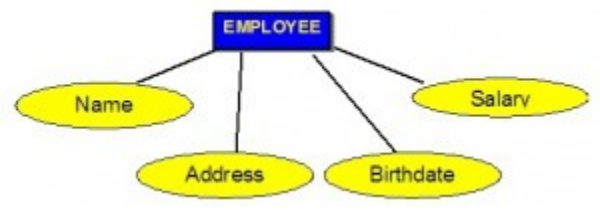


Figure 8.2. How attributes are represented in an ERD.

Types of Attributes

There are a few types of attributes you need to be familiar with. Some of these are to be left as is, but some need to be adjusted to facilitate representation in the relational model. This first section will discuss the types of attributes. Later on we will discuss fixing the attributes to fit correctly into the relational model.

Simple attributes

Simple attributes are those drawn from the atomic value domains; they are also called *single-valued attributes*. In the COMPANY database, an example of this would be: Name = {John} ; Age = {23}

Composite attributes

Composite attributes are those that consist of a hierarchy of attributes. Using our database example, and shown in Figure 8.3, Address may consist of Number, Street and Suburb. So this would be written as → Address = {59 + 'Meek Street' + 'Kingsford'}

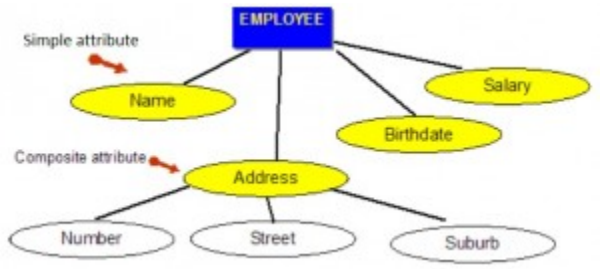


Figure 8.3. An example of composite attributes.

Multivalued attributes

Multivalued attributes are attributes that have a set of values for each entity. An example of a multivalued attribute from the COMPANY database, as seen in Figure 8.4, are the degrees of an employee: BSc, MIT, PhD.

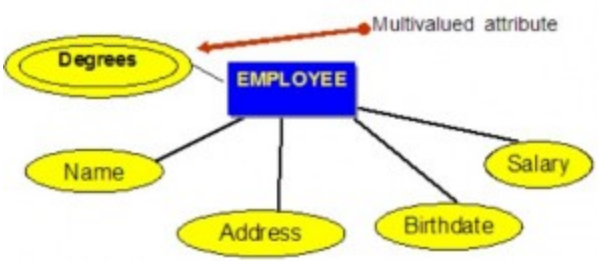


Figure 8.4. Example of a multivalued attribute.

Derived attributes

*Derived attribute*s are attributes that contain values calculated from other attributes. An example of this can be seen in Figure 8.5. Age can be derived from the attribute Birthdate. In this situation, Birthdate is called a *stored attribute,* which is physically saved to the database.

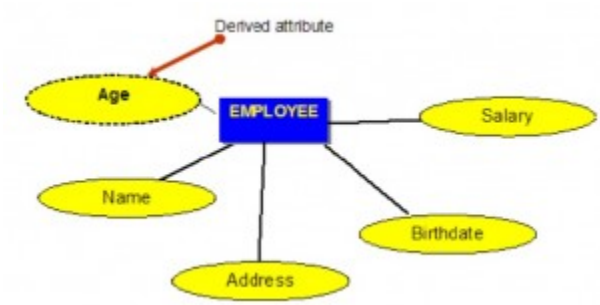


Figure 8.5. Example of a derived attribute.

Keys

An important constraint on an entity is the key. The *key* is an attribute or a group of attributes whose values can be used to uniquely identify an individual entity in an entity set.

Types of Keys

There are several types of keys. These are described below.

Candidate key

A *candidate key* is a simple or composite key that is unique and minimal. It is unique because no two rows in a table may have the same value at any time. It is minimal because every column is necessary in order to attain uniqueness.

From our COMPANY database example, if the entity is**Employee**(EID, First Name, Last Name, *SIN*, Address, Phone, BirthDate, Salary, DepartmentID), possible candidate keys are:

- EID, SIN
- First Name and Last Name – assuming there is no one else in the company with the same name
- Last Name and DepartmentID – assuming two people with the same last name don't work in the same department

Composite key

A *composite key* is composed of two or more attributes, but it must be minimal.

Using the example from the candidate key section, possible composite keys are:

- First Name and Last Name – assuming there is no one else in the company with the same name

- Last Name and Department ID – assuming two people with the same last name don't work in the same department

Primary key

The primary key is a candidate key that is selected by the database designer to be used as an identifying mechanism for the whole entity set. It must uniquely identify tuples in a table and not be null. The primary key is indicated in the ER model by underlining the attribute.

- A candidate key is selected by the designer to uniquely identify tuples in a table. It must not be null.
- A key is chosen by the database designer to be used as an identifying mechanism for the whole entity set. This is referred to as the primary key. This key is indicated by underlining the attribute in the ER model.

In the following example, EID is the primary key:

Employee(EID, First Name, Last Name, SIN, Address, Phone, BirthDate, Salary, DepartmentID)

Secondary key

A *secondary key* is an attribute used strictly for retrieval purposes (can be composite), for example: Phone and Last Name.

Alternate key

Alternate keys are all candidate keys not chosen as the primary key.

Foreign key

A *foreign key (FK)* is an attribute in a table that references the primary key in another table OR it can be null. Both foreign and primary keys must be of the same data type.

In the COMPANY database example below, DepartmentID is the foreign key:

Employee(EID, First Name, Last Name, SIN, Address, Phone, BirthDate, Salary, DepartmentID)

Nulls

A *null* is a special symbol, independent of data type, which means either unknown or inapplicable. It does not mean zero or blank. Features of null include:

- No data entry
- Not permitted in the primary key
- Should be avoided in other attributes
- Can represent
  - An unknown attribute value
  - A known, but missing, attribute value
  - A "not applicable" condition
- Can create problems when functions such as COUNT, AVERAGE and SUM are used
- Can create logical problems when relational tables are linked

NOTE: The result of a comparison operation is null when either argument is null. The result of an arithmetic operation is null when either argument is null (except functions that ignore nulls).

Example of how null can be used

Use the Salary table (Salary_tbl) in Figure 8.6 to follow an example of how null can be used.

Salary_tbl

emp#	jobName	salary	commission
E10	Sales	12500	32090
E11	Null	25000	8000
E12	Sales	44000	0
E13	Sales	44000	Null

Figure 8.6. Salary table for null example, by A. Watt.

To begin, find all employees (emp#) in Sales (under the jobName column) whose salary plus commission are greater than 30,000.

- SELECT emp# FROM Salary_tbl
- WHERE jobName = Sales AND
- (commission + salary) > 30,000 –> E10 and E12

This result does not include E13 because of the null value in the commission column. To ensure that the row with the null value is included, we need to look at the individual fields. By adding commission and salary for employee E13, the result will be a null value. The solution is shown below.

- SELECT emp# FROM Salary_tbl
- WHERE jobName = Sales AND
- (commission > 30000 OR
- salary > 30000 OR
- (commission + salary) > 30,000 –>E10 and E12 and E13

Relationships

Relationships are the glue that holds the tables together. They are used to connect related information between tables.

Relationship strength is based on how the primary key of a related entity is defined. A weak, or non-identifying, relationship exists if the primary key of the related entity does not contain a primary key component of the parent entity. Company database examples include:

- Customer(**CustID**, CustName)
- Order(**OrderID**, CustID, Date)

A strong, or identifying, relationship exists when the primary key of the related entity contains the primary key component of the parent entity. Examples include:

- Course(**CrsCode**, DeptCode, Description)
- Class(**CrsCode, Section**, ClassTime...)

Types of Relationships

Below are descriptions of the various types of relationships.

One to many (1:M) relationship

A one to many (1:M) relationship should be the norm in any relational database design and is found in all relational database environments. For example, one department has many employees. Figure 8.7 shows the relationship of one of these employees to the department.

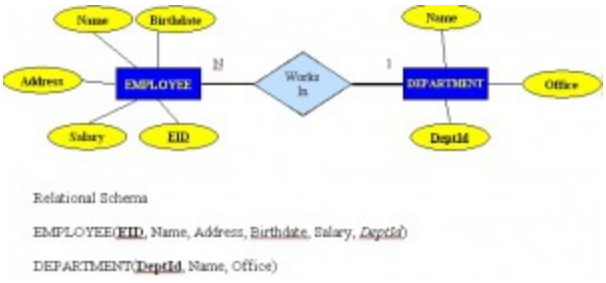


Figure 8.7. Example of a one to many relationship.

One to one (1:1) relationship

A one to one (1:1) relationship is the relationship of one entity to only one other entity, and vice versa. It should be rare in any relational database design. In fact, it could indicate that two entities actually belong in the same table.

An example from the COMPANY database is one employee is associated with one spouse, and one spouse is associated with one employee.

Many to many (M:N) relationships

For a many to many relationship, consider the following points:

- It cannot be implemented as such in the relational model.
- It can be changed into two 1:M relationships.
- It can be implemented by breaking up to produce a set of 1:M relationships.
- It involves the implementation of a composite entity.
- Creates two or more 1:M relationships.
- The composite entity table must contain at least the primary keys of the original tables.
- The linking table contains multiple occurrences of the foreign key values.
- Additional attributes may be assigned as needed.
- It can avoid problems inherent in an M:N relationship by creating a composite entity or bridge entity. For example, an employee can work on many projects OR a project can have many employees working on it, depending on the business rules. Or, a student can have many classes and a class can hold many students.

Figure 8.8 shows another another aspect of the M:N relationship where an employee has different start dates for different projects. Therefore, we need a JOIN table that contains the EID, Code and StartDate.

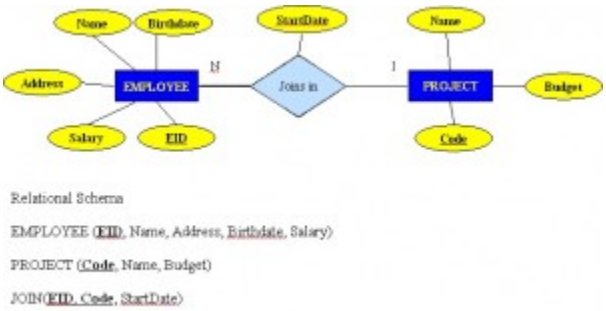


Figure 8.8. Example where employee has different start dates for different projects.

Example of mapping an M:N binary relationship type

- For each M:N binary relationship, identify two relations.
- A and B represent two entity types participating in R.
- Create a new relation S to represent R.
- S needs to contain the PKs of A and B. These together can be the PK in the S table OR these together with another simple attribute in the new table R can be the PK.

- The combination of the primary keys (A and B) will make the primary key of S.

Unary relationship (recursive)

A *unary relationship,* also called *recursive,* is one in which a relationship exists between occurrences of the same entity set. In this relationship, the primary and foreign keys are the same, but they represent two entities with different roles. See Figure 8.9 for an example.

For some entities in a unary relationship, a separate column can be created that refers to the primary key of the same entity set.

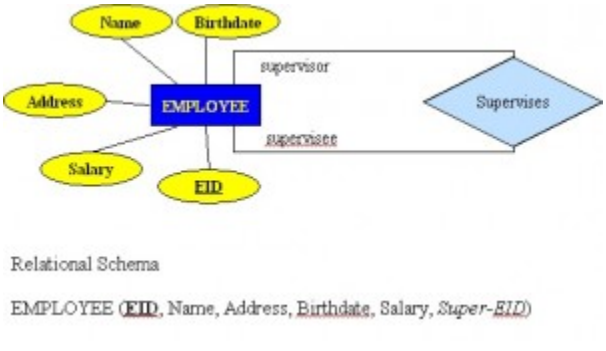


Figure 8.9. Example of a unary relationship.

Ternary Relationships

A *ternary relationship* is a relationship type that involves many to many relationships between three tables.

Refer to Figure 8.10 for an example of mapping a ternary relationship type. Note *n-ary* means multiple tables in a relationship. (Remember, N = many.)

- For each n-ary (> 2) relationship, create a new relation to represent the relationship.
- The primary key of the new relation is a combination of the primary keys of the participating entities that hold the N (many) side.
- In most cases of an n-ary relationship, all the participating entities hold a **many** side.

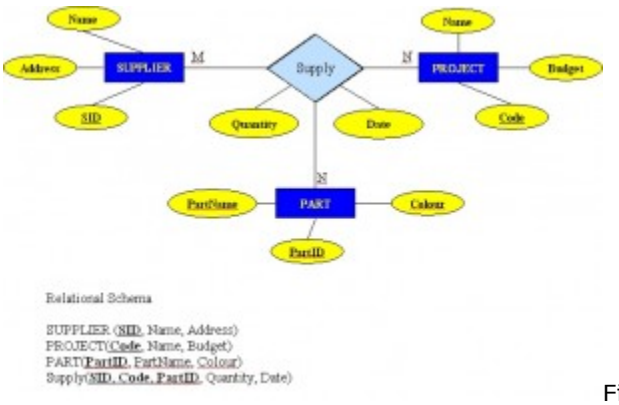


Figure 8.10. Example of a ternary relationship.

Key Terms

alternate key: all candidate keys not chosen as the primary key

candidate key: a simple or composite key that is unique (no two rows in a table may have the same value) and minimal (every column is necessary)

characteristic entities: entities that provide more information about another table

composite attributes: attributes that consist of a hierarchy of attributes

composite key: composed of two or more attributes, but it must be minimal

dependent entities: these entities depend on other tables for their meaning

derived attributes: attributes that contain values calculated from other attributes

derived entities: see *dependent entities*

EID: employee identification (ID)

entity: a thing or object in the real world with an independent existence that can be differentiated from other objects

entity relationship (ER) data model: also called an ER schema, are represented by ER diagrams. These are well suited to data modelling for use with databases.

entity relationship schema : see *entity relationship data model*

entity set:a collection of entities of an entity type at a point of time

entity type: a collection of similar entities

foreign key (FK): an attribute in a table that references the primary key in another table OR it can be null

independent entity: as the building blocks of a database, these entities are what other tables are based on

kernel: see *independent entity*

key: an attribute or group of attributes whose values can be used to uniquely identify an individual entity in an entity set

multivalued attributes: attributes that have a set of values for each entity

n-ary: multiple tables in a relationship

null: a special symbol, independent of data type, which means either unknown or inapplicable; it does not mean zero or blank

recursive relationship: see *unary relationship*

relationships: the associations or interactions between entities; used to connect related information between tables

relationship strength: based on how the primary key of a related entity is defined

secondary key an attribute used strictly for retrieval purposes

simple attributes: drawn from the atomic value domains

SIN: social insurance number

single-valued attributes: see *simple attributes*

stored attribute: saved physically to the database

ternary relationship: a relationship type that involves many to many relationships between three tables.

unary relationship: one in which a relationship exists between occurrences of the same entity set.

Exercises

1. What two concepts are ER modelling based on?

2. The database in Figure 8.11 is composed of two tables. Use this figure to answer questions 2.1 to 2.5.

DIRECTOR

DIRNUM	DIRNAME	DIRDOB
100	J.Broadway	01/08/39
101	J.Namath	11/12/48
102	W.Blake	06/15/44

PLAY

PLAYNO	PLAYNAME	DIRNUM
1001	Cat on a cold bare roof	102
1002	Hold the mayo, pass the bread	101
1003	I never promised you coffee	102
1004	Silly putty goes to Texas	100
1005	See no sound, hear no sight	101
1006	Starstruck in Biloxi	102
1007	Stranger in parrot ice	101

Figure 8.11. Director and Play tables for question 2, by A. Watt.

   1. Identify the primary key for each table.
   2. Identify the foreign key in the PLAY table.
   3. Identify the candidate keys in both tables.
   4. Draw the ER model.
   5. Does the PLAY table exhibit referential integrity? Why or why not?

3. Define the following terms (you may need to use the Internet for some of these):
   schema
   host language
   data sublanguage
   data definition language
   unary relation
   foreign key
   virtual relation
   connectivity
   composite key
   linking table

4. The RRE Trucking Company database includes the three tables in Figure 8.12. Use Figure 8.12 to answer questions 4.1 to 4.5.

TRUCK

TNUM	BASENUM	TYPENUM	TMILES	TBOUGHT	TSERIAL
1001	501	1	5900.2	11/08/90	aa-125
1002	502	2	64523.9	11/08/90	ac-213
1003	501	2	32116.0	09/29/91	ac-215
1004		2	3256.9	01/14/92	ac-315

BASE

BASENUM	BASECITY	BASESTATE	BASEPHON	BASEMGR
501	Dallas	TX	893-9870	J. Jones
502	New York	NY	234-7689	K. Lee

TYPE

TYPENUM	TYPEDESC
1	single box, double axle
2	tandem trailer, single axle

Figure 8.12. Truck, Base and Type tables for question 4, by A. Watt.

1. Identify the primary and foreign key(s) for each table.
2. Does the TRUCK table exhibit entity and referential integrity? Why or why not? Explain your answer.
3. What kind of relationship exists between the TRUCK and BASE tables?
4. How many entities does the TRUCK table contain ?
5. Identify the TRUCK table candidate key(s).

Customer

CustID	CustName	AccntNo.
100	Joe Smith	010839
101	Andy Blake	111248
102	Sue Brown	061544

BookOrders

OrderID	Title	CustID	Price
1001	The Dark Tower	**102**	**12.00**
1002	Incubus Dreams	**101**	**19.99**
1003	Song of Susannah	**102**	**23.00**
1004	The Time Traveler's Wife	**100**	**21.00**
1005	The Dark Tower	**101**	**12.00**
1006	Tanequil	**102**	**15.00**
1007	Song of Susannah	**101**	**23.00**

Figure 8.13. Customer and BookOrders tables for question 5, by A. Watt.

5. Suppose you are using the database in Figure 8.13, composed of the two tables. Use Figure 8.13 to answer questions 5.1 to 5.6.
    1. Identify the primary key in each table.
    2. Identify the foreign key in the BookOrders table.
    3. Are there any candidate keys in either table?
    4. Draw the ER model.
    5. Does the BookOrders table exhibit referential integrity? Why or why not?
    6. Do the tables contain redundant data? If so which table(s) and what is the redundant data?
6. Looking at the student table in Figure 8.14, list all the possible candidate keys. Why did you select these?

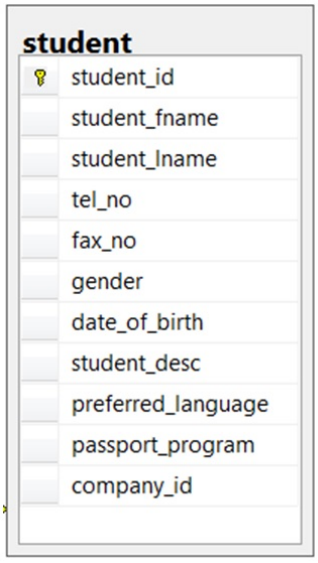


Figure 8.14. Student table for question 6, by A. Watt.

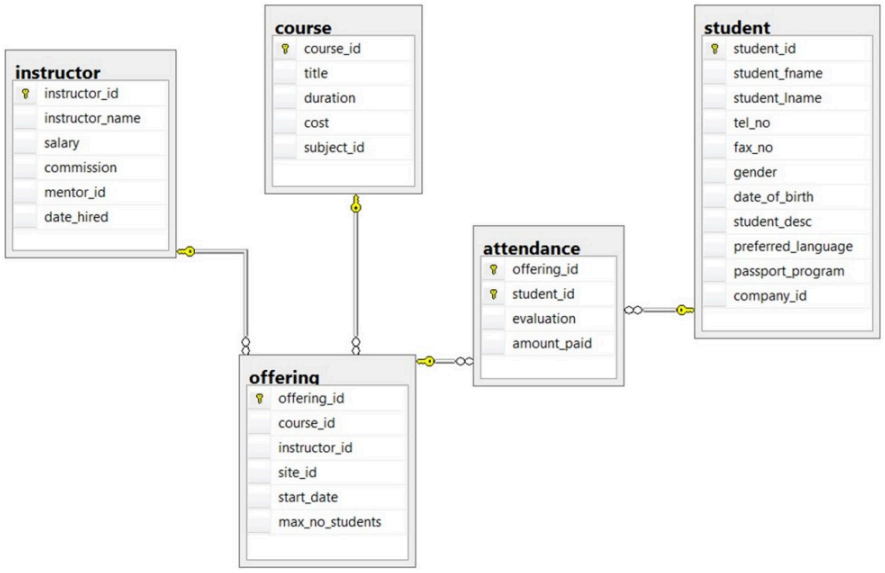


Figure 8.15. ERD of school database for questions 7-10, by A. Watt.

Use the ERD of a school database in Figure 8.15 to answer questions 7 to 10.

7. Identity all the kernels and dependent and characteristic entities in the ERD.
8. Which of the tables contribute to weak relationships? Strong relationships?
9. Looking at each of the tables in the school database in Figure 8.15, which attribute could have a NULL value? Why?
10. Which of the tables were created as a result of many to many relationships?

Also see *Appendix B: Sample ERD Exercises*

Attribution

This chapter of *Database Design* (including images, except as otherwisse noted) is a derivative copy of Data Modeling Using Entity-Relationship Model by Nguyen Kim Anh licensed under Creative Commons Attribution License 3.0 license

The following material was written by Adrienne Watt:

1. Nulls section and example
2. Key Terms
3. Exercises

7.1: Entity, Entity Set and Entity Type

7.2: Types of Attributes

7.3: Keys

7.4: Nulls

7.5: Relationships

7.6: Key Terms and Exercises

This page titled 7: The Entity Relationship Data Model is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

7.1: Entity, Entity Set and Entity Type

Entity, Entity Set and Entity Type

An *entity* is an object in the real world with an independent existence that can be differentiated from other objects. An entity might be

- An object with physical existence (e.g., a lecturer, a student, a car)
- An object with conceptual existence (e.g., a course, a job, a position)

Entities can be classified based on their strength. An entity is considered weak if its tables are existence dependent.

- That is, it cannot exist without a relationship with another entity
- Its primary key is derived from the primary key of the parent entity
    - The Spouse table, in the COMPANY database, is a weak entity because its primary key is dependent on the Employee table. Without a corresponding employee record, the spouse record would not exist.

An entity is considered strong if it can exist apart from all of its related entities.

- Kernels are strong entities.
- A table without a foreign key or a table that contains a foreign key that can contain nulls is a strong entity

Another term to know is *entity type* which defines a collection of similar entities.

An *entity set* is a collection of entities of an entity type at a particular point of time. In an entity relationship diagram (ERD), an entity type is represented by a name in a box. For example, in Figure 8.1, the entity type is EMPLOYEE.

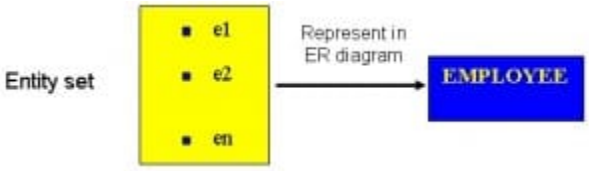


Figure 8.1. ERD with entity type EMPLOYEE.

Existence dependency

An entity's existence is dependent on the existence of the related entity. It is existence-dependent if it has a mandatory foreign key (i.e., a foreign key attribute that cannot be null). For example, in the COMPANY database, a Spouse entity is existence -dependent on the Employee entity.

Kinds of Entities

You should also be familiar with different kinds of entities including independent entities, dependent entities and characteristic entities. These are described below.

Independent entities

Independent entities, also referred to as kernels, are the backbone of the database. They are what other tables are based on. *Kernels* have the following characteristics:

- They are the building blocks of a database.
- The primary key may be simple or composite.
- The primary key is not a foreign key.
- They do not depend on another entity for their existence.

If we refer back to our COMPANY database, examples of an independent entity include the Customer table, Employee table or Product table.

Dependent entities

Dependent entities, also referred to as *derived entities*, depend on other tables for their meaning. These entities have the following characteristics:

- Dependent entities are used to connect two kernels together.
- They are said to be existence dependent on two or more tables.
- Many to many relationships become associative tables with at least two foreign keys.
- They may contain other attributes.
- The foreign key identifies each associated table.
- There are three options for the primary key:
  1. Use a composite of foreign keys of associated tables if unique
  2. Use a composite of foreign keys and a qualifying column
  3. Create a new simple primary key

Characteristic entities

Characteristic entities provide more information about another table. These entities have the following characteristics:

- They represent multivalued attributes.
- They describe other entities.
- They typically have a one to many relationship.
- The foreign key is used to further identify the characterized table.
- Options for primary key are as follows:
  1. Use a composite of foreign key plus a qualifying column
  2. Create a new simple primary key. In the COMPANY database, these might include:
     - Employee (EID, Name, Address, Age, Salary) – EID is the simple primary key.
     - EmployeePhone (EID, Phone) – EID is part of a composite primary key. Here, EID is also a foreign key.

This page titled 7.1: Entity, Entity Set and Entity Type is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

7.2: Types of Attributes

Types of Attributes

There are a few types of attributes you need to be familiar with. Some of these are to be left as is, but some need to be adjusted to facilitate representation in the relational model. This first section will discuss the types of attributes. Later on we will discuss fixing the attributes to fit correctly into the relational model.

Simple attributes

Simple attributes are those drawn from the atomic value domains; they are also called *single-valued attributes*. In the COMPANY database, an example of this would be: Name = {John} ; Age = {23}

Composite attributes

Composite attributes are those that consist of a hierarchy of attributes. Using our database example, and shown in Figure 8.3, Address may consist of Number, Street and Suburb. So this would be written as → Address = {59 + 'Meek Street' + 'Kingsford'}

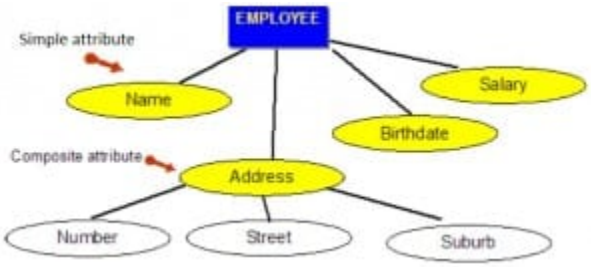


Figure 8.3. An example of composite attributes.

Multivalued attributes

Multivalued attributes are attributes that have a set of values for each entity. An example of a multivalued attribute from the COMPANY database, as seen in Figure 8.4, are the degrees of an employee: BSc, MIT, PhD.

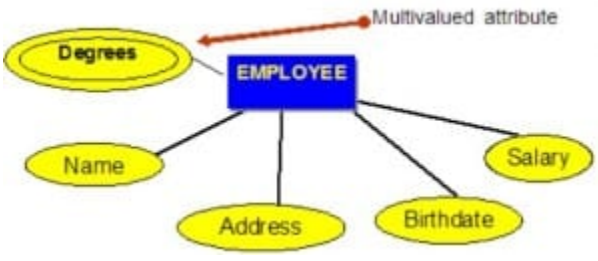


Figure 8.4. Example of a multivalued attribute.

Derived attributes

Derived attributes are attributes that contain values calculated from other attributes. An example of this can be seen in Figure 8.5. Age can be derived from the attribute Birthdate. In this situation, Birthdate is called a *stored attribute,* which is physically saved to the database.

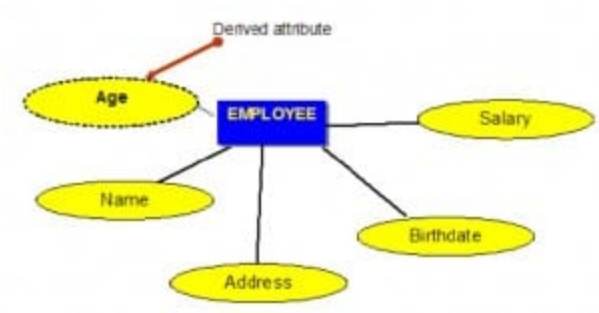


Figure 8.5. Example of a derived attribute.

This page titled 7.2: Types of Attributes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

7.3: Keys

Keys

An important constraint on an entity is the key. The *key* is an attribute or a group of attributes whose values can be used to uniquely identify an individual entity in an entity set.

Types of Keys

There are several types of keys. These are described below.

Candidate key

A *candidate key* is a simple or composite key that is unique and minimal. It is unique because no two rows in a table may have the same value at any time. It is minimal because every column is necessary in order to attain uniqueness.

From our COMPANY database example, if the entity is**Employee**(EID, First Name, Last Name, *SIN*, Address, Phone, BirthDate, Salary, DepartmentID), possible candidate keys are:

- EID, SIN
- First Name and Last Name – assuming there is no one else in the company with the same name
- Last Name and DepartmentID – assuming two people with the same last name don't work in the same department

Composite key

A *composite key* is composed of two or more attributes, but it must be minimal.

Using the example from the candidate key section, possible composite keys are:

- First Name and Last Name – assuming there is no one else in the company with the same name
- Last Name and Department ID – assuming two people with the same last name don't work in the same department

Primary key

The primary key is a candidate key that is selected by the database designer to be used as an identifying mechanism for the whole entity set. It must uniquely identify tuples in a table and not be null. The primary key is indicated in the ER model by underlining the attribute.

- A candidate key is selected by the designer to uniquely identify tuples in a table. It must not be null.
- A key is chosen by the database designer to be used as an identifying mechanism for the whole entity set. This is referred to as the primary key. This key is indicated by underlining the attribute in the ER model.

In the following example, EID is the primary key:

Employee(EID, First Name, Last Name, SIN, Address, Phone, BirthDate, Salary, DepartmentID)

Secondary key

A *secondary key* is an attribute used strictly for retrieval purposes (can be composite), for example: Phone and Last Name.

Alternate key

Alternate keys are all candidate keys not chosen as the primary key.

Foreign key

A *foreign key (FK)* is an attribute in a table that references the primary key in another table OR it can be null. Both foreign and primary keys must be of the same data type.

In the COMPANY database example below, DepartmentID is the foreign key:

Employee(EID, First Name, Last Name, SIN, Address, Phone, BirthDate, Salary, DepartmentID)

This page titled 7.3: Keys is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

7.4: Nulls

Nulls

A *null* is a special symbol, independent of data type, which means either unknown or inapplicable. It does not mean zero or blank. Features of null include:

- No data entry
- Not permitted in the primary key
- Should be avoided in other attributes
- Can represent
  - An unknown attribute value
  - A known, but missing, attribute value
  - A “not applicable” condition
- Can create problems when functions such as COUNT, AVERAGE and SUM are used
- Can create logical problems when relational tables are linked

NOTE: The result of a comparison operation is null when either argument is null. The result of an arithmetic operation is null when either argument is null (except functions that ignore nulls).

Example of how null can be used

Use the Salary table (Salary_tbl) in Figure 8.6 to follow an example of how null can be used.

Salary_tbl

emp#	jobName	salary	commission
E10	Sales	12500	32090
E11	Null	25000	8000
E12	Sales	44000	0
E13	Sales	44000	Null

Figure 8.6. Salary table for null example, by A. Watt.

To begin, find all employees (emp#) in Sales (under the jobName column) whose salary plus commission are greater than 30,000.

- SELECT emp# FROM Salary_tbl
- WHERE jobName = Sales AND
- (commission + salary) > 30,000 –> E10 and E12

This result does not include E13 because of the null value in the commission column. To ensure that the row with the null value is included, we need to look at the individual fields. By adding commission and salary for employee E13, the result will be a null value. The solution is shown below.

- SELECT emp# FROM Salary_tbl
- WHERE jobName = Sales AND
- (commission > 30000 OR
- salary > 30000 OR
- (commission + salary) > 30,000 –>E10 and E12 and E13

This page titled 7.4: Nulls is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

7.5: Relationships

Relationships

Relationships are the glue that holds the tables together. They are used to connect related information between tables.

Relationship strength is based on how the primary key of a related entity is defined. A weak, or non-identifying, relationship exists if the primary key of the related entity does not contain a primary key component of the parent entity. Company database examples include:

- Customer(**CustID**, CustName)
- Order(**OrderID**, CustID, Date)

A strong, or identifying, relationship exists when the primary key of the related entity contains the primary key component of the parent entity. Examples include:

- Course(**CrsCode**, DeptCode, Description)
- Class(**CrsCode, Section**, ClassTime...)

Types of Relationships

Below are descriptions of the various types of relationships.

One to many (1:M) relationship

A one to many (1:M) relationship should be the norm in any relational database design and is found in all relational database environments. For example, one department has many employees. Figure 8.7 shows the relationship of one of these employees to the department.

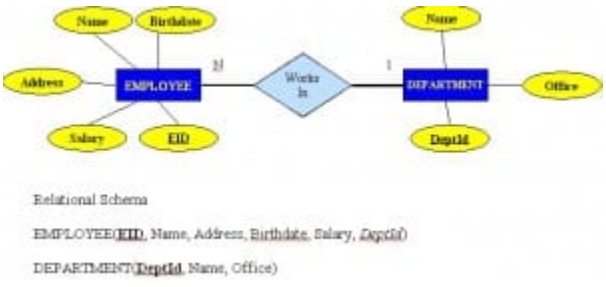


Figure 8.7. Example of a one to many relationship.

One to one (1:1) relationship

A one to one (1:1) relationship is the relationship of one entity to only one other entity, and vice versa. It should be rare in any relational database design. In fact, it could indicate that two entities actually belong in the same table.

An example from the COMPANY database is one employee is associated with one spouse, and one spouse is associated with one employee.

Many to many (M:N) relationships

For a many to many relationship, consider the following points:

- It cannot be implemented as such in the relational model.
- It can be changed into two 1:M relationships.

- It can be implemented by breaking up to produce a set of 1:M relationships.
- It involves the implementation of a composite entity.
- Creates two or more 1:M relationships.
- The composite entity table must contain at least the primary keys of the original tables.
- The linking table contains multiple occurrences of the foreign key values.
- Additional attributes may be assigned as needed.
- It can avoid problems inherent in an M:N relationship by creating a composite entity or bridge entity. For example, an employee can work on many projects OR a project can have many employees working on it, depending on the business rules. Or, a student can have many classes and a class can hold many students.

Figure 8.8 shows another another aspect of the M:N relationship where an employee has different start dates for different projects. Therefore, we need a JOIN table that contains the EID, Code and StartDate.

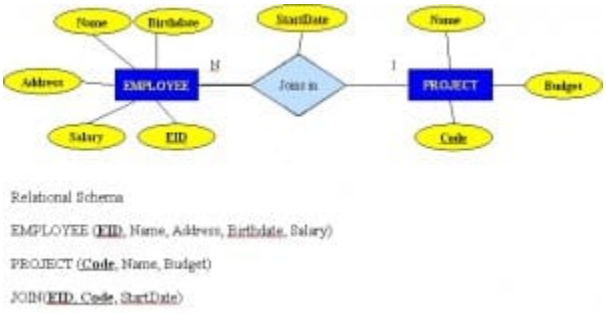


Figure 8.8. Example where employee has different start dates for different projects.

Example of mapping an M:N binary relationship type

- For each M:N binary relationship, identify two relations.
- A and B represent two entity types participating in R.
- Create a new relation S to represent R.
- S needs to contain the PKs of A and B. These together can be the PK in the S table OR these together with another simple attribute in the new table R can be the PK.
- The combination of the primary keys (A and B) will make the primary key of S.

Unary relationship (recursive)

A *unary relationship,* also called *recursive,* is one in which a relationship exists between occurrences of the same entity set. In this relationship, the primary and foreign keys are the same, but they represent two entities with different roles. See Figure 8.9 for an example.

For some entities in a unary relationship, a separate column can be created that refers to the primary key of the same entity set.

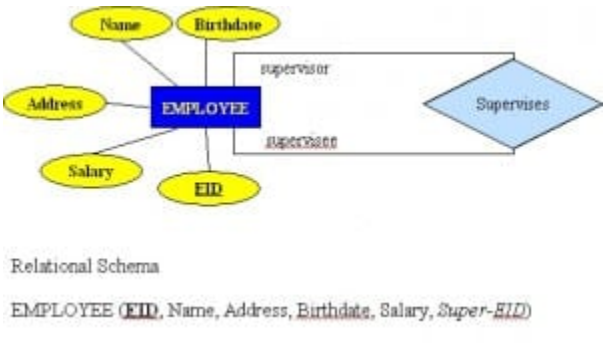


Figure 8.9. Example of a unary relationship.

Ternary Relationships

A *ternary relationship* is a relationship type that involves many to many relationships between three tables.

Refer to Figure 8.10 for an example of mapping a ternary relationship type. Note *n-ary* means multiple tables in a relationship. (Remember, N = many.)

- For each n-ary (> 2) relationship, create a new relation to represent the relationship.
- The primary key of the new relation is a combination of the primary keys of the participating entities that hold the N (many) side.
- In most cases of an n-ary relationship, all the participating entities hold a **many** side.

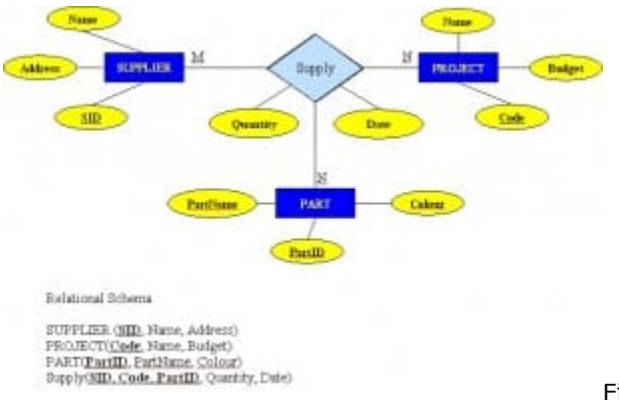


Figure 8.10. Example of a ternary relationship.

This page titled 7.5: Relationships is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

7.6: Key Terms and Exercises

Key Terms

alternate key: all candidate keys not chosen as the primary key

candidate key: a simple or composite key that is unique (no two rows in a table may have the same value) and minimal (every column is necessary)

characteristic entities: entities that provide more information about another table

composite attributes: attributes that consist of a hierarchy of attributes

composite key: composed of two or more attributes, but it must be minimal

dependent entities: these entities depend on other tables for their meaning

derived attributes: attributes that contain values calculated from other attributes

derived entities: see *dependent entities*

EID: employee identification (ID)

entity: a thing or object in the real world with an independent existence that can be differentiated from other objects

entity relationship (ER) data model: also called an ER schema, are represented by ER diagrams. These are well suited to data modelling for use with databases.

entity relationship schema : see *entity relationship data model*

entity set:a collection of entities of an entity type at a point of time

entity type: a collection of similar entities

foreign key (FK): an attribute in a table that references the primary key in another table OR it can be null

independent entity: as the building blocks of a database, these entities are what other tables are based on

kernel: see *independent entity*

key: an attribute or group of attributes whose values can be used to uniquely identify an individual entity in an entity set

multivalued attributes: attributes that have a set of values for each entity

n-ary: multiple tables in a relationship

null: a special symbol, independent of data type, which means either unknown or inapplicable; it does not mean zero or blank

recursive relationship: see *unary relationship*

relationships: the associations or interactions between entities; used to connect related information between tables

relationship strength: based on how the primary key of a related entity is defined

secondary key an attribute used strictly for retrieval purposes

simple attributes: drawn from the atomic value domains

SIN: social insurance number

single-valued attributes: see *simple attributes*

stored attribute: saved physically to the database

ternary relationship: a relationship type that involves many to many relationships between three tables.

unary relationship: one in which a relationship exists between occurrences of the same entity set.

Exercises

1. What two concepts are ER modelling based on?
2. The database in Figure 8.11 is composed of two tables. Use this figure to answer questions 2.1 to 2.5.

DIRECTOR

DIRNUM	DIRNAME	DIRDOB
100	J.Broadway	01/08/39
101	J.Namath	11/12/48
102	W.Blake	06/15/44

PLAY

PLAYNO	PLAYNAME	DIRNUM
1001	**Cat on a cold bare roof**	**102**
1002	**Hold the mayo, pass the bread**	**101**
1003	**I never promised you coffee**	**102**
1004	**Silly putty goes to Texas**	**100**
1005	**See no sound, hear no sight**	**101**
1006	**Starstruck in Biloxi**	**102**
1007	**Stranger in parrot ice**	**101**

Figure 8.11. Director and Play tables for question 2, by A. Watt.

   1. Identify the primary key for each table.
   2. Identify the foreign key in the PLAY table.
   3. Identify the candidate keys in both tables.
   4. Draw the ER model.
   5. Does the PLAY table exhibit referential integrity? Why or why not?
3. Define the following terms (you may need to use the Internet for some of these):
   schema
   host language
   data sublanguage
   data definition language
   unary relation
   foreign key
   virtual relation
   connectivity
   composite key
   linking table
4. The RRE Trucking Company database includes the three tables in Figure 8.12. Use Figure 8.12 to answer questions 4.1

TRUCK

TNUM	BASENUM	TYPENUM	TMILES	TBOUGHT	TSERIAL
1001	501	1	5900.2	11/08/90	aa-125
1002	502	2	64523.9	11/08/90	ac-213
1003	501	2	32116.0	09/29/91	ac-215
1004		2	3256.9	01/14/92	ac-315

BASE

BASENUM	BASECITY	BASESTATE	BASEPHON	BASEMGR
501	Dallas	TX	893-9870	J. Jones
502	New York	NY	234-7689	K. Lee

TYPE

TYPENUM	TYPEDESC
1	single box, double axle
2	tandem trailer, single axle

Figure 8.12. Truck, Base and Type tables for question 4, by A. Watt.

to 4.5.

1. Identify the primary and foreign key(s) for each table.
2. Does the TRUCK table exhibit entity and referential integrity? Why or why not? Explain your answer.
3. What kind of relationship exists between the TRUCK and BASE tables?
4. How many entities does the TRUCK table contain ?
5. Identify the TRUCK table candidate key(s).

Customer

CustID	CustName	AccntNo.
100	Joe Smith	010839
101	Andy Blake	111248
102	Sue Brown	061544

BookOrders

OrderID	Title	CustID	Price
1001	The Dark Tower	102	12.00
1002	Incubus Dreams	101	19.99
1003	Song of Susannah	102	23.00
1004	The Time Traveler's Wife	100	21.00
1005	The Dark Tower	101	12.00
1006	Tanequil	102	15.00
1007	Song of Susannah	101	23.00

Figure 8.13. Customer and BookOrders tables for question 5, by A. Watt.

5. Suppose you are using the database in Figure 8.13, composed of the two tables. Use Figure 8.13 to answer questions 5.1 to 5.6.
    1. Identify the primary key in each table.
    2. Identify the foreign key in the BookOrders table.
    3. Are there any candidate keys in either table?
    4. Draw the ER model.
    5. Does the BookOrders table exhibit referential integrity? Why or why not?
    6. Do the tables contain redundant data? If so which table(s) and what is the redundant data?
6. Looking at the student table in Figure 8.14, list all the possible candidate keys. Why did you select these?

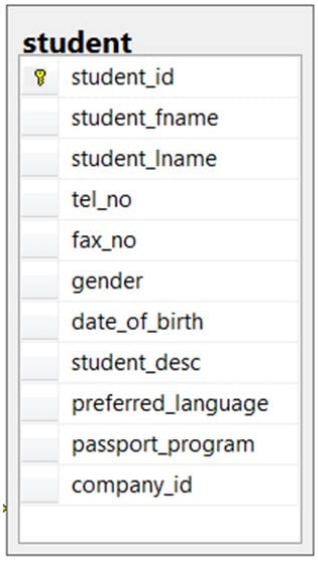


Figure 8.14. Student table for question 6, by A. Watt.

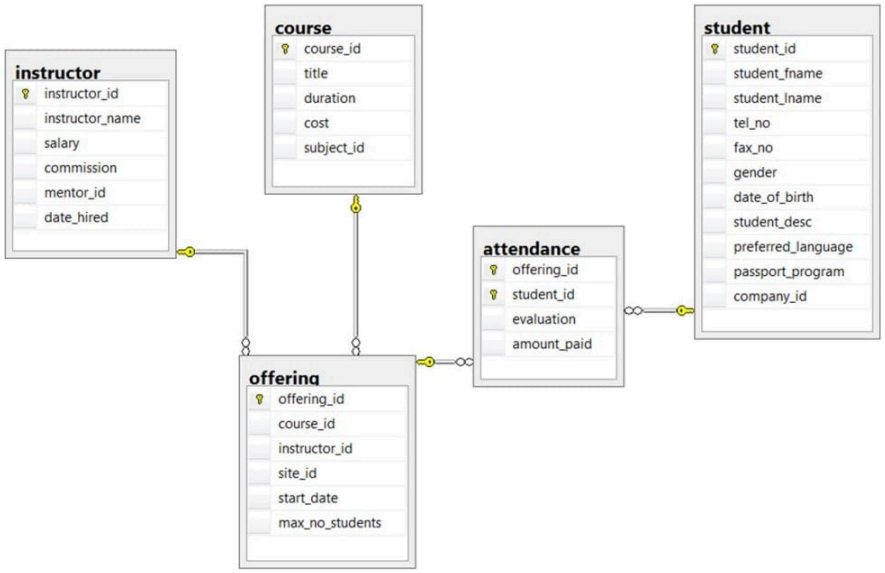


Figure 8.15. ERD of school database for questions 7-10, by A. Watt.

Use the ERD of a school database in Figure 8.15 to answer questions 7 to 10.

7. Identity all the kernels and dependent and characteristic entities in the ERD.
8. Which of the tables contribute to weak relationships? Strong relationships?
9. Looking at each of the tables in the school database in Figure 8.15, which attribute could have a NULL value? Why?
10. Which of the tables were created as a result of many to many relationships?

Also see *Appendix B: Sample ERD Exercises*

Attribution

This chapter of *Database Design* (including images, except as otherwisse noted) is a derivative copy of Data Modeling Using Entity-Relationship Model by Nguyen Kim Anh licensed under Creative Commons Attribution License 3.0 license

The following material was written by Adrienne Watt:

1. Nulls section and example
2. Key Terms

3. Exercises

This page titled 7.6: Key Terms and Exercises is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

8: VISUAL MEDIA

Learning Objectives

- Find the best media to present your message
- Discuss how to most effectively use images in business messages and identify potential sources for these images
- Discuss how to most effectively use charts, diagrams, and other graphics in business messages and identify potential sources for these visual aids
- Discuss the use of video in business messages
- Identify ways to make information more accessible to your audience
- Incorporating meaningful visual media in business messages

8.1: Tables, Charts, and Infographics

8.2: Matching Graphics and Objectives

8.3: Placement, Style, and Coloring

8.4: Captions and Titles

8.5: Introduction to Contemporary Visual Aids

8.6: Video as a Visual Aid

8.7: Designing a Video for Your Needs

8.8: Products and Resources for Creating Videos

8.9: Introduction to Accessible Visual Aids

8.10: Accessibility Matters

8.11: Why It Matters- Visual Media

8.12: Introduction to Using Visuals

8.13: Visuals in a Report

8.14: Increasing Impact with Media

8.15: Revising and Enhancing Visual Media for Impact

8.16: Evaluating the Effectiveness of your Message

8.17: Putting It Together- Visual Media

8.18: Discussion- Visual Media

8.19: Assignment- Visual Media

8.20: Introduction to Media and Your Message

8.21: Using Visual Media Resources

8.22: Visual Design Principles

8.23: Introduction to Images

8.24: Images Overview

8.25: Using Images for Impact

8.26: Using Images Legally

8.27: Introduction to Charts, Diagrams, and Graphic Organizers

This page titled 8: Visual Media is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Freedom Learning Group & Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 15: Recruiting and Selecting New Employees by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.1: Tables, Charts, and Infographics

Learning Outcome

- Discuss the appropriate use of common tables, charts, and infographics

In this age of information, there are huge amounts of data to process and an equally impressive number of graphs and charts you can use to tell the data's story. Here are some of the more commonly used graphs and the kinds of stories they can help you tell.

Numerical Data Charts

Let's talk about a few of the most common types of numerical charts:

- Bar graphs
- Line graphs
- Pie charts
- Stacked comparison charts

Bar Graphs

Bar graphs are used to compare categories. The *x*-axis (the horizontal line at the bottom of a graph) is usually used to show the categories: in this case, countries with universal healthcare expenses for 2015. The *y*-axis (the vertical line) isn't marked here, but it's clearly showing dollars spent in millions. The taller the bar, the more dollars were spent. This chart clearly shows that the United States spends more on health care per citizen than other countries. This is clearly identified on this chart with bars in contrasting colors. You can't help but notice it—it breaks pattern and it's a bold color that draws the eye.

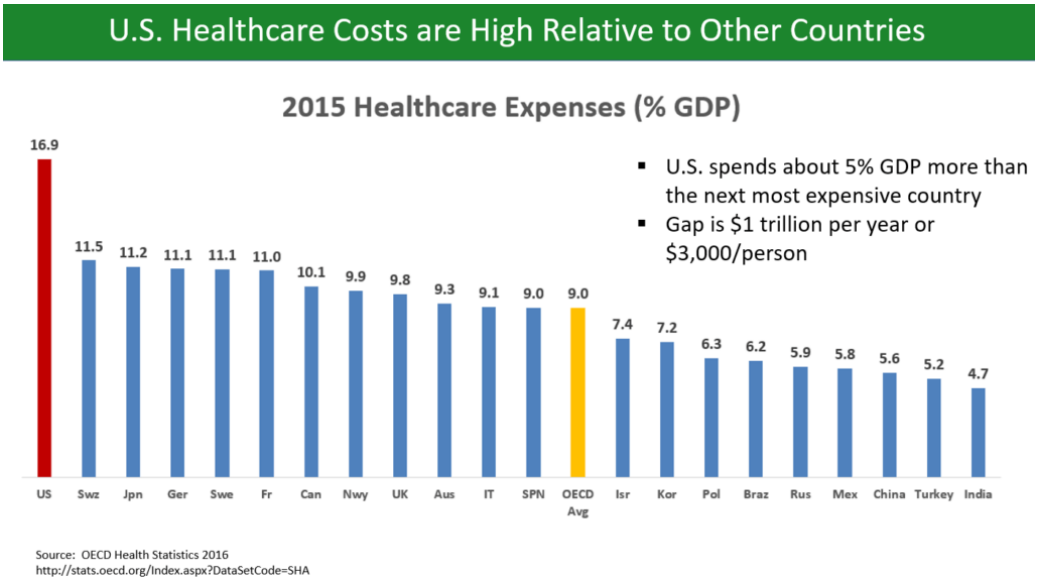
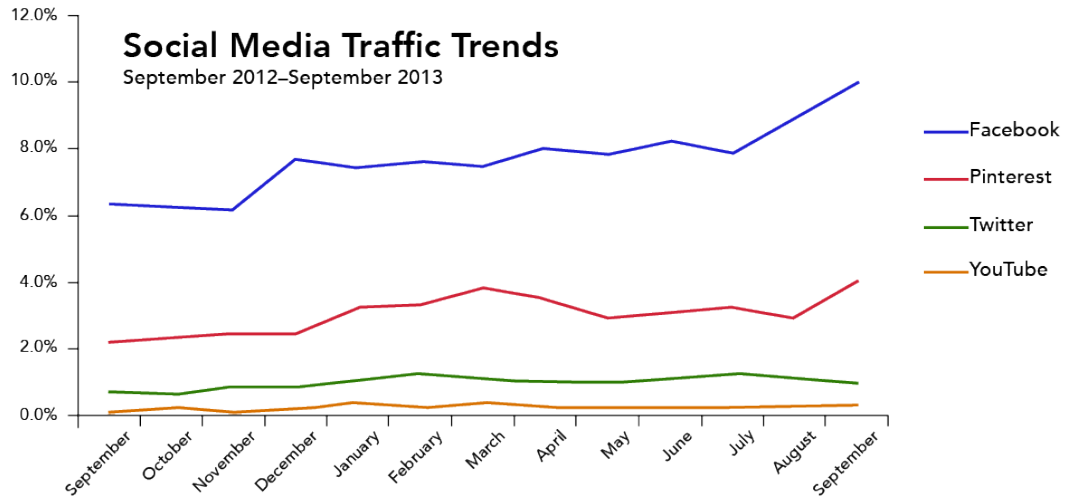


Figure 1. An example of a bar graph

Line Graphs

Like bar graphs, line graphs compare categories, but they're most often used to show trends. With the *x*-axis frequently showing the passage of time, the graphed data points will show an upward or downward trend of the categories in question.



Source: https://socialnewsdaily.com/17728/pinterest-referral-traffic-up-66-facebook-and-twitter-closely-behind/

Figure 2. An example of a line graph

The line graph above illustrates social media traffic trends. Each social media organization is represented by a different colored line. The *x*-axis shows the passage of time, and the *y*-axis shows the percentage of media traffic each organization is capturing. The graph shows that Facebook traffic is trending up, while Pinterest has experienced some ups and downs. Third-place Twitter traffic is relatively flat.

Pie Charts

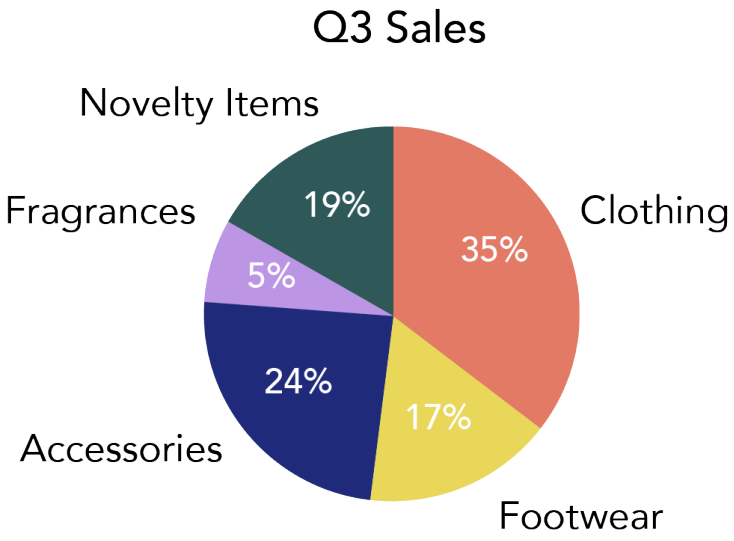


Figure 3. An example of a pie chart

Pie charts show the composition of data, or the pieces of a whole. It can be as simple as “the team here is composed of 50 percent men and 50 percent women” or “Our sales are made up of 30 percent fiction books and 70 percent non-fiction.”

In Figure 3, the whole pie (the whole circle) represents the total products sold at a store; the pieces of that pie show you the percentage of sales each department made. A chart like this makes it very easy to see that the clothing and accessory departments make up the largest section of sales, and fragrances the smallest.

Some comparison charts aren’t shaped like a circle. Sometimes they’re donuts, and other times they’re shown in bars, as we’ll see next.

Stacked Comparison Charts

You can compare categories with a “pie chart” approach, incorporating the composition factor in a variety of ways. The categories in this chart are represented by bars, but the bars themselves are composition charts. Each bar is valued at 100%, and the colored blocks represent different levels of pet ownership within the population.

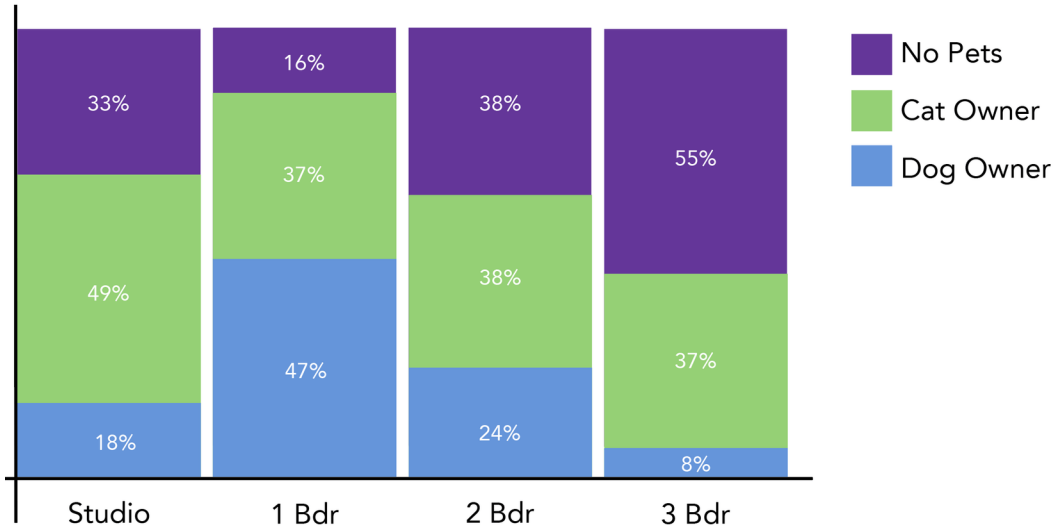


Figure 4. An example of a stacked bar graph

Similarly, the line graph below is “stacked” to show the level of sales based on product type. You’re seeing a trend as well as a portion of a whole – comparing and viewing composition.

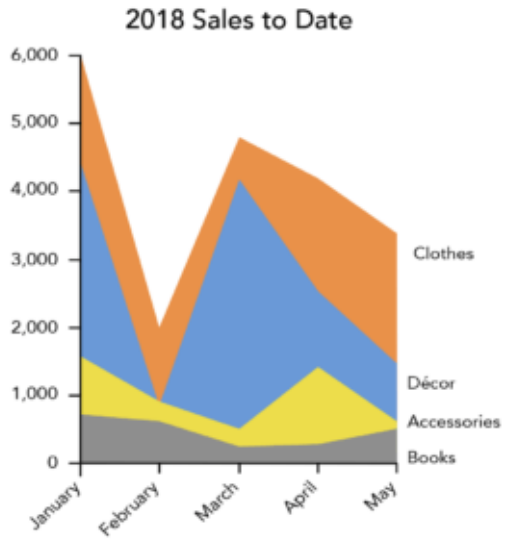


Figure 5. An example of a stacked line graph

Non-Numerical Data Charts

The charts above are very handy when your data is numeric. But there are charts for other types of data. Let’s talk about a few of the most common types of non-numerical charts.

- Venn diagrams
- Flow charts

- Gantt charts
- Organizational charts
- Pictographs
- Infographics

Venn Diagrams

A Venn diagram shows a comparison of two different categories and the items they have in common. The diagram in Figure 6 shows us the different types of image types that can be safely shared on the Internet.

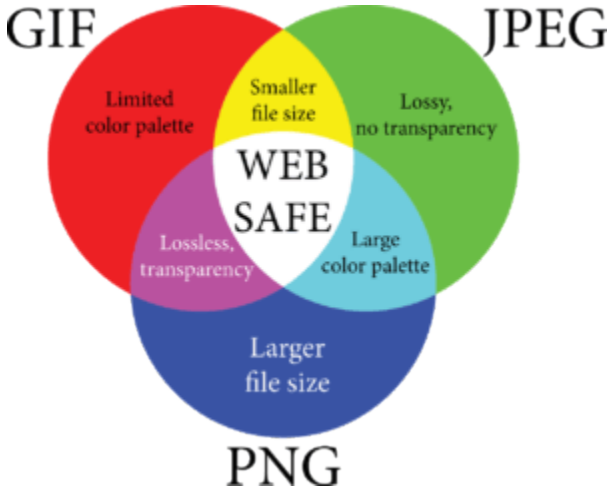


Figure 6. An example of a Venn diagram

Wherever the circles overlap represents a shared characteristic. For example, JPG and PNG files both have a large color palette, so they overlap in that area. “Web safe” is a characteristic shared by all three, so it’s in the middle where all the circles overlap. Areas of each circle that do not overlap represent characteristics unique to each file type—something they don’t share with any other file type represented.

Flow Charts

Flow charts show a process. Flow charts document a sequence of events from start to finish so that the process can be documented, followed, and managed.

The flow chart in Figure 7 shows the process of how theories are created, spread, and accepted.

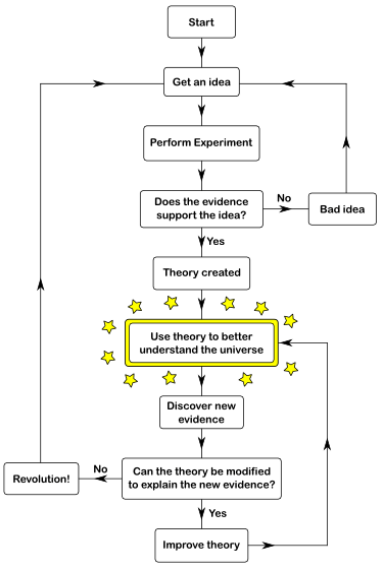


Figure 7. An example of a flow chart

You can see at the top there is an idea that leads to an experiment, and then the results determine which path is taken. If the experiment is successful or unsuccessful, different paths will be chosen. Along the path of the flow chart, all alternatives are presented and choices are made between them. Your eye follows the path from start to finish for every part of the scenario.

Gantt Chart

A Gantt chart is a timeline. Multiple projects can be added to the timeline with start and finish dates, and milestones and deadlines are also reflected. This chart is used to determine how long a project will take, the resources needed, and the order in which tasks need to be completed.

ID	Task Name	Predecessors	Duration
1	Start		0 days
2	a	1	4 days
3	b	1	5.33 days
4	c	2	5.17 days
5	d	2	6.33 days
6	e	3,4	5.17 days
7	f	5	4.5 days
8	g	6	5.17 days
9	Finish	7,8	0 days

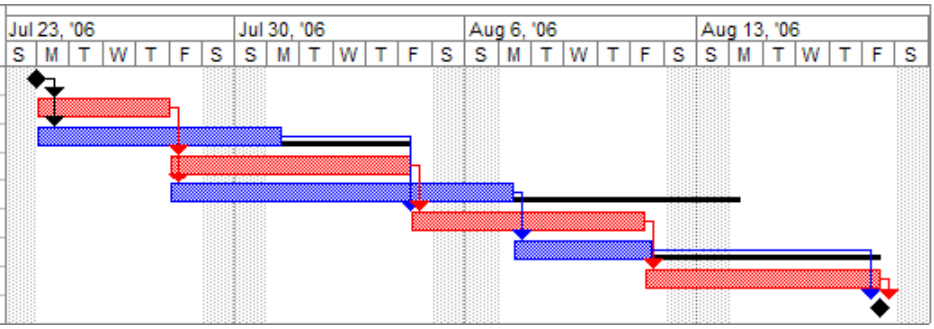


Figure 8. An example of a Gantt chart

In Figure 8, you have two different teams running one project. The Market Team (red) completes the market research and defines specifications by the week of July 23rd (the date of the first milestone). Then, the Planning Team (blue) takes over with the overall architecture and project planning and is responsible for hitting that second milestone on August 6th. As you can see, the Market and Planning Teams have additional work to contribute even after their milestones are hit, and the project is not complete until the end of November.

Organizational Charts

Organizational charts (sometimes call hierarchy charts) show the people in an organization and their reporting relationships. Usually, the organizational chart will have a chairman or CEO at the top, followed by a team of presidents and vice presidents, and then their direct reports, and so on. An organizational chart is usually created and maintained by human resource professionals who wants a visual view of their organization's structure and reporting relationships so they can make better decisions about leveraging the company's talent.

The organizational chart in Figure 8 shows a chairman at the top of the hierarchy and a managing director, quality assurance leader and a secretary reporting directly to the chairman. Each of those direct reports has direct reports of his or her own, and so on.

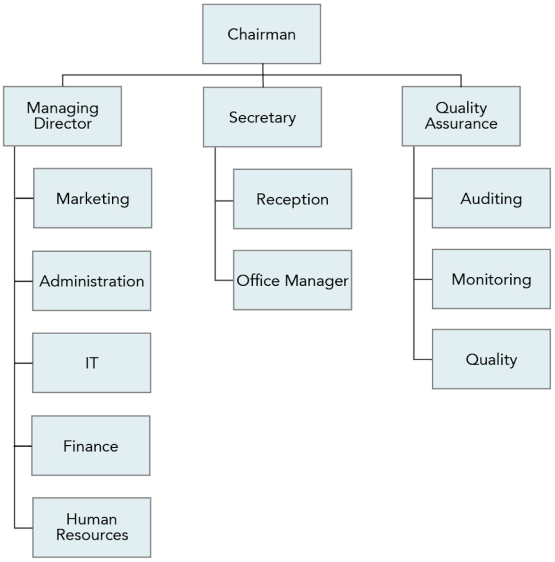


Figure 9. An example of an organizational chart

Pictographs

Pictographs use symbols and images to convey data, information, or ideas. The pictographs in Figure 9 are actually part of a resume and signal what kind of information can be found next to each pictograph (phone number, address, Twitter handle, and email respectively).



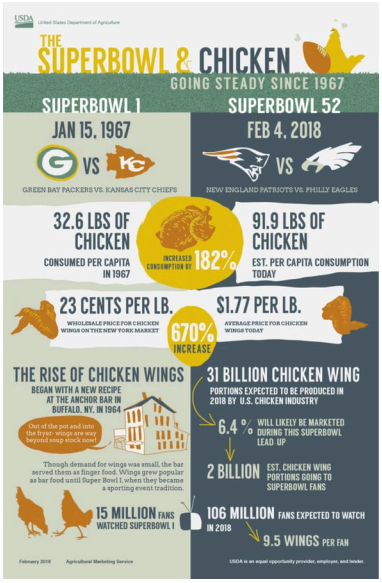
Figure 9. Four examples of pictographs

This makes it easier for a recruiter to view and pick out the necessary contact information. Other pictograms can convey processes (like the instructions to build IKEA furniture) or information (like when the weather app on your phone tells you it's going to rain today by posting a picture of a rain cloud).

Infographics

Similarly, infographics use pictures, but they also incorporate data and words, all to explain a single point. Infographics can be

used to make a complex subject a little simpler to understand, particularly when there are more than two or three factors to consider. They can also be used to compare two or more categories and make your data more interesting and eye catching.



This infographic shows how chicken consumption relates to the Superbowl. You can see how consumption and the price per pound of chicken has increased over time. It also describes how chicken wings in particular have become more popular over time, all with a an eye-catching and entertaining quality that's engaging and interesting to the audience.

Even More Charts and Graphs

This is not an all-inclusive list of the kinds of charts and graphs available. If you work in the science arena, you may find yourself using scatter charts or heat maps. Statisticians might be reaching for a trellis chart or even a function graph. Learn what graphs your industry relies on most and take the time to familiarize yourself with them so you can use these important visual communication tools to your advantage.

Making Charts and Graphs

As we mentioned earlier in this module, all graphs and charts that you see on the internet are someone else's property, and using them in your work is stealing. But there's little reason to use someone else's chart—you're creating your own chart to communicate your own data! Here are a few tools that can help you create something great:

- MS Excel. You've probably stumbled upon the charts and graph tools in your Microsoft program long ago, and if it's simple graphs you want, that's an easy tool to use to create them. Google Sheets and Apple's Pages offer similar options. See Module 4: Research for more on making charts and graphs in Excel.
- Lucidchart. If it's flowcharts you're looking for, Lucidchart is the right place to start. This free, easy-to-use resource allows you to create flowcharts, organizational charts and diagrams that look smart.
- OnlineChartTool. If you're looking for something different in terms of bar charts and graphs, and you're tired of the graphics Microsoft Excel provides, take a look at OnlineChartTool. It might have what you're looking for.
- Beam. This tool allows you to make engaging graphs and charts, even on your mobile phone.
- BeFunky, Visme, and Canva. Don't let infographics intimidate you! Infographics are easily created with the BeFunky, Visme and Canva tools. These tools have free versions and include hundreds of images and templates to help you make a stunning visual.

Tables

A communicator can also use tables to display data. Tables can be formatted for words or for numerical data. They can be used

for comparing data, or when one item has several data points associated with it.

Table 1 is a simple table of three people who are running races. George, Alana, and Sebastián each have four data points associated with them. By using a simple table, you can put all of this data in front of your audience. They can compare best times and review rankings of each runner.

Table 1. Running a Race

Participants	Races Run	Best Time	Average Time	Ranking
George	2	3:23	3:47	3
Alana	3	2:56	3:12	1
Sebastián	2	2:54	3:36	2

Table 2 is a bit more complex, and because of that, the data is not quite as accessible to the reader. But the data is complex as well, and if it's going to be displayed for ease of review, this seems like a decent choice. Table 2 shows the results of a survey where participants were asked to rank eight speakers (a male and a female of different English Language dialects) according to their professionalism, intelligence, education, friendliness, and sociability. The response rates are shown for male and female speakers, and then an average is calculated for each dialect.

Table 2. Average Perceptions of English Speakers*

Standard American English					
Gender	**Professional**	**Intelligent**	**Educated**	**Friendly**	**Extroverted**
Female Speaker	5.83	5.83	5.75	5.42	4.92
Male Speaker	6.92	6.67	6.75	6.42	6.33
Southern American English					
Gender	**Professional**	**Intelligent**	**Educated**	**Friendly**	**Extroverted**
Female Speaker	5.75	5.17	5.00	7.25	7.00
Male Speaker	4.33	4.17	3.75	5.92	6.42
British English					
Gender	**Professional**	**Intelligent**	**Educated**	**Friendly**	**Extroverted**
Female Speaker	7.50	7.33	7.33	5.50	5.25
Male Speaker	6.50	6.25	6.17	5.17	4.92
Australian English					
Gender	**Professional**	**Intelligent**	**Educated**	**Friendly**	**Extroverted**
Female Speaker	7.00	6.92	7.08	6.25	6.42
Male Speaker	6.92	6.92	6.75	6.17	6.00
*Participants in this survey were asked to rate speakers on a scale of 1–10.					

Tables help you manage more complex sets of data. A table can be used if you're looking to display individual values, if values are being compared, or if data is going to be shown and then summarized. They won't convey your story to the reader as quickly as a graph might, but you will still be conveying a large amount of information in an easy-to-understand way.

Contributors and Attributions

CC licensed content, Original

- Tables, Charts, and Infographics: Overview. **Authored by**: Freedom Learning Group. **Provided by**: Lumen Learning. **License**:

CC BY: Attribution
- Complex Table. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Stacked Line Graph. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Stacked Bar Chart. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Pie chart. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Pictographs. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Organizational Hierarchy Chart. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Line Graph. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Bar chart comparing healthcare costs as percentage of GDP across OECD countries. **Authored by**: Farcaster. **Provided by**: Wikimedia. **Located at**: https://commons.wikimedia.org/wiki/File:Healthcare_costs_to_GDP_OECD_2015_v1.png. **License**: *CC BY-SA: Attribution-ShareAlike*
- Venn Diagram. **Authored by**: Dflamini. **Provided by**: Pixabay. **Located at**: pixabay.com/en/infographic-web-safe-venn-2722977/. **License**: *CC0: No Rights Reserved*
- Pert example gantt chart. **Authored by**: Dbsheajr. **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/File:Pert_example_gantt_chart.gif. **License**: *CC BY-SA: Attribution-ShareAlike*
- Super Bowl infographic and chicken wing facts. **Authored by**: U.S. Department of Agriculture. **Provided by**: Flickr. **Located at**: https://www.flickr.com/photos/usdagov/24700496851. **License**: *CC BY: Attribution*

Public domain content

- Flowchart. **Authored by**: Wellington Grey. **Provided by**: Wikimedia. **Located at**: https://commons.wikimedia.org/wiki/File:Scientific_Theory_Flowchart.png. **License**: *Public Domain: No Known Copyright*

This page titled 8.1: Tables, Charts, and Infographics is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.1: Tables, Charts, and Infographics by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.2: Matching Graphics and Objectives

Learning Outcome

- Describe factors in deciding which type of visual aid and graphic will best report your data

It can be difficult to determine what graph or chart to use when. Some are for numbers, some are for words... So many charts, so little time!

The flow chart in Figure 1 can help you in choosing the right kind of chart or graph to support your message.

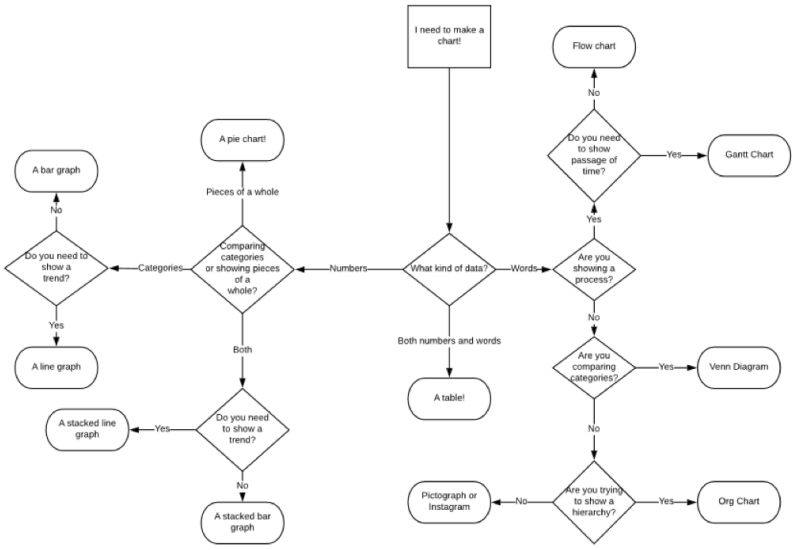


Figure 1. What type of visual aid should I use? Click on the image for a text-only version of this chart.

Now, was that easier than writing it all out?

Contributors and Attributions

CC licensed content, Original

- Matching Graphics and Objectives. **Authored by**: Freedom Learning Group. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 8.2: Matching Graphics and Objectives is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.2: Matching Graphics and Objectives by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.3: Placement, Style, and Coloring

Learning Outcome

- Describe the impact of placement, style, and coloring when incorporating graphics into a message

Once you've decided what graph or chart to use, we need to make sure it fits with our visual media usage standards. It needs to make your message more accessible by being:

- Clean and simple
- Uniform
- Persuasive
- On brand

Let's build a pie chart and apply each of these standards to it as we go along.

Say you want to give a quarter bonus to the best member of each department, and you have settled on giving a $50 gift card for a movie theater. You have surveyed the company to determine which local cinemas are used by your employees to see movies. You have decided to make a pie chart from the results, because you're looking to show the composition of your employees and their movie-going preferences.

At the moment, your pie chart looks like this:

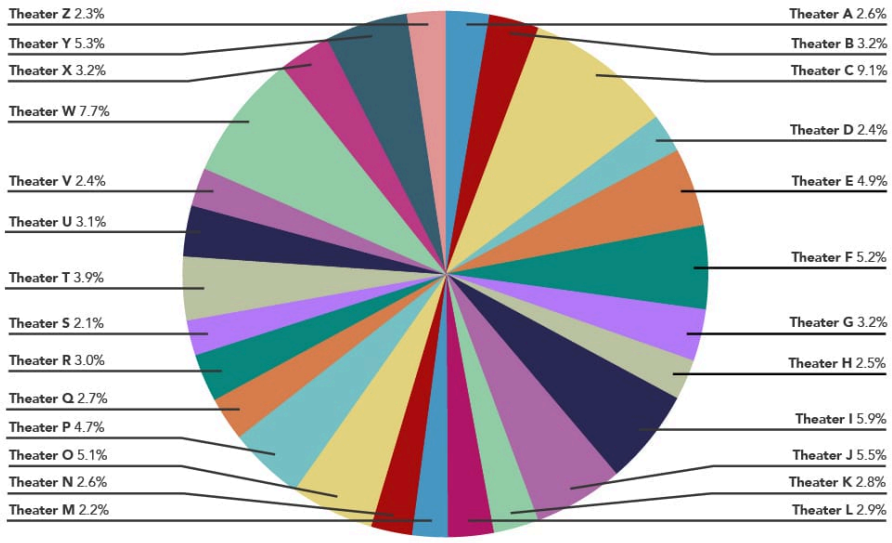


Figure 1. Which employees attend which theaters?

Figure 1 is definitely **not** clean and simple. There is so much to look at here that it's hard to see anything. No one will find this useful. In fact, a rule of thumb for pie charts is that if you have more than ten categories, you should present the information differently. (There's something called an exploding pie chart if you want to check that out.)

So how do we simplify the data? To begin with, we need to narrow down the categories. In this case, you could display theaters by company, rather than location, since all theaters of the same company will take the same gift card. Let's take a look at how the revised chart would look in Figure 2:

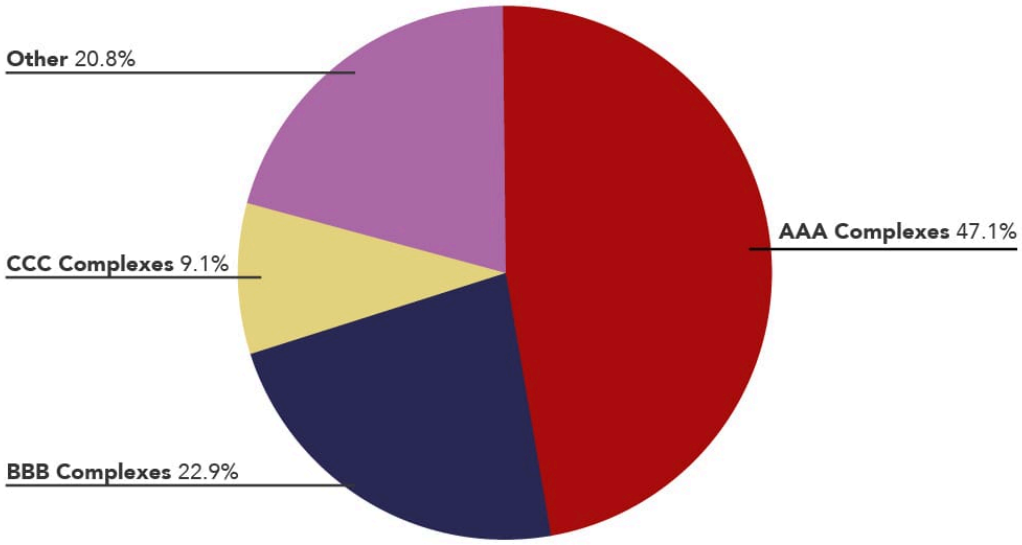


Figure 2. Which employees attend which theaters (by company)?

Now we're getting somewhere! This is a much easier chart to read. We can see at a glance that nearly half of our employees go to a movie theater owned by AAA, and smaller portions see films at the BBB and CCC theaters. This information has become way more accessible for the reader.

We've also chosen some bold, pleasing colors here. We can clearly tell what part of the pie belongs to BBB and what part belongs to the others. This is clean and simple!

Once we've achieved that, we must make it uniform. Do the rest of your charts, tables and graphs use the same bold, pleasing colors we use here? Have we chosen the same font size for our chart key? If the answer is yes, we can move on. We know that we are sharing information without distracting the reader.

Now we ask ourselves, is this data persuasive? Well, that depends on the story you're looking to tell. In this case, your best option is to purchase gift cards for AAA theaters because that is what your data is telling, since most of your employees enjoy seeing movies there. So let's give this a title:

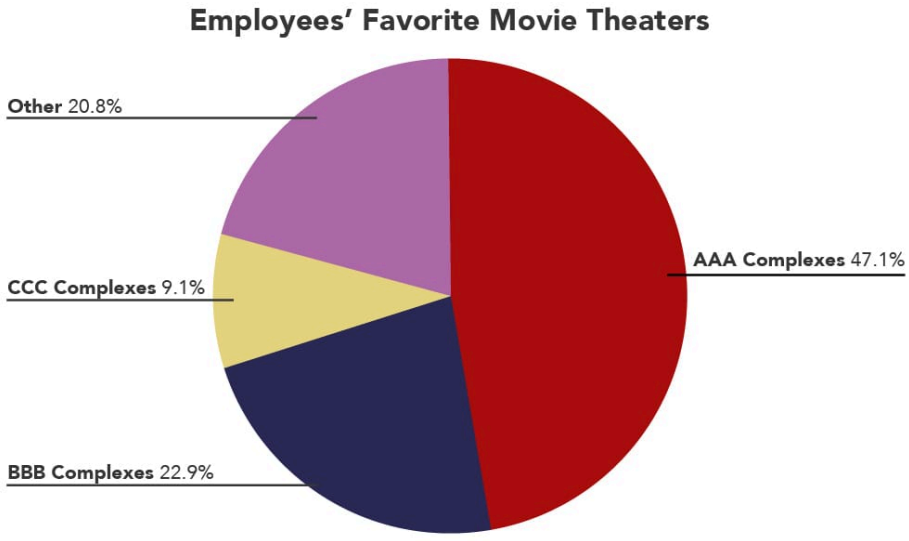


Figure 3: Your completed pie chart

Figure 3 a bit more persuasive. This immediately tells the reader "AAA Complexes are our employees' favorite movie theater."

Now, is your chart on brand? If your company's palette of colors includes blue, red, yellow and green, then yes! We are on brand. If your company uses different colors, go ahead and change them. A lot of companies have very specific color requirements, even for internal projects, so be sure to look out for your company's style requirements! For this project, it's as easy as that.

You've made the point of this communication very easy and accessible by making these changes and following our standards of visual media communication!

Additional Resources

- Misleading Graphs: Real Life Examples
- Good Data, Bad Graphs

This page titled 8.3: Placement, Style, and Coloring is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Freedom Learning Group via source content that was edited to the style and standards of the LibreTexts platform.

- 5.3: Placement, Style, and Coloring by Freedom Learning Group is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.4: Captions and Titles

Learning Outcome

- Describe the impact of descriptive captions and titles when incorporating graphics into a message

Most of the time, putting the data into graph form isn't quite enough. In fact, it's just the start. You created this graph to help you tell the story of your data, and to ensure your message gets across, you need to be clever about the captions and titles you include.

You'll find most communicators title their graphs according to the content they display. It might say "Year-over-year performance" or "Weekly Average Ticket Sales." But if your message is about how weekly average ticket sales are down 10 percent compared to last year, you might consider calling your graph "Average Ticket Sales are Decreasing."

watch it

The video below is a demonstration about how to make a graph tell your story by making it easier to read and making the title active:

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Captions usually indicate the source of information. If your sources and communications are not produced by people within your company, then this is an important step. Captioning the source gives your information credibility and strengthens your story.

You can also use captions to convey other pertinent information. You may want to include the sample size of a survey the graph is illustrating or additional background information about the data (as shown in Figure 1). Using captions in this manner helps the reader draw the right conclusion.

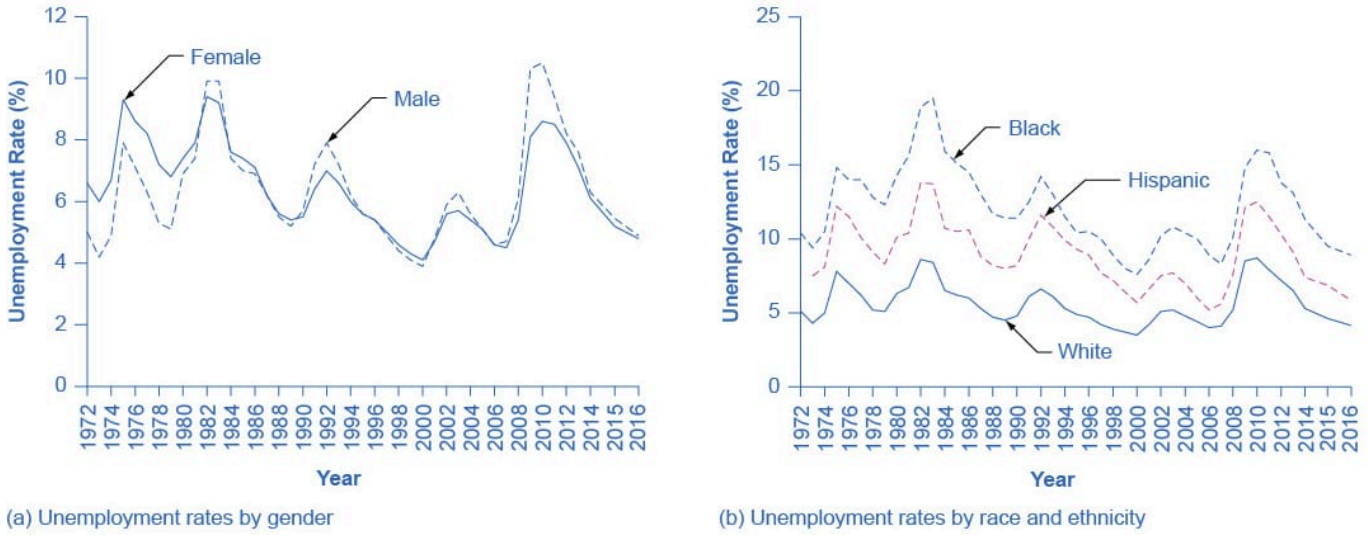


Figure 1. Unemployment Rate by Demographic Group.**(a) By gender, 1972–2016.** Unemployment rates for men used to be lower than unemployment rates for women, but in recent decades, the two rates have been very close, often—and especially during and soon after the Great Recession—with the unemployment rate for men somewhat higher. **(b) By race and ethnicity, 1972–2016.** Although unemployment rates for all groups tend to rise and fall together, the unemployment rate for blacks is typically about twice as high as that for whites, while the unemployment rate for Hispanics is in between. (Source: www.bls.gov)

A caption on a graph can go badly when there is more information captioned than is of interest to your audience, or if information included in the caption would be better displayed elsewhere.

Let's take a look at this line graph displaying the percent of world GDP (Figure 2):

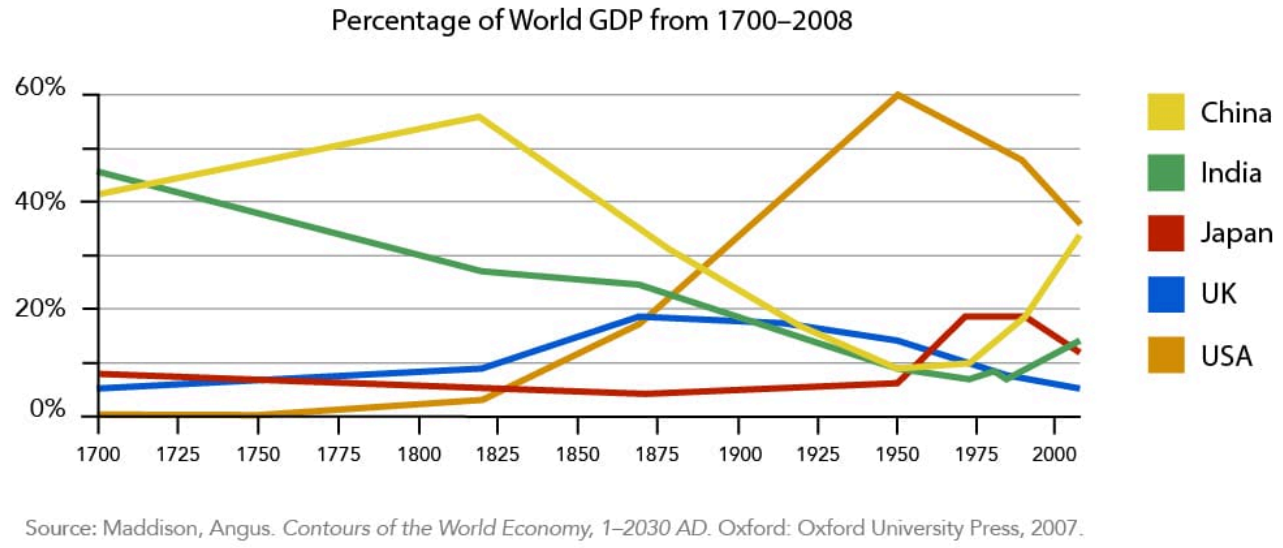
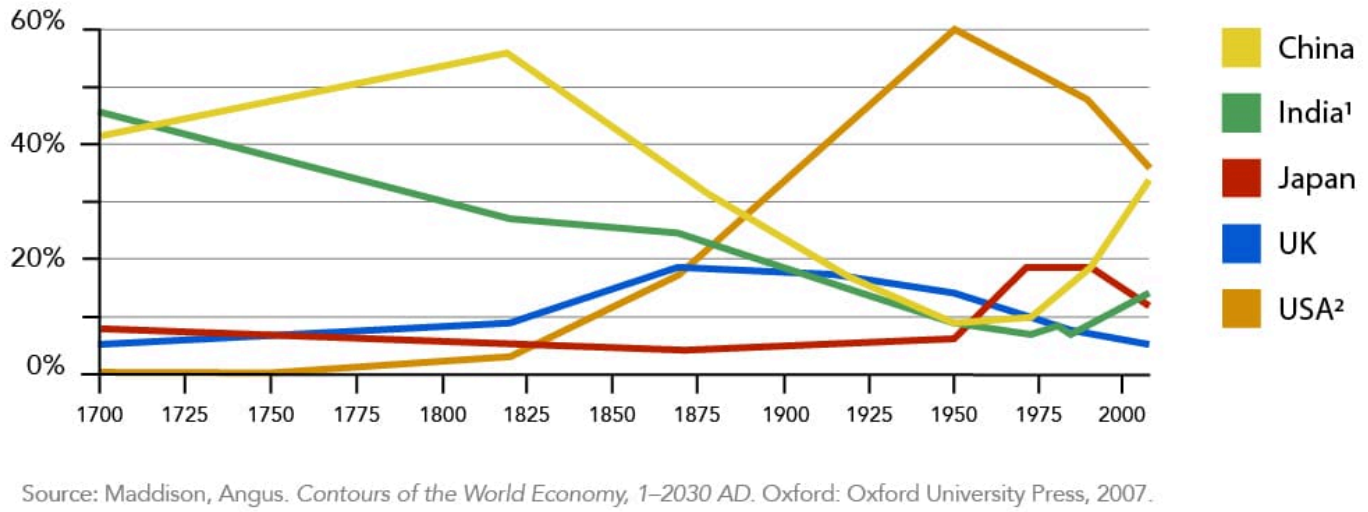


Figure 2. Percentage of World GPD from 1700–2008

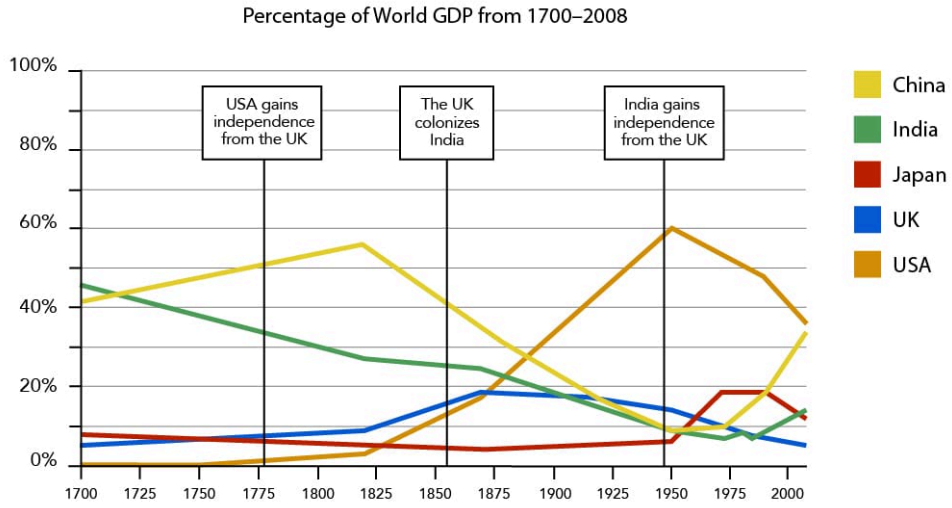
Imagine if the graph above, which shows the percent of the world's gross domestic product from 1700–2000, looked more like Figure 3:



Source: Maddison, Angus. *Contours of the World Economy, 1–2030 AD*. Oxford: Oxford University Press, 2007.

Figure 3. Percentage of World Gross Domestic Product by Country. You will note that (1) India was officially under British rule starting in 1858, when their entire GDP was transferred to the UK. India gained its Independence in 1947 after WWII. (2) The US was colonized by the British and was not its own country until 1776.

Now the title is included in the caption, as well as some information about British colonization. How could we make it a little easier for the audience to absorb that information?



Source: Maddison, Angus. *Contours of the World Economy, 1–2030 AD*. Oxford: Oxford University Press, 2007.

Figure 4. It should be noted that India's GDP was actually considered in the UK's totals from 1858 to 1947.

Figure 4 is even better. Now the graph is titled, and those instances of colonization are marked on the timeline. However, let's imagine the author is focusing on India in their report or presentation. In that case, the note about the colonization of the United States isn't quite relevant and should be trimmed out (Figure 5).

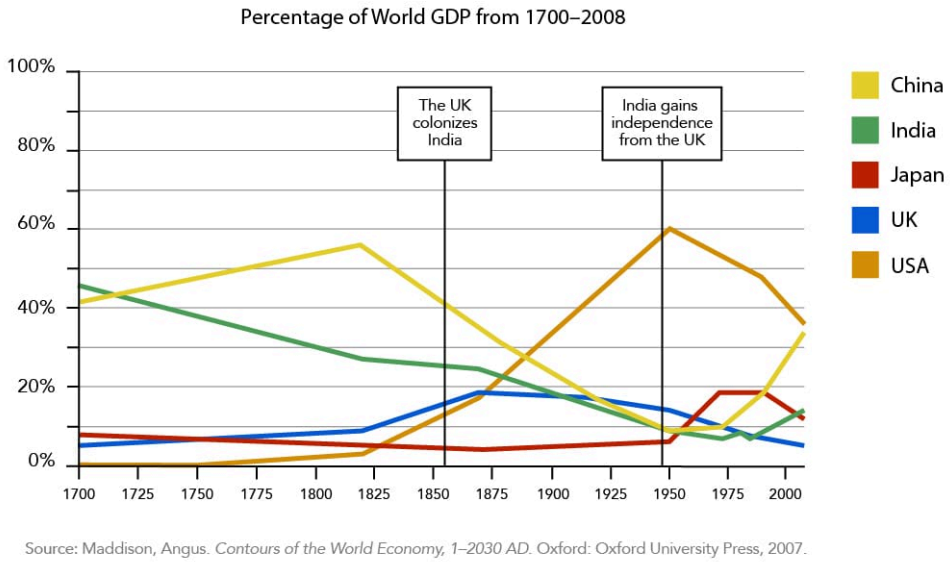


Figure 5. It should be noted that India's GDP was actually considered in the UK's totals from 1858 to 1947.

Much better! Now the graph shows only the information relevant to the point the creator is trying to make.

Visual media should always make a point clearer, so make sure your graph's format, titles and captions are working for you rather than against you.

Contributors and Attributions

CC licensed content, Original

- Captions and Titles. **Authored by**: Freedom Learning Group. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Unemployment graph from Patterns of Unemployment. **Provided by**: OpenStax College. **Located at**: https://cnx.org/contents/vEmOH-_p@4.44:82qMsJce/Patterns-of-Unemployment. **License**: *CC BY: Attribution*. **License Terms**: Download for free at http://cnx.org/donate/download/40061c...cb31@10.49/pdf

This page titled 8.4: Captions and Titles is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.4: Captions and Titles by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.5: Introduction to Contemporary Visual Aids

What you'll learn to do: Discuss the use of video in business messages

Every day, hundreds of hours of video are uploaded to YouTube, and a millions of hours of video are watched. Impressive, to be sure, and while it doesn't mean that your audience is out on YouTube searching for business communications, it does mean that they've grown accustomed to, or have even established a preference for, consuming information delivered on video.

Luckily, videos are amazingly easy and inexpensive to make these days. A communicator can create and deliver a video using nothing more than a phone; or, with minimal investment, purchasing a few microphones and a camera with a decent lens can create a more polished, higher-quality product.

If it's true that facts bore and stories sell, then perhaps your message is better delivered in the format of a video. In this section, we'll talk about the kind of message that works best for a video and how to put together your first blockbuster communication hit.

This page titled 8.5: Introduction to Contemporary Visual Aids is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Freedom Learning Group via source content that was edited to the style and standards of the LibreTexts platform.

- 5.5: Introduction to Contemporary Visual Aids by Freedom Learning Group is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.6: Video as a Visual Aid

Learning Outcome

- Discuss the pros and cons of using videos as a visual aid

Video helps you tell a story in the most visually engaging way possible, giving every employee in the company a face-to-face opportunity with the CEO, or allowing for the broadcast of team meetings and gatherings. Making a video that features the employees who work at your company can "humanize" the company's image, a valuable tool when communicating with investors and other external stakeholders.

Videos are an excellent visual media choice when communicating things like,

- The features of a new facility or office the company has opened
- The details of a new product or service the company has introduced
- Instructions for a new company process, like signing up for benefits or a new 401k plan
- The introduction of a new business idea, plan or merger, especially if the subject is complex or the audience is highly emotional about the announcement
- Webinars and meetings that all attendees might not be able to attend in person

While videos can be easy and inexpensive to produce these days, it can still be costly to create a professional, polished video. A video made on your camera phone likely won't be appropriate for any medium besides a short post on social media. Beyond the cost and talent associated with creating more complex videos, there are a few roadblocks you might encounter, even when creating a short clip. You might want to reconsider video as a choice in any of the following scenarios:

- Your human subjects are visibly uncomfortable in front of a camera and cannot deliver a message effectively in that manner
- Your subject requires the display of a lot of data, and the audience will require time to review, contemplate and study the information
- Your video is longer than fifteen minutes and viewers are likely to tune out after a while
- You're covering a sensitive topic or the topic of discussion shouldn't be made public in any recorded format, written or visual

Video can be used to accompany text, or it can stand alone as its own communication. Consider where your audience will access the video, what information will accompany that video and in what format, and how they'll work together when you start to plan the creation of your video.

watch it

Take a look at this video "What is the Best Explainer Video Style for Your Business?" and the accompanying article *"How Our Explainer Video Got to Rant #1 On Youtube (real case study)"* by Juan Jose Mendez.

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

This page titled 8.6: Video as a Visual Aid is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Freedom Learning Group via source content that was edited to the style and standards of the LibreTexts platform.

- 5.6: Video as a Visual Aid by Freedom Learning Group is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.7: Designing a Video for Your Needs

Learning Outcome

- Describe the process of planning, designing, and producing a video as a visual aid

If circumstances present themselves, you may want to consider creating your own video as a visual aid. This means you might need to wear multiple hats—screenwriter, director, and producer.

Let's imagine a scenario where you work for a pet supply retailer, and you've been asked to put together a video that helps your human resources department recruit the right individuals for positions in the retail stores, distribution center, and corporate office. The video should cover four or five different elements, is only a couple of minutes long, and has a medium level of complexity. Feel free to skip steps as you see fit, based on the complexity of your project. Let's get started!

Writing the Story

Videos are a storytelling tool, so first you need to determine what story you want to tell. You do this by:

- **Interviewing stakeholders**. Always sit down with a subject matter expert (SME) to determine what needs to be covered in the video. In order to determine what you want to advertise about your workplace, sit down with your recruiters and find out why people come to work at your company. In the case of the pet supply company, you find out that people enjoy working for the company because they love animals, they like the people they work with, and they get to learn great things. They also enjoy interacting with their customers and the community. Good information! These ideas and concepts will be the building blocks of your story.
- **Assembling a storyline or storyboard**. If you're going to highlight these qualities in a video, you need to put together the outline of a story. In the video it's a good idea to start out with a couple of broad statements about what it's like to work at the company. Then you'll cover each one of those reasons people work for your company (loving animals, great teams, etc.) a little more in depth. Finally, you'll end with a couple of thoughts that reinforce that your company is a good place to work.
- **Writing a script.** If you have very specific items that you want to capture, this would be the point at which you'd assemble a script. Sticking with the pet supply example, you're going to go out on the road with your camera and interview people who work for the company. It's better to let them tell you in their own words what they like about working at the company, and maybe ask a couple of questions that lead them to talk about loving animals or enjoying their interactions with customers. However, if you want a very polished, smooth video (for sales purposes, let's say) you might want to write a script. If you do, read it through out loud a few times to make sure it sounds as good as it looks.
- **Choosing people to interview**. Once you know what you want to do, send out a message to your field leaders in all areas of the business. Those leaders have a really good idea which of their employees would enjoy the opportunity to talk about their work. They'll assist you in choosing employees who would interview well and have great things to say.

Getting Ready to Film

Now that you have a basic idea of what kind of story you want to tell and how you're going to tell it, you're going to pack up your equipment and get on the road. The things you will need:

- **Your camera**. Today, most decent digital cameras have video capability. Make sure you have a good camera and lens, particularly if your final product is going to be shown on a large screen. A good DSLR (digital single lens reflex) camera is more than adequate to capture high-quality digital video. Most cameras are pretty point-and-shoot friendly, but read the instruction booklet carefully if it's your first time using the camera. Make sure you have a cloth for cleaning the lens, and make sure your batteries are charged!
- **Your microphones**. Good mics, even wired or wireless lapel mics, are fairly inexpensive. Pack a couple to take along with you so you can capture the best sound quality possible. If you're following a group of people around from place to place and trying to capture sound, a horn mic is an excellent option that won't restrict your sound capture. Pack a couple of batteries for each device so you're not caught off guard. (Note: it's always good to have a second source of sound in case a microphone fails. Videographers often use handheld sound capture devices to record while they're filming, to keep as a backup).
- **Your tripod**. Find a good, sturdy tripod that allows you to keep the camera still while shooting, and pan (move) the camera

right, left, up and down. Even though it's a little heavier, purchase a tripod that has some heft so the camera's not easily knocked over. If you want to be able to move around more, look for tripod tools like a dolly (wheels for the tripod) or a monopod, which is a one-legged post/pod that helps stabilize your camera while you're filming but allows you to move more easily and get into tighter spaces.
- **Your script and notes**. Don't forget to write down the questions you want to ask or the script you want your subjects to read/act out. Look it over a few times in advance and familiarize yourself with what needs to happen while you're onsite at the video shoot.

You can also consider things like lighting and make-up if you're going for a more professional look. You can contract with a video company to do the actual filming if you prefer to hand those details off to an "expert." OR...you can head out with your mobile phone and forget all the extras.

Filming Your Video

You've arrived at your filming location, and the person your interviewing is ready to go! You'll need to get your equipment set up and then start your filming. Remember to

- Scope out a quiet but well-lit place to do your filming. Today you're interviewing a store team member, so you may want to film in one of the aisles of the store, but find one that doesn't get a lot of foot traffic!
- Make sure your camera is securely on the tripod
- Test all the equipment to make sure it's working properly. Check battery levels. Do a couple of test shots and play them back to see how they look and sound.

Then, it's time to film. Turn on the camera and start asking your questions. Some good hints to follow:

- Ask your questions and get answers the first time around. When an interviewee is telling a story, he or she is almost always more animated the first time through. If you ask the interviewee to repeat it, the story will probably not be as entertaining! If he or she is nervous, you can tell them that you're practicing, and that usually puts the speaker at ease.
- Coach your interviewees to repeat the question in the answer, so you can edit yourself out of their responses. If you ask, "Why do you like working here?" the interviewee should respond with, "I like working here because..."
- Shoot your subject from the waist up with a lot of background all around him or her. You can always close in on the subject when you edit, but you'll never be able to get a longer shot!
- Capture a little bit of the room's ambient noise when no one is talking. Recording five or ten seconds of silence will help you fill in dead space if you have to edit out a noise or a sound. Every room has its own special noise—an HVAC unit running in the background, a radio playing softly—and when it's missing from the background you notice!
- Film in short spurts. Don't create a 20 minute file. Digital files that are three and four minutes long are much easier to view to determine if there's any useful footage.
- Capture more footage than you think you'll need! Especially when you're shooting footage to compliment the video. That's called "b-roll" in the video business, and it constitutes the action shots you see when someone's being interviewed. The interviewee might be talking about helping customers in the store, and you see the interviewee in action helping a customer while she's talking. That's b-roll. If you know you're going to need b-roll for your video, be patient and film a lot of it. You always need more than you think.

Editing and Post-Production

You've filmed your video, and now you're going to put it together. This is where you add the style and pizzazz that makes your video engaging. Let's get started.

- **Organize your footage**. Take a look at your footage and determine which files you'll be using for your final product. Go back to your notes, review your story structure, and assemble your digital files so that you know what footage you will use in what part of your story.
- **Start placing your chosen footage into an editing program**. Import all the footage into your editing program and start putting your story together. Determine where you can use graphics to break up sections and where you will need to add text to help tell the story or break up the sections. In this video you're making for the pet supply company, you've decided to call out the "pluses" of working there, which include loving animals, a great team of people, opportunities to learn, and so on.
- **Pick out some video music**. This is important and will be an influencing factor in your final product. Often you're going to want to make cuts and add transitions based on the music, so now's the time to pick a song and set the mood of your video!

Check out the next section for some excellent places to buy royalty-free music.

- **Edit**. If you've never edited a video before, please be sure to take some time to play around and see what the program will do. Go online and check out YouTube videos that give hints and tips for the program you're using. But the best way to learn is to just play with the software and experiment. You'll learn by doing!
- **Add company graphics and branding elements**. Make sure you include your company logo and the elements of the brand that make your company identifiable!

Once the editing is complete, circulate it around to your subject matter experts for approval!

Assessing the Final Product

Take a look at this final product and let's determine if it fits our standards of good visual media:

A link to an interactive elements can be found at the bottom of this page.

Let's take a look at each standard separately:

- **Is it clean, clear and simple?** Yes, you've established a theme up front that there are "pluses" to working for Pet Supplies Plus. Your speakers are featured prominently in each of your shots, and you have nicely framed b-roll shots.
- **Is it uniform?** Absolutely! Your titles of "Plus the opportunity to learn" and "Plus the uncommon" teach the audience what to expect right up front.
- **Is it persuasive?** Yes. Not only is the audience getting short, clear soundbites from interviewees talking about the reasons why they should be rushing out to apply for a job there, but they're being emotionally engaged with the kittens and puppies. There was a pig with sunglasses for heaven's sake!
- **Is it on brand?** That's the most important thing! Pet Supplies Plus prides itself on delivering an outstanding customer experience, and it's clear in the theme of this video that there are outstanding experiences happening in their stores daily. Their logo and their tagline are featured at the end.

This video makes the working experience at this company very accessible to prospective employees. Would they have gotten the same interactive experience if they'd been reading a website or an article? Probably not. This video brought the work experience to life. And that's what video can do for you as a visual media: it gives your story life like no other media can.

This page titled 8.7: Designing a Video for Your Needs is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Freedom Learning Group via source content that was edited to the style and standards of the LibreTexts platform.

- 5.7: Designing a Video for Your Needs by Freedom Learning Group is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.8: Products and Resources for Creating Videos

Learning Outcome

- Describe the process of planning, designing, and producing a video as a visual aid

As we mentioned a couple of times earlier in this module, it's not legal to snag things off the internet and add them into your videos without checking to see who owns the rights and whether there's a cost attached. This goes for images, music, and even video. But, rest assured, there are good songs, video and editing tools you can use for free or cheap.

Music

- Melody Loops. Melody Loops is a low-cost source of royalty free music, and they have a huge database of music you can search by type or even "feeling." (Want happy music? Just search "happy"!) You can download samples of the music to test them under your video before you download the actual sound file, so you can make sure it works before you purchase.
- Incompetech. Incompetech is another royalty-free music site with lots of great tunes to choose from. Again, there's a minimal cost for the music. All the songs on the site are written by one person, and he likes to be credited on your work, but he's talented and it's worth it!

Video

- Internet Movie Archives. If you're looking for old film, interesting film clips...well, there's a world of good videos on this site. From old instructional videos to entire feature films, these are all public domain and can be downloaded.
- National Parks Multimedia. The National Parks have an archive of video footage from nearly all their parks that you can download.
- Wikimedia Commons. Wikimedia Commons offers all of its images, video and audio footage to use, but follow their reuse guidelines so you're giving credit where credit is due and following the licensing rules for reuse of the content.
- Vimeo. You can search Vimeo for videos tagged "Creative Commons." There are hundreds of pages of video to choose from.

Editing Programs

There are many video editing programs on the market today. Adobe Premiere and its easier-to-use cousin, Adobe Premiere Elements, are always a good choice in editing programs, and they're not terribly expensive. A serious editor might also check out Final Cut. If you're looking for other alternatives, you can look at these:

- Lightworks. Lightworks is an excellent program, free to everyone even though it's been used to edit feature films like *The King's Speech* and *Road to Perdition*. They likely used the Pro license, which comes with a price tag, but the free version will allow you to do a lot of basic and even professional editing. In fact, it won an editing and technology Emmy Award in 2018.
- Shotcut. Shotcut is a free, open source video editor that works pretty well as far as basic editing goes. It may be challenging to learn up front, but there are many tutorials on YouTube. If you want to produce something that looks really good without buying an editing program, then it's worth the effort of conquering the learning curve with this program.
- Rawshorts. Rawshorts is an animation program. If you have no drawing skills and you think that your message is best delivered in animated form, check out Rawshorts.

This page titled 8.8: Products and Resources for Creating Videos is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Freedom Learning Group via source content that was edited to the style and standards of the LibreTexts platform.

- 5.8: Products and Resources for Creating Videos by Freedom Learning Group is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.9: Introduction to Accessible Visual Aids

What you'll learn to do: Identify ways to make information more accessible to your audience

A good communicator will consider their audience when preparing a written communication **and** when choosing the visual media that will accompany it. In this section we'll take the time learn how to ensure our audience can consume and use visual aids, no matter their physical abilities.

This page titled 8.9: Introduction to Accessible Visual Aids is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Freedom Learning Group via source content that was edited to the style and standards of the LibreTexts platform.

- 5.9: Introduction to Accessible Visual Aids by Freedom Learning Group is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.10: Accessibility Matters

Learning Objectives

- Discuss the importance of making your media accessible to the widest audience possible
- Identify techniques and tools to make visual media more accessible to your audience

When giving a presentation or writing an email you are always trying to convey a specific message to your audience; not just a portion of your audience—all of your audience. Your audience will consist of individuals who might have a variety of needs and abilities. The best way to reach the widest audience possible is by thinking inclusively and creating accessible visual media.

Accessibility means different things to different people, and an effective communicator strives to create documents and presentations that are usable by people with the widest possible range of abilities, operating within the widest possible range of situations. This means making things accessible to all people (whether they have a disability or not).

To be inclusive of all audiences, it is important to keep in mind any factors that could pose as a barrier to understanding the message. Individuals in your audience may have auditory, cognitive, neurological, physical, speech, or visual disabilities or special needs that could adversely impact their comprehension of a message (we will discuss disabilities in Module 13: Social Diversity in the Workplace). Accessibility also benefits people *without* disabilities, for example:

- Older people with changing abilities due to aging
- People with "temporary disabilities" such as a broken leg or lost glasses
- People with "situational limitations" such as trying to read in bright sunlight, external noise such as music or a continuous car alarm, or in an environment where they cannot listen to audio

Guidelines for Making Media Accessible

The first step to making your media accessible is to assess its **purpose** or **value** to your presentation or written message.

Consider the following questions:

1. Does your image or video serve a **functional** purpose? In other words, is it conveying non-text content to the audience? If so, you should:
    - Provide a text alternative that serves the equivalent purpose of the non-text material
    - Not use color as the only visual means of conveying information
2. Does your image serve more of a **decorative** purpose? In other words, is it primarily a design element that does not convey content? If so, you should avoid unnecessary text descriptions.

When putting a presentation together consider what your content page would look like if the images didn't load or someone was too far away to see specific details. One way to reduce losing information or alienating an audience member is to write alternative text for each image, which would work as a replacement and provide the same service as the image. For guidelines on writing alternative text, visit this link on Images from the Accessibility Toolkit.

Another thing to consider when using visual media is what your images would look like if they displayed only in black and white. Would any necessary context or content be lost if the color was "turned off"? For example, have you ever used a black and white laser printer to print a web page? Many details of the images or text are easily lost, and that can affect how your reader interprets your data and whether they can understand the point you are trying to make. Images should not rely on color to convey information, so you must design that PowerPoint slide or image page to work in black and white, otherwise you'll lose your audience if they can't view it in color. If the point you are making depends on color to be understood, you may need to edit your image or formatting so that concepts presented are not lost to those who are color blind or who require high contrast between colors. Your text should also not rely solely on color to make distinctions—use bold, italics, underline or a different font to highlight the important information.

Tools for Accessibility

Including visual media in your presentations, emails, or professional training materials can help engage the audience and help them understand the message more efficiently. Yet visual media, if not presented in conjunction with accessibility techniques and tools, can do the exact opposite and can instead alienate audience members with visual, hearing, or cognitive disabilities. Visual media, such as graphs, charts, photographs, videos, or instructional diagrams, that do not take the needs of impaired individuals into consideration, might deny users with disabilities the opportunity to really understand and feel what you're saying.

Screen readers, refreshable Braille displays and closed captioning tools are among the digital tools out there today that assist those with disabilities. When choosing visual media, a good communicator should keep in mind how their materials will be understood by people with disabilities.

Images, Tables, Charts and Alt-Text

Images, tables and charts can be particularly challenging for sight-impaired audience members. A screen reader is a tool that will read the alt-text included with any visual media. A downside is that a screen reader can only read the verbiage provided, it cannot interpret the visual media on its own. Charts are notorious for having very limited alt-text descriptors. In the video example below the screen reader does not see an X-axis or a Y-axis. The screen reader does not recognize the numbers or data represented in a graph if they're not included as text. It will not recognize trends or provide any kind analysis. Captioning the chart can help get to the underlying point, but it certainly doesn't allow the sight-impaired audience member to study the data and arrive at a conclusion.

Watch It

Watch this video to experience a screen reader ineffectively describing charts and graphs that have not been made accessible. Unfortunately, this is a typical experience for a screen reader user:

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Video

Often, our hearing impaired audience members need additional assistance in comprehending what's going on in a video. Voice-over narration and music are among the things that convey information and feeling to a hearing audience member but will go undetected by the hearing impaired.

Captioning is the first and most obvious way a communicator can make sure a hearing impaired audience member understands a video. It can also benefit a non-native speaker, a viewer who has trouble understanding the dialect or accent of a speaker, or even just a viewer watching the video in a noisy environment.

There are vendors that provide video captioning services. The National Association of the Deaf provides a comprehensive list of vendors who will caption videos. However, a communicator with video editing software and a little bit of savvy can do it on their own. Some captioning standards to keep in mind:

- Captioning should not exceed three lines of text at a time
- Captioning should be synchronized with the spoken word–no faster or slower
- Captions should not cover up text or other important information on the video
- Captions should identify who is speaking when multiple speakers are present (as in an interview)
- Captions should use a sans-serif font such as Arial or Calibri
- Non-speech sounds should be indicated with brackets, like [applause]

Transcripts of videos can also provide hearing-impaired audience members with more accessible information from your video. One tool for creating a quick transcript of your video is to upload your video to YouTube and use its auto-generated subtitles. See the following video for a tutorial on how to use this free feature.

Watch It

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Contributors and Attributions

CC licensed content, Shared previously

- Universal Design. **Authored by**: Amanda Coolidge, Sue Doner, and Tara Robertson. **Located at**: pressbooks-dev.oer.hawaii.edu/accessibilitytoolkit/chapter/images/#Functional. **License**: *CC BY: Attribution*
- Accessibility. **Provided by**: Wikipedia. **Located at**: https://en.Wikipedia.org/wiki/Accessibility. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 8.10: Accessibility Matters is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.10: Accessibility Matters by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.11: Why It Matters- Visual Media

Our prehistoric ancestors struggled, rock in hand, to carve just the right picture into the wall of the cave, hoping that someone would come along and understand the story being told. They may have drawn stick figures with spears, a woolly mammoth in the distance about to become dinner. In the next series of drawings, the mammoth has been struck, the stick figure looming over him. In a third series, a gathering of stick figures at a fire, presented with a feast.




There is beauty and clarity in the simplicity of the prehistoric man's drawings. You understand his hunting victory even though he's used no words. Prehistoric man was on to something with this approach.

Fast forward to today where studies have been done on the impact of visual aids:

- Only 70 percent of people reading labels on medicine containers understand the instructions when they are written in text form, but that understanding increases to 95 percent when the text is accompanied by images.[1]
- People follow written directions 323 percent better when the images accompany the instructions.[2]

From this, we know that visual media can make your communication easier to understand. But does it help you make a point or sell an idea? When listening to an oral presentation, 50 percent of an audience will be persuaded by the speech alone, but that number increases to 67 percent when the speech features visual aids.[3]

Because of this, we know that visual media can make your communication more convincing. In an age in which we humans are asked to process more information than we have ever processed before, let's look at today's trends:

- People are 80 percent more likely to engage with content when it features color visuals.[4]
- Facebook audiences are 651 percent more likely to engage with a post if it includes an image, compared with posts that don't.[5]

This shows us that visuals attract attention and draw our audiences in. Visuals make our communication noticeable.

Communications are more memorable, persuasive and easily understood when visual media is involved. Like prehistoric man, we can use visual media to more effectively deliver a message. In this module, we're going to talk about the uses of various visual media—charts, graphs, images, and even video—and how communicators can leverage them to connect with their colleagues and other professionals.

---

1. Dowse, R. & Ehlers, M. (2005). Medicine labels incorporating pictograms: Do they influence understanding and adherence?,

Patient Education and Counseling, Vol 58, Issue 1.
2. Levie, W. J. & Lentz, R. (1982). Effects of text illustrations: A review of research, Educational Communication and Technology.
3. Wharton School of Business. 'Effectiveness of Visual Language'.
4. Green, R. (1989). The Persuasive Properties of Color, Marketing Communications.
5. Social Intelligence Report, Adobe Digital Index Q4 2013

Contributors and Attributions

CC licensed content, Original

- Why It Matters: Visual Media. **Authored by**: Freedom Learning Group. **Provided by**: Lumen Learning. **License**: *Public Domain: No Known Copyright*

CC licensed content, Shared previously

- Bison Cave Painting . **Authored by**: janeb13. **Provided by**: Pixabay. **Located at**: pixabay.com/en/bison-cave-of-altamira-1171794/. **License**: *CC0: No Rights Reserved*
- Cave Painting Nevada USA. **Authored by**: PDPhotos. **Provided by**: Pixabay. **Located at**: pixabay.com/en/cave-paintings-mural-usa-nevada-3699/. **License**: *CC0: No Rights Reserved*

This page titled 8.11: Why It Matters- Visual Media is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.11: Why It Matters- Visual Media by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.12: Introduction to Using Visuals

What you'll learn to do: Incorporate meaningful visual media in business messages

We've already learned what a great visual looks like: it's clear, clean and simple, uniform, persuasive and on brand. We know we have graphs, charts, images, and even video at our fingertips, all great tools to help us communicate our story. Now, all we need to do is incorporate these visuals into our various methods of communication and let them get to work!

This section will explore how visual media and text relate to one another, how to evaluate effectiveness, and how to reexamine the visual media you're using when it doesn't work. We'll look at how businesses use visual media in their reports, presentations, speeches, and other documents. We'll review several examples and judge whether their visuals were used correctly and, if not, what they could have done better.

This page titled 8.12: Introduction to Using Visuals is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Freedom Learning Group via source content that was edited to the style and standards of the LibreTexts platform.

- 5.12: Introduction to Using Visuals by Freedom Learning Group is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.13: Visuals in a Report

Learning Outcome

- Identify appropriate and professional visual representations of information for a business report

Reports don't end in high school, unfortunately. In fact, businesses that are publicly traded usually prepare an annual report each year for their stockholders and investors. That annual report might be a hundred or more pages of highly detailed information, including strategic plans and financial data, and will certainly include visuals.

Let's take a look at a few annual reports from well-known companies. We're going to study what they've written, the visual media they've chosen to use to support their written word, and if that choice of visual media meets our visual media standards of clarity, consistency, relevancy and persuasion.

Nestle Global: Images in Reports

Nestle is the world's largest food and beverage company, and their mission of "Good Food, Good Life" is to provide consumers with the best tasting, most nutritious choices in a wide range of food and beverage categories and eating occasions, from morning to night.

The cover of Nestle's 2017 Annual Review is subtitled, "Nestle. Enhancing quality of life and contributing to a healthier future." Do the images they choose convey that idea? Here are examples from several pages of the report:







Are these images clean, clear, and simple? They are simple images, with very little to look at except the young people or the products themselves, so we can say they are clean, clear, and simple.

Are they uniform? The two larger shots feature the children shot from about the knees up, with only a small amount of background around them. The smaller group of four images on the middle page feature three products shot at about the same distance, with similar shot compositions (the product and other items used with that product). The fourth picture breaks that pattern–and draws your attention–by showing some young people enjoying a fourth product. Finally, all four images are square with two curved corners on the right and 90 degree corners on the left. These images are consistent and uniform.

Are they on brand and relevant? These images show a series of products and happy young people. If they're trying to convey a good quality of life, they seem to have done so. These children are well dressed, playing sports, enjoying the benefits of friendship. And the fact that these subjects are all young people drives home the idea that Nestle is contributing to the future. So we can say these images are relevant and on brand.

Are they persuasive? We can decide that by measuring how the image supports the message in the text. Let's look at the first one, "Our Strategy." The text says,

Through enhancing quality of life and contributing to a healthier future, we aim to deliver sustainable, industry leading financial performance and earn trust...

In the photo we see two young girls, a symbol of our future. They're worry free and having a great time. One of them is carrying a Nestle product. Does the image help us feel what Nestle is saying in the text? Yes. This image is persuasive.

The second large photo talks about innovation for a changing world. The text reads,

At Nestle, continuous innovation is part of our DNA. Our success is founded on over 150 years of anticipating trends and understanding consumers' needs.

Here again, the future is represented by two young people enjoying a healthy lifestyle of sports and sampling a Nestle product. They're looking at a cell phone, a symbol of modern innovation. Does this image help us feel what Nestle is saying in the text? Yes.

Nestle does an excellent job of incorporating images into their reports.

Target: Charts, Graphs and Tables

"Expect more. Pay less." That's Target's brand promise to its customers. Target is one of the most identifiable brands in the world, and you don't need to look past the cover of Target's 2016 annual report to know exactly what it is and who made it.

Let's take a look page 2 of their report to see how their clean, crisp style is translated into their report's charts and graphs.

The first set of charts shows their financial highlights in a set of four bar graphs that compare five years of sales, EBIT (that's 'earnings before interest and taxes'), net earnings, and diluted EPS (that's 'earnings per share'). The chart is simply labeled, captioned appropriately with earnings and CAGR (that's 'compound annual growth rate'). You can see at a glance that Target's 2016 wasn't quite as profitable as the prior four years. And according to the footnote at the bottom, that's because of the pharmacy sale to CVS.

The second set of charts shows their total segment sales. They've chosen a composition chart to display this information, because they're showing what portion of total sales each department has contributed. Notice that this is actually one pie chart shown five times, each with a different department highlighted.

- **Are these charts clean, clear, and simple?** Very much so. They feature only the information we need to see. The font they chose is easy to read, and the colors stand out.
- **Are they uniform?** Definitely. The styles of the bar charts and the pie charts are the same size, they use the same colors.
- **Are they on brand and relevant?** Most certainly. Target has used their brand colors, they've maintained their clean, crisp style, and their pie charts are the outside ring of the bulls eye. The information is relevant in that, here on page two of the report, they've shown you all the financial data that 90% of readers open the report to find.
- **Are they persuasive?** Yes. This is a display of information, and because it's captioned and footnoted, there's no reason to question it. But where did these numbers come from, and how did they determine which ones to show in the graph? As it turns out, a more in-depth look at their financials is featured on a table on page 4.

All of the information in the financial highlights bar charts on page 2 is featured here in this table, too. This table, which is on brand with its easy to read font and its Target red headers, allows the audience to dig in and really understand the numbers they saw on page 2 of the report. Again, this table follows our visual media standards in that it's clear, consistent, relevant and persuasive.

Target does an excellent job incorporating charts, graphs and tables into its reports.

Microsoft: Video

Technology giant Microsoft's mission is to empower every person and every business on the planet to achieve more. Microsoft's 2017 annual report is online, and because that communication method is so flexible, it allows them to include video right in their letter to shareholders.

The letter opens with a reminder of their mission and proceeds to highlight a variety of businesses, from Boeing to Land o' Lakes to Case Western Reserve University and the Cleveland Clinic. Please try to watch at least three of these videos—they're less than a minute each.

Did you see a few of them? Okay, let's determine if these videos fit in with our visual media standards.

- **Were they uniform?** Each of the videos starts with a picture of the planet and then zooms in on an area. From there, you get a series of visuals that show people interacting with Microsoft products to achieve innovative results. The uniformity of these videos is actually the foundation of their communication effectiveness. Every other element falls in place because of it.
- **Were they clear and simple?** Yes. Videos add a level of complexity by their very nature–there's so much to see! The uniformity of these videos adds an element of predictability, so by the second one, you know right away what to expect and what to look for. That makes the video simple. Add to that the simple audio; no verbal communication is added to these videos, it's just a series of visuals.
- **Were they on brand?** The first image of the planet, followed by the zoom in to the people and businesses using their product, is a visual translation of their mission statement: to empower every person and every business on the planet to achieve more. Their technology is featured throughout, and each video ends with a logo. This is very much on brand and relevant.
- **Were they persuasive?** Absolutely. They visited companies all over the world and showed us amazing things without using one spoken word. Planes were built. Human lives were saved. They have delivered a lot of emotional punch with this series of videos.

Microsoft made terrific use of video in their report and incorporated it very skillfully with the text.

Contributors and Attributions

CC licensed content, Original

- Screenshot of Strategy Visual from Nestle Global. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Screenshot of Products from Nestle Global. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Screenshot of Innovation Message from Nestle Global. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Screenshot of Target Annual Report. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Screenshot of Target Annual Report Numbers. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Screenshot of Target Annual Report Financial Summary. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 8.13: Visuals in a Report is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Freedom Learning Group via source content that was edited to the style and standards of the LibreTexts platform.

- 5.13: Visuals in a Report by Freedom Learning Group is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.14: Increasing Impact with Media

Learning Outcome

- Identify effective use of visual media in presentations, documents, spreadsheets, and messages

Speakers don't need visual media to make a point. Just watch any stand-up comedian, and you'll see that an engaging message can be delivered with nothing but a good story and a funny punchline.

Still, Microsoft PowerPoint and Apple Keynote offer speakers the opportunity to reinforce their messages visually, and, done right, this can have a powerful impact. Still, often, PowerPoint presentations are not done as well as they could be. We'll learn more about making effective slide decks in Module 8: Developing and Delivering Business Presentations.

The PowerPoint Myth

If you Google "PowerPoint bad," you get over ninety million results. It's been fashionable for a while to bash presentation software—especially PowerPoint—as stultifying or boring. But as anyone who's seen the classroom scene in *Ferris Bueller's Day Off* knows, bad presentations didn't start with PowerPoint.

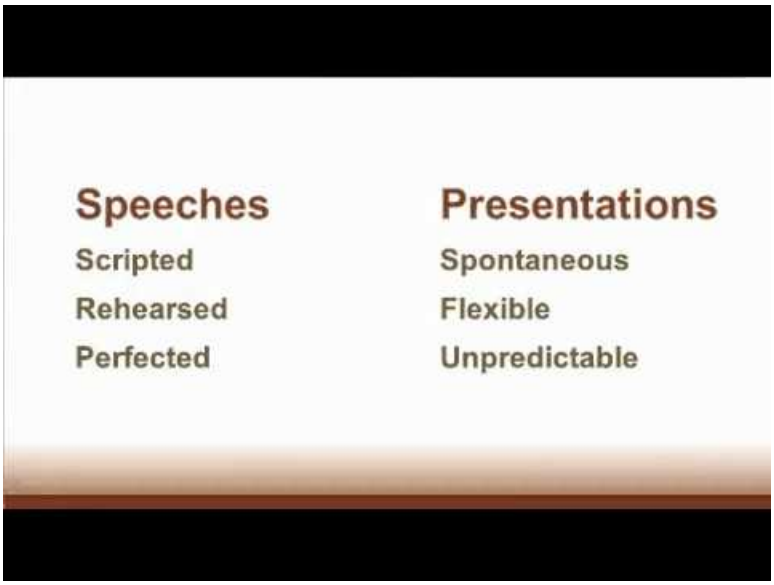
If you try to pound a nail in with the claw side of a hammer—or worse yet, with the handle of a screwdriver—do you blame the tool? No, of course not. Blaming PowerPoint for dull presentations and even duller presenters doesn't get at the core issue. Even when the software could practically construct the presentation for you, you still need to focus on the best ways to present **your** message to **your** audience in the most effective way possible, including relevant visuals as needed.

There are a few key things to remember when creating visuals that you're going to use to present to an audience. If you work on formulating satisfying answers and goals, your presentation should be effective and persuasive.

Everything we've discussed up to now about audience analysis and honing your message applies to the process of creating visuals for your spoken presentations. In addition, you will want to keep two other key points in mind:

1. Unlike the exchange between audience and recipient that happens with an email or report, presentations happen in real time, so you want to be respectful of your audience's time and not waste it.
2. The best thing that can be said about a business presentation is that it was effective and helped everyone in the room do their work. If after the presentation, people are talking about how gorgeous your slides are or how funny you were, rather than about the topic of your presentation, you might not have been focused on the important aspects of your task.

One more critical question to ask yourself is this: are you delivering a speech or giving a business presentation?



A YouTube element has been excluded from this version of the text. You can view it online here: pb.libretexts.org/bcsfm/?p=262

If you are giving a speech—where you expect no back-and-forth conversation with your audience—then you probably don't need visuals at all. Obviously, if you're giving a speech about the details of an ancient Roman sculpture, it helps to show a picture, but speeches are scripted and rehearsed, so you don't absolutely need words or images on your screen for you or for your audience. We'll talk more about speeches in Module 7: Public Speaking.

If you are giving a business presentation, you might need visuals to help both you and your audience stay on track. There are a few rules for this that we can sketch out here, so you'll be prepared for Module 8.

- A brief agenda slide lets the audience know you have a plan in mind. Give them a sense of the big chunks of information you're going to cover so they trust you with their time and believe you know what you're talking about. Deliver the agenda quickly and resist the temptation to elaborate on it. That's what the rest of your presentation is for. Delivering a slide like the one below should take no more than thirty seconds.



- Using images to convey your message can be really effective, especially if the images apply to your content, such as charts, graphs, and pictures of products or displays. Keep in mind that . . .
  - Simple is good. If you need to present only a few data points from a graph that has several, delete or gray out the ones you don't need in order to help your audience focus on what's important.
  - Pictures should be well-edited and as sharp as possible. If you need to show a detail from a larger picture, edit down to

the detail you need and show that. Your audience will mentally check-out if they can't see or follow what you're talking about.
    - A descriptive slide title helps remind your audience of what they're looking at. If an audience member zones out while you introduce the slide, without a slide title, they'll be lost when they come back into focus.
    - Pictures are great for eliciting emotion *if that's what you want to do.* A discussion of how good customer service changed the life of one of your customers is expected to get an emotional response, so use pictures of the customer, of their letter to management, or other images that will get at the emotion you want to generate. However, when you're simply presenting a quarterly earnings report, emotion-generating pictures will be perceived as confusing or manipulative. This doesn't mean you can't have strong feelings about a strong quarter. It just means that the feeling will come through in how you deliver your message rather than through pictures.
- Your slides are for you as well as for your audience. This means that they should help you stay on track and remind you of what you want to say. Therefore,
    - Ignore those who say you shouldn't read your slides. The words you put on a slide should be so brief that your audience doesn't even notice you're reading them. Rather, what's on the slide will remind you of what you want to talk about in detail. As soon as you turn away from the slide and re-engage with your audience to elaborate on your point, they will be there with you—focused and learning. Watch the video following this list for a good explanation of how to read what's on your slides without losing your audience.
    - Putting on your slides the few words you need to stay on track eliminates the need for notes. More than anything, notes get in the way of engaging with your audience and staying on track. Presenters who use notes tend to either get mesmerized by them (especially if they're on a small screen like a laptop or tablet) or they try to toggle between their notes and their audience, which is at best an awkward kind of yo-yoing act that steals focus from the point you're trying to make.
    - **Remember, there is no award given for the longest presentation given with the fewest slides. It is loads better to make more slides with less on each of them. It helps you stay on track, and it gives your audience something new to focus on more frequently.**



A YouTube element has been excluded from this version of the text. You can view it online here: pb.libretexts.org/bcsfm/?p=262

Remember the visual media standards when choosing your supporting images, charts, and video. Images should be very simple and clear, they should still be uniform, and they should support your company's brand and be relevant to the points the speaker is trying to make. Most importantly, they should be effective and persuasive. There's no better way to educate and persuade your audience than to be right in front of them.

If you are going to include images in a file to be sent electronically—a Word doc, a PowerPoint, or even an email—there are a few things to keep in mind. First of all, image files can be so large that they slow down the speed at which your document downloads and/or opens. This can be really annoying for your recipient, especially if they're working on a phone or tablet and using up their data allowance. Sometimes, these files won't even arrive because the system can't handle them. There is a lot of information online about how to reduce image file sizes. Also keep in mind that if your recipient is reading your report or message on a small screen like a tablet or phone, the actual photo dimensions matter. In Module 8: Developing and Delivering Business

Presentations, you can find information on how to resize images. Finally, since images can cause some technical trouble, include them in your documents only if they are necessary to support the points you are making. If they're merely decorative, delete.

This page titled 8.14: Increasing Impact with Media is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Freedom Learning Group via source content that was edited to the style and standards of the LibreTexts platform.

- 5.14: Increasing Impact with Media by Freedom Learning Group is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.15: Revising and Enhancing Visual Media for Impact

Learning Outcome

- Describe the process of revising and enhancing visual media to create impact



Figure 1. Simple and clear visuals can help you communicate complex ideas

During the development of a visual aid, the author will review, change, or amend the visual as a part of the revision process. The end product can benefit from the author's taking the time to pause to analyze whether the visual aid is aligned with the purpose of the message she is trying to convey.This includes the visual aid accommodating the audience's needs and characteristics and providing a persuasive conclusion. The revision process allows for the fine tuning of a draft (or completed project) that will enhance your visual media and ensure the message reaches the intended audience. When revisiting a visual aid, remember the four visual media standards we have explored in this module: Is it clear and simple? Is it consistent and uniform? Is it relevant and on brand? Is it persuasive?

The revision stage is a prime time to receive feedback from someone less familiar with what you have been working on tor receive an outside viewpoint.

Feedback is an important and inevitable part of work, and it comes from the people around you: supervisors, peers, and customers. At some point in your communication career, someone is going to point out some image or graph you chose (or made) say: "I don't like this," "I don't understand this," or "This doesn't match our message."

Whether you are in the development phase or a visual aid has already been published, hearing feedback like this can be frustrating and feel dismissive after spending time and effort on creating a visual aid. If a colleague doesn't like your choice of image or graph, this doesn't mean your choice was wrong. Do not despair because this is an opportunity to make a better visual aid. Asking the right questions to clarify what is missing from your visual aid allows you to use this feedback to revise your visual and potentially communicate your message more successfully to your audience.

When feedback points to how a visual aid isn't working with the text, be curious, not defensive. It is tempting to become defensive or "explain away" the criticism, but resist this natural reflex. Do not debate or try to explain your behavior. Instead, let the other person finish completely and try to listen deeply. Then ask questions with the intent of inquiry:

- If you had to make two suggestions for improving my work, what would they be?
- Do the colors, alignment, or content of the image add or distract from the message or text?
- Is there something confusing or that could be done differently in this visual aid?
- How eye catching or engaging is the visual aid?

- Is it interactive, original, funny, or interesting?
- What is something that works well in this visual aid?
- Which parts of the visual aid are successful unsuccessful and why?

Request examples of what they think a good visual is. Stay curious until you can see how they reached their opinions—even if you don't completely agree. Later, you can decide what you agree or disagree with, but for now, your goal is simply to learn. Reflect thoughtfully on what you've heard.

Just as getting someone to read your writing can make your writing better, getting someone to look at your visual aids will help you create a better product. Asking someone to look over your slides, listen to your presentation, or watch a video you created can seem scary, but mistakes and feedback as *opportunities to grow*, rather than personal failures. Seeing feedback as an opportunity is referred to as a **growth mindset**. Criticism is not an attack on you as a person. It is about something you did. You can't change who you are, but you *can* change what you do.

Additional Resources

- How to give and receive feedback in an agile organization
- Feedback for Teams
- Presenting with Sensory Enhancements

Contributors and Attributions

CC licensed content, Original

- Revising and Enhancing Visual Media for Impact. **Authored by**: Freedom Learning Group. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- How to give and receive feedback in an agile organization. **Authored by**: Manuel Kublbuck. **Provided by**: Quality Software Development with Ease. **Located at**: https://qualityswdev.com/2017/01/22/how-to-give-and-receive-feedback-in-an-agile-organization/. **License**: *CC BY: Attribution*
- Marketing Strategy image. **Authored by**: rawpixel. **Provided by**: Unsplash. **Located at**: unsplash.com/photos/FmuMAvt4Ym8. **License**: *CC0: No Rights Reserved*

This page titled 8.15: Revising and Enhancing Visual Media for Impact is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.15: Revising and Enhancing Visual Media for Impact by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.16: Evaluating the Effectiveness of your Message

Learning Outcome

- Evaluate the effectiveness of a message

If only this were easy! A communicator can produce messages all day long, but they are not effective unless the audience receives them, consumes them, understands them, and (if applicable) provides the requested response.

Trying to pin down evaluative data like that is similar to your cat trying to catch the spot of light generated by a laser pointer. But you can actually gain some understanding through data collection with the right tools.

- **Did the audience receive the message sent?** Companies like ContactMonkey or Politemail provide email tracking services to help a communicator determine how many of your readers opened your message. If you use a collaborative intranet platform to share your messages, a web tracking program like Google Analytics can give you myriad insights as to who is "landing" on your message.
- **Did the audience consume any of the message?** If your message is encouraging a reader to click through to a video or webpage, this can help you understand better if the reader engaged with the content. How many readers clicked through to watch the video?
- **Did the audience understand the message?** This is where it gets dicey. Certainly if you've requested a behavior of your audience and they've responded accordingly, then you know your message was understood. For instance, if you requested that customers update their passwords, you can tell how many audience members understood the message by measuring how many customers changed their passwords. However, if your message was about company strategy, about strengthening employee engagement, or about increasing customer confidence in the company, it's not as easy to gauge the effectiveness of these more nebulous messages. Long-term, you're likely to see the results of these messages (whether positive or negative), but you likely want to know relatively quickly if your message was successful. You can use tools like surveys to determine if you're on the right track.

Contributors and Attributions

CC licensed content, Original

- Evaluating the Effectiveness of your Message. **Authored by**: Freedom Learning Group. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 8.16: Evaluating the Effectiveness of your Message is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.16: Evaluating the Effectiveness of your Message by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.17: Putting It Together- Visual Media

Statistics from different professional fields tell us that visuals support and can even improve communication—that's a compelling reason to use visual media in all your communication. Visual media makes your message more accessible. It helps your audience process the information faster, it can even transcend cultural and language differences. With visual media, your reach is wider and your message more powerful.

You have some different kinds of visual media to choose from:

- Images
- Charts
- Graphs
- Tables
- Video

And you have some uncomplicated standards by which to select just the right type to make your message more accessible:

- Clear, clean and simple
- Consistent and uniform
- Relevant and on brand
- Persuasive

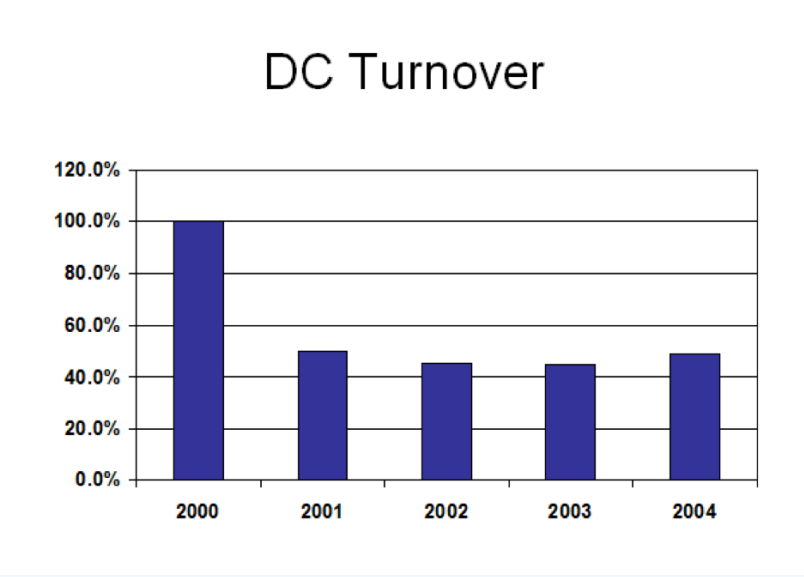
Now it's time to put all of these tools to work for you.

Reducing Turnover

Imagine that your human resources department is celebrating the fact that they have curbed the high turnover rate in your distribution center. Once at almost 100 percent employee turnover, they've been under 50 percent for the last three years. It's time to report to the board of directors how the HR team is doing, and this should be a part of their success story.

In your report you write, "The HR department is pleased to report that turnover has been under 50 percent for the last three years after significant development of items to create value of employment."

But on it's own, this statement seems very flat and one dimensional. To emphasize this amazing success story, you decide to add a chart like this to support the message and give it more impact.



Leverage visual media to tell your story in the most memorable, impactful way possible.

Contributors and Attributions

CC licensed content, Original

- Putting It Together: Visual Media. **Authored by**: Freedom Learning Group. **License**: *CC BY: Attribution*

This page titled 8.17: Putting It Together- Visual Media is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.17: Putting It Together- Visual Media by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.18: Discussion- Visual Media

Effective images function by communicating the intended message of the brand or mission of a company. A visual analysis can help you explore the effectiveness of images used in a business context to begin to see how they might be interpreted differently by customers with different backgrounds.

Part 1: Visual Essay

Your task is to create a visual essay to explore how a topic or idea is communicated through images. A visual essay is a type of visual analysis that explores how images (logos, advertisements, etc.) convey certain messages. These ideas can be at surface level/literal or at a deeper level of meaning that requires the viewer to use context clues and design features to infer the intended message of the image.

What you will do is:

1. Choose one of the following topics or buzzwords: “green” living, fresh, local, savings/discounts, “box store,” online shopping, social media—or choose a retail buzzword you want to explore (be sure to ask your instructor if the term is acceptable).
2. Use the skills you learned in Using Images Legally to search the images tagged with this term.
3. Choose 8–10 images that you feel best convey the search term you chose and express a point of view on the topic.
4. Create a PowerPoint with a title slide and the 8–10 images on the rest of the slides (one image per slide). The images you use should be able to express a point of view on a particular topic all by themselves.

You will present your visual essay to the class. Please email the instructor a copy **before class** on the due date. You will also post your presentation in the discussion board for your classmates to analyze.

Part 2: Peer Feedback

Choose at least two of your classmates’ sets of slides to analyze. Write down the topic you believe they chose and the main argument or message you believe the slides attempt to make. Then, write down support for your argument. Be sure to explain the use of visual design components (contrast, color, repetition, alignment, proximity, audience, purpose, context, etc.) when supporting your argument.

Grading Rubric

Criteria	Not Evident	Developing	Exemplary	Points
Submit your slide deck	**0 pts** No slide deck submitted	**5 pts** Slide deck does not have at least 8 slides	**10 pts** Slide deck has 8–10 slides focused on a central theme	10 pts
Present to the class	**0 pts** No presentation completed	**5 pts** Presentation is disjointed or unprofessional	**10 pts** Presentation is professional and feels cohesive	10 pts
Analyze at least two peers’ presentations	**0 pts** No response to peers	**2 pts** Responded to only one peer	**5 pts** Responded to two peers	5 pts
			Total:	**25 pts**

This page titled 8.18: Discussion- Visual Media is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.18: Discussion- Visual Media by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.19: Assignment- Visual Media

Everyday we encounter hundreds to thousands of visual images, but why do some images stick in our minds while others just get glossed over? Color, positioning, patterns, and repetition are all features that can can be engineered to convey a message within an image. Images play an important role in persuading an audience of a particular message, but a downside to this is that images can be interpreted in various ways based on people's history and experiences.

Breaking down the design components of an image is also a helpful when you need to design or create an image or logo to convey a message for your own business needs.

Creating a Personal Brand

If you were represented as a logo, what would you look like? Your task is to design a logo using your initials or parts of your name. The logo must visually communicate something about you or represent you. Ask yourself: What do I want to communicate about me? What shapes, forms, fonts, etc. will help to visually communicate my 'brand'? You can use any digital design tool or hand draw your logo.

You will turn in two things:

1. A short summary (minimum 200 words) describing your process of completing this project and to analyze the choices you made in the final version. You will discuss why you chose a particular logo shape, use of colors, and the reasons for any revisions.
2. Two copies of your logo: a color and a grayscale version.

Assignment Resources

- Watch What makes a truly great logo? (4:44 minutes)
- Read How to Create a Personal Logo That Makes Your Friends Jealous by Vladimir Gendelman and the accompanying infographic: How to Create a Personal Logo that Makes Your Friends Jealous.
- Lists of digital tools and tutorials

Grading Rubric

Criteria	Not Evident	Developing	Exemplary	Points
Turn in short summary essay	**0 pts** Essay not provided	**5 pts** Essay does not meet minimum word requirements	**10 pts** Essay is at least 200 words and describes the design process of the logo	10 pts
Submit color copy of logo	**0 pts** Color logo not provided	**5 pts** Logo does not communicate the ideas described in the essay	**10 pts** Logo communicates ideas described in the essay and is simple enough to be a logo	10 pts
Submit grayscale copy of logo	**0 pts** Grayscale logo not provided	**2 pts** Logo does not communicate the ideas described in the essay	**5 pts** Logo communicates ideas described in the essay and is simple enough to be a logo	5 pts
			Total:	**25 pts**

Contributors and Attributions

CC licensed content, Original

- Assignment: Visual Media. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Project Logo Design. **Authored by**: Rachael Sullivan. **Provided by**: Saint Joseph's University, Department of Communication and Digital Media. **Located at**: courses.rachaelsullivan.com/202fall17/assignments/project-1-logo-design/. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 8.19: Assignment- Visual Media is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.19: Assignment- Visual Media by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.20: Introduction to Media and Your Message

What you'll learn to do: Find the best media to present your message

I hear and I forget. I see and I understand. I do and I remember.

—Confucius



Confucius had the essence of business communication defined back in 500 BC. Your organization is full of people who hear and forget; however, if they see, they will understand.

Visual media improves a business communicator's chances of being understood by presenting information with a unique combination of logic and creativity. In this section, we will learn what visual media resources are available to business communicators and how to choose those that will strengthen your messages. We will learn about basic visual design principles and the messages that these principles send to your audience.

Contributors and Attributions

CC licensed content, Original

- Introduction to Media and Your Message. **Authored by**: Freedom Learning Group. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

Public domain content

- Confucius Chinese philosopher. **Provided by**: Wikimedia. **Located at**: https://commons.wikimedia.org/wiki/File:Konfuzius-1770.jpg. **License**: *Public Domain: No Known Copyright*

This page titled 8.20: Introduction to Media and Your Message is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.20: Introduction to Media and Your Message by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.21: Using Visual Media Resources

Learning Outcome

- Describe professional standards for using visual media resources for business purposes

As we learned in Module 1: Communicating in Business, in all business communications, we should have:

- Clarity
- Conciseness
- Objectivity
- Consistency
- Completeness
- Relevancy
- Understanding of audience knowledge

If you keep these seven communication principles in mind when you're crafting your message, you improve the chances that your message will be effective. Now we are going to add visual media to our communication toolbox; visual media increases your message's impact.

How does a visual element give your message more impact? Think about the last time you asked directions to a friend's house. If you are one of those "direction-impaired" people whose eyes glaze over at the first "go north" or "head east," then you know how helpful it is when someone draws you a map. It's a simple visual aid that takes the place of all of those words you're not quite sure what to do with. That map has made your friend's communication of directions easier to remember. There are examples of this all around you. The word "STOP" is a fairly easy word, but a red light or octagonal sign removes the need of any words at all. The weather report shows you a rain cloud, and you grab your umbrella without a second thought.

If your message can be more clearly understood by incorporating visual media, then by all means, you should do it. And to incorporate visuals effectively, you should understand that you're going to heed all seven principles of communication, particularly these:

- **Clarity:** your visuals should be clear, clean, and simple
- **Consistency:** your visuals should all maintain a uniform look and feel
- **Relevancy:** your visuals should make sense as a part of the whole communication and be on-brand

And then we're going to add a whole new principle to the mix, that your visuals should be **persuasive**: your visuals should inspire an emotional bond or a new level of understanding.

Let's look at each one of these principles separately.

Clarity

When considering the use of visual media, make sure that it's easy for the reader to glean the information he needs. Some general rules we'll keep in mind as we go through this module include,

- **Use bold, contrasting colors**. If you have a pie chart or a graph, it helps to make one piece red, another yellow, and yet another blue. This makes the chart easier to read than if each piece is a different shade of green. Make sure your chart pops with color, and there will be no question which piece is which.
- **Use easy-to-read fonts**. Loopy letters and heavy calligraphy strokes slow your reader down. Choose a font that's easy to read, like one of the many discussed by author John Wood in his blog for the American Writers & Artists, Inc. website.
- **Use only pertinent information**. If the point of your communication is to show that sales have gone up 22 percent over last year, your graph should feature that information—and nothing else. If you throw expenses, employee turnover and gross margin on that same chart, your reader will miss your message. Point out the information you need to highlight, and let everything else fade into the background.

Consistency

This module on visual media looks a lot like the other modules in this business communication course, doesn't it? Same kind of headers, same colors. Uniformity helps the reader understand what to expect and better prepares them to take in your message. Here are a few things we'll try to do in this module as we study different types of visual media:

- **Stick to the format of your charts and graphs wherever you're able.** If you start out with that bar chart showing annual sales, don't make it a line graph in the next section and a stacked bar chart in the section after that. Using different charts to show the same information slows your reader down unnecessarily. Keep in mind that if you found the best visual scheme to explain the data in the first place, there's no reason to change it to a less-effective one just for the sake of variety.
- **Stick to the color scheme and fonts you've already established.** If you show sales on your graphs in red, always show them in red. If you've chosen one easy-to-read font for all your tables, or a similar style of photo for all the sections of your annual report, there's no need to deviate.
- **Use pictures that are visually similar.** If you're using a series of head shots, the heads should all be about the same size. If four of the pictures show a person's head and shoulders, the fifth one should not be showing a person from the waist up.

Relevancy

If your message is communicating annual sales, your charts and graphs shouldn't be dealing with employee turnover rates. If your message is about your company's efforts to reduce waste, that message should not feature a photo of a cute puppy. That's relevancy at a very basic level.

Keeping communication "on brand" takes relevancy to a whole new level. Companies rely on visual media as much as the written word to deliver their brand message, and as a communicator, you need to keep your choices in visual media relevant to your company's mission and promise to its customers.

Colors in Branding: Target and Disney

If you ever visit Target's website, you may notice there's a whole lot of red. That didn't happen by accident. Target has a series of colors and images that coincide with the way they identify themselves as a company and the promises they make to their customers. You see a lot of crisp clean backgrounds with vivid pictures showing style and value—and a lot of red fonts and bulls eyes.

Conversely, Disney's main color is blue, but Sleeping Beauty's castle and that cartoon mouse are even more closely associated with the company's mission and promise to customers. Disney photos always depict happy families interacting with characters and enjoying the entertainment.

Large companies usually have a set of brand guidelines or a brand "style guide" that communicators can consult to familiarize themselves with the company's preferred color palette, fonts and image standards. All of your business communications represent an opportunity to reinforce and reflect your company's brand, and it's your duty as a communicator to do so.

Persuasive

Your visual media choice should help you tell your story. Even if your data is perfect, it's no guarantee your audience is going to jump on board with you. Your use of image, chart, or video should indicate it is from a reliable source, be simple to read, and allow you to show the audience exactly how you drew the conclusion you've drawn. If you're communicating in aid of a cause, it doesn't hurt to choose an image that invokes a little emotion.

This doesn't mean that you should sacrifice the principle of objectivity when you employ persuasiveness in your visual media. But your communication strategy on the whole is an engagement tool, and your choice of visuals should strengthen that engagement.

Contributors and Attributions

CC licensed content, Original

- Using Visual Media Resources. **Authored by**: Freedom Learning Group. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 8.21: Using Visual Media Resources is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.21: Using Visual Media Resources by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.22: Visual Design Principles

Learning Outcome

- Describe basic visual design principles

Business communicators don't always have access to a graphic artist. In the event that you as communicator find yourself needing to create visuals that dazzle without the help of a graphic artist, here are a few basic principles of visual design you can keep in your back pocket.

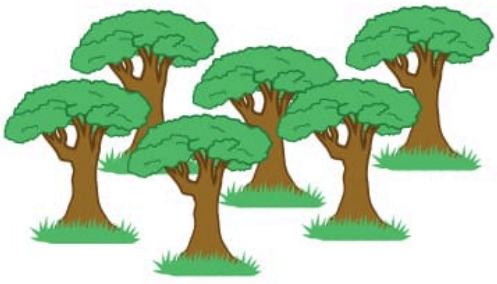
Contrast

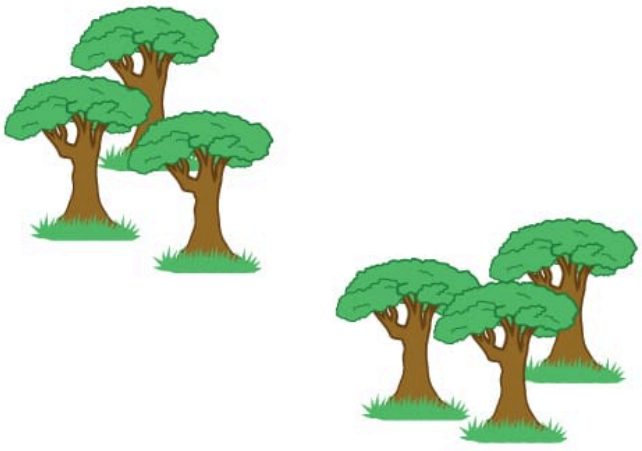
Contrast is when two aspects of an image are strikingly different from one another, like dark and light. Contrast is an important principle in visual design and helps highlight the important part of the image. It adds "weight" to your design and guides the viewer's eye to what you want them to see.



Alignment

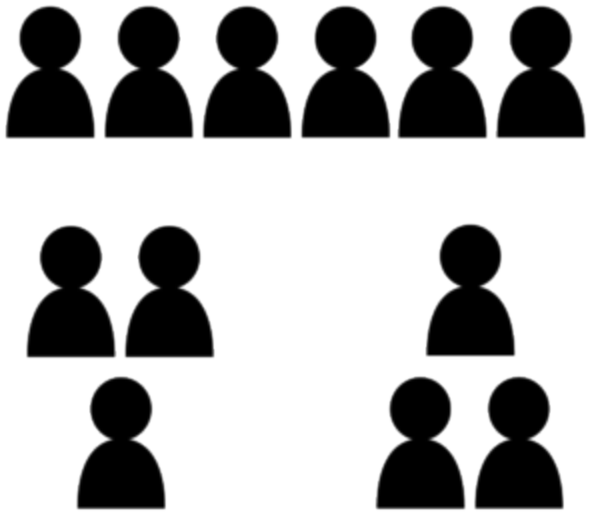
Alignment creates a sharp, linear order to the elements of your visual, so they all have a connection to each other. If objects are closer together, the viewer assumes that they're related. In the first image of trees below, we see six trees that are in two rows even though they're not precisely linear. In the second image, we perceive two groups of three.





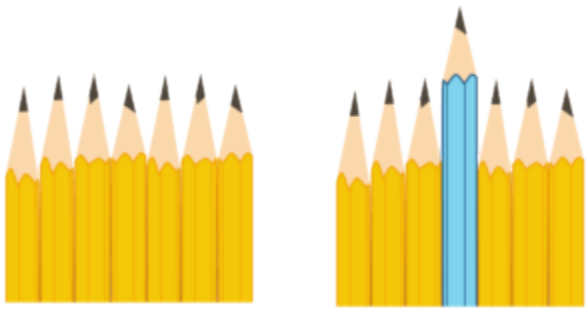
Hierarchy

If there are multiple elements in a design, more visual "weight" should be given to the most important part of the graphic. Establish the most essential part of the graphic first, and then fill in the rest with the less important parts.



Repetition and Pattern

Repetition strengthens the overall design and ties together elements to make them more consistent. This technique is often used in branding to make items more recognizable.



Color

Color is an important choice in visual communication because each color has a meaning. If you're following brand guidelines, your colors will reinforce your brand, but if not, you might want to consider some of the universal associations that go along with each color. Green tends to conjure images of the environment, while red symbolizes anger, and yellow, happiness. Which of these roses looks cold to you?



Xerox dug deep into why color is important in communication. Check out their two-page cheat sheet to learn more about how to leverage color for the best effect in your presentation.

Balance and Space

Keeping the elements of your design balanced gives the design some form and stability. Even spacing makes it look professional

and attractive, but that doesn't mean it needs to be symmetrical. It can be asymmetrical, with larger items in the upper left corner balancing out smaller ones in the lower right, and so on.

Leaving open or "negative" space ensures that your visual isn't cluttered and can highlight the important parts of a design. As we mentioned in our adopted standards above, simplicity is your friend!

Font

Design doesn't stop at the picture. Fonts have everything to do with your audience's engagement with your communication. Take a look at this font and decide if it's easy to read:

The quick brown fox jumps over the lazy dog

You can tell what it says; however, reading this font for too long could get taxing, especially on a screen. Is this next font easier to read?

The quick brown fox jumps over the lazy dog

Your audience won't continue to read your communication if you've chosen a font that's difficult to read.

In addition to legibility, there's a question of style. How do you feel about these lines of text and how they work together?



They're just words, but they're very visual; the use of color and different fonts draws your attention to the words "dream it" and "do it."

watch it

Graphic artists use a variety of rules to choose fonts and lay them out in a graphic design. This video shows you—very visually—how graphic artists make fonts work as a component of visual media.

A link to an interactive elements can be found at the bottom of this page.

Business communicators aren't necessarily graphic artists, but a good command of a graphic designer's visual design techniques will help you evaluate your visual media and decide if it's going to support your message. These aren't all the visual design principles a graphic designer employs, of course, but for our purposes, they're a good place to start.

Additional Resources

- Visual Design Principles by Joel Marsh
- Beginning Graphic Design: Fundamentals of Design from Goodwill Community Foundation, Inc.

Contributors and Attributions

CC licensed content, Original

- Visual Design Principles. **Authored by**: Freedom Learning Group. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Roses - Contrast. **Authored by**: Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Trees - Alignment. **Authored by**: Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- figures - Hierarchy. **Authored by**: Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Pencils - Repetition and Pattern. **Authored by**: Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Roses - color. **Authored by**: Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Typography examples. **Authored by**: Lumen Learning. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 8.22: Visual Design Principles is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.22: Visual Design Principles by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.23: Introduction to Images

What you'll learn to do: Discuss how to most effectively use images in business messages and identify potential sources for these images

We've already looked at compelling evidence that visual media helps your audience better understand your message. Visual media can help you capture your audience's attention and even persuade them to understand your point of view. So let's use some visual media!

Before we jump right in, we want to learn how to use visual media correctly. In this section, we'll discuss images—that is photos, drawings, and other multimedia pictures—and how they are best incorporated into your communications. We'll use our visual communication plans to put those images to work for us and get the best results. And then we'll talk about some resources to find images for your communication projects.

Contributors and Attributions

CC licensed content, Original

- Introduction to Images. **Authored by**: Freedom Learning Group. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 8.23: Introduction to Images is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.23: Introduction to Images by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.24: Images Overview

Learning Outcome

- Identify types of images used for business messages

When creating a message take a moment to you ask yourself, “Would including an image strengthen or support the message more than text alone?” Studies show that content that includes images get up to 94% more views than content without images. Users are 40 times more likely to share visual content on social media, and consumers are 80% more likely to read a piece of content if it contains colorful visuals.[1]

In short, visual images can greatly increase the comprehensibility and understandably of a message.

But how do you choose the right picture to match your content? We should apply the visual media standards we discussed earlier:

- The image should be clear and simple
- The image should have the same look and feel as the other images in the document
- The image should persuade the reader (or at least capture some feeling)
- The image should fit with your company’s brand.

Let’s take a look at each of these by checking out the websites of some companies you’re familiar with.

Clear and Simple Images

A company that has a clear and strong visual identity is Apple. Apple is a technology company committed to bringing the best computing experiences to its customers. The clear, clean look of Apple’s website conveys a message to the customer: Our technology is sleek, bold, easy to use.



Figure 1. Apple’s iPhone X website

The iPhone X promotional website screen (Figure 1) features an eye-catching smear of colors, something that draws the viewer’s attention but doesn’t compete with the product itself. Imagine if the iPhone X screen featured a photo of a child, or a cute pet. Would you be looking at the product then, or the content on the phone? Clean and simple images help you convey a singular idea.

Conversely, Figure 2 shows a website that sells new and gently used electronics. Considering this company is trying to sell multiple types of products on one platform, what type of impact factor does it have? How would you compare the impact of this

website to Figure 1?

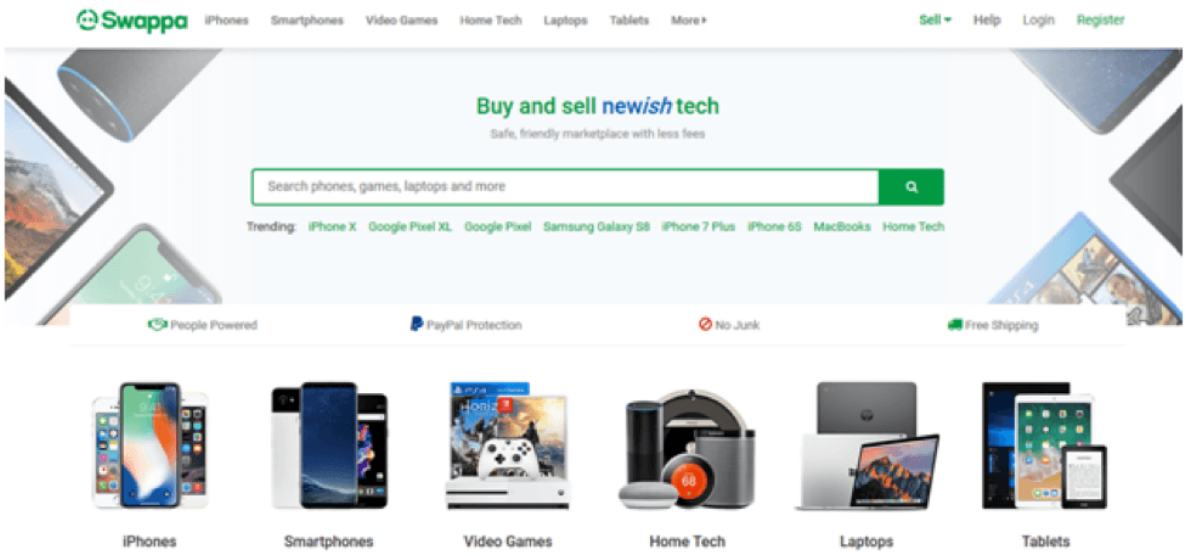


Figure 2. Swappa's website

Uniform Images

Chrysler is an automobile maker founded on the principle of "design with purpose," and their mission is to build cars people will enjoy driving and want to buy again. Figure 3 shows a screenshot of the homepage of Chrysler's website; here the company is using the layout and composition of the images to showcase a wide selection of items to explore here—all while doing it in a way that's interesting to the eye. In the top row we have three of their top-selling vehicles, shown in a uniform format: all shot in the light of the afternoon from about the same distance, all of them positioned at an angle with the front of the car pointing to the right, all of them suggesting the car is in motion.

The next row features two of their vehicles in the center of human interaction – one with a dog and one with other humans, and a black box beneath each to frame written content. Note that the third box is uniform in that it's a black box for text, but it instead features they Kelly's Blue Book logo. A break in the pattern! Did you notice that box first?

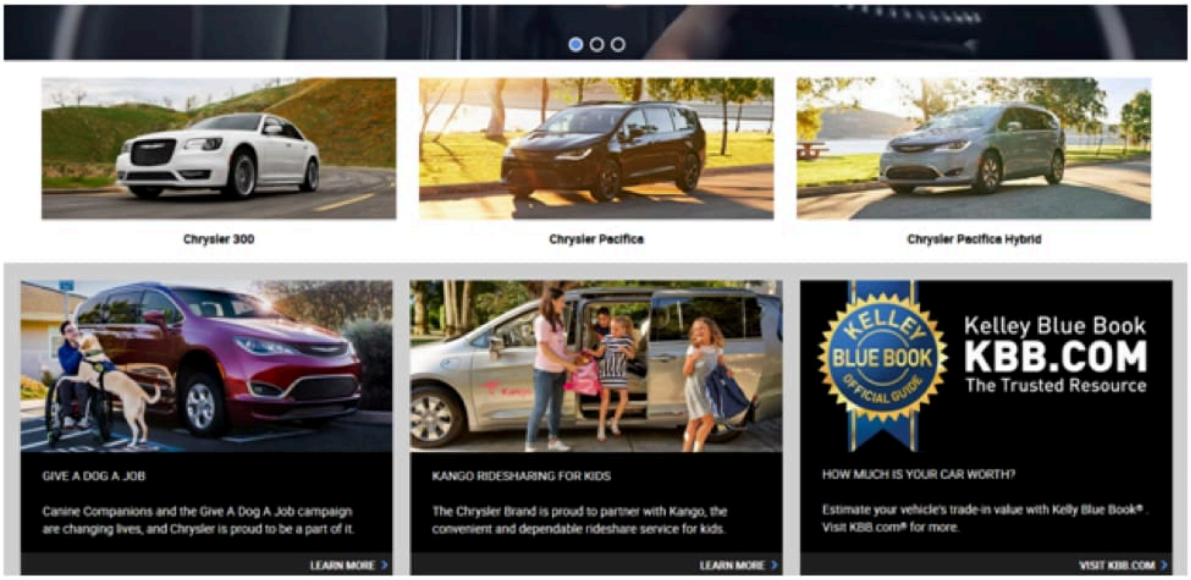


Figure 3. Chrysler's website

Figure 3 is actually the bottom of Chrysler's webpage, and the black also adds that "weight" to the design, drawing the viewers'

eyes down to the bottom of the page, adding some insurance that they'll read the whole page. If the images in Figure 3 weren't essentially "alike" it's likely the viewer would avoid looking that far. If you're using multiple images in your communication, take the time to make sure they're visually similar.

Now let's take a look at a website that doesn't use similar images or uniform layout on its front page (Figure 4). How does the visual representation of products impact your desire to purchase a product from the company? In Figure 4, no two images are alike, and the audience doesn't really know where to look—not a very successful use of images.



Figure 4. Arngren.net's website

Persuasive Images

As a nonprofit organization, UNICEF relies on donations to work towards their mission of ending preventable child deaths. This puts them in a unique position where their main business messages must be persuasive enough to inspire people to donate money to supporting their quest. In UNICEF's main homepage (Figure 5) the viewer is presented with a compelling image. Here you have have four young children depicted as happy little kids, just as they should be. The message this image is trying to project is that donor's money put to good use and that by supporting this organization UNICEF creates results.



Figure 5. Unicef's homepage

Other design elements to note are the colors on their site (Figure 5). The entire photo, with the exception of the children, the image has been digitally altered so that the building has been re-tinted blue. The altered image now aligns with the colors associated with the UNICEF brand. The brand color is then contrasted by the children's bright yellow raincoats. Not only do the yellow raincoats stand out and catch your eye, but the color yellow is most associated with happiness. The effects of color on an image can have a large impact on how a brand, or image connected to a message, is perceived by viewers.

Now look at the homepage from another nonprofit organization—Ferndale Cat Shelter website (Figure 6). Take a close look at the main image from their homepage and think about what underlying message the image is trying to convey and which emotions is it trying to evoke.

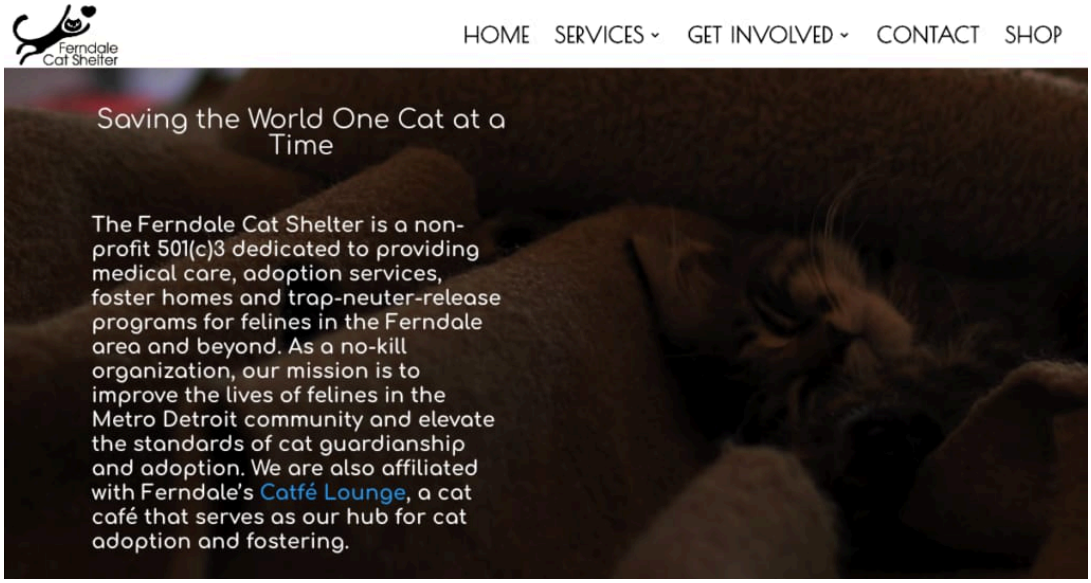


Figure 6. Ferndale Cat Shelter's website

In Figure 6, you see the image of a cat looking sleepy and content against a brown blanket. Due to the darker or shadowed nature of the photo it makes the viewer struggle to connect a message to the image and can make it difficult to feel a particular emotion towards the felines at the shelter. While you might have difficulty seeing the cat, the overall monochromatic use of brown and tan presents a stable, comfortable, or safe space for rescued cats. With a different image or use different colors or tones to make the subject of the photo stand out this non-profit could have made a stronger visual image to project their message and persuade the audience to offer their monetary support.

On Brand Images

"What can brown do for you?" That's the retired slogan United Parcel Service (UPS) used to show its customers that their company can solve client's shipping problems. The current slogan, "We (heart) logistics," similar to the "What brown can do for you?" slogan still needs something more doesn't it? In fact, it requires visual media to support their message.

On the homepage of UPS United Sates' website (Figure 7), the top of the page is headed with their logo and that trademark UPS brown. The image they've chosen is a clean and simple photo of two people in a "small business" situation. One is reviewing the contents of a box, perhaps getting some items ready to ship. The other is at the computer, perhaps checking shipping rates right here on the UPS website. The composition of the photo is a complimentary brown, with stand-out colors in the subjects' clothing that match UPS's secondary color palette. The background of the photo suggests a small business, but it's out of focus so your attention is drawn to the two individuals and their activities.

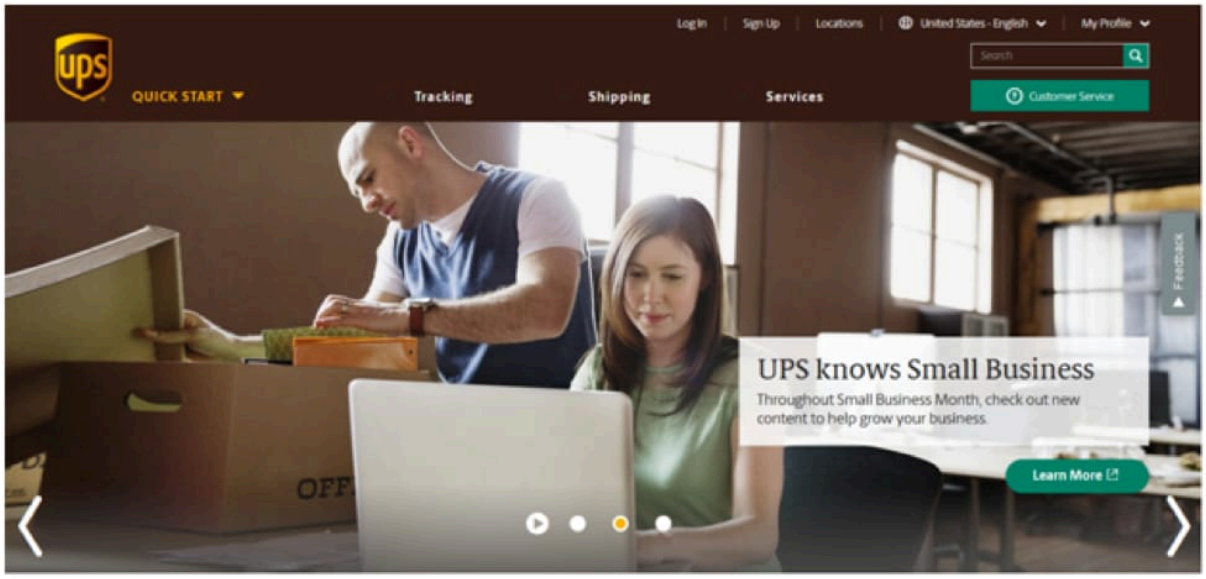


Figure 7. UPS's homepage

This could be considered a successful representation of the UPS brand. The image suggests that UPS is supporting the work of a small business. The image is composed of all the right colors so it's not distracting and supports the overall brand idea. This image is saying, "This is what brown can do for you."

On the other hand, take a look at a website for a children's juice manufacturer Penny Juice (Figure 8). Upon first impression this homepage does not convey any clear message about its brand promise or what the company stands for:

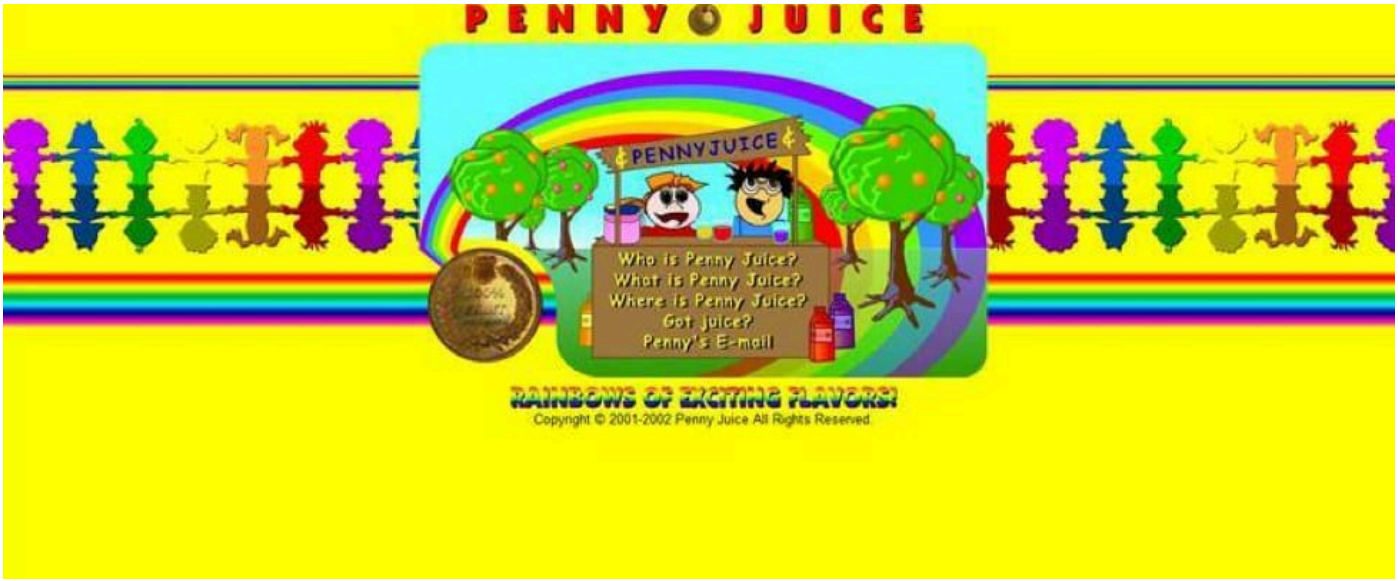


Figure 8. Penny Juice's website

In fact, the cartoon figures and garish use of colors might lead the audience to have a difficult time identifying the product. As a viewer you might have ideas or suggestions for how to convey the message of "We sell juice for kids!" in a more effective manner. Possibly a cartoon of a smiling child drinking a sippy cup full of juice or images of fruit would be a better hint. This current visual does not efficiently communicate brand.

When you measure your images against the visual media standards we put in place in section one of this module, you increase the chance that your message will be effective.

---

1. Desmarais, Ellen. "Make Your Content Engagement Skyrocket." *Contently*. 21 Mar 2018. Web. 28 Jun 2018.

Contributors and Attributions

CC licensed content, Original

- Images Overview. **Authored by**: Freedom Learning Group. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Screenshot Apple iPhone X. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Screenshot Swappa home page. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Screenshot Fiat Chrysler home page. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Screenshot Arngen.net home page. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Screenshot UNICEF home page. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Screenshot Ferndale Cat Shelter home page. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Screenshot United Parcel Service home page. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Screenshot Penny Juice website. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 8.24: Images Overview is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.24: Images Overview by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.25: Using Images for Impact

Learning Outcome

- Describe how to determine when to use an image and when to use text

If a picture were really worth a thousand words, the issue of image versus text would be easily decided: use an image. That's a thousand less words you have to write. Yet in the real world, the winner of "image versus text" isn't as easy to determine. Both play a vital role in your communication efforts.

When to Use an Image

Use an image if the information is presented verbally. Are you preparing a PowerPoint presentation for a speaker? If so, you'll want to stick with mostly images in your presentation. While you always want the audience to listen to the speaker, images actually help drive home the point the speaker makes and increases the memorability of that point as well. Therefore, you should use images that support the subject matter. Steve Jobs was famous for his image-based presentations. Watch "Steve Jobs Unveils the Original iPhone" for an example of this presentation style. Most TED Talk speakers follow Steve Jobs's speaking style with regard to visuals. For example, watch Dave Brain's TED Talk "What a Planet Needs to Sustain Life."

Use an image if the information is complicated and can be better explained in a visual format. You may be charged with explaining the difference between nuclear fission and nuclear fusion. You can prepare an enormous amount of text to review the process of each and hope that your audience understands, or you can reduce the text and use the photo in Figure 1.

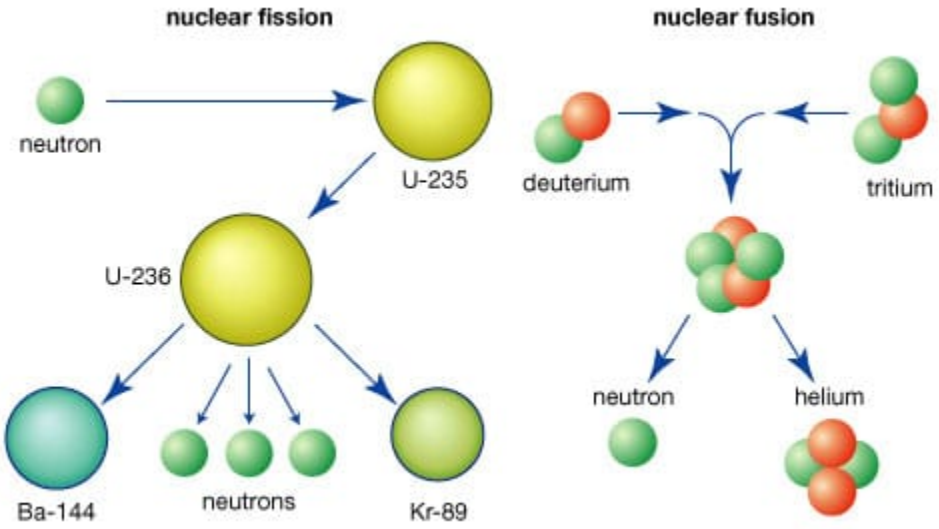


Figure 1. Nuclear fission compared to nuclear fusion

Use an image when it conveys something words just cannot. You may try to describe the northern lights to your reader, but he won't truly understand the beauty of the *aurora borealis* until he sees it for himself. Images can also convey the emotion of a situation in ways words often can't. For example, you can tell people to adopt shelter dogs, but a sweet puppy face looking out beseechingly from a cage will grab people's emotional attention in a way the idea on its own can't.

When to Use Text

Use text when you want to make a powerful point. If you want to leave your audience with a thought like "When the whole world is silent, even one voice becomes powerful." (Malala Yousafzai), using text has far more impact than just showing a picture.

Use text when you're creating a list or a mnemonic. Is your company a fan of SMART (Specific, Measurable, Achievable, Relevant, and Timely) goals, or do they employ the STAR (Situation, Task, Action, Result) candidate interview technique? It's difficult to even discuss these without text.

Use text when an image just isn't enough. And that's a lot of the time! Prehistoric man used images to tell his story on the walls of caves, but words were invented for a reason, and that reason is that images can be limiting. Complex processes may be better served by a simplifying image, but in just about every other instance, complexity is better served with words.

Contributors and Attributions

CC licensed content, Original

- Using Images for Impact. **Authored by**: Freedom Learning Group. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Nuclear Fission and Fusion. **Authored by**: Padsohot1. **Provided by**: Wikimedia. **Located at**: https://commons.wikimedia.org/wiki/File:%E0%B8%9F%E0%B8%B4%E0%B8%A7%E0%B8%8A%E0%B8%B1%E0%B9%88%E0%B8%99.jpg. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 8.25: Using Images for Impact is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.25: Using Images for Impact by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.26: Using Images Legally

Learning Outcome

- Discuss how to legally source images for your communications

All of the images you see on the internet are someone's property, and if you copy them and paste them into your communication, you're most likely stealing (unless the images are royalty-free or free, as we'll discuss below). Additionally, the Fair Use Act makes it okay to use an image "for purposes such as criticism, comment, news reporting, teaching (including multiple copies for classroom use), scholarship, or research...." If your work falls into one of those categories, you're safe. Otherwise, your use of the image is considered a violation of copyright law.

Royalty-Free Images

Luckily, there are royalty-free sources of images, videos, and other visual media out there. (A royalty is a payment made to the copyright owner for each use that doesn't fall under the Fair Use law.) A word of caution: "royalty-free" doesn't necessarily mean "free." On the contrary, sites like Getty Images or Shutterstock often charge licensing fees and even instructions by which you give the original artist credit for his or her work. But once that transaction is complete, the visual media is yours to use as you wish. There's no need to pay royalties for copies sold or time of use.

Open Images

There are plenty of free sources out there, though. Some authors and artists have "released" their works under an open license. This means their works are free to use, modify, and share.

Many of them work with a Creative Commons license. Creative Commons is a US not-for-profit organization that's devoted to expanding the range of creative works available to the general public, and it does so by providing copyright licenses to creators allowing them to express which rights they want to reserve and which they waive. These licenses replace individual negotiations between artist and user. In many cases, images covered by the Creative Commons license can be used free of charge if you simply credit the artist. This might mean captioning your photo with "Joe Photographer/Getty Images" or "Credit: Joe Photographer." You can learn more about the different Creative Commons Licenses on their site.

All of the images and text in this course are licensed under a Creative Commons license. You can check out the "Licenses and Attributions" link on any page of this course to see examples of proper attribution of Creative Commons works.



Figure 1. Example of attributing an author's work

Finding Open Images

- Pixabay:There are over a million free stock photos to choose from on this site, all high quality and high resolution.
- Unsplash:This site has a great selection of business photos to choose from. This website has the option of signing up for an email service to get ten free photos in your inbox every ten days.
- Flickr:There's a section of free use photos here as well, about 415 million of them under the Creative Common license. You simply need to filter your Flickr search by license (Figure 2).

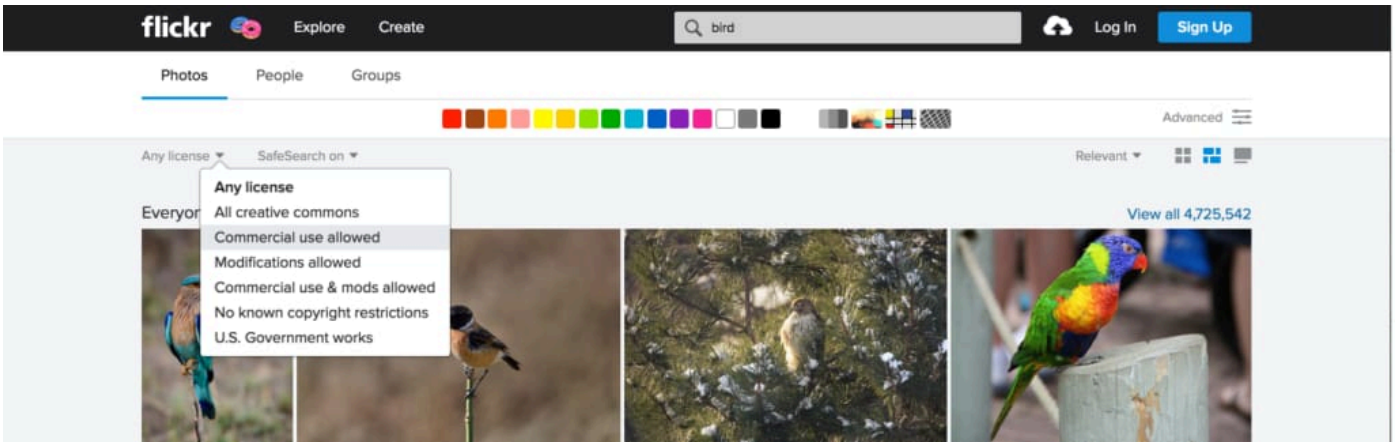


Figure 2. Filtering Flickr for permissions

- Google Images. A simple Google image search now has the function of searching for image based on usage rights. Click on "Tools" and then the "Usage rights" will appear for you to choose the type of license that fits your needs.

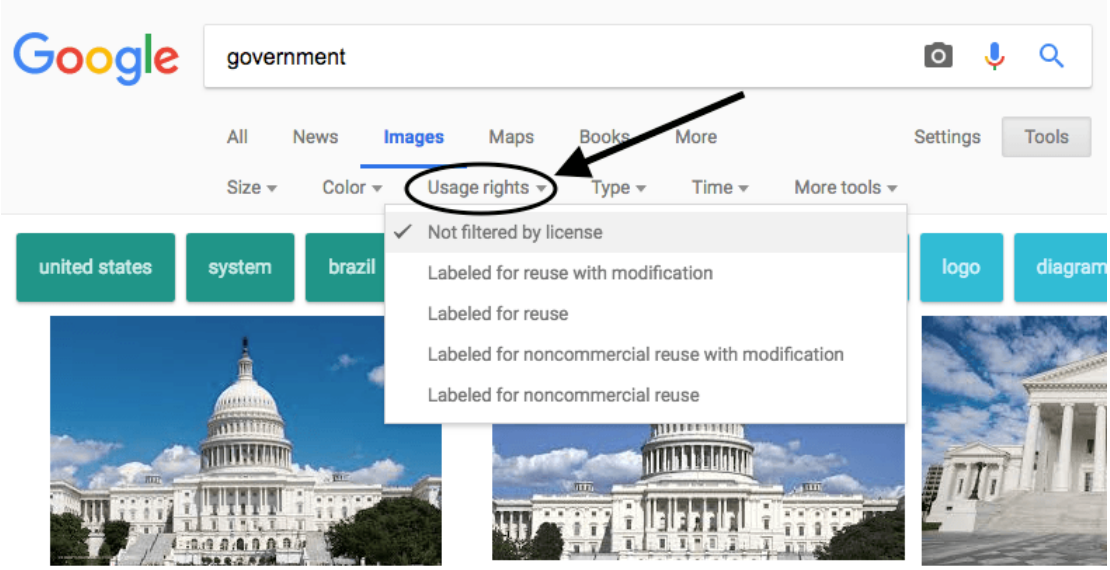


Figure 3. Filtering a Google Image search by license

Contributors and Attributions

CC licensed content, Original

- Products and Resources for Finding Images. **Authored by**: Freedom Learning Group. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Screenshot of Licenses and Attributions. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 8.26: Using Images Legally is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.26: Using Images Legally by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

8.27: Introduction to Charts, Diagrams, and Graphic Organizers

What you'll learn to do: Discuss how to most effectively use charts, diagrams, and other graphics in business messages and identify potential sources for these visual aids

When you communicate data, you can't just throw a whole bunch of numbers on a page and expect that readers will understand what you want to say. Charts, tables, and graphs help communicators organize that data in a way that helps their audiences understand the story the data tells and, hopefully, interpret it correctly.

In this section, we'll discuss the types of charts, graphs, and diagrams available to help you show off your data in ways that make it accessible to your audience. We'll look at specific communication challenges and determine which kind of chart or graph best illustrates your message, and finally, we'll discuss how to format your chart so that your story is easily and quickly understood.

Contributors and Attributions

CC licensed content, Original

- Introduction to Charts, Diagrams and Graphic Organizers. **Authored by**: Freedom Learning Group. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 8.27: Introduction to Charts, Diagrams, and Graphic Organizers is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 5.27: Introduction to Charts, Diagrams, and Graphic Organizers by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

9: RESEARCH AND STATISTICS

> To call in statisticians after the experiment is done may be no more than asking them to perform a post-mortem examination: They may be able to say what the experiment died of.
>
> — *Sir Ronald Fisher*

9.1: Introduction

9.2: Measures of Center

9.3: Measures of Variation

9.4: There's More to Research Methods than Statistics

9.5: A brief Introduction to Research Design

9.6: Introduction to Psychological Measurement

9.7: Scales of measurement

9.8: Assessing the Reliability of a Measurement

9.9: The role of variables — predictors and outcomes

9.10: Experimental and non-experimental research

9.11: Assessing the validity of a study

9.12: Confounds, Artifacts and other Threats to Validity

9.13: Videos

9.14: References

This page titled 9: Research and Statistics is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

9.1: Introduction



Figure 3.1 Statistics and data science play significant roles in stock market analysis, offering insights into market trends and risk assessment for better financial decision-making. (credit: modification of work “That was supposed to be going up, wasn't it?” by Rafael Matsunaga/Flickr, CC BY 2.0)

Statistical analysis is the science of collecting, organizing, and interpreting data to make decisions. Statistical analysis lies at the core of data science, with applications ranging from consumer analysis (e.g., credit scores, retirement planning, and insurance) to government and business concerns (e.g., predicting inflation rates) to medical and engineering analysis.

Statistical analysis is an essential aspect of data science, involving the systematic collection, organization, and interpretation of data for decision-making. It serves as the foundation for various applications in consumer analysis such as credit scoring, retirement planning, and insurance as well as in government and business decision-making processes such as inflation rate prediction and marketing strategies. As a consumer, statistical analysis plays a significant role in various decision-making processes. For instance, when considering a large financial decision such as purchasing a house, the probability of interest rate fluctuations and their impact on mortgage financing must be taken into account.

Part of statistical analysis involves descriptive statistics, which refers to the collection, organization, summarization, and presentation of data using various graphs and displays. Once data is collected and summarized using descriptive statistics methods, the next step is to analyze the data using various probability tools and probability distributions in order to come to conclusions about the dataset and formulate predictions that will be useful for planning and estimation purposes.

Descriptive statistics includes measures of the center and dispersion of data, measures of position, and generation of various graphical displays. For example, a human resources administrator might be interested in generating some statistical measurements for the distribution of salaries at a certain company. This might involve calculating the mean salary, the median salary, and standard deviation, among others. The administrator might want to present the data to other employees, so generating various graphical displays such as histograms, box plots, and scatter plots might also be appropriate.

The human resources administrator might also want to derive estimates and predictions about the average salary in the company five years into the future, taking into account inflation effects, or they might want to create a model to predict an individual employee’s salary based on the employee's years of experience in the field.

Such analyses are based on the statistical methods and techniques we will discuss in this chapter, building on the introduction to statistical analysis presented in What Are Data and Data Science? and Collecting and Preparing Data. Statistical analysis utilizes a variety of technological tools to automate statistical calculations and generate graphical displays. This chapter also will demonstrate the use of Python to generate results and graphs. The supplementary material at the end of this book will show the same using Excel (Appendix A) and R (Appendix B).

In this chapter, you will also study probability concepts and probability distributions. Probability theory helps us deal with quantifying uncertainty, which is always inherent in real-world data. We will see that in real-world datasets we often have to deal with noise and randomness that will be analyzed using statistical analyses.

Probability analysis provides the tools to model, understand, and quantify uncertainties, allowing data scientists to make informed decisions from data. Probability theory also forms the basis for different types of analyses, such as confidence intervals and hypothesis testing, which will be discussed further in Inferential Statistics and Regression Analysis. Such methods rely on probability models to make predictions and detect patterns. In addition, machine learning (discussed in Time Series and Forecasting) uses probabilistic models to help represent uncertainty and make predictions based on the collected data. Probability analysis helps data scientists to select data models, interpret the results, and assess the reliability of these conclusions. For a more detailed review of statistical concepts, please refer to Introductory Statistics 2e.

This page titled 9.1: Introduction is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

9.2: Measures of Center

Learning Objectives

By the end of this section, you should be able to:

- 3.1.1 Define and calculate mean, trimmed mean, median, and mode for a dataset.
- 3.1.2 Determine the effect of outliers on the mean and the median.
- 3.1.3 Use Python to calculate measures of center for a dataset.

Measures of center are statistical measurements that provide a central, or typical, representation of a dataset. These measures can help indicate where the bulk of the data is concentrated and are often called the data's central tendency. The most widely used measures of the center of a dataset are the *mean* (average), the *median*, and the *mode*.

Mean and Trimmed Mean

The mean, or average, (sometimes referred to as the *arithmetic mean*) is the most commonly used measure of the center of a dataset. Sometimes the mean can be skewed by the presence of outliers, or data values that are significantly different as compared to the remainder of the dataset. In these instances, the *trimmed mean* is often used to provide a more representative measure of the center of the dataset, as we will discuss in the following section.

Mean

To calculate the mean, add the values of all the items in a dataset and divide by the number of items. For example, if the scores on your last three exams were 87, 92, and 73, then the mean score would be 87+92+733=8487+92+733=84. If you had a large number of data values, you would proceed in the same way. For example, to calculate the mean value of 5050 exam scores, add the 5050 scores together and divide by 5050. If the 5050 scores add up to 4,0504,050, for example, the mean score is 4050504050504050, or 81.

In data science applications, you will encounter two types of datasets: sample data and population data. Population data represents all the outcomes or measurements that are of interest. Sample data represents outcomes or measurements collected from a subset, or part, of the population of interest. Of course in many applications, collecting data from an entire population is not practical or feasible, and so we often rely on sample data.

The notation x–x– is used to indicate the sample mean, where the mean is calculated based on data taken from a sample. The notation ∑x∑x is used to denote the sum of the data values, and nn is used to indicate the number of data values in the sample, also known as the sample size.

The sample mean can be calculated using the following formula:

x–=∑xnx–=∑xn

The notation μμ is used to indicate the population mean, where the mean is calculated based on data taken from the entire population, and NN is used to indicate the number of data values in the population, also known as the population size. The population mean can be calculated using the following formula:

μ=∑xNμ=∑xN

The mean can also be determined by its frequency distribution. For every unique data value in the dataset, the frequency distribution gives the number of times, or frequency, that this unique value appears in the dataset. In this type of situation, the mean can be calculated by multiplying each distinct value by its frequency, summing these values, and then dividing this sum by the total number of data values. Here is the corresponding formula for the sample mean using the frequency distribution:

x–=∑x·fnx–=∑x·fn

When all the values in the dataset are unique, this reduces to the previous formula given for the sample mean.

Example 3.1

During a clinical trial, a sample is taken of 10 patients and pulse rates are measured in beats per minute:

68, 92, 76, 51, 65, 83, 94, 72, 88, 59

Calculate the mean pulse rate for this sample.

Answer available at the end of this section.

Example 3.2

A college professor records the ages of 25 students in a data science class as shown:

Student Age	Number of Students (Frequency)
19	3
20	4
21	8
22	6
23	2
27	1
31	1
Total	**25**

Calculate the mean age for this sample of students.

Answer

Substitute the values from the table in the following formula:

$x - = \sum x \cdot f n = 19 \cdot 3 + 20 \cdot 4 + 21 \cdot 8 + 22 \cdot 6 + 23 \cdot 2 + 27 \cdot 1 + 31 \cdot 1 25 = 541 25 = 21.64$ $x - = \sum x \cdot f n = 19 \cdot 3 + 20 \cdot 4 + 21 \cdot 8 + 22 \cdot 6 + 23 \cdot 2 + 27 \cdot 1 + 31 \cdot 1 25 = 541 25 = 21.64$

Trimmed Mean

A trimmed mean helps mitigate the effects of outliers, which are data values that are significantly different from most of the other data values in the dataset. In the dataset given in Example 3.1, a pulse rate of 35 or 120 would be considered outlier data values since these pulse rates are significantly different as compared to the rest of the values in Example 3.1. We will see that there is a formal method for determining outliers, and in fact there are several methods to identify outliers in a dataset.

The presence of outlier data values tends to disproportionally skew the mean and produce a potentially misleading result for the mean.

To calculate the trimmed mean for a dataset, first sort the data in ascending order (from smallest to largest). Then decide on a certain percentage of data values to be deleted from the lower and upper ends of the dataset. This might represent the extent of outliers in the dataset; trimmed mean percentages of 10% and 20% are common. Then delete the specified percentage of data values from both the lower end and upper end of the dataset. Then find the mean for the remaining undeleted data values.

As an example, to calculate a 10% trimmed mean, first sort the data values from smallest to largest. Then delete the lower 10% of the data values and delete the upper 10% of the data values. Then calculate the mean for the resulting dataset. Any outliers would tend to be deleted as part of the trimmed mean calculation, and thus the trimmed mean would then be a more representative measure of the center of the data for datasets containing outliers.

Example 3.3

A real estate agent collects data on a sample of recently sold homes in a certain neighborhood, and the data are shown in the following dataset:

397900, 452600, 507400, 488300, 623400, 573200, 1689300, 403890, 612300, 599000, 2345800, 499000, 525000, 675000, 385000

1. Calculate the mean of the dataset.
2. Calculate a 20% trimmed mean rate for the dataset.

Answer available at the end of this section.

Median

The median provides another measure of central tendency for a dataset. The median is generally a better measure of the central tendency when there are outliers (extreme values) in the dataset. Since the median focuses on the middle value of the ordered dataset, the median is preferred when outliers are present because the median is not affected by the numerical values of the outliers.

To determine the median of a dataset, first order the data from smallest to largest, and then find the middle value in the ordered dataset. For example, to find the median value of 50 exam scores, find the score that splits the data into two equal parts. The exam scores for 25 students will be below the median, and 25 students will have exam scores above the median.

If there is an odd number of data values in the dataset, then there will be one data value that represents the middle value, and this is the median. If there is an even number of data values in the dataset, then to find the median, add the two middle values together and divide by 2 (this is essentially finding the mean of the two middle values in the dataset).

Example 3.4

The same dataset of pulse rates from Example 3.1 is:

68, 92, 76, 51, 65, 83, 94, 72, 88, 59

Calculate the median pulse rate for this sample.

Answer available at the end of this section.

You can also quickly find the sample median of a dataset as follows.

Let nn represent the number of data values in the sample.

- If nn is odd, then the median is the data value in position n+12n+12.
- If nn is even, the median is the mean of the observations in position n2n2 and position n2+1n2+1.

For example, let's say a dataset has 25 data values. Since nn is odd, to identify the position of the median, calculate n+12n+12, which is 25+1225+12, or 13. This indicates that the median is located in the 13th data position.

As another example, let's say a dataset has 100 data values. Since nn is even, to identify the position of the median, calculate

n2n2, which is 100210002, which is 50, and also calculate n2+1n2+1, which is 50+150+1, which is 51. This indicates that the median is calculated as the mean of the 50th and 51st data values.

Mode

Another measure of center is the mode. The mode is the data value that occurs with the greatest frequency. If there are no repeating data values in a dataset, then there is no mode. If two data values occur with same greatest frequency, then there are two modes, and we say the data is bimodal. For example, assume that the weekly closing stock price for a technology stock, in dollars, is recorded for 20 consecutive weeks as follows:

50, 53, 59, 59, 63, 63, 72, 72, 72, 72, 72, 76, 78, 81, 83, 84, 84, 84, 90, 93

To find the mode, determine the most frequent score, which is 72, which occurs five times. Thus, the mode of this dataset is 72.

The mode can also be applied to non-numeric (qualitative) data, whereas the mean and the median can only be applied for numeric (quantitative) data. For example, a restaurant manager might want to determine the mode for responses to customer surveys on the quality of the service of a restaurant, as shown in Table 3.1.

Table 3.1 Customer Survey Results for Customer Survey Rating

Customer Service Rating	Number of Respondents
Excellent	267
Very Good	410
Good	392
Fair	107
Poor	18

Based on the survey responses, the mode is the Customer Service Rating of “Very Good,” since this is the data value with the greatest frequency.

Influence of Outliers on Measures of Center

As mentioned earlier, when outliers are present in a dataset, the mean may not represent the center of the dataset, and the median will provide a better measure of center. The reason is that the median focuses on the middle value of the ordered dataset. Thus, any outliers at the lower end of the dataset or any outliers at the upper end of the dataset will not affect the median. Note: A formal method for identifying outliers is presented in Measures of Position when measures of position are discussed. The following example illustrates the point that the median is a better measure of central tendency when potential outliers are present.

Example 3.5

Suppose that in a small company of 40 employees, one person earns a salary of $3 million per year, and the other 39 individuals each earn $40,000. Which is the better measure of center: the mean or the median?

Answer available at the end of this section.

Using Python for Measures of Center

We learned in What Are Data and Data Science? how the `DataFrame.describe()` method is used to summarize data. Recall that the method `describe()` is defined for a DataFrame type object so should be called upon a DataFrame type variable (e.g. given a DataFrame `d`, use `d.describe()`).

Figure 3.2 shows the output of `DataFrame.describe()` on the "Movie Profit" dataset we used in What Are Data and Data Science?, movie_profit.csv. The mean and 50% quartile show the average and median of each column. For example, the average worldwide gross earnings are $410.14 million, and the median earnings are $309.35 million. Note that average and/or median of some columns are not as meaningful as the others. The first column—Unnamed: 0—was simply used as an identifier of each item in the dataset, so the average and mean of this column is not quite useful. `DataFrame.describe()` still computes the values because it can (and it does *not* care which column is meaningful to do so or not).

	Unnamed: 0	Rating	Duration	US_Gross_Million	Worldwide_Gross_Million
count	966.00000	966.000000	966.000000	966.000000	966.000000
mean	483.50000	6.814286	117.506211	156.158975	410.140600
std	279.00448	0.894383	21.615612	110.629617	294.758791
min	1.00000	3.300000	69.000000	0.010000	176.600000
25%	242.25000	6.200000	101.250000	90.832500	223.277500
50%	483.50000	6.800000	116.000000	129.245000	309.345000
75%	724.75000	7.400000	130.000000	187.090000	472.645000
max	966.00000	9.200000	238.000000	936.660000	2847.400000

Figure 3.2 The Output of

`DataFrame.describe()`

with the Movie Profit Dataset

Exploring Further

Working with Python

See the Python website for more details on using, installing, and working with Python. See this additional documentation, for more specific information on the statistics module.

Answers

Add the 10 data values, and the sum is 748. Divide this sum by the number of data values, which is 10. The result is:

x – = ∑ x n = 748 10 = 74.8 x – = ∑ x n = 748 10 = 74.8

1. For the mean, add the 15 data values, and the sum is 10,777,090. Divide this sum by the number of data values, which is 15. The result is:
   x–=∑xn=10,777,09015=718,472.70x–=∑xn=10,777,09015=718,472.70
2. For the trimmed mean, first order the data from smallest to largest. The sorted dataset is:

   385000, 397900, 403890, 452600, 488300, 499000, 507400, 525000, 573200, 599000, 612300, 623400, 675000, 1689300, 2345800

   Twenty percent of 15 data values is 3, and this indicates that 3 data values are to be deleted from each of the lower end and upper end of the dataset. The resulting 9 undeleted data values are:

   452600, 488300, 499000, 507400, 525000, 573200, 599000, 612300, 623400

   Then find the mean for the remaining data values. The sum of these 9 data values is 4,880,200. Divide this sum by the number of data values (9). The result is:
   x–=∑xn=4,880,2009=542,244.40x–=∑xn=4,880,2009=542,244.40

Notice how the mean calculated in Part (a) is significantly larger as compared to the trimmed mean calculated in Part (b). The reason is the presence of several large outlier home prices. Once these outlier data values are removed by the trimmed mean calculation, the resulting trimmed mean is more representative of the typical home price in this neighborhood as compared to the mean.

First, order the 10 data values from smallest to largest. Divide this sum by the number of data values, which is 10. The result is:

51, 59, 65, 68, 72, 76, 83, 88, 92, 94

Since there is an even number of data values, add the two middle values together and divide by 2.
The two middle values are 72 and 76.

Median = 72 + 76 2 = 148 2 = 74 Median = 72 + 76 2 = 148 2 = 74

The mean, in dollars, would be arrived at mathematically as follows:

x – = 3,000,000 + 39(40,000) 40 = 114,000 x – = 3,000,000 + 39(40,000) 40 = 114,000

However, the median would be $40,000 since this is the middle data value in the ordered dataset. There are 39 people who earn $40,000 and one person who earns $3,000,000.

Notice that the mean is not representative of the typical value in the dataset since $114,000 is not reflective of the average salary for most employees (who are earning $40,000). The median is a much better measure of the “average” than the mean in this case because 39 of the values are $40,000 and one is $3,000,000. The data value of $3,000,000 is an outlier. The median result of $40,000 gives us a better sense of the center of the dataset.

This page titled 9.2: Measures of Center is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

9.3: Measures of Variation

Learning Objectives

By the end of this section, you should be able to:

- 3.2.1 Define and calculate the range, the variance, and the standard deviation for a dataset.
- 3.2.2 Use Python to calculate measures of variation for a dataset.

Providing some measure of the spread, or *variation*, in a dataset is crucial to a comprehensive summary of the dataset. Two datasets may have the same mean but can exhibit very different spread, and so a measure of dispersion for a dataset is very important. While measures of central tendency (like mean, median, and mode) describe the center or average value of a distribution, measures of dispersion give insights into how much individual data points deviate from this central value.

The following two datasets are the exam scores for a group of three students in a biology course and in a statistics course.

Dataset A: Exam scores for students in a biology course: 40, 70, 100
Dataset B: Exam scores for students in a statistics course: 69, 70, 71

Notice that the mean score for both Dataset A and Dataset B is 70.

However, the datasets are significantly different from one another:

Dataset A has larger variability where one student scored 30 points below the mean and another student scored 30 points above the mean.
Dataset B has smaller variability where the exam scores are much more tightly clustered around the mean of 70.

This example illustrates that publishing the mean of a dataset is often inadequate to fully communicate the characteristics of the dataset. Instead, data scientists will typically include a measure of variation as well.

The three primary measures of variability are range, variance, and standard deviation, and these are described next.

Range

Range is a measure of dispersion for a dataset that is calculated by subtracting the minimum from the maximum of the dataset:

Range=Max−MinRange=Max−Min

Range is a straightforward calculation but makes use of only two of the data values in a dataset. The range can also be affected by outliers.

Example 3.6

Calculate the range for Dataset A and Dataset B:

Dataset A: Exam scores for students in a biology course: 40, 70, 100
Dataset B: Exam scores for students in a statistics course: 69, 70, 71

Answer available at the end of this section.

One drawback to the use of the range is that it doesn't take into account every data value. The range only uses two data values from the dataset: the minimum (min) and the maximum (max). Also the range is influenced by outliers since an outlier might appear as a minimum or maximum data value and thus skew the results. For these reasons, we typically use other measures of

variation, such as variance or standard deviation.

Variance

The variance provides a measure of the spread of data values by using the squared deviations from the mean. The more the individual data values differ from the mean, the larger the variance.

A financial advisor might use variance to determine the volatility of an investment and therefore help guide financial decisions. For example, a more cautious investor might opt for investments with low volatility.

The formula used to calculate variance also depends on whether the data is collected from a sample or a population. The notation s2s2 is used to represent the *sample variance*, and the notation σ2σ2 is used to represent the *population variance*.

Formula for the sample variance:

s2=∑(x–x–)2n–1s2=∑(x–x–)2n–1

Formula for the population variance:

σ2=∑(x–μ)2Nσ2=∑(x–μ)2N

In these formulas:
xx represents the individual data values
x–x– represents the sample mean
nn represents the sample size
μμ represents the population mean
NN represents the population size

Alternate Formula for Variance

An alternate formula for the variance is available. It is sometimes used for more efficient computations:

σ2=∑x2N–μ2σ2=∑x2N–μ2

In the formulas for sample variance and population variance, notice the denominator for the sample variance is n–1n–1, whereas the denominator for the population variance is NN. The use of n–1n–1 in the denominator of the sample variance is used to provide the best estimate for the population variance, in the sense that if repeated samples of size nn are taken and the sample mean computed each time, then the average of those sample means will tend to the population mean as the number of repeated samples increase.

It is important to note that in many data science applications, population data is unavailable, and so we typically calculate the sample variance. For example, if a researcher wanted to estimate the percentage of smokers for all adults in the United States, it would be impractical to collect data from every adult in the United States.

Notice that the sample variance is a sum of squares. Its units of measurement are squares of the units of measurement of the original data. Since these square units are different than the units in the original data, this can be confusing. By contrast, standard deviation is measured in the same units as the original dataset, and thus the standard deviation is more commonly used to measure the spread of a dataset.

Standard Deviation

The standard deviation of a dataset provides a numerical measure of the overall amount of variation *in a dataset in the same units as the data*; it can be used to determine whether a particular data value is close to or far from the mean, relative to the typical distance from the mean.

The standard deviation is always positive or zero. It is small when the data values are all concentrated close to the mean,

exhibiting little variation, or spread. It is larger when the data values are spread out more from the mean, exhibiting more variation. A smaller standard deviation implies less variability in a dataset, and a larger standard deviation implies more variability in a dataset.

Suppose that we are studying the variability of two companies (A and B) with respect to employee salaries. The average salary for both companies is $60,000. For Company A, the standard deviation of salaries is $8,000, whereas the standard deviation for salaries for Company B is $19,000. Because Company B has a higher standard deviation, we know that there is more variation in the employee salaries for Company B as compared to Company A.

There are two different formulas for calculating standard deviation. Which formula to use depends on whether the data represents a sample or a population. The notation ss is used to represent the sample standard deviation, and the notation σσ is used to represent the population standard deviation. In the formulas shown, x–x– is the sample mean, μμ is the population mean, nn is the sample size, and NN is the population size.

Formula for the sample standard deviation:

s=∑(x–x–)2n–1s=∑(x–x–)2n–1

Formula for the population standard deviation:

σ=∑(x–μ)2Nσ=∑(x–μ)2N

Notice that the sample standard deviation is calculated as the square root of the variance. This means that once the sample variance has been calculated, the sample standard deviation can then be easily calculated as the square root of the sample variance, as in Example 3.7.

Example 3.7

A biologist calculates that the sample variance for the amount of plant growth for a sample of plants is 8.7 cm^2. Calculate the sample standard deviation.

Answer available at the end of this section.

Example 3.8

Assume the sample variance (s2s2) for a dataset is calculated as 42.2. Based on this, calculate the sample standard deviation.

Answer available at the end of this section.

Notice that the sample variance is the square of the sample standard deviation, so if the sample standard deviation is known, the sample variance can easily be calculated.

Use of Technology for Calculating Measures of Variability

Due to the complexity of calculating variance and standard deviation, technology is typically utilized to calculate these measures of variability. For example, refer to the examples shown in Coefficient of Variation on using Python for measures of variation.

Coefficient of Variation

A data scientist might be interested in comparing variation with different units of measurement of different means, and in these scenarios the coefficient of variation (CV) can be used. The coefficient of variation measures the variation of a dataset

by calculating the standard deviation as a percentage of the mean. Note: coefficient of variation is typically expressed in a percentage format.

CV = σ μ × 100 % Sample CV = s x – × 100 % CV = σ μ × 100 % Sample CV = s x – × 100 %

Example 3.9

Compare the relative variability for Company A versus Company B using the coefficient of variation, based on the following sample data:

Company A: Sample Mean=$68,000, Sample Standard Deviation=$9,200Sample Mean=$68,000, Sample Standard Deviation=$9,200

Company B: Sample Mean=$71,000, Sample Standard Deviation=$6,400Sample Mean=$71,000, Sample Standard Deviation=$6,400

Answer available at the end of this section.

Using Python for Measures of Variation

`DataFrame.describe()` computes standard deviation as well on each column of a dataset. The **std** lists the standard deviation of each column (See Figure 3.3).

	Unnamed: 0	Rating	Duration	US_Gross_Million	Worldwide_Gross_Million
count	966.00000	966.000000	966.000000	966.000000	966.000000
mean	483.50000	6.814286	117.506211	156.158975	410.140600
std	279.00448	0.894383	21.615612	110.629617	294.758791
min	1.00000	3.300000	69.000000	0.010000	176.600000
25%	242.25000	6.200000	101.250000	90.832500	223.277500
50%	483.50000	6.800000	116.000000	129.245000	309.345000
75%	724.75000	7.400000	130.000000	187.090000	472.645000
max	966.00000	9.200000	238.000000	936.660000	2847.400000

Figure 3.3 The Output of

`DataFrame.describe()`

with the Movie Profit Dataset

Answers

For Dataset A, the maximum data value is 100 and the minimum data value is 40.
The range is then calculated as:

Range = Max – Min Range = 100 – 40 Range = 60 Range = Max – Min Range = 100 – 40 Range = 60

For Dataset B, the maximum data value is 71 and the minimum data value is 69.
The range is then calculated as:

Range = Max – Min Range = 71 – 69 Range = 2 Range = Max – Min Range = 71 – 69 Range = 2

The range clearly indicates that there is much less spread in Dataset B as compared to Dataset A.

The sample standard deviation (ss) is calculated as the square root of the variance.

s = s 2 = 8.7 = 2.9 cm s = s 2 = 8.7 = 2.9 cm

The sample standard deviation (ss) is calculated as the square root of the variance.

s = s 2 = 42.2 = 6.5 years s = s 2 = 42.2 = 6.5 years

This result indicates that the standard deviation is about 6.5 years.

Calculate the coefficient of variation for each company:

CV for Company A = s x – × 100 % = 9,200 68,000 × 100 % = 13.5 % CV for Company B = s x – × 100 % = 6,400 71,000 × 100 % = 9.0 % CV for Company A = s x – × 100 % = 9,200 68,000 × 100 % = 13.5 % CV for Company B = s x – × 100 % = 6,400 71,000 × 100 % = 9.0 %

Company A exhibits more variability relative to the mean as compared to Company B.

This page titled 9.3: Measures of Variation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

9.4: There's More to Research Methods than Statistics

So far, most of what I've talked about is statistics, and so you'd be forgiven for thinking that statistics is all I care about in life. To be fair, you wouldn't be far wrong, but research methodology is a broader concept than statistics. So most research methods courses will cover a lot of topics that relate much more to the pragmatics of research design, and in particular the issues that you encounter when trying to do research with humans. However, about 99% of student *fears* relate to the statistics part of the course, so I've focused on the stats in this discussion, and hopefully I've convinced you that statistics matters, and more importantly, that it's not to be feared. That being said, it's pretty typical for introductory research methods classes to be very stats-heavy. This is not (usually) because the lecturers are evil people. Quite the contrary, in fact. Introductory classes focus a lot on the statistics because you almost always find yourself needing statistics before you need the other research methods training. Why? Because almost all of your assignments in other classes will rely on statistical training, to a much greater extent than they rely on other methodological tools. It's not common for undergraduate assignments to require you to design your own study from the ground up (in which case you would need to know a lot about research design), but it *is* common for assignments to ask you to analyse and interpret data that were collected in a study that someone else designed (in which case you need statistics). In that sense, from the perspective of allowing you to do well in all your other classes, the statistics is more urgent.

But note that "urgent" is different from "important" – they both matter. I really do want to stress that research design is just as important as data analysis, and this book does spend a fair amount of time on it. However, while statistics has a kind of universality, and provides a set of core tools that are useful for most types of psychological research, the research methods side isn't quite so universal. There are some general principles that everyone should think about, but a lot of research design is very idiosyncratic, and is specific to the area of research that you want to engage in. To the extent that it's the details that matter, those details don't usually show up in an introductory stats and research methods class.

This page titled 9.4: There's More to Research Methods than Statistics is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

9.5: A brief Introduction to Research Design

In this chapter, we're going to start thinking about the basic ideas that go into designing a study, collecting data, checking whether your data collection works, and so on. It won't give you enough information to allow you to design studies of your own, but it will give you a lot of the basic tools that you need to assess the studies done by other people. However, since the focus of this book is much more on data analysis than on data collection, I'm only giving a very brief overview. Note that this chapter is "special" in two ways. Firstly, it's much more psychology-specific than the later chapters. Secondly, it focuses much more heavily on the scientific problem of research methodology, and much less on the statistical problem of data analysis. Nevertheless, the two problems are related to one another, so it's traditional for stats textbooks to discuss the problem in a little detail. This chapter relies heavily on Campbell and Stanley (1963) for the discussion of study design, and Stevens (1946) for the discussion of scales of measurement. Later versions will attempt to be more precise in the citations.

This page titled 9.5: A brief Introduction to Research Design is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

9.6: Introduction to Psychological Measurement

The first thing to understand is data collection can be thought of as a kind of **measurement**. That is, what we're trying to do here is measure something about human behaviour or the human mind. What do I mean by "measurement"?

Some thoughts about psychological measurement

Measurement itself is a subtle concept, but basically it comes down to finding some way of assigning numbers, or labels, or some other kind of well-defined descriptions to "stuff". So, any of the following would count as a psychological measurement:

- My **age** is *33 years*.
- I *do not* **like anchovies**.
- My **chromosomal gender** is *male*.
- My **self-identified gender** is *male*.

In the short list above, the **bolded part** is "the thing to be measured", and the *italicized part* is "the measurement itself". In fact, we can expand on this a little bit, by thinking about the set of possible measurements that could have arisen in each case:

- My **age** (in years) could have been *0, 1, 2, 3* ..., etc. The upper bound on what my age could possibly be is a bit fuzzy, but in practice you'd be safe in saying that the largest possible age is *150*, since no human has ever lived that long.
- When asked if I **like anchovies**, I might have said that *I do*, or *I do not*, or *I have no opinion*, or *I sometimes do*.
- My **chromosomal gender** is almost certainly going to be *male (XY)* or *female (XX)*, but there are a few other possibilities. I could also have *Klinfelter's syndrome (XXY)*, which is more similar to male than to female. And I imagine there are other possibilities too.
- My **self-identified gender** is also very likely to be *male* or *female*, but it doesn't have to agree with my chromosomal gender. I may also choose to identify with *neither*, or to explicitly call myself *transgender*.

As you can see, for some things (like age) it seems fairly obvious what the set of possible measurements should be, whereas for other things it gets a bit tricky. But I want to point out that even in the case of someone's age, it's much more subtle than this. For instance, in the example above, I assumed that it was okay to measure age in years. But if you're a developmental psychologist, that's way too crude, and so you often measure age in *years and months* (if a child is 2 years and 11 months, this is usually written as "2;11"). If you're interested in newborns, you might want to measure age in *days since birth*, maybe even *hours since birth*. In other words, the way in which you specify the allowable measurement values is important.

Looking at this a bit more closely, you might also realise that the concept of "age" isn't actually all that precise. In general, when we say "age" we implicitly mean "the length of time since birth". But that's not always the right way to do it. Suppose you're interested in how newborn babies control their eye movements. If you're interested in kids that young, you might also start to worry that "birth" is not the only meaningful point in time to care about. If Baby Alice is born 3 weeks premature and Baby Bianca is born 1 week late, would it really make sense to say that they are the "same age" if we encountered them "2 hours after birth"? In one sense, yes: by social convention, we use birth as our reference point for talking about age in everyday life, since it defines the amount of time the person has been operating as an independent entity in the world, but from a scientific perspective that's not the only thing we care about. When we think about the biology of human beings, it's often useful to think of ourselves as organisms that have been growing and maturing since conception, and from that perspective Alice and Bianca aren't the same age at all. So you might want to define the concept of "age" in two different ways: the length of time since conception, and the length of time since birth. When dealing with adults, it won't make much difference, but when dealing with newborns it might.

Moving beyond these issues, there's the question of methodology. What specific "measurement method" are you going to use to find out someone's age? As before, there are lots of different possibilities:

- You could just ask people "how old are you?" The method of self-report is fast, cheap and easy, but it only works with people old enough to understand the question, and some people lie about their age.
- You could ask an authority (e.g., a parent) "how old is your child?" This method is fast, and when dealing with kids it's not all that hard since the parent is almost always around. It doesn't work as well if you want to know "age since conception", since a lot of parents can't say for sure when conception took place. For that, you might need a different authority (e.g., an obstetrician).

- You could look up official records, like birth certificates. This is time consuming and annoying, but it has its uses (e.g., if the person is now dead).

Operationalization: defining your measurement

All of the ideas discussed in the previous section all relate to the concept of **operationalization**. To be a bit more precise about the idea, operationalization is the process by which we take a meaningful but somewhat vague concept, and turn it into a precise measurement. The process of operationalization can involve several different things:

- Being precise about what you are trying to measure: For instance, does "age" mean "time since birth" or "time since conception" in the context of your research?
- Determining what method you will use to measure it: Will you use self-report to measure age, ask a parent, or look up an official record? If you're using self-report, how will you phrase the question?
- Defining the set of the allowable values that the measurement can take: Note that these values don't always have to be numerical, though they often are. When measuring age, the values are numerical, but we still need to think carefully about what numbers are allowed. Do we want age in years, years and months, days, hours? Etc. For other types of measurements (e.g., gender), the values aren't numerical. But, just as before, we need to think about what values are allowed. If we're asking people to self-report their gender, what options to we allow them to choose between? Is it enough to allow only "male" or "female"? Do you need an "other" option? Or should we not give people any specific options, and let them answer in their own words? And if you open up the set of possible values to include all verbal response, how will you interpret their answers?

Operationalization is a tricky business, and there's no "one, true way" to do it. The way in which you choose to operationalize the informal concept of "age" or "gender" into a formal measurement depends on what you need to use the measurement for. Often you'll find that the community of scientists who work in your area have some fairly well-established ideas for how to go about it. In other words, operationalization needs to be thought through on a case by case basis. Nevertheless, while there a lot of issues that are specific to each individual research project, there are some aspects to it that are pretty general.

Before moving on, I want to take a moment to clear up our terminology, and in the process introduce one more term. Here are four different things that are closely related to each other:

- **A theoretical construct**. This is the thing that you're trying to take a measurement of, like "age", "gender" or an "opinion". A theoretical construct can't be directly observed, and often they're actually a bit vague.
- **A measure**. The measure refers to the method or the tool that you use to make your observations. A question in a survey, a behavioural observation or a brain scan could all count as a measure.
- **An operationalization**. The term "operationalization" refers to the logical connection between the measure and the theoretical construct, or to the process by which we try to derive a measure from a theoretical construct.
- **A variable**. Finally, a new term. A variable is what we end up with when we apply our measure to something in the world. That is, variables are the actual "data" that we end up with in our data sets.

In practice, even scientists tend to blur the distinction between these things, but it's very helpful to try to understand the differences.

This page titled 9.6: Introduction to Psychological Measurement is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

9.7: Scales of measurement

As the previous section indicates, the outcome of a psychological measurement is called a variable. But not all variables are of the same qualitative type, and it's very useful to understand what types there are. A very useful concept for distinguishing between different types of variables is what's known as **scales of measurement**.

Nominal scale

A **nominal scale** variable (also referred to as a **categorical** variable) is one in which there is no particular relationship between the different possibilities: for these kinds of variables it doesn't make any sense to say that one of them is "bigger" or "better" than any other one, and it absolutely doesn't make any sense to average them. The classic example for this is "eye colour". Eyes can be blue, green and brown, among other possibilities, but none of them is any "better" than any other one. As a result, it would feel really weird to talk about an "average eye colour". Similarly, gender is nominal too: male isn't better or worse than female, neither does it make sense to try to talk about an "average gender". In short, nominal scale variables are those for which the only thing you can say about the different possibilities is that they are different. That's it.

Let's take a slightly closer look at this. Suppose I was doing research on how people commute to and from work. One variable I would have to measure would be what kind of transportation people use to get to work. This "transport type" variable could have quite a few possible values, including: "train", "bus", "car", "bicycle", etc. For now, let's suppose that these four are the only possibilities, and suppose that when I ask 100 people how they got to work today, and I get this:

Transportation	Number of people
(1) Train	12
(2) Bus	30
(3) Car	48
(4) Bicycle	10

So, what's the average transportation type? Obviously, the answer here is that there isn't one. It's a silly question to ask. You can say that travel by car is the most popular method, and travel by train is the least popular method, but that's about all. Similarly, notice that the order in which I list the options isn't very interesting. I could have chosen to display the data like this and nothing really changes.

Transportation	Number of people
(3) Car	48
(1) Train	12
(4) Bicycle	10
(2) Bus	30

Ordinal scale

Ordinal scale variables have a bit more structure than nominal scale variables, but not by a lot. An ordinal scale variable is one in which there is a natural, meaningful way to order the different possibilities, but you can't do anything else. The usual example given of an ordinal variable is "finishing position in a race". You *can* say that the person who finished first was faster than the person who finished second, but you *don't* know how much faster. As a consequence we know that 1st $>$ 2nd, and we know that 2nd $>$ 3rd, but the difference between 1st and 2nd might be much larger than the difference between 2nd and 3rd.

Here's an more psychologically interesting example. Suppose I'm interested in people's attitudes to climate change, and I ask them to pick one of these four statements that most closely matches their beliefs:

1. Temperatures are rising, because of human activity
2. Temperatures are rising, but we don't know why
3. Temperatures are rising, but not because of humans
4. Temperatures are not rising

Notice that these four statements actually do have a natural ordering, in terms of "the extent to which they agree with the current science". Statement 1 is a close match, statement 2 is a reasonable match, statement 3 isn't a very good match, and statement 4 is in strong opposition to the science. So, in terms of the thing I'm interested in (the extent to which people endorse the science), I can order the items as $1 > 2 > 3 > 4$. Since this ordering exists, it would be very weird to list the options like this...

3. Temperatures are rising, but not because of humans
4. Temperatures are rising, because of human activity
5. Temperatures are not rising
6. Temperatures are rising, but we don't know why

...because it seems to violate the natural "structure" to the question.

So, let's suppose I asked 100 people these questions, and got the following answers:

	Number
(1) Temperatures are rising, because of human activity	51
(2) Temperatures are rising, but we don't know why	20
(3) Temperatures are rising, but not because of humans	10
(4) Temperatures are not rising	19

When analysing these data, it seems quite reasonable to try to group (1), (2) and (3) together, and say that 81 of 100 people were willing to *at least partially* endorse the science. And it's *also* quite reasonable to group (2), (3) and (4) together and say that 49 of 100 people registered *at least some disagreement* with the dominant scientific view. However, it would be entirely bizarre to try to group (1), (2) and (4) together and say that 90 of 100 people said...what? There's nothing sensible that allows you to group those responses together at all.

That said, notice that while we *can* use the natural ordering of these items to construct sensible groupings, what we *can't* do is average them. For instance, in my simple example here, the "average" response to the question is 1.97. If you can tell me what that means, I'd love to know. Because that sounds like gibberish to me!

Interval scale

In contrast to nominal and ordinal scale variables, **interval scale** and ratio scale variables are variables for which the numerical value is genuinely meaningful. In the case of interval scale variables, the *differences* between the numbers are interpretable, but the variable doesn't have a "natural" zero value. A good example of an interval scale variable is measuring temperature in degrees celsius. For instance, if it was 15° yesterday and 18° today, then the 3° difference between the two is genuinely meaningful. Moreover, that 3° difference is *exactly the same* as the 3° difference between 7° and 10°. In short, addition and subtraction are meaningful for interval scale variables.

However, notice that the 0° does not mean "no temperature at all": it actually means "the temperature at which water freezes", which is pretty arbitrary. As a consequence, it becomes pointless to try to multiply and divide temperatures. It is wrong to say that 20° is *twice as hot* as 10°, just as it is weird and meaningless to try to claim that 20° is negative two times as hot as -10°.

Again, lets look at a more psychological example. Suppose I'm interested in looking at how the attitudes of first-year university students have changed over time. Obviously, I'm going to want to record the year in which each student started. This is an interval scale variable. A student who started in 2003 did arrive 5 years before a student who started in 2008. However, it would be completely insane for me to divide 2008 by 2003 and say that the second student started "1.0024 times later" than the first one. That doesn't make any sense at all.

Ratio scale

The fourth and final type of variable to consider is a **ratio scale** variable, in which zero really means zero, and it's okay to multiply and divide. A good psychological example of a ratio scale variable is response time (RT). In a lot of tasks it's very common to record the amount of time somebody takes to solve a problem or answer a question, because it's an indicator of how difficult the task is. Suppose that Alan takes 2.3 seconds to respond to a question, whereas Ben takes 3.1 seconds. As with an interval scale variable, addition and subtraction are both meaningful here. Ben really did take $3.1 - 2.3 = 0.8$ seconds longer than Alan did. However, notice that multiplication and division also make sense here too: Ben took $3.1/2.3 = 1.35$ times as long as Alan did to answer the question. And the reason why you can do this is that, for a ratio scale variable such as RT, "zero seconds" really does mean "no time at all".

Continuous versus discrete variables

There's a second kind of distinction that you need to be aware of, regarding what types of variables you can run into. This is the distinction between continuous variables and discrete variables. The difference between these is as follows:

- A **continuous variable** is one in which, for any two values that you can think of, it's always logically possible to have another value in between.
- A **discrete variable** is, in effect, a variable that isn't continuous. For a discrete variable, it's sometimes the case that there's nothing in the middle.

These definitions probably seem a bit abstract, but they're pretty simple once you see some examples. For instance, response time is continuous. If Alan takes 3.1 seconds and Ben takes 2.3 seconds to respond to a question, then it's possible for Cameron's response time to lie in between, by taking 3.0 seconds. And of course it would also be possible for David to take 3.031 seconds to respond, meaning that his RT would lie in between Cameron's and Alan's. And while in practice it might be impossible to measure RT that precisely, it's certainly possible in principle. Because we can always find a new value for RT in between any two other ones, we say that RT is continuous.

Discrete variables occur when this rule is violated. For example, nominal scale variables are always discrete: there isn't a type of transportation that falls "in between" trains and bicycles, not in the strict mathematical way that 2.3 falls in between 2 and 3. So transportation type is discrete. Similarly, ordinal scale variables are always discrete: although "2nd place" does fall between "1st place" and "3rd place", there's nothing that can logically fall in between "1st place" and "2nd place". Interval scale and ratio scale variables can go either way. As we saw above, response time (a ratio scale variable) is continuous. Temperature in degrees celsius (an interval scale variable) is also continuous. However, the year you went to school (an interval scale variable) is discrete. There's no year in between 2002 and 2003. The number of questions you get right on a true-or-false test (a ratio scale variable) is also discrete: since a true-or-false question doesn't allow you to be "partially correct", there's nothing in between 5/10 and 6/10. The table summarizes the relationship between the scales of measurement and the discrete/continuity distinction. Cells with a tick mark correspond to things that are possible. I'm trying to hammer this point home, because (a) some textbooks get this wrong, and (b) people very often say things like "discrete variable" when they mean "nominal scale variable". It's very unfortunate.

The relationship between the scales of measurement and the discrete/continuity distinction. Cells with an x correspond to things that are possible.

	continuous	discrete
nominal		x
ordinal		x
interval	x	x
ratio	x	x

Some complexities

Okay, I know you're going to be shocked to hear this, but ...the real world is much messier than this little classification scheme suggests. Very few variables in real life actually fall into these nice neat categories, so you need to be kind of careful not to treat the scales of measurement as if they were hard and fast rules. It doesn't work like that: they're guidelines, intended to help you

think about the situations in which you should treat different variables differently. Nothing more.

So let's take a classic example, maybe *the* classic example, of a psychological measurement tool: the **Likert scale**. The humble Likert scale is the bread and butter tool of all survey design. You yourself have filled out hundreds, maybe thousands of them, and odds are you've even used one yourself. Suppose we have a survey question that looks like this:

> Which of the following best describes your opinion of the statement that "all pirates are freaking awesome"
>
> ...

and then the options presented to the participant are these:

> (1) Strongly disagree
> (2) Disagree
> (3) Neither agree nor disagree
> (4) Agree
> (5) Strongly agree

This set of items is an example of a 5-point Likert scale: people are asked to choose among one of several (in this case 5) clearly ordered possibilities, generally with a verbal descriptor given in each case. However, it's not necessary that all items be explicitly described. This is a perfectly good example of a 5-point Likert scale too:

> (1) Strongly disagree
> (2)
> (3)
> (4)
> (5) Strongly agree

Likert scales are very handy, if somewhat limited, tools. The question is, what kind of variable are they? They're obviously discrete, since you can't give a response of 2.5. They're obviously not nominal scale, since the items are ordered; and they're not ratio scale either, since there's no natural zero.

But are they ordinal scale or interval scale? One argument says that we can't really prove that the difference between "strongly agree" and "agree" is of the same size as the difference between "agree" and "neither agree nor disagree". In fact, in everyday life it's pretty obvious that they're not the same at all. So this suggests that we ought to treat Likert scales as ordinal variables. On the other hand, in practice most participants do seem to take the whole "on a scale from 1 to 5" part fairly seriously, and they tend to act as if the differences between the five response options were fairly similar to one another. As a consequence, a lot of researchers treat Likert scale data as if it were interval scale. It's not interval scale, but in practice it's close enough that we usually think of it as being **quasi-interval scale**.

This page titled 9.7: Scales of measurement is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

9.8: Assessing the Reliability of a Measurement

At this point we've thought a little bit about how to operationalize a theoretical construct and thereby create a psychological measure; and we've seen that by applying psychological measures we end up with variables, which can come in many different types. At this point, we should start discussing the obvious question: is the measurement any good? We'll do this in terms of two related ideas: *reliability* and *validity*. Put simply, the **reliability** of a measure tells you how *precisely* you are measuring something, whereas the validity of a measure tells you how *accurate* the measure is.

Reliability is actually a very simple concept: it refers to the repeatability or consistency of your measurement. The measurement of my weight by means of a "bathroom scale" is very reliable: if I step on and off the scales over and over again, it'll keep giving me the same answer. Measuring my intelligence by means of "asking my mom" is very unreliable: some days she tells me I'm a bit thick, and other days she tells me I'm a complete moron. Notice that this concept of reliability is different to the question of whether the measurements are correct (the correctness of a measurement relates to it's validity). If I'm holding a sack of potatos when I step on and off of the bathroom scales, the measurement will still be reliable: it will always give me the same answer. However, this highly reliable answer doesn't match up to my true weight at all, therefore it's wrong. In technical terms, this is a *reliable but invalid* measurement. Similarly, while my mom's estimate of my intelligence is a bit unreliable, she might be right. Maybe I'm just not too bright, and so while her estimate of my intelligence fluctuates pretty wildly from day to day, it's basically right. So that would be an *unreliable but valid* measure. Of course, to some extent, notice that if my mum's estimates are too unreliable, it's going to be very hard to figure out which one of her many claims about my intelligence is actually the right one. To some extent, then, a very unreliable measure tends to end up being invalid for practical purposes; so much so that many people would say that reliability is necessary (but not sufficient) to ensure validity.

Okay, now that we're clear on the distinction between reliability and validity, let's have a think about the different ways in which we might measure reliability:

- **Test-retest reliability**. This relates to consistency over time: if we repeat the measurement at a later date, do we get a the same answer?
- **Inter-rater reliability**. This relates to consistency across people: if someone else repeats the measurement (e.g., someone else rates my intelligence) will they produce the same answer?
- **Parallel forms reliability**. This relates to consistency across theoretically-equivalent measurements: if I use a different set of bathroom scales to measure my weight, does it give the same answer?
- **Internal consistency reliability**. If a measurement is constructed from lots of different parts that perform similar functions (e.g., a personality questionnaire result is added up across several questions) do the individual parts tend to give similar answers.

Not all measurements need to possess all forms of reliability. For instance, educational assessment can be thought of as a form of measurement. One of the subjects that I teach, *Computational Cognitive Science*, has an assessment structure that has a research component and an exam component (plus other things). The exam component is *intended* to measure something different from the research component, so the assessment as a whole has low internal consistency. However, within the exam there are several questions that are intended to (approximately) measure the same things, and those tend to produce similar outcomes; so the exam on its own has a fairly high internal consistency. Which is as it should be. You should only demand reliability in those situations where you want to be measure the same thing!

This page titled 9.8: Assessing the Reliability of a Measurement is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

9.9: The role of variables — predictors and outcomes

Okay, I've got one last piece of terminology that I need to explain to you before moving away from variables. Normally, when we do some research we end up with lots of different variables. Then, when we analyse our data we usually try to explain some of the variables in terms of some of the other variables. It's important to keep the two roles "thing doing the explaining" and "thing being explained" distinct. So let's be clear about this now. Firstly, we might as well get used to the idea of using mathematical symbols to describe variables, since it's going to happen over and over again. Let's denote the "to be explained" variable Y, and denote the variables "doing the explaining" as X_1, X_2, etc.

Now, when we doing an analysis, we have different names for X and Y, since they play different roles in the analysis. The classical names for these roles are **independent variable** (IV) and **dependent variable** (DV). The IV is the variable that you use to do the explaining (i.e., X) and the DV is the variable being explained (i.e., Y). The logic behind these names goes like this: if there really is a relationship between X and Y then we can say that Y depends on X, and if we have designed our study "properly" then X isn't dependent on anything else. However, I personally find those names horrible: they're hard to remember and they're highly misleading, because (a) the IV is never actually "independent of everything else" and (b) if there's no relationship, then the DV doesn't actually depend on the IV. And in fact, because I'm not the only person who thinks that IV and DV are just awful names, there are a number of alternatives that I find more appealing.

For example, in an experiment the IV refers to the **manipulation**, and the DV refers to the **measurement**. So, we could use **manipulated variable** (independent variable) and **measured variable** (dependent variable).

The terminology used to distinguish between different roles that a variable can play when analysing a data set.

role of the variable	classical name	modern name
"to be explained"	dependent variable (DV)	Measurement
"to do the explaining"	independent variable (IV)	Manipulation

We could also use **predictors** and **outcomes**. The idea here is that what you're trying to do is use X (the predictors) to make guesses about Y (the outcomes). This is summarized in the table:

The terminology used to distinguish between different roles that a variable can play when analysing a data set.

role of the variable	classical name	modern name
"to be explained"	dependent variable (DV)	outcome
"to do the explaining"	independent variable (IV)	predictor

This page titled 9.9: The role of variables — predictors and outcomes is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

9.10: Experimental and non-experimental research

One of the big distinctions that you should be aware of is the distinction between "experimental research" and "non-experimental research". When we make this distinction, what we're really talking about is the degree of control that the researcher exercises over the people and events in the study.

Experimental research

The key features of **experimental research** is that the researcher controls all aspects of the study, especially what participants experience during the study. In particular, the researcher manipulates or varies something (IVs), and then allows the outcome variable (DV) to vary naturally. The idea here is to deliberately vary the something in the world (IVs) to see if it has any causal effects on the outcomes. Moreover, in order to ensure that there's no chance that something other than the manipulated variable is causing the outcomes, everything else is kept constant or is in some other way "balanced" to ensure that they have no effect on the results. In practice, it's almost impossible to *think* of everything else that might have an influence on the outcome of an experiment, much less keep it constant. The standard solution to this is **randomization**: that is, we randomly assign people to different groups, and then give each group a different treatment (i.e., assign them different values of the predictor variables). We'll talk more about randomization later in this course, but for now, it's enough to say that what randomization does is minimize (but not eliminate) the chances that there are any systematic difference between groups.

Let's consider a very simple, completely unrealistic and grossly unethical example. Suppose you wanted to find out if smoking causes lung cancer. One way to do this would be to find people who smoke and people who don't smoke, and look to see if smokers have a higher rate of lung cancer. This is *not* a proper experiment, since the researcher doesn't have a lot of control over who is and isn't a smoker. And this really matters: for instance, it might be that people who choose to smoke cigarettes also tend to have poor diets, or maybe they tend to work in asbestos mines, or whatever. The point here is that the groups (smokers and non-smokers) actually differ on lots of things, not *just* smoking. So it might be that the higher incidence of lung cancer among smokers is caused by something else, not by smoking per se. In technical terms, these other things (e.g. diet) are called "confounds", and we'll talk about those in just a moment.

In the meantime, let's now consider what a proper experiment might look like. Recall that our concern was that smokers and non-smokers might differ in lots of ways. The solution, as long as you have no ethics, is to *control* who smokes and who doesn't. Specifically, if we randomly divide participants into two groups, and force half of them to become smokers, then it's very unlikely that the groups will differ in any respect other than the fact that half of them smoke. That way, if our smoking group gets cancer at a higher rate than the non-smoking group, then we can feel pretty confident that (a) smoking does cause cancer and (b) we're murderers.

Non-experimental research

Non-experimental research is a broad term that covers "any study in which the researcher doesn't have quite as much control as they do in an experiment". Obviously, control is something that scientists like to have, but as the previous example illustrates, there are lots of situations in which you can't or shouldn't try to obtain that control. Since it's grossly unethical (and almost certainly criminal) to force people to smoke in order to find out if they get cancer, this is a good example of a situation in which you really shouldn't try to obtain experimental control. But there are other reasons too. Even leaving aside the ethical issues, our "smoking experiment" does have a few other issues. For instance, when I suggested that we "force" half of the people to become smokers, I must have been talking about *starting* with a sample of non-smokers, and then forcing them to become smokers. While this sounds like the kind of solid, evil experimental design that a mad scientist would love, it might not be a very sound way of investigating the effect in the real world. For instance, suppose that smoking only causes lung cancer when people have poor diets, and suppose also that people who normally smoke do tend to have poor diets. However, since the "smokers" in our experiment aren't "natural" smokers (i.e., we forced non-smokers to become smokers; they didn't take on all of the other normal, real life characteristics that smokers might tend to possess) they probably have better diets. As such, in this silly example they wouldn't get lung cancer, and our experiment will fail, because it violates the structure of the "natural" world (the technical name for this is an "artifactual" result; see later).

One distinction worth making between two types of non-experimental research is the difference between **quasi-experimental research** and **case studies**. The example I discussed earlier – in which we wanted to examine incidence of lung cancer among

smokers and non-smokers, without trying to control who smokes and who doesn't – is a quasi-experimental design. That is, it's the same as an experiment, but we don't control the predictors (IVs). We can still use statistics to analyse the results, it's just that we have to be a lot more careful.

The alternative approach, case studies, aims to provide a very detailed description of one or a few instances. In general, you can't use statistics to analyse the results of case studies, and it's usually very hard to draw any general conclusions about "people in general" from a few isolated examples. However, case studies are very useful in some situations. Firstly, there are situations where you don't have any alternative: neuropsychology has this issue a lot. Sometimes, you just can't find a lot of people with brain damage in a specific area, so the only thing you can do is describe those cases that you do have in as much detail and with as much care as you can. However, there's also some genuine advantages to case studies: because you don't have as many people to study, you have the ability to invest lots of time and effort trying to understand the specific factors at play in each case. This is a very valuable thing to do. As a consequence, case studies can complement the more statistically-oriented approaches that you see in experimental and quasi-experimental designs. We won't talk much about case studies in these lectures, but they are nevertheless very valuable tools!

This page titled 9.10: Experimental and non-experimental research is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

9.11: Assessing the validity of a study

More than any other thing, a scientist wants their research to be "valid". The conceptual idea behind **validity** is very simple: can you trust the results of your study? If not, the study is invalid. However, while it's easy to state, in practice it's much harder to check validity than it is to check reliability. And in all honesty, there's no precise, clearly agreed upon notion of what validity actually is. In fact, there's lots of different kinds of validity, each of which raises it's own issues, and not all forms of validity are relevant to all studies. I'm going to talk about five different types:

- Internal validity
- External validity
- Construct validity
- Face validity
- Ecological validity

To give you a quick guide as to what matters here...(1) Internal and external validity are the most important, since they tie directly to the fundamental question of whether your study really works. (2) Construct validity asks whether you're measuring what you think you are. (3) Face validity isn't terribly important except insofar as you care about "appearances". (4) Ecological validity is a special case of face validity that corresponds to a kind of appearance that you might care about a lot.

Internal validity

Internal validity refers to the extent to which you are able draw the correct conclusions about the causal relationships between variables. It's called "internal" because it refers to the relationships between things "inside" the study. Let's illustrate the concept with a simple example. Suppose you're interested in finding out whether a university education makes you write better. To do so, you get a group of first year students, ask them to write a 1000 word essay, and count the number of spelling and grammatical errors they make. Then you find some third-year students, who obviously have had more of a university education than the first-years, and repeat the exercise. And let's suppose it turns out that the third-year students produce fewer errors. And so you conclude that a university education improves writing skills. Right? Except... the big problem that you have with this experiment is that the third-year students are older, and they've had more experience with writing things. So it's hard to know for sure what the causal relationship is: Do older people write better? Or people who have had more writing experience? Or people who have had more education? Which of the above is the true *cause* of the superior performance of the third-years? Age? Experience? Education? You can't tell. This is an example of a failure of internal validity, because your study doesn't properly tease apart the *causal* relationships between the different variables.

External validity

External validity relates to the **generalizability** of your findings. That is, to what extent do you expect to see the same pattern of results in "real life" as you saw in your study. To put it a bit more precisely, any study that you do in psychology will involve a fairly specific set of questions or tasks, will occur in a specific environment, and will involve participants that are drawn from a particular subgroup. So, if it turns out that the results don't actually generalize to people and situations beyond the ones that you studied, then what you've got is a lack of external validity.

The classic example of this issue is the fact that a very large proportion of studies in psychology will use undergraduate psychology students as the participants. Obviously, however, the researchers don't care *only* about psychology students; they care about people in general. Given that, a study that uses only psych students as participants always carries a risk of lacking external validity. That is, if there's something "special" about psychology students that makes them different to the general populace in some *relevant* respect, then we may start worrying about a lack of external validity.

That said, it is absolutely critical to realize that a study that uses only psychology students does not necessarily have a problem with external validity. I'll talk about this again later, but it's such a common mistake that I'm going to mention it here. The external validity is threatened by the choice of population if (a) the population from which you sample your participants is very narrow (e.g., psych students), and (b) the narrow population that you sampled from is systematically different from the general population, *in some respect that is relevant to the psychological phenomenon that you intend to study*. The italicized part is the bit that lots of people forget: it is true that psychology undergraduates differ from the general population in lots of ways, and so a study

that uses only psych students *may* have problems with external validity. However, if those differences aren't very relevant to the phenomenon that you're studying, then there's nothing to worry about. To make this a bit more concrete, here's two extreme examples:

- You want to measure "attitudes of the general public towards psychotherapy", but all of your participants are psychology students. This study would almost certainly have a problem with external validity.
- You want to measure the effectiveness of a visual illusion, and your participants are all psychology students. This study is very unlikely to have a problem with external validity

Having just spent the last couple of paragraphs focusing on the choice of participants (since that's the big issue that everyone tends to worry most about), it's worth remembering that external validity is a broader concept. The following are also examples of things that might pose a threat to external validity, depending on what kind of study you're doing:

- People might answer a "psychology questionnaire" in a manner that doesn't reflect what they would do in real life.
- Your lab experiment on (say) "human learning" has a different structure to the learning problems people face in real life.

Construct validity

Construct validity is basically a question of whether you're measuring what you want to be measuring. A measurement has good construct validity if it is actually measuring the correct theoretical construct, and bad construct validity if it doesn't. To give very simple (if ridiculous) example, suppose I'm trying to investigate the rates with which university students cheat on their exams. And the way I attempt to measure it is by asking the cheating students to stand up in the lecture theatre so that I can count them. When I do this with a class of 300 students, 0 people claim to be cheaters. So I therefore conclude that the proportion of cheaters in my class is 0%. Clearly this is a bit ridiculous. But the point here is not that this is a very deep methodological example, but rather to explain what construct validity is. The problem with my measure is that while I'm *trying* to measure "the proportion of people who cheat" what I'm actually measuring is "the proportion of people stupid enough to own up to cheating, or bloody minded enough to pretend that they do". Obviously, these aren't the same thing! So my study has gone wrong, because my measurement has very poor construct validity.

Face validity

Face validity simply refers to whether or not a measure "looks like" it's doing what it's supposed to, nothing more. If I design a test of intelligence, and people look at it and they say "no, that test doesn't measure intelligence", then the measure lacks face validity. It's as simple as that. Obviously, face validity isn't very important from a pure scientific perspective. After all, what we care about is whether or not the measure *actually* does what it's supposed to do, not whether it *looks like* it does what it's supposed to do. As a consequence, we generally don't care very much about face validity. That said, the concept of face validity serves three useful pragmatic purposes:

- Sometimes, an experienced scientist will have a "hunch" that a particular measure won't work. While these sorts of hunches have no strict evidentiary value, it's often worth paying attention to them. Because often times people have knowledge that they can't quite verbalize, so there might be something to worry about even if you can't quite say why. In other words, when someone you trust criticizes the face validity of your study, it's worth taking the time to think more carefully about your design to see if you can think of reasons why it might go awry. Mind you, if you don't find any reason for concern, then you should probably not worry: after all, face validity really doesn't matter much.
- Often (very often), completely uninformed people will also have a "hunch" that your research is crap. And they'll criticize it on the internet or something. On close inspection, you'll often notice that these criticisms are actually focused entirely on how the study "looks", but not on anything deeper. The concept of face validity is useful for gently explaining to people that they need to substantiate their arguments further.
- Expanding on the last point, if the beliefs of untrained people are critical (e.g., this is often the case for applied research where you actually want to convince policy makers of something or other) then you *have* to care about face validity. Simply because – whether you like it or not – a lot of people will use face validity as a proxy for real validity. If you want the government to change a law on scientific, psychological grounds, then it won't matter how good your studies "really" are. If they lack face validity, you'll find that politicians ignore you. Of course, it's somewhat unfair that policy often depends more on appearance than fact, but that's how things go.

Ecological validity

Ecological validity is a different notion of validity, which is similar to external validity, but less important. The idea is that, in order to be ecologically valid, the entire set up of the study should closely approximate the real world scenario that is being investigated. In a sense, ecological validity is a kind of face validity – it relates mostly to whether the study "looks" right, but with a bit more rigour to it. To be ecologically valid, the study has to look right in a fairly specific way. The idea behind it is the intuition that a study that is ecologically valid is more likely to be externally valid. It's no guarantee, of course. But the nice thing about ecological validity is that it's much easier to check whether a study is ecologically valid than it is to check whether a study is externally valid. An simple example would be eyewitness identification studies. Most of these studies tend to be done in a university setting, often with fairly simple array of faces to look at rather than a line up. The length of time between seeing the "criminal" and being asked to identify the suspect in the "line up" is usually shorter. The "crime" isn't real, so there's no chance that the witness being scared, and there's no police officers present, so there's not as much chance of feeling pressured. These things all mean that the study *definitely* lacks ecological validity. They might (but might not) mean that it also lacks external validity.

This page titled 9.11: Assessing the validity of a study is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

9.12: Confounds, Artifacts and other Threats to Validity

If we look at the issue of validity in the most general fashion, the two biggest worries that we have are *confounds* and *artifact*. These two terms are defined in the following way:

- **Confound**: A confound is an additional, often unmeasured variable that turns out to be related to both the predictors and the outcomes. The existence of confounds threatens the internal validity of the study because you can't tell whether the predictor causes the outcome, or if the confounding variable causes it, etc.
- **Artifact**: A result is said to be "artifactual" if it only holds in the special situation that you happened to test in your study. The possibility that your result is an artifact describes a threat to your external validity, because it raises the possibility that you can't generalize your results to the actual population that you care about.

As a general rule confounds are a bigger concern for non-experimental studies, precisely because they're not proper experiments: by definition, you're leaving lots of things uncontrolled, so there's a lot of scope for confounds working their way into your study. Experimental research tends to be much less vulnerable to confounds: the more control you have over what happens during the study, the more you can prevent confounds from appearing.

However, there's always swings and roundabouts, and when we start thinking about artifacts rather than confounds, the shoe is very firmly on the other foot. For the most part, artifactual results tend to be a concern for experimental studies than for non-experimental studies. To see this, it helps to realize that the reason that a lot of studies are non-experimental is precisely because what the researcher is trying to do is examine human behaviour in a more naturalistic context. By working in a more real-world context, you lose experimental control (making yourself vulnerable to confounds) but because you tend to be studying human psychology "in the wild" you reduce the chances of getting an artifactual result. Or, to put it another way, when you take psychology out of the wild and bring it into the lab (which we usually have to do to gain our experimental control), you always run the risk of accidentally studying something different than you wanted to study: which is more or less the definition of an artifact.

Be warned though: the above is a rough guide only. It's absolutely possible to have confounds in an experiment, and to get artifactual results with non-experimental studies. This can happen for all sorts of reasons, not least of which is researcher error. In practice, it's really hard to think everything through ahead of time, and even very good researchers make mistakes. But other times it's unavoidable, simply because the researcher has ethics (e.g., see "differential attrition").

Okay. There's a sense in which almost any threat to validity can be characterized as a confound or an artifact: they're pretty vague concepts. So let's have a look at some of the most common examples...

History effects

History effects refer to the possibility that specific events may occur during the study itself that might influence the outcomes. For instance, something might happen in between a pre-test and a post-test. Or, in between testing participant 23 and participant 24. Alternatively, it might be that you're looking at an older study, which was perfectly valid for its time, but the world has changed enough since then that the conclusions are no longer trustworthy. Examples of things that would count as history effects:

- You're interested in how people think about risk and uncertainty. You started your data collection in December 2010. But finding participants and collecting data takes time, so you're still finding new people in February 2011. Unfortunately for you (and even more unfortunately for others), the Queensland floods occurred in January 2011, causing billions of dollars of damage and killing many people. Not surprisingly, the people tested in February 2011 express quite different beliefs about handling risk than the people tested in December 2010. Which (if any) of these reflects the "true" beliefs of participants? I think the answer is probably both: the Queensland floods genuinely changed the beliefs of the Australian public, though possibly only temporarily. The key thing here is that the "history" of the people tested in February is quite different to people tested in December.
- You're testing the psychological effects of a new anti-anxiety drug. So what you do is measure anxiety before administering the drug (e.g., by self-report, and taking physiological measures, let's say), then you administer the drug, and then you take the same measures afterwards. In the middle, however, because your labs are in Los Angeles, there's an earthquake, which increases the anxiety of the participants.

Maturation effects

As with history effects, **maturational effects** are fundamentally about change over time. However, maturation effects aren't in response to specific events. Rather, they relate to how people change on their own over time: we get older, we get tired, we get bored, etc. Some examples of maturation effects:

- When doing developmental psychology research, you need to be aware that children grow up quite rapidly. So, suppose that you want to find out whether some educational trick helps with vocabulary size among 3 year olds. One thing that you need to be aware of is that the vocabulary size of children that age is growing at an incredible rate (multiple words per day), all on its own. If you design your study without taking this maturational effect into account, then you won't be able to tell if your educational trick works.
- When running a very long experiment in the lab (say, something that goes for 3 hours), it's very likely that people will begin to get bored and tired, and that this maturational effect will cause performance to decline, regardless of anything else going on in the experiment

Repeated testing effects

An important type of history effect is the effect of **repeated testing**. Suppose I want to take two measurements of some psychological construct (e.g., anxiety). One thing I might be worried about is if the first measurement has an effect on the second measurement. In other words, this is a history effect in which the "event" that influences the second measurement is the first measurement itself! This is not at all uncommon. Examples of this include:

- *Learning and practice*: e.g., "intelligence" at time 2 might appear to go up relative to time 1 because participants learned the general rules of how to solve "intelligence-test-style" questions during the first testing session.
- *Familiarity with the testing situation*: e.g., if people are nervous at time 1, this might make performance go down; after sitting through the first testing situation, they might calm down a lot precisely because they've seen what the testing looks like.
- *Auxiliary changes caused by testing*: e.g., if a questionnaire assessing mood is boring, then mood at measurement at time 2 is more likely to become "bored", precisely because of the boring measurement made at time 1.

Selection bias

Selection bias is a pretty broad term. Suppose that you're running an experiment with two groups of participants, where each group gets a different "treatment", and you want to see if the different treatments lead to different outcomes. However, suppose that, despite your best efforts, you've ended up with a gender imbalance across groups (say, group A has 80% females and group B has 50% females). It might sound like this could never happen, but trust me, it can. This is an example of a selection bias, in which the people "selected into" the two groups have different characteristics. If any of those characteristics turns out to be relevant (say, your treatment works better on females than males) then you're in a lot of trouble.

Differential attrition

One quite subtle danger to be aware of is called **differential attrition**, which is a kind of selection bias that is caused by the study itself. Suppose that, for the first time ever in the history of psychology, I manage to find the perfectly balanced and representative sample of people. I start running "Dan's incredibly long and tedious experiment" on my perfect sample, but then, because my study is incredibly long and tedious, lots of people start dropping out. I can't stop this: as we'll discuss later in the chapter on research ethics, participants absolutely have the right to stop doing any experiment, any time, for whatever reason they feel like, and as researchers we are morally (and professionally) obliged to remind people that they do have this right. So, suppose that "Dan's incredibly long and tedious experiment" has a very high drop out rate. What do you suppose the odds are that this drop out is random? Answer: zero. Almost certainly, the people who remain are more conscientious, more tolerant of boredom etc than those that leave. To the extent that (say) conscientiousness is relevant to the psychological phenomenon that I care about, this attrition can decrease the validity of my results.

When thinking about the effects of differential attrition, it is sometimes helpful to distinguish between two different types. The first is **homogeneous attrition**, in which the attrition effect is the same for all groups, treatments or conditions. In the example I gave above, the differential attrition would be homogeneous if (and only if) the easily bored participants are dropping out of all of the conditions in my experiment at about the same rate. In general, the main effect of homogeneous attrition is likely to be that it makes your sample unrepresentative. As such, the biggest worry that you'll have is that the generalisability of the results

decreases: in other words, you lose external validity.

The second type of differential attrition is **heterogeneous attrition**, in which the attrition effect is different for different groups. This is a much bigger problem: not only do you have to worry about your external validity, you also have to worry about your internal validity too. To see why this is the case, let's consider a very dumb study in which I want to see if insulting people makes them act in a more obedient way. Why anyone would actually want to study that I don't know, but let's suppose I really, deeply cared about this. So, I design my experiment with two conditions. In the "treatment" condition, the experimenter insults the participant and then gives them a questionnaire designed to measure obedience. In the "control" condition, the experimenter engages in a bit of pointless chitchat and then gives them the questionnaire. Leaving aside the questionable scientific merits and dubious ethics of such a study, let's have a think about what might go wrong here. As a general rule, when someone insults me to my face, I tend to get much less co-operative. So, there's a pretty good chance that a lot more people are going to drop out of the treatment condition than the control condition. And this drop out isn't going to be random. The people most likely to drop out would probably be the people who don't care all that much about the importance of obediently sitting through the experiment. Since the most bloody minded and disobedient people all left the treatment group but not the control group, we've introduced a confound: the people who actually took the questionnaire in the treatment group were *already* more likely to be dutiful and obedient than the people in the control group. In short, in this study insulting people doesn't make them more obedient: it makes the more disobedient people leave the experiment! The internal validity of this experiment is completely shot.

Non-response bias

Non-response bias is closely related to selection bias, and to differential attrition. The simplest version of the problem goes like this. You mail out a survey to 1000 people, and only 300 of them reply. The 300 people who replied are almost certainly not a random subsample. People who respond to surveys are systematically different to people who don't. This introduces a problem when trying to generalize from those 300 people who replied, to the population at large; since you now have a very non-random sample. The issue of non-response bias is more general than this, though. Among the (say) 300 people that did respond to the survey, you might find that not everyone answers every question. If (say) 80 people chose not to answer one of your questions, does this introduce problems? As always, the answer is maybe. If the question that wasn't answered was on the last page of the questionnaire, and those 80 surveys were returned with the last page missing, there's a good chance that the missing data isn't a big deal: probably the pages just fell off. However, if the question that 80 people didn't answer was the most confrontational or invasive personal question in the questionnaire, then almost certainly you've got a problem. In essence, what you're dealing with here is what's called the problem of **missing data**. If the data that is missing was "lost" randomly, then it's not a big problem. If it's missing systematically, then it can be a big problem.

Regression to the mean

Regression to the mean is a curious variation on selection bias. It refers to any situation where you select data based on an extreme value on some measure. Because the measure has natural variation, it almost certainly means that when you take a subsequent measurement, that later measurement will be less extreme than the first one, purely by chance.

Here's an example. Suppose I'm interested in whether a psychology education has an adverse effect on very smart kids. To do this, I find the 20 psych I students with the best high school grades and look at how well they're doing at university. It turns out that they're doing a lot better than average, but they're not topping the class at university, even though they did top their classes at high school. What's going on? The natural first thought is that this must mean that the psychology classes must be having an adverse effect on those students. However, while that might very well be the explanation, it's more likely that what you're seeing is an example of "regression to the mean". To see how it works, let's take a moment to think about what is required to get the best mark in a class, regardless of whether that class be at high school or at university. When you've got a big class, there are going to be *lots* of very smart people enrolled. To get the best mark you have to be very smart, work very hard, and be a bit lucky. The exam has to ask just the right questions for your idiosyncratic skills, and you have to not make any dumb mistakes (we all do that sometimes) when answering them. And that's the thing: intelligence and hard work are transferrable from one class to the next. Luck isn't. The people who got lucky in high school won't be the same as the people who get lucky at university. That's the very definition of "luck". The consequence of this is that, when you select people at the very extreme values of one measurement (the top 20 students), you're selecting for hard work, skill and luck. But because the luck doesn't transfer to the second measurement (only the skill and work), these people will all be expected to drop a little bit when you measure them a second time (at university). So their scores fall back a little bit, back towards everyone else. This is regression to the mean.

Regression to the mean is surprisingly common. For instance, if two very tall people have kids, their children will tend to be taller than average, but not as tall as the parents. The reverse happens with very short parents: two very short parents will tend to have

short children, but nevertheless those kids will tend to be taller than the parents. It can also be extremely subtle. For instance, there have been studies done that suggested that people learn better from negative feedback than from positive feedback. However, the way that people tried to show this was to give people positive reinforcement whenever they did good, and negative reinforcement when they did bad. And what you see is that after the positive reinforcement, people tended to do worse; but after the negative reinforcement they tended to do better. But! Notice that there's a selection bias here: when people do very well, you're selecting for "high" values, and so you should *expect* (because of regression to the mean) that performance on the next trial should be worse, regardless of whether reinforcement is given. Similarly, after a bad trial, people will tend to improve all on their own. The apparent superiority of negative feedback is an artifact caused by regression to the mean (Kahneman and Tversky 1973).

Experimenter bias

Experimenter bias can come in multiple forms. The basic idea is that the experimenter, despite the best of intentions, can accidentally end up influencing the results of the experiment by subtly communicating the "right answer" or the "desired behaviour" to the participants. Typically, this occurs because the experimenter has special knowledge that the participant does not – either the right answer to the questions being asked, or knowledge of the expected pattern of performance for the condition that the participant is in, and so on. The classic example of this happening is the case study of "Clever Hans", which dates back to 1907, Pfungst (1911; Hothersall 2004). Clever Hans was a horse that apparently was able to read and count, and perform other human like feats of intelligence. After Clever Hans became famous, psychologists started examining his behaviour more closely. It turned out that – not surprisingly – Hans didn't know how to do maths. Rather, Hans was responding to the human observers around him. Because they did know how to count, and the horse had learned to change its behaviour when people changed theirs.

The general solution to the problem of experimenter bias is to engage in double blind studies, where neither the experimenter nor the participant knows which condition the participant is in, or knows what the desired behaviour is. This provides a very good solution to the problem, but it's important to recognize that it's not quite ideal, and hard to pull off perfectly. For instance, the obvious way that I could try to construct a double blind study is to have one of my Ph.D. students (one who doesn't know anything about the experiment) run the study. That feels like it should be enough. The only person (me) who knows all the details (e.g., correct answers to the questions, assignments of participants to conditions) has no interaction with the participants, and the person who does all the talking to people (the Ph.D. student) doesn't know anything. Except, that last part is very unlikely to be true. In order for the Ph.D. student to run the study effectively, they need to have been briefed by me, the researcher. And, as it happens, the Ph.D. student also knows me, and knows a bit about my general beliefs about people and psychology (e.g., I tend to think humans are much smarter than psychologists give them credit for). As a result of all this, it's almost impossible for the experimenter to avoid knowing a little bit about what expectations I have. And even a little bit of knowledge can have an effect: suppose the experimenter accidentally conveys the fact that the participants are expected to do well in this task. Well, there's a thing called the "Pygmalion effect": if you expect great things of people, they'll rise to the occasion; but if you expect them to fail, they'll do that too. In other words, the expectations become a self-fulfilling prophesy.

Demand effects and reactivity

When talking about experimenter bias, the worry is that the experimenter's knowledge or desires for the experiment are communicated to the participants, and that these effect people's behaviour Rosenthal (1966). However, even if you manage to stop this from happening, it's almost impossible to stop people from knowing that they're part of a psychological study. And the mere fact of knowing that someone is watching/studying you can have a pretty big effect on behaviour. This is generally referred to as **reactivity** or **demand effects**. The basic idea is captured by the Hawthorne effect: people alter their performance because of the attention that the study focuses on them. The effect takes its name from a the "Hawthorne Works" factory outside of Chicago (Adair 1984). A study done in the 1920s looking at the effects of lighting on worker productivity at the factory turned out to be an effect of the fact that the workers knew they were being studied, rather than the lighting.

To get a bit more specific about some of the ways in which the mere fact of being in a study can change how people behave, it helps to think like a social psychologist and look at some of the *role*s that people might adopt during an experiment, but might not adopt if the corresponding events were occurring in the real world:

- The *good participant* tries to be too helpful to the researcher: he or she seeks to figure out the experimenter's hypotheses and confirm them.
- The *negative participant* does the exact opposite of the good participant: he or she seeks to break or destroy the study or the hypothesis in some way.
- The *faithful participant* is unnaturally obedient: he or she seeks to follow instructions perfectly, regardless of what might have

happened in a more realistic setting.
- The *apprehensive participant* gets nervous about being tested or studied, so much so that his or her behaviour becomes highly unnatural, or overly socially desirable.

Placebo effects

The **placebo effect** is a specific type of demand effect that we worry a lot about. It refers to the situation where the mere fact of being treated causes an improvement in outcomes. The classic example comes from clinical trials: if you give people a completely chemically inert drug and tell them that it's a cure for a disease, they will tend to get better faster than people who aren't treated at all. In other words, it is people's belief that they are being treated that causes the improved outcomes, not the drug.

Situation, measurement and subpopulation effects

In some respects, these terms are a catch-all term for "all other threats to external validity". They refer to the fact that the choice of subpopulation from which you draw your participants, the location, timing and manner in which you run your study (including who collects the data) and the tools that you use to make your measurements might all be influencing the results. Specifically, the worry is that these things might be influencing the results in such a way that the results won't generalize to a wider array of people, places and measures.

Fraud, deception and self-deception

> *It is difficult to get a man to understand something, when his salary depends on his not understanding it.*
>
> – Upton Sinclair

One final thing that I feel like I should mention. While reading what the textbooks often have to say about assessing the validity of the study, I couldn't help but notice that they seem to make the assumption that the researcher is honest. I find this hilarious. While the vast majority of scientists are honest, in my experience at least, some are not. Not only that, as I mentioned earlier, scientists are not immune to belief bias – it's easy for a researcher to end up deceiving themselves into believing the wrong thing, and this can lead them to conduct subtly flawed research, and then hide those flaws when they write it up. So you need to consider not only the (probably unlikely) possibility of outright fraud, but also the (probably quite common) possibility that the research is unintentionally "slanted". I opened a few standard textbooks and didn't find much of a discussion of this problem, so here's my own attempt to list a few ways in which these issues can arise are:

- **Data fabrication**. Sometimes, people just make up the data. This is occasionally done with "good" intentions. For instance, the researcher believes that the fabricated data do reflect the truth, and may actually reflect "slightly cleaned up" versions of actual data. On other occasions, the fraud is deliberate and malicious. Some high-profile examples where data fabrication has been alleged or shown include Cyril Burt (a psychologist who is thought to have fabricated some of his data), Andrew Wakefield (who has been accused of fabricating his data connecting the MMR vaccine to autism) and Hwang Woo-suk (who falsified a lot of his data on stem cell research).
- **Hoaxes**. Hoaxes share a lot of similarities with data fabrication, but they differ in the intended purpose. A hoax is often a joke, and many of them are intended to be (eventually) discovered. Often, the point of a hoax is to discredit someone or some field. There's quite a few well known scientific hoaxes that have occurred over the years (e.g., Piltdown man) some of were deliberate attempts to discredit particular fields of research (e.g., the Sokal affair).
- **Data misrepresentation**. While fraud gets most of the headlines, it's much more common in my experience to see data being misrepresented. When I say this, I'm not referring to newspapers getting it wrong (which they do, almost always). I'm referring to the fact that often, the data don't actually say what the researchers think they say. My guess is that, almost always, this isn't the result of deliberate dishonesty, it's due to a lack of sophistication in the data analyses. For instance, think back to the example of Simpson's paradox that I discussed in the beginning of these notes. It's very common to see people present "aggregated" data of some kind; and sometimes, when you dig deeper and find the raw data yourself, you find that the aggregated data tell a different story to the disaggregated data. Alternatively, you might find that some aspect of the data is being hidden, because it tells an inconvenient story (e.g., the researcher might choose not to refer to a particular variable). There's a lot of variants on this; many of which are very hard to detect.
- **Study "misdesign"**. Okay, this one is subtle. Basically, the issue here is that a researcher designs a study that has built-in flaws, and those flaws are never reported in the paper. The data that are reported are completely real, and are correctly analysed, but they are produced by a study that is actually quite wrongly put together. The researcher really wants to find a

particular effect, and so the study is set up in such a way as to make it "easy" to (artifactually) observe that effect. One sneaky way to do this – in case you're feeling like dabbling in a bit of fraud yourself – is to design an experiment in which it's obvious to the participants what they're "supposed" to be doing, and then let reactivity work its magic for you. If you want, you can add all the trappings of double blind experimentation etc. It won't make a difference, since the study materials themselves are subtly telling people what you want them to do. When you write up the results, the fraud won't be obvious to the reader: what's obvious to the participant when they're in the experimental context isn't always obvious to the person reading the paper. Of course, the way I've described this makes it sound like it's always fraud: probably there are cases where this is done deliberately, but in my experience the bigger concern has been with unintentional misdesign. The researcher *believes* ...and so the study just happens to end up with a built in flaw, and that flaw then magically erases itself when the study is written up for publication.

- **Data mining & post hoc hypothesising**. Another way in which the authors of a study can more or less lie about what they found is by engaging in what's referred to as "data mining". As we'll discuss later in the class, if you keep trying to analyse your data in lots of different ways, you'll eventually find something that "looks" like a real effect but isn't. This is referred to as "data mining". It used to be quite rare because data analysis used to take weeks, but now that everyone has very powerful statistical software on their computers, it's becoming very common. Data mining per se isn't "wrong", but the more that you do it, the bigger the risk you're taking. The thing that is wrong, and I suspect is very common, is *unacknowledged* data mining. That is, the researcher run every possible analysis known to humanity, finds the one that works, and then pretends that this was the only analysis that they ever conducted. Worse yet, they often "invent" a hypothesis after looking at the data, to cover up the data mining. To be clear: it's not wrong to change your beliefs after looking at the data, and to reanalyse your data using your new "post hoc" hypotheses. What is wrong (and, I suspect, common) is failing to acknowledge that you've done so. If you acknowledge that you did it, then other researchers are able to take your behaviour into account. If you don't, then they can't. And that makes your behaviour deceptive. Bad!
- **Publication bias & self-censoring**. Finally, a pervasive bias is "non-reporting" of negative results. This is almost impossible to prevent. Journals don't publish every article that is submitted to them: they prefer to publish articles that find "something". So, if 20 people run an experiment looking at whether reading *Finnegans Wake* causes insanity in humans, and 19 of them find that it doesn't, which one do you think is going to get published? Obviously, it's the one study that did find that *Finnegans Wake* causes insanity. This is an example of a *publication bias*: since no-one ever published the 19 studies that didn't find an effect, a naive reader would never know that they existed. Worse yet, most researchers "internalize" this bias, and end up *self-censoring* their research. Knowing that negative results aren't going to be accepted for publication, they never even try to report them. As a friend of mine says "for every experiment that you get published, you also have 10 failures". And she's right. The catch is, while some (maybe most) of those studies are failures for boring reasons (e.g. you stuffed something up) others might be genuine "null" results that you ought to acknowledge when you write up the "good" experiment. And telling which is which is often hard to do. A good place to start is a paper by Ioannidis (2005) with the depressing title "Why most published research findings are false". I'd also suggest taking a look at work by Kühberger, Fritz, and Scherndl (2014) presenting statistical evidence that this actually happens in psychology.

There's probably a lot more issues like this to think about, but that'll do to start with. What I really want to point out is the blindingly obvious truth that real world science is conducted by actual humans, and only the most gullible of people automatically assumes that everyone else is honest and impartial. Actual scientists aren't usually *that* naive, but for some reason the world likes to pretend that we are, and the textbooks we usually write seem to reinforce that stereotype.

This page titled 9.12: Confounds, Artifacts and other Threats to Validity is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

9.13: Videos

Terms of Statistics

(https://www.youtube.com/watch?v=HgtnpOtk350)

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

This page titled 9.13: Videos is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

- 1.3: Basic Concepts and Categories of US Positive Law is licensed CC BY-NC-SA 4.0.

9.14: References

Adair, G. 1984. “The Hawthorne Effect: A Reconsideration of the Methodological Artifact.” *Journal of Applied Psychology* 69: 334–45.

Bickel, P. J., E. A. Hammel, and J. W. O’Connell. 1975. “Sex Bias in Graduate Admissions: Data from Berkeley.” *Science* 187: 398–404.

Campbell, D. T., and J. C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Boston, MA: Houghton Mifflin.

Evans, J. St. B. T., J. L. Barston, and P. Pollard. 1983. “On the Conflict Between Logic and Belief in Syllogistic Reasoning.” *Memory and Cognition* 11: 295–306.

Hothersall, D. 2004. *History of Psychology*. McGraw-Hill.

Ioannidis, John P. A. 2005. “Why Most Published Research Findings Are False.” *PLoS Med* 2 (8): 697–701.

Kahneman, D., and A. Tversky. 1973. “On the Psychology of Prediction.” *Psychological Review* 80: 237–51.

Kühberger, A, A Fritz, and T. Scherndl. 2014. “Publication Bias in Psychology: A Diagnosis Based on the Correlation Between Effect Size and Sample Size.” *Public Library of Science One* 9: 1–8.

Pfungst, O. 1911. *Clever Hans (the Horse of Mr. Von Osten): A Contribution to Experimental Animal and Human Psychology*. Translated by C. L. Rahn. New York: Henry Holt.

Rosenthal, R. 1966. *Experimenter Effects in Behavioral Research*. New York: Appleton.

Stevens, S. S. 1946. “On the Theory of Scales of Measurement.” *Science* 103: 677–80.

This page titled 9.14: References is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

- 3.8: Relations with Lawyers is licensed CC BY-NC-SA 4.0.

10: DATA VALIDATION

In computer science, **data validation** is the process of ensuring data has undergone data cleansing to ensure they have data quality, that is, that they are both correct and useful. It uses routines, often called "validation rules", "validation constraints", or "check routines", that check for correctness, meaningfulness, and security of data that are input to the system. The rules may be implemented through the automated facilities of a data dictionary, or by the inclusion of explicit application program validation logic of the computer and its application.

This is distinct from formal verification, which attempts to prove or disprove the correctness of algorithms for implementing a specification or property.

Overview

Data validation is intended to provide certain well-defined guarantees for fitness and consistency of data in an application or automated system. Data validation rules can be defined and designed using various methodologies, and be deployed in various contexts.[1] Their implementation can use declarative data integrity rules, or procedure-based business rules.[2]

Note that the guarantees of data validation do not necessarily include accuracy, and it is possible for data entry errors such as misspellings to be accepted as valid. Other clerical and/or computer controls may be applied to reduce inaccuracy within a system.

Different kinds

In evaluating the basics of data validation, generalizations can be made regarding the different kinds of validation according to their scope, complexity, and purpose.

For example:

- Data type validation;
- Range and constraint validation;
- Code and cross-reference validation;
- Structured validation; and
- Consistency validation

Data-type check

Data type validation is customarily carried out on one or more simple data fields.

The simplest kind of data type validation verifies that the individual characters provided through user input are consistent with the expected characters of one or more known primitive data types as defined in a programming language or data storage and retrieval mechanism.

For example, an integer field may require input to use only characters 0 through 9.

Simple range and constraint check

Simple range and constraint validation may examine input for consistency with a minimum/maximum range, or consistency with a test for evaluating a sequence of characters, such as one or more tests against regular expressions. For example, a counter value may be required to be a non-negative integer, and a password may be required to meet a minimum length and contain characters from multiple categories.

Code and cross-reference check

Code and cross-reference validation includes operations to verify that data is consistent with one or more possibly-external rules, requirements, or collections relevant to a particular organization, context or set of underlying assumptions. These additional validity constraints may involve cross-referencing supplied data with a known look-up table or directory information service such as LDAP.

For example, a user-provided country code might be required to identify a current geopolitical region.

Structured check

Structured validation allows for the combination of other kinds of validation, along with more complex processing. Such complex processing may include the testing of conditional constraints for an entire complex data object or set of process operations within a system.

Consistency check

Consistency validation ensures that data is logical. For example, the delivery date of an order can be prohibited from preceding its shipment date.

Example

Multiple kinds of data validation are relevant to 10-digit pre-2007 ISBNs (the 2005 edition of ISO 2108 required ISBNs to have 13 digits from 2007 onwards[3]).

- Size. A pre-2007 ISBN must consist of 10 digits, with optional hyphens or spaces separating its four parts.
- Format checks. Each of the first 9 digits must be 0 through 9, and the 10th must be either 0 through 9 or an *X*.
- Check digit. To detect transcription errors in which digits have been altered or transposed, the last digit of a pre-2007 ISBN must match the result of a mathematical formula incorporating the other 9 digits (ISBN-10 check digits).

Validation Types

Allowed character checks
: Checks to ascertain that only expected characters are present in a field. For example a numeric field may only allow the digits 0–9, the decimal point and perhaps a minus sign or commas. A text field such as a personal name might disallow characters used for markup. An e-mail address might require at least one @ sign and various other structural details. Regular expressions can be effective ways to implement such checks.

Batch totals
: Checks for missing records. Numerical fields may be added together for all records in a batch. The batch total is entered and the computer checks that the total is correct, e.g., add the 'Total Cost' field of a number of transactions together.

Cardinality check
: Checks that record has a valid number of related records. For example, if a contact record is classified as "customer" then it must have at least one associated order (cardinality > 0). This type of rule can be complicated by additional conditions. For example, if a contact record in a payroll database is classified as "former employee" then it must not have any associated salary payments after the separation date (cardinality = 0).

Check digits
: Used for numerical data. To support error detection, an extra digit is added to a number which is calculated from the other digits.

Consistency checks
: Checks fields to ensure data in these fields correspond, e.g., if expiration date is in the past then status is not "active".

Cross-system consistency checks
: Compares data in different systems to ensure it is consistent. Systems may represent the same data differently, in which case comparison requires transformation (e.g., one system may store customer name in a single Name field as 'Doe, John Q', while another uses First_Name 'John' and Last_Name 'Doe' and Middle_Name 'Quality').

Data type checks
: Checks input conformance with typed data. For example, an input box accepting numeric data may reject the letter 'O'.

File existence check
: Checks that a file with a specified name exists. This check is essential for programs that use file handling.

Format check
: Checks that the data is in a specified format (template), e.g., dates have to be in the format YYYY-MM-DD. Regular expressions may be used for this kind of validation.

Presence check
: Checks that data is present, e.g., customers may be required to have an email address.

Range check
: Checks that the data is within a specified range of values, e.g., a probability must be between 0 and 1.

Referential integrity
: Values in two relational database tables can be linked through foreign key and primary key. If values in the foreign key field are not constrained by internal mechanisms, then they should be validated to ensure that the referencing table always refers to a row in the referenced table.

Spelling and grammar check
: Looks for spelling and grammatical errors.

Uniqueness check
: Checks that each value is unique. This can be applied to several fields (i.e. Address, First Name, Last Name).

Table look up check
: A table look up check compares data to a collection of allowed values.

Source - Wikipedia

10.1: Importance of Statistics and Experimental Measurements

This page titled 10: Data Validation is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

10.1: Importance of Statistics and Experimental Measurements

Here you can find pre-recorded lecture videos that cover this topic here: https://youtube.com/playlist?list=PL...EANc_k_3z5Y5n8

1.1 Performing Error and Data Analysis in the Lab:

Here are some common comments that I will typically overhear when students are in lab and talking about data analysis

1. **How good is your data?**
2. **How well does your experimental data fit the theory?**
3. **What is the R^2 value?**
4. **What is the percent error?**

While well intentioned, these are **not the questions an experimentalist should ask**. And before any data set can be utilized for an application the **quality** of the data must first be established.

For example in the first question is the student assessing the **accuracy? Precision? How well it fits with theory? We must be specific and careful with our language.** In terms of the second question, data is **not necessarily good** just because it agrees with theory. The quality of the data must be assessed before any conclusions can be drawn. **What is the reason the theory and experiments disagree? Is that theory applicable in this situation?** We must ask these questions!

You all probably have experience with R^2 as a **measure of the goodness of a fit** but in this course we will do a **much more thorough error analysis**, we will go much deeper than R^2 so we will not stop there.

Finally, **never ever ever ever ever include percent error in any lab anal-ysis** that we perform in this class. It is a meaningless error analysis. We never know the actual value or true value of a measurement so will not perform this error analysis.

When designing an experiment it is crucial that do the following

• Clearly defined goals and ideally graphs/figures for study

• Experimental plan, apparatus, and DAQ for collecting data

• Perform appropriate statistical analysis to effectively extract critical information from raw data

• Interpret results and draw conclusions that are statistically significant

Additionally, when we speak of quality what we mean is to find the **actual or true value** of the physical quantity being measured. And often this is different than the **measured value** of the physical quantity. The difference between the **measured value** and the **actual or true value** of a measurement is the **error**.

Here we encounter a problem as we cannot calculate the error exactly unless we know the true value of the quantity being measured. Instead we **estimate the bounds or the likelihood that the error exceeds a specific value**. For example say that 95% (19 out of 20) of the yield strength measurements will have an error less than 1 MPa. **Thus the reading has an uncertainty of 1 MPa at confidence level of 95%**. We will calculate confidence levels or confidence intervals when quantifying error.

As experimentalists we often work with large data sets with a large number of measurements. The **error or uncertainty can be estimated with statistical tools when a large number of measurements are taken**. Later on in the class we will learn how to use some software tools like **Matlab, R, Python, and/or Mathematica** to perform error analysis. Much more on this later, but let's get to some basics of experimentation and measurements.

1.2 Measurements:

A **measurement** is a **quantitative** comparison between a predefined **standard** and a **measurand**.

What are some examples of different types of measurements we can make?

• Temperature

• Length

• Time

• Stress

• Strain

• Viscosity

• Force

• Torque

• ....

Measurand: the particular physical parameter begin observed and quantified, i.e., the **input quantity** of the **measuring process**

Standard: has a defined relationship to a unit of measurement for a given measurand. Typically a standard will be recognized by the NIST (National Institute of Standards and Technology)

What are some examples of standards?

• Tape Measure

• Balance

• Speed of Light

• Stopwatch

• Extensometer

• ...

2.1.1 What is the significance of measurements, why do we experimentally measure parameters?

• Obtain quantitative information on the actual state of physical variables and processes (i.e. T, P, V, etc.), this is critical especially for thermodynamics and finding thermodynamic equilibrium

• Vital for control processes which require measured discrepancy between actual and desired output (i.e. temperature control on sheet annealing line)

• Daily operations (i.e. temperature controls, heating and cooling, tire gauge, oil gauge, etc.)

• Establishing costs, Ti $140/lb precision is important when dealing with tons of Ti.

• Temperature in bioreactor can influence cell growth

• Thermal conductivity, acceleration, material properties

2.2 Uncertainty in Measurements: Measurements Are Useless without Uncertainty Bounds or Error Bars

Measurements are only useful to us as engineers if we can quantify the **uncertainty** of a measurement. **No measurement is meaningful** or useful if they do not have an **associated error bar**. In this class and moving forward in your career as an engineer be sure to place error bars on all your reported values. Error bars are critical because no measurement is perfect. If you measure the length of a tensile specimen with a micrometer your measurement is likely to be different than that of your colleague. Thus, there will be some **uncertainty** in any measurement that one makes. In this section of this course we will develop a comprehensive framework for evaluating uncertainty and placing **tolerances and confidence intervals** on measured values. Much more on this later....

2.3 Using Correct Terminology When Assessing Measurement Quality/ Apparatus Performance

• **Accuracy:** Difference between measured and true values. Manufacturer will specify maximum error but not confidence intervals typically.

• **Precision:** Difference between reported values during repeated measurements of same quantity.

• **Resolution:** Smallest increment of change in the measured value that can be determined by the readout scale. Often the same as precision.

• **Sensitivity:** The change of an instrument's output per unit change in the measured quantity. Higher sensitivity often indicates finer resolution, better precision, and higher accuracy.

2.4 Error and Uncertainty:

Error is the difference between the *measured* and *true* value of the quantity being measured. Unfortunately, the true value can never really be known so instead we quantify the **uncertainty** of a measurement at specific levels of confidence. Error can be categorized into two basic categories either **bias** or **precision** error/uncertainty. We will cover uncertainty and error analysis much more in-depth in the next several lectures...but never use percent error in this course.

Error

We can begin with our discussion of error by defining the following equation

$$Error = x_m - x_{true} \tag{10.1.1}$$

where x_m is the measured value and x_{true} is the true value.

When designing an experimental system or executing an experiment the experimentalist's primary objective should be to minimize the error. After the experiment is completed we need to estimate a bound on the error with a degree of certainty or a confidence interval (i.e. 90%, 95%, 99%), hopefully you have encountered confidence intervals before but if not no worries we will discuss this in depth.

There are many different sources of inaccuracy or error however most errors can be classified as:

(1) **Precision**

(2) **Bias**

(3) **Illegitimate**

(4) **Errors that are sometimes bias or sometimes precision errors**

In this course we will spend a considerable amount of time discussing **1. Precision** and **2. Bias** error in great detail. But we will discuss all these sources of error and how to account for them when designing experiments or presenting your results. Let's start

with **Precision Error.**

2.4.1 Precision or Random Errors

I really do not like the term random error but you will occasionally see that term in literature but in this course we will use the term precision primarily.

Precision errors are **different for each successive measurement**. Thus if you create a histogram, which shows the probability of encountering or measuring a specific measurand, the width of your distribution (hopefully it is Gaussian) will be due to precision error if that is the only source of error in your experiment. From this distribution, **statistical analysis can be utilized to estimate the size of the precision error.** Typically bias and precision errors occur simultaneously and contributes to the **total error**, more on this fun calculation later, hope you remember Taylor expansions....

Why Do We Have Precision Error?

(1) Disturbances to the equipment

(2) Fluctuating experimental conditions

(3) Insufficient measuring-system sensitivity

(4) Stochastic (Random) nature of measurements even in static conditions

Examples of Precision Error:

- Measuring temperature with thermocouple as temperature fluctuates
- Measure the length of a beam with a micrometer you will get slightly different measurements than classmates
- Changes in temperature conditions can affect swimming of bacteria

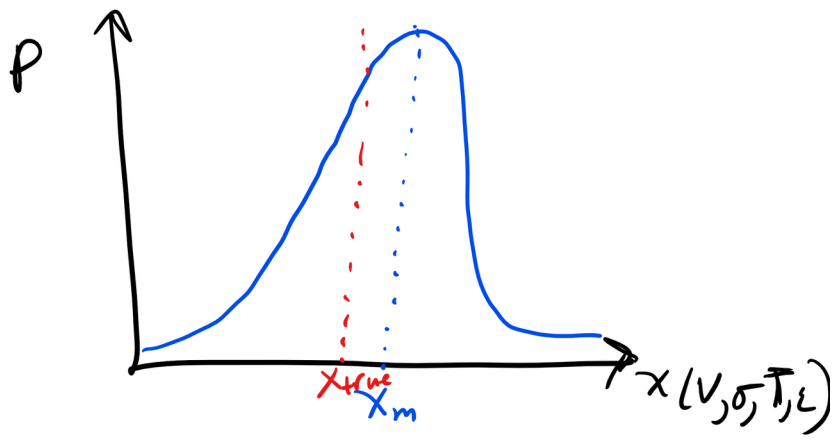


Figure 10.1.1: Precision Error Distribution.

2.4.2 Bias or Systematic Errors

I like the term bias and systematic much better than random. However, in this course we will primarily use the term Bias error.

Bias errors will occur the same way each time a measurement is made. For example a balance that consistently reads 0.5 mg heavier. In some literature you might see this termed as a fixed offset error. The key thing to note about bias errors is they are

fixed and thus **they do not show a distribution**. This is in stark contrast to precision errors which show a distribution, typically Gaussian. Therefore if there is no distribution we cannot apply any **statistical analysis**.

There are multiple types of bias errors:

(1) Calibration errors

- This is the most common error
- Typically calibration errors are **zero-offset errors** which cause all read- ing to be offset by constant amount or *scale errors*
- I.e. scale constantly light by 1lb or clock runs fast by 1 minute
- Can be reduced via calibration procedures and proving the system via comparison with *known* standard

(2) Consistently recurring human errors

- Ex. experimenter who consistently jumps the gun when recording synchronized readings either time, voltage, etc.

(3) Errors caused by defective equipment

(4) Loading errors

- i.e. not putting a DSC pan at the center of the DSC will skew temperature measurements

(5) System resolution limits

- i.e. Bias error for a ruler with cm tick marks will be cm, compared to a micrometer with 1μm resolution, that is 4 order of magnitude difference.

3. Illegitimate Error

Illegitimate error can and most likely will occur to you at some point as you perform experiments as an engineer. Illegitimate errors include

- Blunders and mistakes
- Computational or calculation errors
- Analysis errors

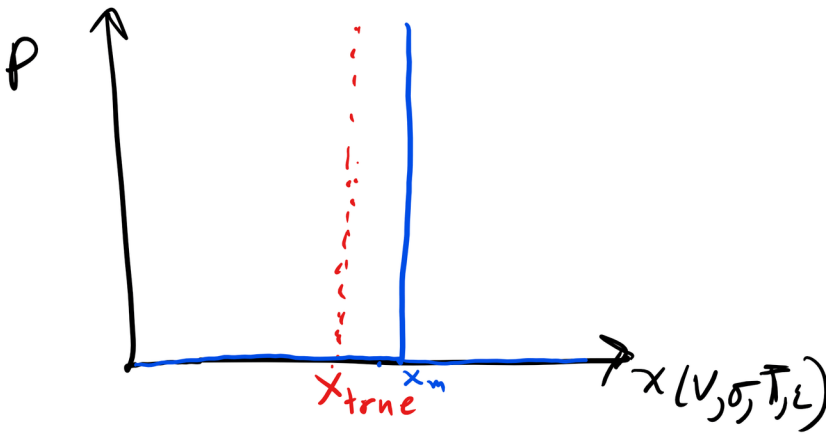


Figure 10.1.2: Bias Error Distribution

While these things can occur in lab **you must always perform the experiment again**. You cannot simply present your results in either a presentation or a form of technical writing and attribute error to illegitimate error. You must perform the experiment again.

4. Errors That Are Sometimes Bias and Sometimes Precision Error:

These are the most difficult sources of error that we must analyze as it is difficult to determine what type of error analysis must be done and which error has a distribution and what does not.

- **Instrument Hysteresis**

– Hysteresis error can be seen in stress strain measurements and is path dependent

- **Calibration drift and variation in test or environmental conditions**

– Drift occurs when the response varies over time often due to sensitivity to temperature or humidity

- **Variations in procedure or definition among experimenters**

Quantifying Total Uncertainty, Comprehensive Error Analysis

Again typically we are concerned with **bias** and **precision** error when evaluating experimental data. And this data comes from two classes of experiments: i) *single- sample* and ii) *repeated-sample* experiments.

A **sample** refers to an individual measurement of a specific quantity. Ex. Measuring strain in a material several times under identical loading is a repeated-sample experiment. A single measurement of strain is a single-sample experiment and we cannot extract the distribution of precision error in this case. In this course and in your career as an engineer you will most likely be taking multiple measurements,

i.e. repeated sample experiments. Single sample experiments are more uncommon and typically occur when the measurement you are making is extremely difficult and multiple experiments are non realistic to conduct, typically due to economic or practicality reasons.

The total uncertainty U_x in a measurement of *x* is

$$U_x = \sqrt{B_x^2 + P_x^2} \tag{10.1.2}$$

where B_x if the bias uncertainty in a measurement of *x* and P_x is the precision uncertainty. Note that *x* is just a variable and we can talk about uncertainty in measuring volume (V), pressure (P), stress (σ), etc.

This is a largely empirically derived formula however it does depend on the **assump-tion that the bias and precision uncertainty are independent sources of error.** Also the bias and precision uncertainty should have the **same confidence level** when calculating the **total uncertainty.** We will come back to this a bit later in excruciating detail for this calculation so please be patient.

Minimizing Error in Designing Experiments

Best time to **minimize experimental error** is in the **design stage**. You should **perform a single-sample uncertainty analysis** of the proposed experimental arrangement **prior to building the apparatus**. This allows one to make a decision on whether the expected uncertainty is acceptable and to identify the sources of uncertainty. So when designing an experiment:

• **Avoid approaches** that require two large numbers to be measured to determine the small difference between them

• **Design experiments or sensors that amplify the signal strength** to improve sensitivity

• Build *null designs* in which the output is measured as a change from zero rather than as a change in nonzero value

• Avoid experiments with large *correction factors* must be applied as part of the data-reduction procedure

• Attempt to **minimize the influence of the measuring system on the measured variable**

• **Calibrate** entire system to minimize calibration related bias errors.

Once we have established the error and uncertainty with a measurement we also have to **report these results** and **explain our analysis**. **Technical writing** is a very distinct class of writing with many different formats and audiences. It can be difficult to strike the right **tone** and **level of technical information** that must be provided at times. In this class we will focus on **Laboratory Reports** but as you can see below there are many different form of technical writing.

Forms of Technical Writing:

• **Executive summary**

– Audience is a supervisor. Summary should be concise, usually only a few paragraphs, and highlight key aspects of work in first paragraph.

– Consequences can be drastic as in the case of the Challenger Shuttle tragedy.

• **Laboratory note or technical memo**

– Limited audience, immediate supervisor

– Sufficient information should be included so the experimentalist can re- construct the experiment

• **Progress report**

– Audience is supervisor of sponsoring agency

– Can vary in terms of length, structure, timelines, etc.

• **Full technical report**

– Wider audience typically not people closely associated with the work

– Relates all facts pertinent to project

– Contains enough detail so another professional can repeat experiments

• **Full technical report**

– Wider audience typically not people closely associated with the work

– Relates all facts pertinent to project

– Contains enough detail so another professional can repeat experiments

Line Fitting: Identify Outliers and Physical Interpretation

When you are fitting a curve or a line to a data set **one must be careful to select a fit that is appropriate for the experiment being conducted**. Often researchers are too consumed with obtaining a **high** R^2 **value**. While the goodness of a fit is important you must make sure that the equation that you are **fitting matches with the physical experiment that you are conducting**. For example if you are measuring how an electrochemical cell voltage changes as a function of temperature you must use the Nernst Equation. Looking at the data perhaps a 12th degree polynomial fit will give you a better R^2 value but that equation has no physical meaning in terms of your analysis. So always keep that in mind.

When identifying outliers and a particular point is in doubt, temporarily exclude the data point and fit a new line through the remaining data. **If the fit is better the point can possibly be dropped however it could indicate that the relation- ship is not linear it may indicate another scaling regime or a transition in behavior.**

We can also apply this same error analysis procedure and to optimize your experimental apparatus and minimize error/total uncertainty. One can identify what component or part of the measuring process is generating the most error and then one can re-design the apparatus to minimize that error.

This page titled 10.1: Importance of Statistics and Experimental Measurements is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

11: DESCRIBING DATA

Chapter by Matthew Crump

> Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.
>
> — John W. Tukey

This chapter is about **descriptive statistics**. These are tools for describing data. Some things to keep in mind as we go along are:

1. There are lots of different ways to describe data
2. There is more than one "correct" way, and you get to choose the most "useful" way for the data that you are describing
3. It is possible to invent new ways of describing data, all of the ways we discuss were previously invented by other people, and they are commonly used because they are useful.
4. Describing data is necessary because there is usually too much of it, so it doesn't make any sense by itself.

11.1: This is what too many numbers looks like

11.2: Look at the data

11.3: Important Ideas - Distribution, Central Tendency, and Variance

11.4: Measures of Central Tendency (Sameness)

11.5: Measures of Variation (Differentness)

11.6: Using Descriptive Statistics with data

11.7: Rolling your own descriptive statistics

11.8: Remember to look at your data

11.9: Videos

11.10: References

This page titled 11: Describing Data is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

11.1: This is what too many numbers looks like

Let's say you wanted to know how happy people are. So, you ask thousands of people on the street how happy they are. You let them pick any number they want from negative infinity to positive infinity. Then you record all the numbers. Now what?

Well, how about you look at the numbers and see if that helps you determine anything about how happy people are. What could the numbers look like. Perhaps something like this:

73	594	-22	-20	-547	162	-90	312	235	-511
-337	85	552	377	241	-382	241	-439	264	-292
-136	-262	432	835	73	-180	-93	218	597	419
-500	-120	588	-96	-412	502	1058	761	549	-320
14	-869	338	935	531	339	83	37	820	544
50	-397	203	-374	-186	518	530	1320	816	1293
580	-741	-102	-56	933	-228	-347	656	162	714
440	569	-431	557	-502	-331	-281	73	311	459
-143	-348	136	-624	55	-790	374	-988	-1102	-408
-666	671	660	452	1299	717	369	158	679	411
-593	-364	115	379	56	-440	505	-370	-102	-1020
610	-86	-181	-143	75	-188	502	606	443	74
181	-355	40	551	-362	414	-307	415	-930	-302
1416	-387	437	-126	-407	28	466	-25	-413	-286
106	257	459	703	3	1592	1042	-124	102	-578
550	-605	-41	167	-581	830	-17	200	98	472
242	-30	94	-619	-885	424	320	241	193	121
-373	-478	-398	1035	425	-199	-350	189	-394	346
-161	-355	108	-685	-668	-667	893	-623	19	879
-430	119	830	-236	-527	61	313	265	453	-565
-523	9	-413	-705	-527	237	-341	80	349	891
181	555	371	-623	-107	859	-673	855	4	117
-1225	317	279	266	24	-387	368	567	-717	717
-110	706	-40	-836	-882	48	307	1150	-917	-236
-669	-401	-274	-465	-178	104	517	635	86	186
-357	356	932	118	-51	62	-111	-154	-409	852
-91	-568	640	-48	-349	-481	511	-544	254	-641
654	-127	-563	-340	30	-293	-100	292	220	41
312	640	-628	335	-808	105	77	-674	108	-1177
-804	-318	608	954	-350	606	-394	-68	-226	161
-580	174	622	-433	-758	-49	949	496	802	-271

745	184	-41	281	-318	-323	634	-53	-307	446
245	368	163	-489	-124	-258	-463	357	-465	-321
628	1055	-11	-177	-28	139	-531	134	-400	-182
-298	153	-206	946	534	295	543	350	184	-311
1109	-174	1169	-175	88	804	-555	-269	-376	1199
-463	1078	-384	-804	2	-29	219	-467	375	503
1717	264	-177	-222	1125	-738	569	-335	581	364
-36	-523	847	-1189	-379	-704	-654	51	-136	303
609	-200	675	286	353	67	-993	-181	1198	-508
77	58	-53	-510	-343	657	1303	-300	804	-376
421	73	-165	-238	409	470	648	127	347	-296
659	280	1397	-715	979	-793	565	-102	510	333
-848	571	-297	630	286	-512	275	468	-314	-246
-212	603	-152	-474	428	-315	-38	-53	-324	-225
-46	-89	316	341	516	-655	613	249	334	94
-66	-688	101	-128	-422	424	326	-287	417	-605
357	-959	-149	387	-39	-104	-596	55	-25	-26
-533	-667	280	863	215	-182	397	333	-56	36
-118	-329	44	-1	354	-545	630	460	458	30

Now, what are you going to with that big pile of numbers? Look at it all day long? When you deal with data, it will deal so many numbers to you that you will be overwhelmed by them. That is why we need ways to describe the data in a more manageable fashion.

The complete description of the data is always the data itself. **Descriptive statistics** and other tools for describing data go one step further to summarize aspects of the data. Summaries are a way to compress the important bits of a thing down to a useful and manageable tidbit. It's like telling your friends why they should watch a movie: you don't replay the entire movie for them, instead you hit the highlights. Summarizing the data is just like a movie preview, only for data.

This page titled 11.1: This is what too many numbers looks like is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

11.2: Look at the data

We already tried one way of looking at the numbers, and it wasn't useful. Let's look at some other ways of looking at the numbers, using graphs.

Stop, plotting time (o o oh) U can plot this

Let's turn all of the numbers into dots, then show them in a graph. Note, when we do this, we have not yet summarized anything about the data. Instead, we just look at all of the data in a visual format, rather than looking at the numbers.

```
happiness<-rnorm(500,100,500)
plot(happiness)
```

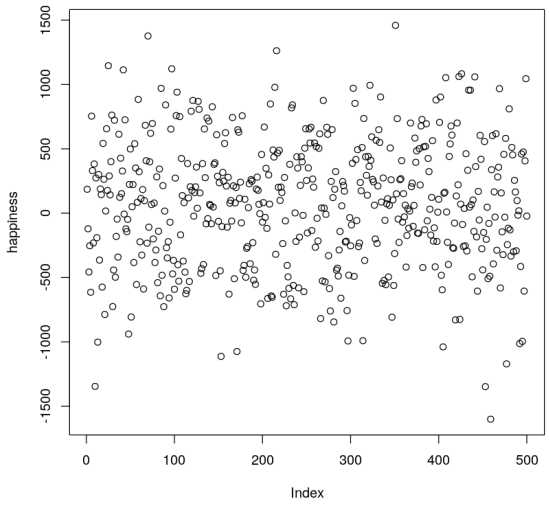


Figure 11.2.1: Pretend happiness ratings from 500 people.

Figure 11.2.1 shows 500 measurements of happiness. The graph has two axes. The horizontal **x-axis**, going from left to right is labeled "Index". The vertical **y-axis**, going up and down, is labelled "happiness". Each dot represents one measurement of every person's happiness from our pretend study. Before we talk about what we can and cannot see about the data, it is worth mentioning that the way you plot the data will make some things easier to see and some things harder to see. So, what can we now see about the data?

There are lots of dots everywhere. It looks like there are 500 of them because the index goes to 500. It looks like some dots go as high as 1000-1500 and as low as -1500. It looks like there are more dots in the middle-ish area of the plot, sort of spread about 0.

> Take home: we can see all the numbers at once by putting them in a plot, and that is much easier and more helpful than looking at the raw numbers.

OK, so if these dots represent how happy 500 people are, what can we say about those people? First, the dots are kind of all over the place, so different people have different levels of happiness. Are there any trends? Are more people happy than unhappy, or vice-versa? It's hard to see that in the graph, so let's make a different one, called a **histogram**

Histograms

Making a histogram will be our first act of officially summarizing something about the data. We will no longer look at the individual bits of data, instead we will see how the numbers group together. Let's look at a histogram of the happiness data, and then explain

it.

```
happiness<-rnorm(500,100,500)
hist(happiness)
```

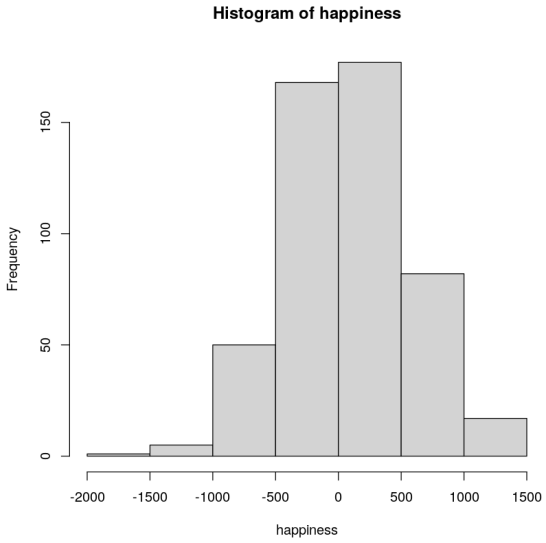


Figure 11.2.2: A histogram of the happiness ratings.

The dots have disappeared, and now we some bars. Each bar is a summary of the dots, representing the number of dots (frequency count) inside a particular range of happiness, also called **bins**. For example, how many people gave a happiness rating between 0 and 500? The fifth bar, the one between 0 and 500 on the x-axis, tells you how many. Look how tall that bar is. How tall is it? The height is shown on the y-axis, which provides a frequency count (the number of dots or data points). It looks like around 150 people said their happiness was between 0-500.

More generally, we see there are many bins on the x-axis. We have divided the data into bins of 500. Bin #1 goes from -2000 to -1500, bin #2 goes from -1500 to -1000, and so on until the last bin. To make the histogram, we just count up the number of data points falling inside each bin, then plot those frequency counts as a function of the bins. Voila, a histogram.

What does the histogram help us see about the data? First, we can see the **shape** of data. The shape of the histogram refers to how it goes up and down. The shape tells us where the data is. For example, when the bars are low we know there isn't much data there. When the bars are high, we know there is more data there. So, where is most of the data? It looks like it's mostly in the middle two bins, between -500 and 500. We can also see the **range** of the data. This tells us the minimums and the maximums of the data. Most of the data is between -1500 and +1500, so no infinite sadness or infinite happiness in our data-set.

When you make a histogram you get to choose how wide each bar will be. For example, below are four different histograms of the very same happiness data. What changes is the width of the bins.

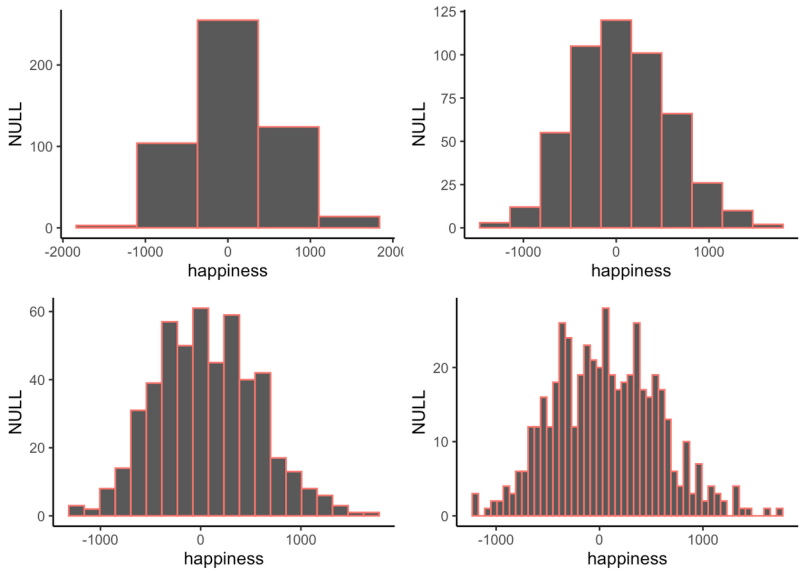


Figure 11.2.3: Four histograms of the same data using different bin widths.

All of the histograms have roughly the same overall shape: From left to right, the bars start off small, then go up, then get small again. In other words, as the numbers get closer to zero, they start to occur more frequently. We see this general trend across all the histograms. But, some aspects of the trend fall apart when the bars get really narrow. For example, although the bars generally get taller when moving from -1000 to 0, there are some exceptions and the bars seem to fluctuate a little bit. When the bars are wider, there are less exceptions to the general trend. How wide or narrow should your histogram be? It's a Goldilocks question. Make it just right for your data.

This page titled 11.2: Look at the data is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

11.3: Important Ideas - Distribution, Central Tendency, and Variance

Let's introduce three important terms we will use a lot, **distribution**, **central tendency**, and **variance**. These terms are similar to their everyday meanings (although I suspect most people don't say central tendency very often).

Distribution. When you order something from Amazon, where does it come from, and how does it get to your place? That stuff comes from one of Amazon's distribution centers. They distribute all sorts of things by spreading them around to your doorstep. "To Distribute"" is to spread something. Notice, the data in the histogram is distributed, or spread across the bins. We can also talk about a distribution as a noun. The histogram is a distribution of the frequency counts across the bins. Distributions are **very, very, very, very, very** important. They can have many different shapes. They can describe data, like in the histogram above. And as we will learn in later chapters, they can **produce** data. Many times we will be asking questions about where our data came from, and this usually means asking what kind of distribution could have created our data (more on that later.)

Central Tendency is all about sameness: What is common about some numbers? For example, is there anything similar about all of the numbers in the histogram? Yes, we can say that most of them are near 0. There is a tendency for most of the numbers to be centered near 0. Notice we are being cautious about our generalization about the numbers. We are not saying they are all 0. We are saying there is a tendency for many of them to be near zero. There are lots of ways to talk about the central tendency of some numbers. There can even be more than one kind of tendency. For example, if lots of the numbers were around -1000, and a similar large amount of numbers were grouped around 1000, we could say there was two tendencies.

Variance is all about differentness: What is different about some numbers?. For example, is there anything different about all of the numbers in the histogram? YES!!! The numbers are not all the same! When the numbers are not all the same, they must vary. So, the variance in the numbers refers to how the numbers are different. There are many ways to summarize the amount of variance in the numbers, and we discuss these very soon.

This page titled 11.3: Important Ideas - Distribution, Central Tendency, and Variance is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

11.4: Measures of Central Tendency (Sameness)

We've seen that we can get a sense of data by plotting dots in a graph, and by making a histogram. These tools show us what the numbers look like, approximately how big and small they are, and how similar and different they are from another. It is good to get a feeling about the numbers in this way. But, these visual sensitudes are not very precise. In addition to summarizing numbers with graphs, we can summarize numbers using numbers (NO, please not more numbers, we promise numbers can be your friend).

From many numbers to one

Measures of central have one important summary goal: to reduce a pile of numbers to a single number that we can look at. We already know that looking at thousands of numbers is hopeless. Wouldn't it be nice if we could just look at one number instead? We think so. It turns out there are lots of ways to do this. Then, if your friend ever asks the frightening question, "hey, what are all these numbers like?". You can say they are like this one number right here.

But, just like in Indiana Jones and the Last Crusade (highly recommended movie), you must choose your measure of central tendency wisely.

Mode

The **mode** is the most frequently occurring number in your measurement. That is it. How do you find it? You have to count the number of times each number appears in your measure, then whichever one occurs the most, is the mode.

> Example: 1 1 1 2 3 4 5 6

The mode of the above set is 1, which occurs three times. Every other number only occurs once.

OK fine. What happens here:

> Example: 1 1 1 2 2 2 3 4 5 6

Hmm, now 1 and 2 both occur three times each. What do we do? We say there are two modes, and they are 1 and 2.

Why is the mode a measure of central tendency? Well, when we ask, "what are my numbers like", we can say, "most of the number are, like a 1 (or whatever the mode is)".

Is the mode a good measure of central tendency? That depends on your numbers. For example, consider these numbers

> 1 1 2 3 4 5 6 7 8 9

Here, the mode is 1 again, because there are two 1s, and all of the other numbers occur once. But, are most of the numbers like, a 1. No, they are mostly not 1s.

"Argh, so should I or should I not use the mode? I thought this class was supposed to tell me what to do?". There is no telling you what to do. Every time you use a tool in statistics you have to think about what you are doing and justify why what you are doing makes sense. Sorry.

Median

The **median** is the exact middle of the data. After all, we are asking about central tendency, so why not go to the center of the data and see where we are. What do you mean middle of the data? Let's look at these numbers:

> 1 5 4 3 6 7 9

Umm, OK. So, three is in the middle? Isn't that kind of arbitrary. Yes. Before we can compute the median, we need to order the

numbers from smallest to largest.

1 3 4 **5** 6 7 9

Now, 5 is in the middle. And, by middle we mean in the middle. There are three numbers to the left of 5, and three numbers to the right. So, five is definitely in the middle.

OK fine, but what happens when there aren't an even number of numbers? Then the middle will be missing right? Let's see:

1 2 3 4 5 6

There is no number between 3 and 4 in the data, the middle is empty. In this case, we compute the median by figuring out the number in between 3 and 4. So, the median would be 3.5.

Is the median a good measure of central tendency? Sure, it is often very useful. One property of the median is that it stays in the middle even when some of the other numbers get really weird. For example, consider these numbers:

1 2 3 4 4 4 **5** 6 6 6 7 7 1000

Most of these numbers are smallish, but the 1000 is a big old weird number, very different from the rest. The median is still 5, because it is in the middle of these ordered numbers. We can also see that five is pretty similar to most of the numbers (except for 1000). So, the median does a pretty good job of representing most of the numbers in the set, and it does so even if one or two of the numbers are very different from the others.

Finally, **outlier** is a term will we use to describe numbers that appear in data that are very different from the rest. 1000 is an outlier, because it lies way out there on the number line compared to the other numbers. What to do with outliers is another topic we discuss sometimes throughout this course.

Mean

Have you noticed this is a textbook about statistics that hasn't used a formula yet? That is about to change, but for those of you with formula anxiety, don't worry, we will do our best to explain them.

The **mean** is also called the average. And, we're guessing you might already now what the average of a bunch of numbers is? It's the sum of the numbers, divided by the number of number right? How do we express that idea in a formula? Just like this:

$$\text{Mean} = \bar{X} = \frac{\sum_{i=1}^{n} x_i}{N} \tag{11.4.1}$$

"That looks like Greek to me". Yup. The $\sum$ symbol is called **sigma**, and it stands for the operation of summing. The little "i" on the bottom, and the little "n" on the top refers to all of the numbers in the set, from the first number "i" to the last number "n". The letters are just arbitrary labels, called **variables** that we use for descriptive purposes. The x_i refers to individual numbers in the set. We sum up all of the numbers, then divide the sum by N, which is the total number of numbers. Sometimes you will see $\bar{X}$ to refer to the mean of all of the numbers.

In plain English, the formula looks like:

$$\text{Mean} = \frac{\text{Sum of my numbers}}{\text{Count of my numbers}}$$

"Well, why didn't you just say that?". We just did in Equation 11.4.1.

Let's compute the mean for these five numbers:

3 7 9 2 6

Add em up:

3+7+9+2+6 = 27

Count em up:

i_1 = 3, i_2 = 7, i_3 = 9, i_4 = 2, i_5 = 6; N=5, because i went from 1 to 5

Divide em:

mean = 27 / 5 = 5.4

Or, to put the numbers in the formula, it looks like this:

$$\text{Mean} = \bar{X} = \frac{\sum_{i=1}^{n} x_i}{N} = \frac{3+7+9+2+6}{5} = \frac{27}{5} = 5.4$$

OK fine, that is how to compute the mean. But, like we imagined, you probably already knew that, and if you didn't that's OK, now you do. What's next?

Is the mean a good measure of central tendency? By now, you should know: it depends.

What does the mean mean?

It is not enough to know the formula for the mean, or to be able to use the formula to compute a mean for a set of numbers. We believe in your ability to add and divide numbers. What you really need to know is what the mean really "means". This requires that you know what the mean does, and not just how to do it. Puzzled? Let's explain.

Can you answer this question: What happens when you divide a sum of numbers by the number of numbers? What are the consequences of doing this? What is the formula doing? What kind of properties does the result give us? FYI, the answer is not that we compute the mean.

OK, so what happens when you divide any number by another number? Of course, the key word here is divide. We literally carve the number up top in the numerator into pieces. How many times do we split the top number? That depends on the bottom number in the denominator. Watch:

$$\frac{12}{3} = 4$$

So, we know the answer is 4. But, what is really going on here is that we are slicing and dicing up 12 aren't we. Yes, and we slicing 12 into three parts. It turns out the size of those three parts is 4. So, now we are thinking of 12 as three different pieces $12 = 4 + 4 + 4$. I know this will be obvious, but what kind of properties do our pieces have? You mean the fours? Yup. Well, obviously they are all fours. Yes. The pieces are all the same size. They are all equal. So, division equalizes the numerator by the denominator...

"Umm, I think I learned this in elementary school, what does this have to do with the mean?". The number on top of the formula for the mean is just another numerator being divided by a denominator isn't it. In this case, the numerator is a sum of all the values in your data. What if it was the sum of all of the 500 happiness ratings? The sum of all of them would just be a single number adding up all the different ratings. If we split the sum up into equal parts representing one part for each person's happiness what would we get? We would get 500 identical and equal numbers for each person. It would be like taking all of the happiness in the world, then dividing it up equally, then to be fair, giving back the same equal amount of happiness to everyone in the world. This would make some people more happy than they were before, and some people less happy right. Of course, that's because it would be equalizing the distribution of happiness for everybody. This process of equalization by dividing something into equal parts is what the **mean** does. See, it's more than just a formula. It's an idea. This is just the beginning of thinking about these kinds of ideas. We will come back to this idea about the mean, and other ideas, in later chapters.

> Pro tip: The mean is the one and only number that can take the place of every number in the data, such that when you add up all the equal parts, you get back the original sum of the data.

All together now

Just to remind ourselves of the mode, median, and mean, take a look at the next histogram. We have overlaid the location of the mean (red), median (green), and mode (blue). For this dataset, the three measures of central tendency all give different answers.

The mean is the largest because it is influenced by large numbers, even if they occur rarely. The mode and median are insensitive to large numbers that occur infrequently, so they have smaller values.

```
library(ggplot2)
my_mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
exp_num<-round(rexp(1000,.15),digits=1)
qplot(exp_num, col=I("grey"), fill=I("white"), bins=30)+
  geom_vline(xintercept=mean(exp_num), color="red", size=1.5)+
  geom_vline(xintercept=median(exp_num), color="green", size=1.5)+
  geom_vline(xintercept=my_mode(exp_num), color="blue", size=1.5)+
  theme_classic()
```

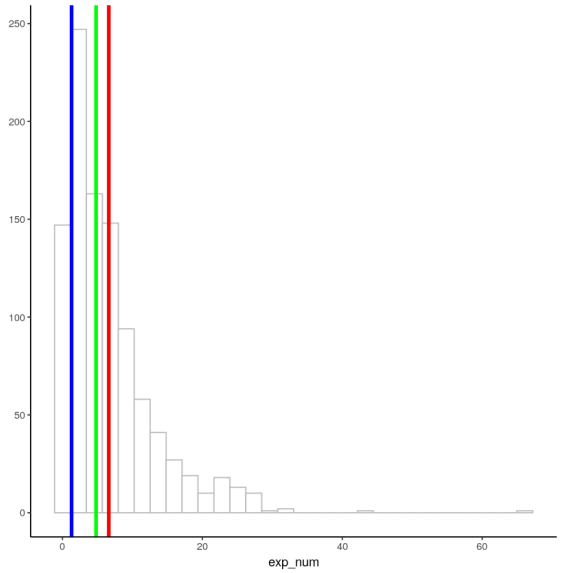


Figure 11.4.1: A histogram with the mean (red), the median (green), and the mode (blue)

This page titled 11.4: Measures of Central Tendency (Sameness) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

11.5: Measures of Variation (Differentness)

What did you do when you wrote essays in high school about a book you read? Probably compare and contrast something right? When you summarize data, you do the same thing. Measures of central tendency give us something like comparing does, they tell us stuff about what is the same. Measures of variation give us something like contrasting does, they tell us stuff about what is different.

First, we note that whenever you see a bunch of numbers that aren't the same, you already know there are some differences. This means the numbers vary, and there is variation in the size of the numbers.

The Range

Consider these 10 numbers, that I already ordered from smallest to largest for you:

1 3 4 5 5 6 7 8 9 24

The numbers have variation, because they are not all the same. We can use the range to describe the width of the variation. The range refers to the **minimum** (smallest value) and **maximum** (largest value) in the set. So, the range would be 1 and 24.

The range is a good way to quickly summarize the boundaries of your data in just two numbers. By computing the range we know that none of the data is larger or smaller than the range. And, it can alert you to outliers. For example, if you are expecting your numbers to be between 1 and 7, but you find the range is 1 - 340,500, then you know you have some big numbers that shouldn't be there, and then you can try to figure out why those numbers occurred (and potentially remove them if something went wrong).

The Difference Scores

It would be nice to summarize the amount of differentness in the data. Here's why. If you thought that raw data (lots of numbers) is too big to look at, then you will be frightened to contemplate how many differences there are to look at. For example, these 10 numbers are easy to look at:

1 3 4 5 5 6 7 8 9 24

But, what about the difference between the numbers, what do those look like? We can compute the difference scores between each number, then put them in a matrix like the one below:

```
numbers<-c(1, 3, 4, 5, 5, 6, 7, 8, 9, 24)
mat<-matrix(rep(numbers,10),ncol=10)
differences<-t(mat)-numbers
row.names(differences)<-numbers
colnames(differences)<-numbers
knitr::kable(differences,row.names=T)
```

```
|   |   1|   3|   4|   5|   5|   6|   7|   8|   9| 24|
|:--|---:|---:|---:|---:|---:|---:|---:|---:|---:|--:|
|1  |   0|   2|   3|   4|   4|   5|   6|   7|   8| 23|
|3  |  -2|   0|   1|   2|   2|   3|   4|   5|   6| 21|
|4  |  -3|  -1|   0|   1|   1|   2|   3|   4|   5| 20|
|5  |  -4|  -2|  -1|   0|   0|   1|   2|   3|   4| 19|
|5  |  -4|  -2|  -1|   0|   0|   1|   2|   3|   4| 19|
|6  |  -5|  -3|  -2|  -1|  -1|   0|   1|   2|   3| 18|
|7  |  -6|  -4|  -3|  -2|  -2|  -1|   0|   1|   2| 17|
|8  |  -7|  -5|  -4|  -3|  -3|  -2|  -1|   0|   1| 16|
|9  |  -8|  -6|  -5|  -4|  -4|  -3|  -2|  -1|   0| 15|
|24 | -23| -21| -20| -19| -19| -18| -17| -16| -15|  0|
```

We are looking at all of the possible differences between each number and every other number. So, in the top left, the difference

between 1 and itself is 0. One column over to the right, the difference between 3 and 1 (3-1) is 2, etc. As you can see, this is a 10x10 matrix, which means there are 100 differences to look at. Not too bad, but if we had 500 numbers, then we would have 500*500 = 250,000 differences to look at (go for it if you like looking at that sort of thing).

Pause for a simple question. What would this matrix look like if all of the 10 numbers in our data were the same number? It should look like a bunch of 0s right? Good. In that case, we could easily see that the numbers have no variation.

But, when the numbers are different, we can see that there is a very large matrix of difference scores. How can we summarize that? How about we apply what we learned from the previous section on measures of central tendency. We have a lot of differences, so we could ask something like, what is the average difference that we have? So, we could just take all of our differences, and compute the mean difference right? What do you think would happen if we did that?

Let's try it out on these three numbers:

1 2 3

	1	2	3
1	0	1	2
2	-1	0	1
3	-2	-1	0

You might already guess what is going to happen. Let's compute the mean:

$$\text{mean of difference scores} = \frac{0+1+2-1+0+1-2-1+0}{9} = \frac{0}{9} = 0$$

Uh oh, we get zero for the mean of the difference scores. This will always happen whenever you take the mean of the difference scores. We can see that there are some differences between the numbers, so using 0 as the summary value for the variation in the numbers doesn't make much sense.

Furthermore, you might also notice that the matrices of difference scores are redundant. The diagonal is always zero, and numbers on one side of the diagonal are the same as the numbers on the other side, except their signs are reversed. So, that's one reason why the difference scores add up to zero.

These are little problems that can be solved by computing the **variance** and the **standard deviation**. For now, the standard deviation is a just a trick that we use to avoid getting a zero. But, later we will see it has properties that are important for other reasons.

The Variance

Variability, variation, variance, vary, variable, varying, variety. Confused yet? Before we describe **the variance**, we want to you be OK with how this word is used. First, don't forget the big picture. We know that variability and variation refers to the big idea of differences between numbers. We can even use the word variance in the same way. When numbers are different, they have variance.

> **Note**
>
> The formulas for variance and standard deviation depend on whether you think your data represents an entire population of numbers, or is sample from the population. We discuss this issue in later on. For now, we divide by N, later we discuss why you will often divide by N-1 instead.

The word **variance** also refers to a specific summary statistic, the sum of the squared deviations from the mean. Hold on what? Plain English please. The variance is the sum of the squared difference scores, where the difference scores are computed between each score and the mean. What are these scores? The scores are the numbers in the data set. Let's see the formula in English first:

$$variance = \frac{\text{Sum of squared difference scores}}{\text{Number of Scores}}$$

Deviations from the mean, Difference scores from the mean

We got a little bit complicated before when we computed the difference scores between all of the numbers in the data. Let's do it again, but in a more manageable way. This time, we calculate the difference between each score and the mean. The idea here is

1. We can figure out how similar our scores are by computing the mean
2. Then we can figure out how different our scores are from the mean

This could tell us, 1) something about whether our scores are really all very close to the mean (which could help us know if the mean is good representative number of the data), and 2) something about how much differences there are in the numbers.

Take a look at this table:

scores	values	mean	Difference_from_Mean
1	1	4.5	-3.5
2	6	4.5	1.5
3	4	4.5	-0.5
4	2	4.5	-2.5
5	6	4.5	1.5
6	8	4.5	3.5
Sums	27	27	0
Means	4.5	4.5	0

The first column shows we have 6 scores in the data set, and the `value` columns shows each score. The sum of the values, and the mean is presented on the last two rows. The sum and the mean were obtained by:

$$\frac{1+6+4+2+6+8}{6} = \frac{27}{6} = 4.5$$

The third column `mean`, appears a bit silly. We are just listing the mean once for every score. If you think back to our discussion about the meaning of the mean, then you will remember that it equally distributes the total sum across each data point. We can see that here, if we treat each score as the mean, then every score is a 4.5. We can also see that adding up all of the means for each score gives us back 27, which is the sum of the original values. Also, we see that if we find the mean of the mean scores, we get back the mean (4.5 again).

All of the action is occurring in the fourth column, `Difference_from_Mean`. Here, we are showing the difference scores from the mean, using $X_i - \bar{X}$. In other words, we subtracted the mean from each score. So, the first score, 1, is -3.5 from the mean, the second score, 6, is +1.5 from the mean, and so on.

Now, we can look at our original scores and we can look at their differences from the mean. Notice, we don't have a matrix of raw difference scores, so it is much easier to look at out. But, we still have a problem:

We can see that there are non-zero values in the difference scores, so we know there are a differences in the data. But, when we add them all up, we still get zero, which makes it seem like there are a total of zero differences in the data...Why does this happen...and what to do about it?

The mean is the balancing point in the data

One brief pause here to point out another wonderful property of the mean. It is the balancing point in the data. If you take a pen or pencil and try to balance it on your figure so it lays flat what are you doing? You need to find the center of mass in the pen, so that half of it is on one side, and the other half is on the other side. That's how balancing works. One side = the other side.

We can think of data as having mass or weight to it. If we put our data on our bathroom scale, we could figure out how heavy it was by summing it up. If we wanted to split the data down the middle so that half of the weight was equal to the other half, then we could balance the data on top of a pin. The mean of the data tells you where to put the pin. It is the location in the data, where the numbers on the one side add up to the same sum as the numbers on the other side.

If we think this through, it means that the sum of the difference scores from the mean will always add up to zero. This is because the numbers on one side of the mean will always add up to -x (whatever the sum of those numbers is), and the numbers of the other side of the mean will always add up to +x (which will be the same value only positive). And:

$-x + x = 0$, right.

Right.

The squared deviations

Some devious someone divined a solution to the fact that differences scores from the mean always add to zero. Can you think of any solutions? For example, what could you do to the difference scores so that you could add them up, and they would weigh something useful, that is they would not be zero?

The devious solution is to square the numbers. Squaring numbers converts all the negative numbers to positive numbers. For example, $2^2 = 4$, and $-2^2 = 4$. Remember how squaring works, we multiply the number twice: $2^2 = 2 * 2 = 4$, and $-2^2 = -2 * -2 = 4$. We use the term **squared deviations** to refer to differences scores that have been squared. Deviations are things that move away from something. The difference scores move away from the mean, so we also call them **deviations**.

Let's look at our table again, but add the squared deviations.

scores	values	mean	Difference_from_Mean	Squared_Deviations
1	1	4.5	-3.5	12.25
2	6	4.5	1.5	2.25
3	4	4.5	-0.5	0.25
4	2	4.5	-2.5	6.25
5	6	4.5	1.5	2.25
6	8	4.5	3.5	12.25
Sums	27	27	0	35.5
Means	4.5	4.5	0	5.91666666666667

OK, now we have a new column called `squared_deviations`. These are just the difference scores squared. So, $-3.5^2 = 12.25$, etc. You can confirm for yourself with your cellphone calculator.

Now that all of the squared deviations are positive, we can add them up. When we do this we create something very special called the sum of squares (SS), also known as the sum of the squared deviations from the mean. We will talk at length about this SS later on in the ANOVA chapter. So, when you get there, remember that you already know what it is, just some sums of some squared deviations, nothing fancy.

Finally, the variance

Guess what, we already computed the variance. It already happened, and maybe you didn't notice. "Wait, I missed that, what happened?".

First, see if you can remember what we are trying to do here. Take a pause, and see if you can tell yourself what problem we are trying solve.

pause

Without further ado, we are trying to get a summary of the differences in our data. There are just as many difference scores from the mean as there are data points, which can be a lot, so it would be nice to have a single number to look at, something like a mean, that would tell us about the average differences in the data.

If you look at the table, you can see we already computed the mean of the squared deviations. First, we found the sum (SS), then below that we calculated the mean = 5.916 repeating. This is **the variance**. The variance is the mean of the sum of the squared deviations:

$variance = \frac{SS}{N}$, where SS is the sum of the squared deviations, and N is the number of observations.

OK, now what. What do I do with the variance? What does this number mean? Good question. The variance is often an unhelpful number to look at. Why? Because it is not in the same scale as the original data. This is because we squared the difference scores before taking the mean. Squaring produces large numbers. For example, we see a 12.25 in there. That's a big difference, bigger than any difference between any two original values. What to do? How can we bring the numbers back down to their original unsquared size?

If you are thinking about taking the square root, that's a ding ding ding, correct answer for you. We can always unsquare anything by taking the square root. So, let's do that to 5.916. $\sqrt{5.916} = 2.4322829$.

The Standard Deviation

Oops, we did it again. We already computed the standard deviation, and we didn't tell you. The standard deviation is the square root of the variance...At least, it is right now, until we complicate matters for you in the next chapter.

Here is the formula for the standard deviation:

$$\text{standard deviation} = \sqrt{Variance} = \sqrt{\frac{SS}{N}}$$

We could also expand this to say:

$$\text{standard deviation} = \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{N}}$$

Don't let those big square root signs put you off. Now, you know what they are doing there. Just bringing our measure of the variance back down to the original size of the data. Let's look at our table again:

scores	values	mean	Difference_from_Mean	Squared_Deviations
1	1	4.5	-3.5	12.25
2	6	4.5	1.5	2.25
3	4	4.5	-0.5	0.25
4	2	4.5	-2.5	6.25
5	6	4.5	1.5	2.25
6	8	4.5	3.5	12.25
Sums	27	27	0	35.5
Means	4.5	4.5	0	5.91666666666667

We measured the standard deviation as 2.4322829. Notice this number fits right in the with differences scores from the mean. All of the scores are kind of in and around + or - 2.4322829. Whereas, if we looked at the variance, 5.916 is just too big, it doesn't summarize the actual differences very well.

What does all this mean? Well, if someone told they had some number with a mean of 4.5 (like the values in our table), and a standard deviation of 2.4322829, you would get a pretty good summary of the numbers. You would know that many of the numbers

are around 4.5, and you would know that not all of the numbers are 4.5. You would know that the numbers spread around 4.5. You also know that the spread isn't super huge, it's only + or - 2.4322829 on average. That's a good starting point for describing numbers.

If you had loads of numbers, you could reduce them down to the mean and the standard deviation, and still be pretty well off in terms of getting a sense of those numbers.

This page titled 11.5: Measures of Variation (Differentness) is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

11.6: Using Descriptive Statistics with data

Remember, you will be learning how to compute descriptive statistics using software in the labs. Check out the lab manual exercises for descriptives to see some examples of working with real data.

This page titled 11.6: Using Descriptive Statistics with data is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

11.7: Rolling your own descriptive statistics

We spent many paragraphs talking about variation in numbers, and how to use calculate the **variance** and **standard deviation** to summarize the average differences between numbers in a data set. The basic process was to 1) calculate some measure of the differences, then 2) average the differences to create a summary. We found that we couldn't average the raw difference scores, because we would always get a zero. So, we squared the differences from the mean, then averaged the squared differences differences. Finally, we square rooted our measure to bring the summary back down to the scale of the original numbers.

Perhaps you haven't heard, but there is more than one way to skin a cat, but we prefer to think of this in terms of petting cats, because some of us love cats. Jokes aside, perhaps you were also thinking that the problem of summing differences scores (so that they don't equal zero), can be solved in more than one way. Can you think of a different way, besides squaring?

Absolute deviations

How about just taking the absolute value of the difference scores. Remember, the absolute value converts any number to a positive value. Check out the following table:

scores	values	mean	Difference_from_Mean	Absolute_Deviations
1	1	4.5	-3.5	3.5
2	6	4.5	1.5	1.5
3	4	4.5	-0.5	0.5
4	2	4.5	-2.5	2.5
5	6	4.5	1.5	1.5
6	8	4.5	3.5	3.5
Sums	27	27	0	13
Means	4.5	4.5	0	2.16666666666667

This works pretty well too. By converting the difference scores from the mean to positive values, we can now add them up and get a non-zero value (if there are differences). Then, we can find the mean of the sum of the absolute deviations. If we were to map the terms sum of squares (SS), variance and standard deviation onto these new measures based off of the absolute deviation, how would the mapping go? For example, what value in the table corresponds to the SS? That would be the sum of absolute deviations in the last column. How about the variance and standard deviation, what do those correspond to? Remember that the variance is mean (SS/N), and the standard deviation is a square-rooted mean ($\sqrt{SS/N}$). In the table above we only have one corresponding mean, the mean of the sum of the absolute deviations. So, we have a **variance** measure that does not need to be square rooted. We might say the mean absolute deviation, is doing double-duty as a variance and a standard-deviation. Neat.

Other sign-inverting operations

In principle, we could create lots of different summary statistics for variance that solve the summing to zero problem. For example, we could raise every difference score to any even numbered power beyond 2 (which is the square). We could use, 4, 6, 8, 10, etc. There is an infinity of even numbers, so there is an infinity of possible variance statistics. We could also use odd numbers as powers, and then take their absolute value. Many things are possible. The important aspect to any of this is to have a reason for what you are doing, and to choose a method that works for the data-analysis problem you are trying to solve. Note also, we bring up this general issue because we want you to understand that statistics is a creative exercise. We invent things when we need them, and we use things that have already been invented when they work for the problem at hand.

This page titled 11.7: Rolling your own descriptive statistics is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

11.8: Remember to look at your data

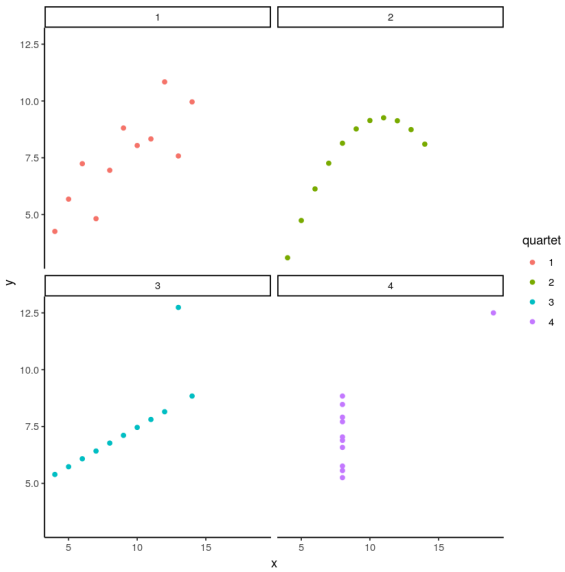
Descriptive statistics are great and we will use them a lot in the course to describe data. You may suspect that descriptive statistics also have some short-comings. This is very true. They are compressed summaries of large piles of numbers. They will almost always be unable to represent all of the numbers fairly. There are also different kinds of descriptive statistics that you could use, and it sometimes not clear which one's you should use.

Perhaps the most important thing you can do when using descriptives is to use them in combination with looking at the data in a graph form. This can help you see whether or not your descriptives are doing a good job of representing the data.

Anscombe's Quartet

To hit this point home, and to get you thinking about the issues we discuss in the next chapter, check this out. It's called Anscombe's Quartet, because these interesting graphs and numbers and numbers were produced by Anscombe (1973). You are looking at pairs of measurements. Each graph has an X and Y axis, and each point represents two measurements. Each of the graphs looks very different, right?

```
library(data.table)
library(ggplot2)
ac <- fread("https://stats.libretexts.org/@api/deki/files/10478/anscombe.txt")
ac<-as.data.frame(ac)
ac_long<-data.frame(x=c(ac[,1],
                        ac[,3],
                        ac[,5],
                        ac[,7]),
                    y=c(ac[,2],
                        ac[,4],
                        ac[,6],
                        ac[,8]),
                    quartet = as.factor(rep(1:4,each=11))
                        )
ggplot(ac_long, aes(x=x, y=y, color=quartet))+
  geom_point()+
  theme_classic()+
  facet_wrap(~quartet)
```



Well, would you be surprised if I told that the descriptive statistics for the numbers in these graphs are exactly the same? It turns

out they do have the same descriptive statistics. In the table below I present the mean and variance for the x-values in each graph, and the mean and the variance for the y-values in each graph.

quartet	mean_x	var_x	mean_y	var_y
1	9	11	7.500909	4.127269
2	9	11	7.500909	4.127629
3	9	11	7.500000	4.122620
4	9	11	7.500909	4.123249

The descriptives are all the same! Anscombe put these special numbers together to illustrate the point of graphing your numbers. If you only look at your descriptives, you don't know what patterns in the data they are hiding. If you look at the graph, then you can get a better understanding.

Datasaurus Dozen

If you thought that Anscombe's quartet was neat, you should take a look at the Datasaurus Dozen (Matejka and Fitzmaurice 2017). Scroll down to see the examples. You will be looking at dot plots. The dot plots show many different patterns, including dinosaurs! What's amazing is that all of the dots have very nearly the same descriptive statistics. Just another reminder to look at your data, it might look like a dinosaur!

This page titled 11.8: Remember to look at your data is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

11.9: Videos

Measures of center: Mode

(https://www.youtube.com/watch?v=hQ2p-QQpGso)

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Measures of center: Median and Mean

(https://www.youtube.com/watch?v=BopmCXCjq08)

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Standard deviation part I

(https://www.youtube.com/watch?v=8Yguf93s5dI)

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Standard deviation part II

(https://www.youtube.com/watch?v=KodmsOXScBc)

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

This page titled 11.9: Videos is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

- 1.4: Sources of Law and Their Priority is licensed CC BY-NC-SA 4.0.

11.10: References

Anscombe, F. J. 1973. “Graphs in Statistical Analysis.” *American Statistician* 27: 17–21.

Matejka, Justin, and George Fitzmaurice. 2017. “Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics Through Simulated Annealing.” In *Proceedings of the 2017 Chi Conference on Human Factors in Computing Systems*, 1290–4. ACM.

This page titled 11.10: References is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

- 29.4: Case is licensed CC BY-NC-SA 4.0.

12: CORRELATION

Chapter by Matthew Crump

> Correlation does not equal causation
>
> — Every Statistics and Research Methods Instructor Ever

In the last chapter we had some data. It was too much too look at and it didn't make sense. So, we talked about how to look at the data visually using plots and histograms, and we talked about how to summarize lots of numbers so we could determine their central tendencies (sameness) and variability (differentness). And, all was well with the world.

Let's not forget the big reason why we learned about descriptive statistics. The big reason is that we are interested in getting answers to questions using data.

If you are looking for a big theme to think about while you take this course, the theme is: how do we ask and answer questions using data?

For every section in this book, you should be connecting your inner monologue to this question, and asking yourself: How does what I am learning about help me answer questions with data? Advance warning: we know it is easy to forget this stuff when we dive into the details, and we will try to throw you a rope to help you out along the way...remember, we're trying to answer questions with data.

We started Chapter two with some fake data on human happiness, remember? We imagined that we asked a bunch of people to tell us how happy they were, then we looked at the numbers they gave us. Let's continue with this imaginary thought experiment.

What do you get when you ask people to use a number to describe how happy they are? A bunch of numbers. What kind of questions can you ask about those numbers? Well, you can look at the numbers and estimate their general properties as we already did. We would expect those numbers tell us some things we already know. There are different people, and different people are different amounts of happy. You've probably met some of those of really happy people, and really unhappy people, and you yourself probably have some amount of happiness. "Great, thanks Captain Obvious".

Before moving on, you should also be skeptical of what the numbers might mean. For example, if you force people to give a number between 0-100 to rate their happiness, does this number truly reflect how happy that person is? Can a person know how happy they are? Does the question format bias how they give their answer? Is happiness even a real thing? These are all good questions about the **validity** of the construct (happiness itself) and the measure (numbers) you are using to quantify it. For now, though, we will side-step those very important questions, and assume that, happiness is a thing, and our measure of happiness measures something about how happy people are.

OK then, after we have measured some happiness, I bet you can think of some more pressing questions. For example, what causes happiness to go up or down. If you knew the causes of happiness what could you do? How about increase your own happiness; or, help people who are unhappy; or, better appreciate why Eeyore from Winnie the Pooh is unhappy; or, present valid scientific arguments that argue against incorrect claims about what causes happiness. A causal theory and understanding of happiness could be used for all of those things. How can we get there?

Imagine you were an alien observer. You arrived on earth and heard about this thing called happiness that people have. You want to know what causes happiness. You also discover that planet earth has lots of other things. Which of those things, you wonder, cause happiness? How would your alien-self get started on this big question.

As a person who has happiness, you might already have some hunches about what causes changes in happiness. For example things like: weather, friends, music, money, education, drugs, books, movies, beliefs, personality, color of your shoes, eyebrow length, number of cat's you see per day, frequency of subway delay, a lifetime supply of chocolate, etcetera etcetera (as Willy Wonka would say), might all contribute to happiness in someway. There could be many different causes of happiness.

12.1: If something caused something else to change, what would that look like?

12.2: Pearson's r
12.3: Turning the numbers into a measure of co-variance
12.4: Examples with Data
12.5: Regression — A mini intro
12.6: Interpreting Correlations
12.7: Summary
12.8: References

This page titled 12: Correlation is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

12.1: If something caused something else to change, what would that look like?

Before we go around determining the causes of happiness, we should prepare ourselves with some analytical tools so that we could identify what causation looks like. If we don't prepare ourselves for what we might find, then we won't know how to interpret our own data. Instead, we need to anticipate what the data could look like. Specifically, we need to know what data would look like when one thing does not cause another thing, and what data would look like when one thing does cause another thing. This chapter does some of this preparation. Fair warning: we will find out some tricky things. For example, we can find patterns that look like one thing is causing another, even when that one thing DOES NOT CAUSE the other thing. Hang in there.

Charlie and the Chocolate factory

Let's imagine that a person's supply of chocolate has a causal influence on their level of happiness. Let's further imagine that, like Charlie, the more chocolate you have the more happy you will be, and the less chocolate you have, the less happy you will be. Finally, because we suspect happiness is caused by lots of other things in a person's life, we anticipate that the relationship between chocolate supply and happiness won't be perfect. What do these assumptions mean for how the data should look?

Our first step is to collect some imaginary data from 100 people. We walk around and ask the first 100 people we meet to answer two questions:

1. how much chocolate do you have, and
2. how happy are you.

For convenience, both the scales will go from 0 to 100. For the chocolate scale, 0 means no chocolate, 100 means lifetime supply of chocolate. Any other number is somewhere in between. For the happiness scale, 0 means no happiness, 100 means all of the happiness, and in between means some amount in between.

Here is some sample data from the first 10 imaginary subjects.

subject	chocolate	happiness
1	1	1
2	1	1
3	2	2
4	2	4
5	4	5
6	4	5
7	7	5
8	8	5
9	8	6
10	9	6

We asked each subject two questions so there are two scores for each subject, one for their chocolate supply, and one for their level of happiness. You might already notice some relationships between amount of chocolate and level of happiness in the table. To make those relationships even more clear, let's plot all of the data in a graph.

Scatter plots

When you have two measurements worth of data, you can always turn them into dots and plot them in a scatter plot. A scatter

plot has a horizontal x-axis, and a vertical y-axis. You get to choose which measurement goes on which axis. Let's put chocolate supply on the x-axis, and happiness level on the y-axis. The plot below shows 100 dots for each subject.

```
library(ggplot2)
subject<-1:100
chocolate<-round(1:100*runif(100,.5,1))
happiness<-round(1:100*runif(100,.5,1))
the_df_CC<-data.frame(subject,chocolate,happiness)
ggplot(the_df_CC,aes(x=chocolate,y=happiness))+
  geom_point()+
  theme_classic()
```

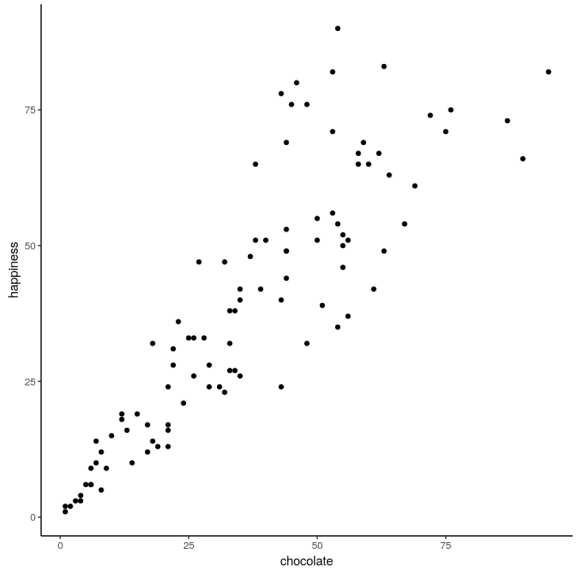


Figure 12.1.1: Imaginary data showing a positive correlation between amount of chocolate and amount happiness.

You might be wondering, why are there only 100 dots for the data. Didn't we collect 100 measures for chocolate, and 100 measures for happiness, shouldn't there be 200 dots? Nope. Each dot is for one subject, there are 100 subjects, so there are 100 dots.

What do the dots mean? Each dot has two coordinates, an x-coordinate for chocolate, and a y-coordinate for happiness. The first dot, all the way on the bottom left is the first subject in the table, who had close to 0 chocolate and close to zero happiness. You can look at any dot, then draw a straight line down to the x-axis: that will tell you how much chocolate that subject has. You can draw a straight line left to the y-axis: that will tell you how much happiness the subject has.

Now that we are looking at the scatter plot, we can see many things. The dots are scattered around a bit aren't they, hence **scatter plot**. Even when the dot's don't scatter, they're still called scatter plots, perhaps because those pesky dots in real life have so much scatter all the time. More important, the dots show a relationship between chocolate supply and happiness. Happiness is lower for people with smaller supplies of chocolate, and higher for people with larger supplies of chocolate. It looks like the more chocolate you have the happier you will be, and vice-versa. This kind of relationship is called a **positive correlation**.

Positive, Negative, and No-Correlation

Seeing as we are in the business of imagining data, let's imagine some more. We've already imagined what data would look like if larger chocolate supplies increase happiness. We'll show that again in a bit. What do you imagine the scatter plot would look like if the relationship was reversed, and larger chocolate supplies decreased happiness. Or, what do you imagine the scatter plot would look like if there was no relationship, and the amount of chocolate that you have doesn't do anything to your happiness. We invite your imagination to look at these graphs:

```
library(ggplot2)
subject_x<-1:100
chocolate_x<-round(1:100*runif(100,.5,1))
```

```
happiness_x<-round(1:100*runif(100,.5,1))
df_positive<-data.frame(subject_x,chocolate_x,happiness_x)
subject_x<-1:100
chocolate_x<-round(1:100*runif(100,.5,1))
happiness_x<-round(100:1*runif(100,.5,1))
df_negative<-data.frame(subject_x,chocolate_x,happiness_x)
subject_x<-1:100
chocolate_x<-round(runif(100,0,100))
happiness_x<-round(runif(100,0,100))
df_random<-data.frame(subject_x,chocolate_x,happiness_x)
all_data<-rbind(df_positive,df_negative,df_random)
all_data<-cbind(all_data,correlation=rep(c("positive","negative","random"),each=100))
ggplot(all_data,aes(x=chocolate_x,y=happiness_x))+
  geom_point()+
  theme_classic()+
  facet_wrap(~correlation)+
  xlab("chocolate supply")+
  ylab("happiness")
```

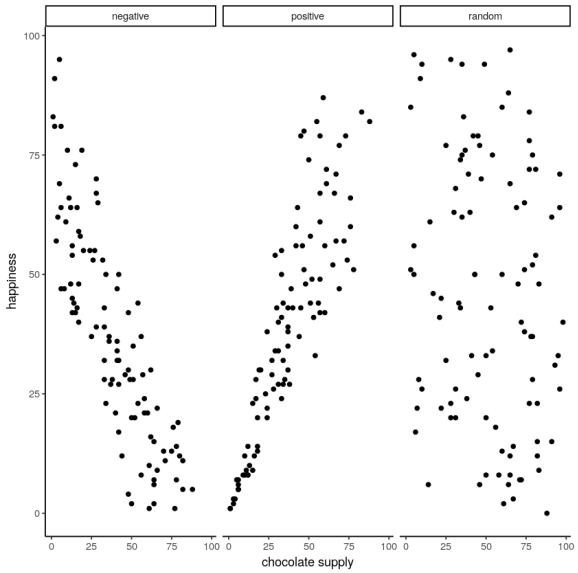


Figure 12.1.2: Three scatterplots showing negative, positive, and zero correlation.

The first panel shows a **negative correlation**. Happiness goes down as chocolate supply increases. Negative correlation occurs when one thing goes up and the other thing goes down; or, when more of X is less of Y, and vice-versa. The second panel shows a **positive correlation**. Happiness goes up as chocolate as chocolate supply increases. Positive correlation occurs when both things go up together, and go down together: more of X is more of Y, and vice-versa. The third panel shows **no correlation**. Here, there doesn't appear to be any obvious relationship between chocolate supply and happiness. The dots are scattered all over the place, the truest of the scatter plots.

Note

We are wading into the idea that measures of two things can be related, or correlated with one another. It is possible for the relationships to be more complicated than just going up, or going down. For example, we could have a relationship that where the dots go up for the first half of X, and then go down for the second half.

Zero correlation occurs when one thing is not related in any way to another things: changes in X do not relate to any changes in Y, and vice-versa.

This page titled 12.1: If something caused something else to change, what would that look like? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

12.2: Pearson's r

If Beyoncé was a statistician, she might look at these scatter plots and want to "put a number on it". We think this is a good idea too. We've already learned how to create descriptive statistics for a single measure, like chocolate, or happiness (i.e., means, variances, etc.). Is it possible to create a descriptive statistic that summarized the relationship between two measures, all in one number? Can it be done? Karl Pearson to the rescue.

> **Note**
>
> The stories about the invention of various statistics are very interesting, you can read more about them in the book, "The Lady Tasting Tea" (Salsburg 2001)

There's a statistic for that, and Karl Pearson invented it. Everyone now calls it, "Pearson's r". We will find out later that Karl Pearson was a big-wig editor at Biometrika in the 1930s. He took a hating to another big-wig statistician, Sir Ronald Fisher (who we learn about later), and they had some stats fights...why can't we all just get along in statistics.

How does Pearson's r work? Let's look again at the first 10 subjects in our fake experiment:

subject	chocolate	happiness
1	1	1
2	2	2
3	2	3
4	3	3
5	3	3
6	5	5
7	4	6
8	5	5
9	9	5
10	6	9
Sums	40	42
Means	4	4.2

What could we do to these numbers to produce a single summary value that represents the relationship between the chocolate supply and happiness?

The idea of co-variance

"Oh please no, don't use the word variance again". Yes, we're doing it, we're going to use the word variance again, and again, until it starts making sense. Remember what variance means about some numbers. It means the numbers have some change in them, they are not all the same, some of them are big, some are small. We can see that there is variance in chocolate supply across the 10 subjects. We can see that there is variance in happiness across the 10 subjects. We also saw in the scatter plot, that happiness increases as chocolate supply increases; which is a positive relationship, a positive correlation. What does this have to do with variance? Well, it means there is a relationship between the variance in chocolate supply, and the variance in happiness levels. The two measures vary together don't they? When we have two measures that vary together, they are like a happy couple who share their variance. This is what co-variance refers to, the idea that the pattern of varying numbers in one measure is shared by the pattern of varying numbers in another measure.

Co-variance is **very, very, very ,very** important. We suspect that the word co-variance is initially confusing, especially if you are not yet fully comfortable with the meaning of variance for a single measure. Nevertheless, we must proceed and use the idea of co-variance over and over again to firmly implant it into your statistical mind (we already said, but redundancy works, it's a thing).

> Pro tip: Three-legged race is a metaphor for co-variance. Two people tie one leg to each other, then try to walk. It works when they co-vary their legs together (positive relationship). They can also co-vary in an unhelpful way, when one person tries to move forward exactly when the other person tries to move backward. This is still co-variance (negative relationship). Funny random walking happens when there is no co-variance. This means one person does whatever they want, and so does the other person. There is a lot of variance, but the variance is shared randomly, so it's just a bunch of legs moving around accomplishing nothing.
>
> Pro tip #2: Succesfully playing paddycake occurs when two people coordinate their actions so they have postively shared co-variance.

This page titled 12.2: Pearson's r is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

12.3: Turning the numbers into a measure of co-variance

“OK, so if you are saying that co-variance is just another word for correlation or relationship between two measures, I’m good with that. I suppose we would need some way to measure that.” Correct, back to our table...notice anything new?

subject	chocolate	happiness	Chocolate_X_Happiness
1	1	1	1
2	2	2	4
3	2	3	6
4	3	3	9
5	3	3	9
6	5	5	25
7	4	6	24
8	5	5	25
9	9	5	45
10	6	9	54
Sums	40	42	202
Means	4	4.2	20.2

We’ve added a new column called “Chocolate_X_Happiness”, which translates to Chocolate scores multiplied by Happiness scores. Each row in the new column, is the product, or multiplication of the chocolate and happiness score for that row. Yes, but why would we do this?

Last chapter we took you back to Elementary school and had you think about division. Now it’s time to do the same thing with multiplication. We assume you know how that works. One number times another, means taking the first number, and adding it as many times as the second says to do,

$2 * 2 = 2 + 2 = 4$

$2 * 6 = 2 + 2 + 2 + 2 + 2 + 2 = 12$, or $6 + 6 = 12$, same thing.

Yes, you know all that. But, can you bend multiplication to your will, and make it do your bidding when need to solve a problem like summarizing co-variance? Multiplication is the droid you are looking for.

We know how to multiple numbers, and all we have to next is think about the consequences of multiplying sets of numbers together. For example, what happens when you multiply two small numbers together, compared to multiplying two big numbers together? The first product should be smaller than the second product right? How about things like multiplying a small number by a big number? Those products should be in between right?.

Then next step is to think about how the products of two measures sum together, depending on how they line up. Let’s look at another table:

scores	X	Y	A	B	XY	AB
1	1	1	1	10	1	10
2	2	2	2	9	4	18
3	3	3	3	8	9	24
4	4	4	4	7	16	28

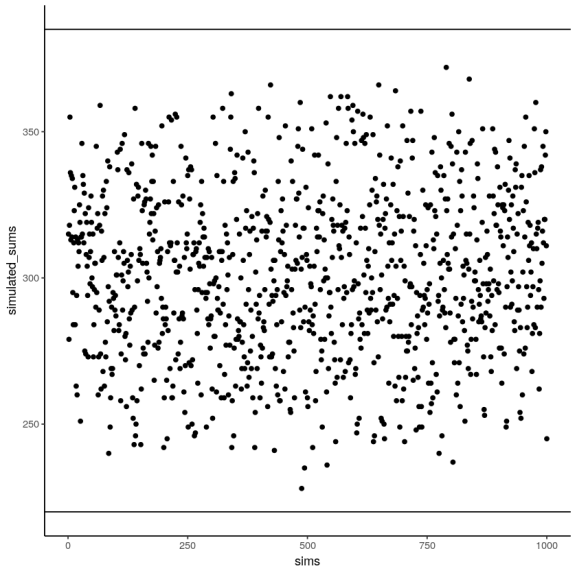
scores	X	Y	A	B	XY	AB
5	5	5	5	6	25	30
6	6	6	6	5	36	30
7	7	7	7	4	49	28
8	8	8	8	3	64	24
9	9	9	9	2	81	18
10	10	10	10	1	100	10
Sums	55	55	55	55	385	220
Means	5.5	5.5	5.5	5.5	38.5	22

Look at the X and Y column. The scores for X and Y perfectly co-vary. When X is 1, Y is 1; when X is 2, Y is 2, etc. They are perfectly aligned. The scores for A and B also perfectly co-vary, just in the opposite manner. When A is 1, B is 10; when A is 2, B is 9, etc. B is a reversed copy of A.

Now, look at the column XY. These are the products we get when we multiply the values of X across with the values of Y. Also, look at the column AB. These are the products we get when we multiply the values of A across with the values of B. So far so good.

Now, look at the Sums for the XY and AB columns. Not the same. The sum of the XY products is 385, and the sum of the AB products is 220. For this specific set of data, the numbers 385 and 220 are very important. They represent the biggest possible sum of products (385), and the smallest possible sum of products (220). There is no way of re-ordering the numbers 1 to 10, say for X, and the numbers 1 to 10 for Y, that would ever produce larger or smaller numbers. Don't believe me? Check this out:

```
library(ggplot2)
simulated_sums<-length(0)
for(sim in 1:1000){
        X<-sample(1:10)
        Y<-sample(1:10)
        simulated_sums[sim]<-sum(X*Y)
}
sim_df<-data.frame(sims=1:1000,simulated_sums)
ggplot(sim_df,aes(x=sims,y=simulated_sums))+
  geom_point()+
  theme_classic()+
  geom_hline(yintercept = 385)+
  geom_hline(yintercept = 220)
```



The above graph shows 1000 computer simulations. I convinced my computer to randomly order the numbers 1 to 10 for X, and randomly order the numbers 1 to 10 for Y. Then, I multiplied X and Y, and added the products together. I did this 1000 times. The dots show the sum of the products for each simulation. The two black lines show the maximum possible sum (385), and the minimum possible sum (220), for this set of numbers. Notice, how all of the dots are in between the maximum and minimum possible values. Told you so.

"OK fine, you told me so...So what, who cares?". We've been looking for a way to summarize the co-variance between two measures right? Well, for these numbers, we have found one, haven't we. It's the sum of the products. We know that when the sum of the products is 385, we have found a perfect, positive correlation. We know, that when the sum of the products is 220, we have found a perfect negative correlation. What about the numbers in between. What could we conclude about the correlation if we found the sum of the products to be 350. Well, it's going to be positive, because it's close to 385, and that's perfectly positive. If the sum of the products was 240, that's going to be negative, because it's close to the perfectly negatively correlating 220. What about no correlation? Well, that's going to be in the middle between 220 and 385 right.

We have just come up with a data-specific summary measure for the correlation between the numbers 1 to 10 in X, and the numbers 1 to 10 in Y, it's the sum of the products. We know the maximum (385) and minimum values (220), so we can now interpret any product sum for this kind of data with respect to that scale.

> Pro tip: When the correlation between two measures increases in the positive direction, the sum of their products increases to its maximum possible value. This is because the bigger numbers in X will tend to line up with the bigger numbers in Y, creating the biggest possible sum of products. When the correlation between two measures increases in the negative direction, the sum of their products decreases to its minimum possible value. This is because the bigger numbers in X will tend to line up with the smaller numbers in Y, creating the smallest possible sum of products. When there is no correlation, the big numbers in X will be randomly lined up with the big and small numbers in Y, making the sum of the products, somewhere in the middle.

Co-variance, the measure

We took some time to see what happens when you multiply sets of numbers together. We found that $big * big = \text{bigger}$ and $small * small = \text{still small}$, and $big * small = \text{in the middle}$. The purpose of this was to give you some conceptual idea of how the co-variance between two measures is reflected in the sum of their products. We did something very straightforward. We just multiplied X with Y, and looked at how the product sums get big and small, as X and Y co-vary in different ways.

Now, we can get a little bit more formal. In statistics, **co-variance** is not just the straight multiplication of values in X and Y. Instead, it's the multiplication of the deviations in X from the mean of X, and the deviation in Y from the mean of Y. Remember those difference scores from the mean we talked about last chapter? They're coming back to haunt you know, but in a good way like Casper the friendly ghost.

Let's see what this look like in a table:

subject	chocolate	happiness	C_d	H_d	Cd_x_Hd
1	1	1	-3	-3.2	9.6
2	2	2	-2	-2.2	4.4
3	2	3	-2	-1.2	2.4
4	3	3	-1	-1.2	1.2
5	3	3	-1	-1.2	1.2
6	5	5	1	0.8	0.8
7	4	6	0	1.8	0
8	5	5	1	0.8	0.8
9	9	5	5	0.8	4
10	6	9	2	4.8	9.6
Sums	40	42	0	0	34
Means	4	4.2	0	0	3.4

We have computed the deviations from the mean for the chocolate scores (column C_d), and the deviations from the mean for the happiness scores (column H_d). Then, we multiplied them together (last column). Finally, you can see the mean of the products listed in the bottom right corner of the table, the official **the covariance**.

The formula for the co-variance is:

$$cov(X,Y) = \frac{\sum_i^n (x_i - \bar{X})(y_i - \bar{Y})}{N}$$

OK, so now we have a formal single number to calculate the relationship between two variables. This is great, it's what we've been looking for. However, there is a problem. Remember when we learned how to compute just the plain old **variance**. We looked at that number, and we didn't know what to make of it. It was squared, it wasn't in the same scale as the original data. So, we square rooted the **variance** to produce the **standard deviation**, which gave us a more interpretable number in the range of our data. The **co-variance** has a similar problem. When you calculate the co-variance as we just did, we don't know immediately know its scale. Is a 3 big? is a 6 big? is a 100 big? How big or small is this thing?

From our prelude discussion on the idea of co-variance, we learned the sum of products between two measures ranges between a maximum and minimum value. The same is true of the co-variance. For a given set of data, there is a maximum possible positive value for the co-variance (which occurs when there is perfect positive correlation). And, there is a minimum possible negative value for the co-variance (which occurs when there is a perfect negative correlation). When there is zero co-variation, guess what happens. Zeroes. So, at the very least, when we look at a co-variation statistic, we can see what direction it points, positive or negative. But, we don't know how big or small it is compared to the maximum or minimum possible value, so we don't know the relative size, which means we can't say how strong the correlation is. What to do?

Pearson's r we there yet

Yes, we are here now. Wouldn't it be nice if we could force our measure of co-variation to be between -1 and +1?

-1 would be the minimum possible value for a perfect negative correlation. +1 would be the maximum possible value for a perfect positive correlation. 0 would mean no correlation. Everything in between 0 and -1 would be increasingly large negative correlations. Everything between 0 and +1 would be increasingly large positive correlations. It would be a fantastic, sensible, easy to interpret system. If only we could force the co-variation number to be between -1 and 1. Fortunately, for us, this episode is brought to you by Pearson's r, which does precisely this wonderful thing.

Let's take a look at a formula for Pearson's r:

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{cov(X,Y)}{SD_X SD_Y}$$

We see the symbol σ here, that's more Greek for you. σ is often used as a symbol for the standard deviation (SD). If we read out the formula in English, we see that r is the co-variance of X and Y, divided by the product of the standard deviation of X and the standard deviation of Y. Why are we dividing the co-variance by the product of the standard deviations. This operation has the effect of **normalizing** the co-variance into the range -1 to 1.

Note

But, we will fill this part in as soon as we can...promissory note to explain the magic. FYI, it's not magic. Brief explanation here is that dividing each measure by its standard deviation ensures that the values in each measure are in the same range as one another.

For now, we will call this mathematical magic. It works, but we don't have space to tell you why it works right now.

> It's worth saying that there are loads of different formulas for computing Pearson's r. You can find them by Googling them. We will probably include more of them here, when we get around to it. However, they all give you the same answer. And, they are all not as pretty as each other. Some of them might even look scary. In other statistics textbook you will often find formulas that are easier to use for calculation purposes. For example, if you only had a pen and paper, you might use one or another formula because it helps you compute the answer faster by hand. To be honest, we are not very interested in teaching you how to plug numbers into formulas. We give one lesson on that here: Put the numbers into the letters, then compute the answer. Sorry to be snarky. Nowadays you have a computer that you should use for this kind of stuff. So, we are more interested in teaching you what the calculations mean, rather than how to do them. Of course, every week we are showing you how to do the calculations in lab with computers, because that is important to.

Does Pearson's r really stay between -1 and 1 no matter what? It's true, take a look at the following simulation. Here I randomly ordered the numbers 1 to 10 for an X measure, and did the same for a Y measure. Then, I computed Pearson's r, and repeated this process 1000 times. As you can see all of the dots are between -1 and 1. Neat huh.

```
library(ggplot2)
simulated_sums <- length(0)
for(sim in 1:1000){
  X <- sample(1:10)
  Y <- sample(1:10)
  simulated_sums[sim] <- cor(X,Y)
}
sim_df <- data.frame(sims=1:1000,simulated_sums)
ggplot(sim_df, aes(x = sims, y = simulated_sums))+
  geom_point()+
  theme_classic()+
  geom_hline(yintercept = -1)+
  geom_hline(yintercept = 1)+
  ggtitle("Simulation of 1000 r values")
```

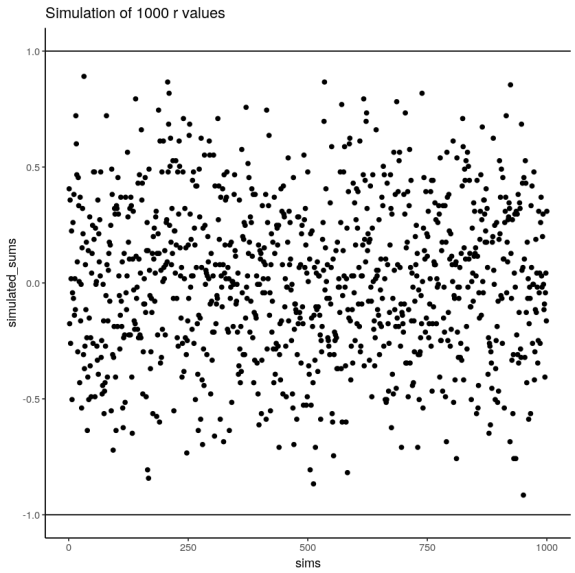


Figure 12.3.1: A simulation of of correlations. Each dot represents the r-value for the correlation between an X and Y variable that each contain the numbers 1 to 10 in random orders. The figure ilustrates that many r-values can be obtained by this random process.

This page titled 12.3: Turning the numbers into a measure of co-variance is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

12.4: Examples with Data

In the lab for correlation you will be shown how to compute correlations in real data-sets using software. To give you a brief preview, let's look at some data from the world happiness report (2018). This report measured various attitudes across people from different countries. For example, one question asked about how much freedom people thought they had to make life choices. Another question asked how confident people were in their national government. Here is a scatterplot showing the relationship between these two measures. Each dot represents means for different countries.

```
library(data.table)
library(ggplot2)
suppressPackageStartupMessages(library(dplyr))
whr_data <- fread("https://stats.libretexts.org/@api/deki/files/10477/WHR2018.csv")
# select DVs and filter for NAs
smaller_df <- whr_data %>%
              dplyr::select(country,
                     `Freedom to make life choices`,
                     `Confidence in national government`) %>%
              dplyr::filter(!is.na(`Freedom to make life choices`),
                     !is.na(`Confidence in national government`))
# plot the data with best fit line
ggplot(smaller_df, aes(x=`Freedom to make life choices`,
                       y=`Confidence in national government`))+
  geom_point(alpha=.5)+
  geom_smooth(method=lm, se=FALSE, formula=y ~ x)+
  theme_classic()
```

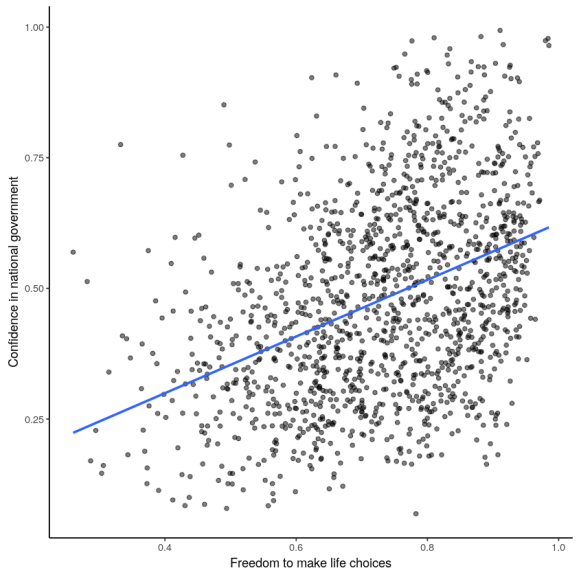


Figure 12.4.1: Relationship between freedom to make life choices and confidence in national government. Data from the world happiness report for 2018.

We put a blue line on the scatterplot to summarize the positive relationship. It appears that as “freedom to make life choices goes up”, so to does confidence in national government. It’s a positive correlation.

The actual correlation, as measured by Pearson’s r is:

```
library(data.table)
suppressPackageStartupMessages(library(dplyr))
whr_data <- fread("https://stats.libretexts.org/@api/deki/files/10477/WHR2018.csv")
# select DVs and filter for NAs
smaller_df <- whr_data %>%
```

```
                dplyr::select(country,
                       `Freedom to make life choices`,
                       `Confidence in national government`) %>%
                dplyr::filter(!is.na(`Freedom to make life choices`),
                       !is.na(`Confidence in national government`))
    # calculate correlation
    cor(smaller_df$`Freedom to make life choices`,
        smaller_df$`Confidence in national government`)

    0.408096292505333
```

You will do a lot more of this kind of thing in the lab. Looking at the graph you might start to wonder: Does freedom to make life choices cause changes how confident people are in their national government? Our does it work the other way? Does being confident in your national government give you a greater sense of freedom to make life choices? Or, is this just a random relationship that doesn't mean anything? All good questions. These data do not provide the answers, they just suggest a possible relationship.

This page titled 12.4: Examples with Data is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

12.5: Regression — A mini intro

We're going to spend the next little bit adding one more thing to our understanding of correlation. It's called **linear regression**. It sounds scary, and it really is. You'll find out much later in your Statistics education that everything we will be soon be talking about can be thought of as a special case of regression. But, we don't want to scare you off, so right now we just introduce the basic concepts.

First, let's look at a linear regression. This way we can see what we're trying to learn about. Here's some scatter plots, same one's you've already seen. But, we've added something new! Lines.

```
library(ggplot2)
subject_x<-1:100
chocolate_x<-round(1:100*runif(100,.5,1))
happiness_x<-round(1:100*runif(100,.5,1))
df_positive<-data.frame(subject_x,chocolate_x,happiness_x)
subject_x<-1:100
chocolate_x<-round(1:100*runif(100,.5,1))
happiness_x<-round(100:1*runif(100,.5,1))
df_negative<-data.frame(subject_x,chocolate_x,happiness_x)
subject_x<-1:100
chocolate_x<-round(runif(100,0,100))
happiness_x<-round(runif(100,0,100))
df_random<-data.frame(subject_x,chocolate_x,happiness_x)
all_data<-rbind(df_positive,df_negative,df_random)
all_data<-cbind(all_data,correlation=rep(c("positive","negative","random"),each=100))
ggplot(all_data,aes(x=chocolate_x,y=happiness_x))+
  geom_point()+
  theme_classic()+
  geom_smooth(method="lm",se=F,formula=y ~ x)+
  facet_wrap(~correlation)+
  xlab("chocolate supply")+
  ylab("happiness")
```

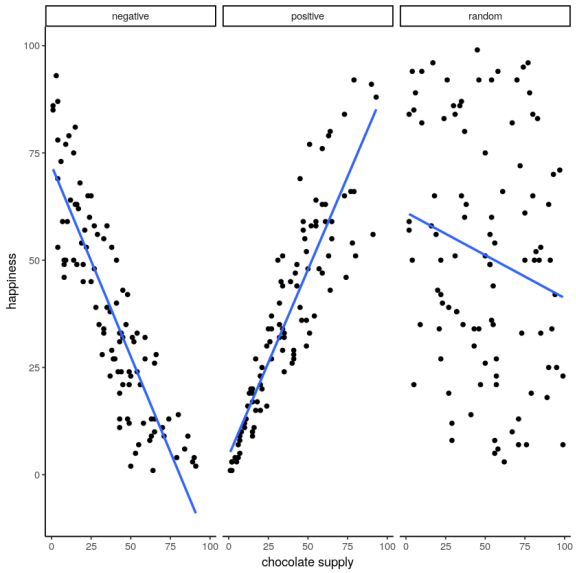


Figure 12.5.1: Three scatterplots showing negative, positive, and a random correlation (where the r-value is expected to be 0), along with the best fit regression line

The best fit line

Notice anything about these blue lines? Hopefully you can see, at least for the first two panels, that they go straight through the data, just like a kebab skewer. We call these lines **best fit** lines, because according to our definition (soon we promise) there are no other lines that you could draw that would do a better job of going straight throw the data.

One big idea here is that we are using the line as a kind of mean to describe the relationship between the two variables. When we only have one variable, that variable exists on a single dimension, it's 1D. So, it is appropriate that we only have one number, like the mean, to describe it's central tendency. When we have two variables, and plot them together, we now have a two-dimensional space. So, for two dimensions we could use a bigger thing that is 2d, like a line, to summarize the central tendency of the relationship between the two variables.

What do we want out of our line? Well, if you had a pencil, and a printout of the data, you could draw all sorts of straight lines any way you wanted. Your lines wouldn't even have to go through the data, or they could slant through the data with all sorts of angles. Would all of those lines be very good a describing the general pattern of the dots? Most of them would not. The best lines would go through the data following the general shape of the dots. Of the best lines, however, which one is the best? How can we find out, and what do we mean by that? In short, the best fit line is the one that has the least error.

> **Note**
>
> R code for plotting residuals thanks to Simon Jackson's blog post: https://drsimonj.svbtle.com/visualising-residuals

Check out this next plot, it shows a line through some dots. But, it also shows some teeny tiny lines. These lines drop down from each dot, and they land on the line. Each of these little lines is called a **residual**. They show you how far off the line is for different dots. It's measure of error, it shows us just how wrong the line is. After all, it's pretty obvious that not all of the dots are on the line. This means the line does not actually represent all of the dots. The line is wrong. But, the best fit line is the least wrong of all the wrong lines.

```
library(ggplot2)
d <- mtcars
fit <- lm(mpg ~ hp, data = d)
d$predicted <- predict(fit)   # Save the predicted values
d$residuals <- residuals(fit) # Save the residual values
ggplot(d, aes(x = hp, y = mpg)) +
  geom_smooth(method = "lm", se = FALSE,
              color = "lightblue", formula=y ~ x) +  # Plot regression slope
  geom_segment(aes(xend = hp, yend = predicted,
                   color="red"), alpha = .5) +  # alpha to fade lines
  geom_point() +
  geom_point(aes(y = predicted), shape = 1) +
  theme_classic()+
  theme(legend.position="none")+
  xlab("X")+ylab("Y")
# Quick look at the actual, predicted, and residual values
#library(dplyr)
#d %>% select(mpg, predicted, residuals) %>% head()
```

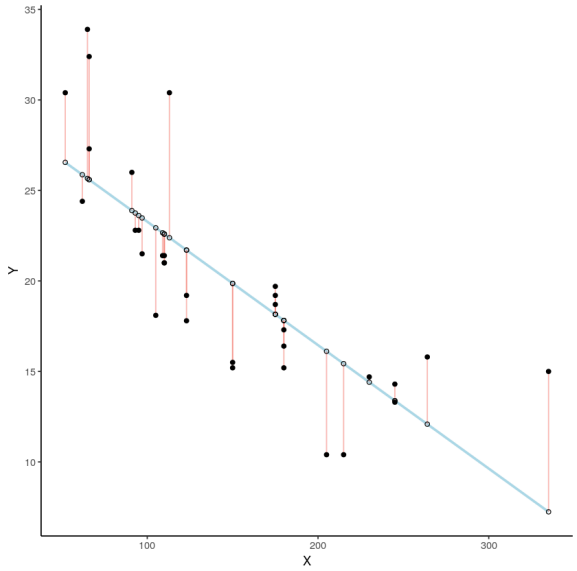


Figure 12.5.2: Black dots represent data points. The blue line is the best fit regression line. The white dots are repesent the predicted location of each black dot. The red lines show the error between each black dot and the regression line. The blue line is the best fit line because it minimizes the error shown by the red lines.

There's a lot going on in this graph. First, we are looking at a scatter plot of two variables, an X and Y variable. Each of the black dots are the actual values from these variables. You can see there is a negative correlation here, as X increases, Y tends to decrease. We drew a regression line through the data, that's the blue line. There's these little white dots too. This is where the line thinks the black dots should be. The red lines are the important residuals we've been talking about. Each black dot has a red line that drops straight down, or straight up from the location of the black dot, and lands directly on the line. We can already see that many of the dots are not on the line, so we already know the line is "off" by some amount for each dot. The red line just makes it easier to see exactly how off the line is.

The important thing that is happening here, is that the the blue line is drawn is such a way, that it minimizes the total length of the red lines. For example, if we wanted to know how wrong this line was, we could simply gather up all the red lines, measure how long they are, and then add all the wrongness together. This would give us the total amount of wrongness. We usually call this the error. In fact, we've already talked about this idea before when we discussed standard deviation. What we will actually be doing with the red lines, is computing the sum of the squared deviations from the line. That sum is the total amount of error. Now, this blue line here minimizes the sum of the squared deviations. Any other line would produce a larger total error.

Here's an animation to see this in action. The animations compares the best fit line in blue, to some other possible lines in black. The black line moves up and down. The red lines show the error between the black line and the data points. As the black line moves toward the best fit line, the total error, depicted visually by the grey area shrinks to it's minimum value. The total error expands as the black line moves away from the best fit line.

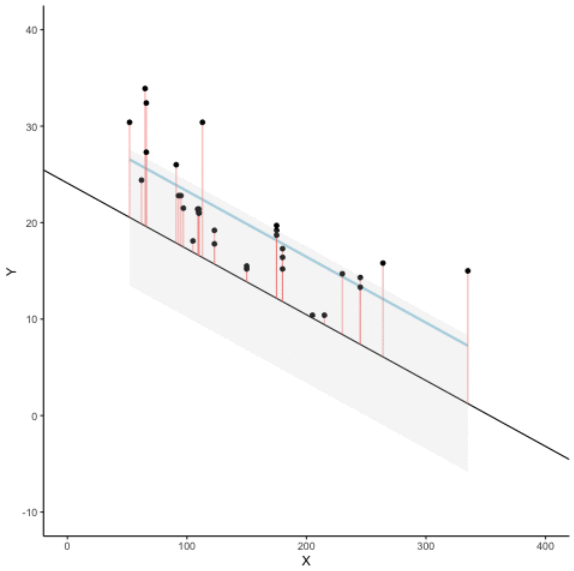


Figure 12.5.3: The blue line is the best fit regression line explaining the co-variation among the black dots. The black line moves up and down showing alternative lines that could be drawn. The red lines show the amount of error between each data point and the black line. The total amount of error is depicted by the shaded grey area. The size of the grey area expands as the black line moves away from the best fit line, and shrinks to a minimum as the black line moves toward the best fit line.

Whenever the black line does not overlap with the blue line, it is worse than the best fit line. The blue regression line is like Goldilocks, it's just right, and it's in the middle.

This next graph shows a little simulation of how the sum of squared deviations (the sum of the squared lengths of the red lines) behaves as we move the line up and down. What's going on here is that we are computing a measure of the total error as the black line moves through the best fit line. This represents the sum of the squared deviations. In other words, we square the length of each red line from the above animation, then we add up all of the squared red lines, and get the total error (the total sum of the squared deviations). The graph below shows what the total error looks like as the black line approaches then moves away from the best fit line. Notice, the dots in this graph start high on the left side, then they swoop down to a minimum at the bottom middle of the graph. When they reach their minimum point, we have found a line that minimizes the total error. This is the best fit regression line.

```
library(ggplot2)
d <- mtcars
fit <- lm(mpg ~ hp, data = d)
d$predicted <- predict(fit)   # Save the predicted values
d$residuals <- residuals(fit) # Save the residual values
coefs<-coef(lm(mpg ~ hp, data = mtcars))
#coefs[1]
#coefs[2]
x<-d$hp
move_line<-seq(-5,5,.5)
total_error<-c(length(move_line))
cnt<-0
for(i in move_line){
  cnt<-cnt+1
  predicted_y <- coefs[2]*x + coefs[1]+i
  error_y <- (predicted_y-d$mpg)^2
  total_error[cnt]<-sum(error_y)
}
df<-data.frame(move_line,total_error)
ggplot(df,aes(x=move_line,y=total_error))+
  geom_point()+
  theme_classic()+
  ylab("sum of squared deviations")+
  xlab("change to y-intercept")
```

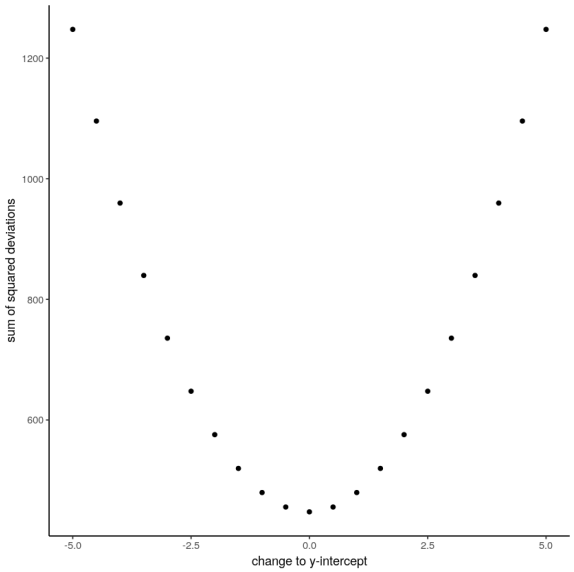


Figure 12.5.4: A plot of the sum of the squared deviations for different lines moving up and down, through the best fit line. The best fit line occurs at the position that minimizes the sum of the sqaured deviations.

OK, so we haven't talked about the y-intercept yet. But, what this graph shows us is how the total error behaves as we move the line up and down. The y-intercept here is the thing we change that makes our line move up and down. As you can see the dots go up when we move the line down from 0 to -5, and the dots go up when we move the line up from 0 to +5. The best line, that minimizes the error occurs right in the middle, when we don't move the blue regression line at all.

Lines

OK, fine you say. So, there is one magic line that will go through the middle of the scatter plot and minimize the sum of the squared deviations. How do I find this magic line? We'll show you. But, to be completely honest, you'll almost never do it the way we'll show you here. Instead, it's much easier to use software and make your computer do it for. You'll learn how to that in the labs.

Before we show you how to find the regression line, it's worth refreshing your memory about how lines work, especially in 2 dimensions. Remember this?

$y = ax + b$, or also $y = mx + b$ (sometimes a or m is used for the slope)

This is the formula for a line. Another way of writing it is:

$$y = slope * x + \text{y-intercept}$$

The slope is the slant of the line, and the y-intercept is where the line crosses the y-axis. Let's look at some lines:

```
library(ggplot2)
ggplot()+
  geom_abline(slope=1,intercept=5,color="blue")+
  geom_abline(slope=-1, intercept=15, color="red")+
  lims(x = c(1,20), y = c(0,20))+
  theme_classic()
```

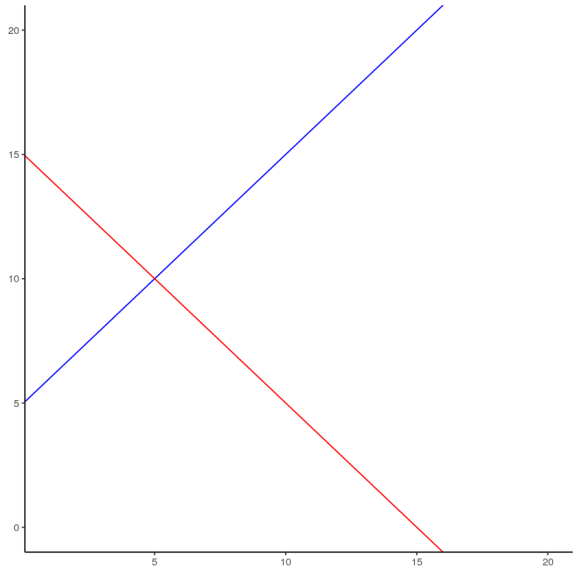


Figure 12.5.5: Two different lines with different y-intercepts (where the line crosses the y-axis), and different slopes. A positive slope makes the line go up from left to right. A negative slope makes the line go down from left to right.

So there is two lines. The formula for the blue line is $y = 1 * x + 5$. Let's talk about that. When x = 0, where is the blue line on the y-axis? It's at five. That happens because 1 times 0 is 0, and then we just have the five left over. How about when x = 5? In that case y =10. You just need the plug in the numbers to the formula, like this:

$$y = 1 * x + 5$$

$$y = 1 * 5 + 5 = 5 + 5 = 10$$

The point of the formula is to tell you where y will be, for any number of x. The slope of the line tells you whether the line is going to go up or down, as you move from the left to the right. The blue line has a positive slope of one, so it goes up as x goes up. How much does it go up? It goes up by one for everyone one of x! If we made the slope a 2, it would be much steeper, and go up faster. The red line has a negative slope, so it slants down. This means y goes down, as x goes up. When there is no slant, and we want to make a perfectly flat line, we set the slope to 0. This means that y doesn't go anywhere as x gets bigger and smaller.

That's lines.

Computing the best fit line

If you have a scatter plot showing the locations of scores from two variables, the real question is how can you find the slope and the y-intercept for the best fit line? What are you going to do? Draw millions of lines, add up the residuals, and then see which one was best? That would take forever. Fortunately, there are computers, and when you don't have one around, there's also some handy formulas.

> **Note**
>
> It's worth pointing out just how much computers have changed everything. Before computers everyone had to do these calculations by hand, such a chore! Aside from the deeper mathematical ideas in the formulas, many of them were made for convenience, to speed up hand calculations, because there were no computers. Now that we have computers, the hand calculations are often just an exercise in algebra. Perhaps they build character. You decide.

We'll show you the formulas. And, work through one example by hand. It's the worst, we know. By the way, you should feel sorry for me as I do this entire thing by hand for you.

Here are two formulas we can use to calculate the slope and the intercept, straight from the data. We won't go into why these formulas do what they do. These ones are for "easy" calculation.

$$intercept = b = \frac{\sum y \sum x^2 - \sum x \sum xy}{n\sum x^2 - (\sum x)^2}$$

$$slope = m = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

In these formulas, the x and the y refer to the individual scores. Here's a table showing you how everything fits together.

```
suppressPackageStartupMessages(library(dplyr))
scores<-c(1,2,3,4,5,6,7)
x<-c(1,4,3,6,5,7,8)
y<-c(2,5,1,8,6,8,9)
x_squared<-x^2
y_squared<-y^2
xy<-x*y
all_df<-data.frame(scores,x,y,x_squared,y_squared,xy)
all_df <- all_df %>%
  rbind(c("Sums",colSums(all_df[1:7,2:6])))
slope=((sum(y)*sum(x_squared))-(sum(x)*sum(xy)))/((7*sum(x_squared))-sum(x)^2)
intercept=(7*sum(xy)-sum(x)*sum(y))/(7*sum(x_squared)-sum(x)^2)
knitr::kable(all_df)
```

scores	x	y	x_squared	y_squared	xy
1	1	2	1	4	2
2	4	5	16	25	20
3	3	1	9	1	3
4	6	8	36	64	48
5	5	6	25	36	30
6	7	8	49	64	56
7	8	9	64	81	72
Sums	34	39	200	275	231

We see 7 sets of scores for the x and y variable. We calculated x^2 by squaring each value of x, and putting it in a column. We calculated y^2 by squaring each value of y, and putting it in a column. Then we calculated xy, by multiplying each x score with each y score, and put that in a column. Then we added all the columns up, and put the sums at the bottom. These are all the number we need for the formulas to find the best fit line. Here's what the formulas look like when we put numbers in them:

$$intercept = b = \frac{\sum y \sum x^2 - \sum x \sum xy}{n\sum x^2 - (\sum x)^2} = \frac{39 * 200 - 34 * 231}{7 * 200 - 34^2} = -.221$$

$$slope = m = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{7 * 231 - 34 * 39}{7 * 275 - 34^2} = 1.19$$

Great, now we can check our work, let's plot the scores in a scatter plot and draw a line through it with slope = 1.19, and a y-intercept of -.221. It should go through the middle of the dots.

```
library(ggplot2)
x<-c(1,4,3,6,5,7,8)
y<-c(2,5,1,8,6,8,9)
plot_df<-data.frame(x,y)
ggplot(plot_df,aes(x=x,y=y))+
  geom_point()+
  geom_smooth(method="lm",se=FALSE, formula=y ~ x)+
  theme_classic()
```

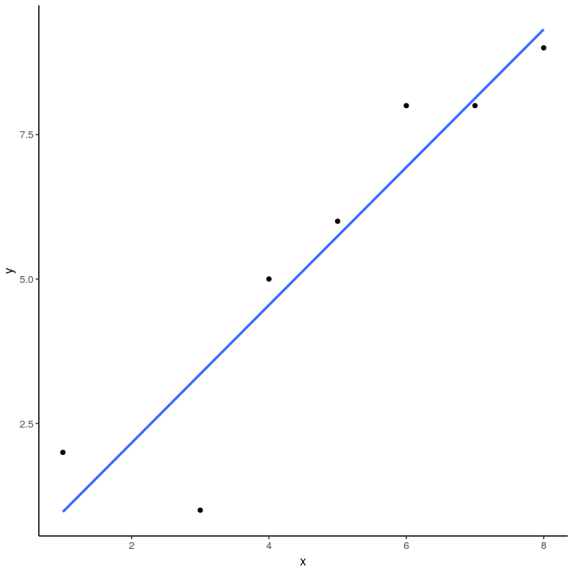


Figure 12.5.6: An example regression line with confidence bands going through a few data points in a scatterplot.

```
x<-c(1,4,3,6,5,7,8)
y<-c(2,5,1,8,6,8,9)
plot_df<-data.frame(x,y)
coef(lm(y~x,plot_df))
```

(Intercept): -0.221311475409834 **x:** 1.19262295081967

This page titled 12.5: Regression — A mini intro is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

12.6: Interpreting Correlations

What does the presence or the absence of a correlation between two measures mean? How should correlations be interpreted? What kind of inferences can be drawn from correlations? These are all very good questions. A first piece of advice is to use caution when interpreting correlations. Here's why.

Correlation does not equal causation

Perhaps you have heard that correlation does not equal causation. Why not? There are lots of reasons why not. However, before listing some of the reasons let's start with a case where we would expect a causal connection between two measurements. Consider, buying a snake plant for your home. Snake plants are supposed to be easy to take care of because you can mostly ignore them.

Like most plants, snake plants need some water to stay alive. However, they also need just the right amount of water. Imagine an experiment where 1000 snake plants were grown in a house. Each snake plant is given a different amount of water per day, from zero teaspoons of water per day to 1000 teaspoons of water per day. We will assume that water is part of the causal process that allows snake plants to grow. The amount of water given to each snake plant per day can also be one of our measures. Imagine further that every week the experimenter measures snake plant growth, which will be the second measurement. Now, can you imagine for yourself what a scatter plot of weekly snake plant growth by tablespoons of water would look like?

Even when there is causation, there might not be obvious correlation

The first plant given no water at all would have a very hard time and eventually die. It should have the least amount of weekly growth. How about the plants given only a few teaspoons of water per day. This could be just enough water to keep the plants alive, so they will grow a little bit but not a lot. If you are imagining a scatter plot, with each dot being a snake plant, then you should imagine some dots starting in the bottom left hand corner (no water & no plant growth), moving up and to the right (a bit of water, and a bit of growth). As we look at snake plants getting more and more water, we should see more and more plant growth, right? "Sure, but only up to a point". Correct, there should be a trend for a positive correlation with increasing plant growth as amount of water per day increases. But, what happens when you give snake plants too much water? From personal experience, they die. So, at some point, the dots in the scatter plot will start moving back down again. Snake plants that get way too much water will not grow very well.

The imaginary scatter plot you should be envisioning could have an upside U shape. Going from left to right, the dot's go up, they reach a maximum, then they go down again reaching a minimum. Computing Pearson's r for data like this can give you r values close to zero. The scatter plot could look something like this:

```
library(ggplot2)
water<-seq(0,999,1)
growth<-c(seq(0,10,(10/499)),seq(10,0,-(10/499)))
noise<-runif(1000,-2,2)
growth<-growth+noise
snake_df<-data.frame(growth,water)
ggplot(snake_df, aes(x=water,y=growth))+
  geom_point()+
  theme_classic()+
  xlab("Water (teaspoons)")+
  ggtitle("Imaginary snake plant growth \n as a function of water")
```

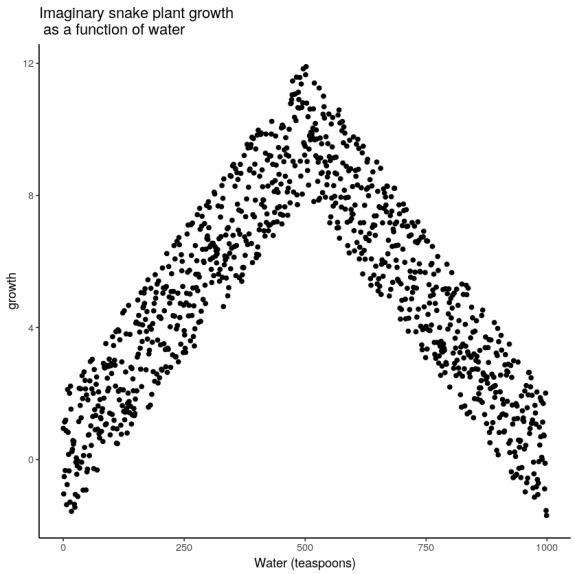


Figure 12.6.1: Illustration of a possible relationship between amount of water and snake plant growth. Growth goes up with water, but eventually goes back down as too much water makes snake plants die.

Granted this looks more like an inverted V, than an inverted U, but you get the picture right? There is clearly a relationship between watering and snake plant growth. But, the correlation isn't in one direction. As a result, when we compute the correlation in terms of Pearson's r, we get a value suggesting no relationship.

```
water<-seq(0,999,1)
growth<-c(seq(0,10,(10/499)),seq(10,0,-(10/499)))
noise<-runif(1000,-2,2)
growth<-growth+noise
cor(growth,water)

-0.00051489425461363
```

What this really means is there is no linear relationship that can be described by a single straight line. When we need lines or curves going in more than one direction, we have a nonlinear relationship.

This example illustrates some conundrums in interpreting correlations. We already know that water is needed for plants to grow, so we are rightly expecting there to be a relationship between our measure of amount of water and plant growth. If we look at the first half of the data we see a positive correlation, if we look at the last half of the data we see a negative correlation, and if we look at all of the data we see no correlation. Yikes. So, even when there is a causal connection between two measures, we won't necessarily obtain clear evidence of the connection just by computing a correlation coefficient.

> Pro Tip: This is one reason why plotting your data is so important. If you see an upside U shape pattern, then a correlation analysis is probably not the best analysis for your data.

Confounding variable, or Third variable problem

Anybody can correlate any two things that can be quantified and measured. For example, we could find a hundred people, ask them all sorts of questions like:

1. how happy are you
2. how old are you
3. how tall are you
4. how much money do you make per year
5. how long are your eyelashes
6. how many books have you read in your life
7. how loud is your inner voice

Let's say we found a positive correlation between yearly salary and happiness. Note, we could have just as easily computed the same correlation between happiness and yearly salary. If we found a correlation, would you be willing to infer that yearly salary causes happiness? Perhaps it does play a small part. But, something like happiness probably has a lot of contributing causes. Money could directly cause some people to be happy. But, more likely, money buys people access to all sorts of things, and some of those things might contribute happiness. These "other" things are called **third** variables. For example, perhaps people living in nicer places in more expensive houses are more happy than people in worse places in cheaper houses. In this scenario, money isn't causing happiness, it's the places and houses that money buys. But, even is this were true, people can still be more or less happy in lots of different situations.

The lesson here is that a correlation can occur between two measures because of a third variable that is not directly measured. So, just because we find a correlation, does not mean we can conclude anything about a causal connection between two measurements.

Correlation and Random chance

Another very important aspect of correlations is the fact that they can be produced by random chance. This means that you can find a positive or negative correlation between two measures, even when they have absolutely nothing to do with one another. You might have hoped to find zero correlation when two measures are totally unrelated to each other. Although this certainly happens, unrelated measures can accidentally produce **spurious** correlations, just by chance alone.

Let's demonstrate how correlations can occur by chance when there is no causal connection between two measures. Imagine two participants. One is at the North pole with a lottery machine full of balls with numbers from 1 to 10. The other is at the south pole with a different lottery machine full of balls with numbers from 1 to 10. There are an endless supply of balls in the machine, so every number could be picked for any ball. Each participant randomly chooses 10 balls, then records the number on the ball. In this situation we will assume that there is no possible way that balls chosen by the first participant could causally influence the balls chosen by the second participant. They are on the other side of the world. We should assume that the balls will be chosen by chance alone.

Here is what the numbers on each ball could look like for each participant:

```
Ball<-1:10
North_pole<-round(round(runif(10,1,10)))
South_pole<-round(round(runif(10,1,10)))
the_df_balls<-data.frame(Ball,North_pole,South_pole)
#the_df_balls <- the_df_balls %>%
#  rbind(c("Sums",colSums(the_df_balls[1:10,2:3]))) %>%
#  rbind(c("Means",colMeans(the_df_balls[1:10,2:3])))
knitr::kable(the_df_balls)
```

Ball	North_pole	South_pole
1	3	1
2	7	7
3	8	8
4	6	9
5	4	6
6	10	10
7	2	2
8	5	8
9	2	5
10	2	3

In this one case, if we computed Pearson's r, we would find that $r =$

```
North_pole<-round(round(runif(10,1,10)))
South_pole<-round(round(runif(10,1,10)))
cor(North_pole,South_pole)
```

```
0.0803444730711034
```

But, we already know that this value does not tell us anything about the relationship between the balls chosen in the north and south pole. We know that relationship should be completely random, because that is how we set up the game.

The better question here is to ask what can random chance do? For example, if we ran our game over and over again thousands of times, each time choosing new balls, and each time computing the correlation, what would we find? First, we will find fluctuation. The r value will sometimes be positive, sometimes be negative, sometimes be big and sometimes be small. Second, we will see what the fluctuation looks like. This will give us a window into the kinds of correlations that chance alone can produce. Let's see what happens.

Monte-carlo simulation of random correlations

It is possible to use a computer to simulate our game as many times as we want. This process is often termed **monte-carlo simulation**.

Below is a script written for the programming language R. We won't go into the details of the code here. However, let's briefly explain what is going on. Notice, the part that says `for(sim in 1:1000)`. This creates a loop that repeats our game 1000 times. Inside the loop there are variables named `North_pole` and `South_pole`. During each simulation, we sample 10 random numbers (between 1 to 10) into each variable. These random numbers stand for the numbers that would have been on the balls from the lottery machine. Once we have 10 random numbers for each, we then compute the correlation using `cor(North_pole,South_pole)`. Then, we save the correlation value and move on to the next simulation. At the end, we will have 1000 individual Pearson r values.

```
library(ggplot2)
simulated_correlations <- length(0)
for(sim in 1:1000){
  North_pole <- runif(10,1,10)
  South_pole <- runif(10,1,10)
  simulated_correlations[sim] <- cor(North_pole,South_pole)
}
sim_df <- data.frame(sims=1:1000,simulated_correlations)
ggplot(sim_df, aes(x = sims, y = simulated_correlations))+
  geom_point()+
  theme_classic()+
  geom_hline(yintercept = -1)+
  geom_hline(yintercept = 1)+
  ggtitle("Simulation of 1000 r values")
```

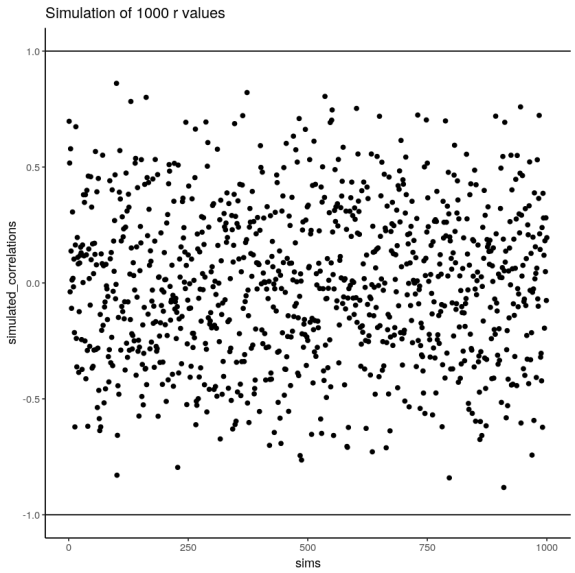


Figure 12.6.2: Another figure showing a range of r-values that can be obtained by chance.

Let's take a look at all of the 1000 Pearson r values. Does the figure below look familiar to you? It should, we have already conducted a similar kind of simulation before. Each dot in the scatter plot shows the Pearson r for each simulation from 1 to 1000. As you can see the dots are all over of the place, in between the range -1 to 1. The important lesson here is that random chance produced all of these correlations. This means we can find "correlations" in the data that are completely meaningless, and do not reflect any causal relationship between one measure and another.

Let's illustrate the idea of finding "random" correlations one more time, with a little movie. This time, we will show you a scatter plot of the random values sampled for the balls chosen from the North and South pole. If there is no relationship we should see dots going everywhere. If there happens to be a positive relationship (purely by chance), we should see the dots going from the bottom left to the top right. If there happens to be a negative relationship (purely by chance), we should see the dots going from the top left down to the bottom right.

On more thing to prepare you for the movie. There are three scatter plots below, showing negative, positive, and zero correlations between two variables. You've already seen this graph before. We are just reminding you that the blue lines are helpful for seeing the correlation.Negative correlations occur when a line goes down from the top left to bottom right. Positive correlations occur when a line goes up from the bottom left to the top right. Zero correlations occur when the line is flat (doesn't go up or down).

```
library(ggplot2)
subject_x<-1:100
chocolate_x<-round(1:100*runif(100,.5,1))
happiness_x<-round(1:100*runif(100,.5,1))
df_positive<-data.frame(subject_x,chocolate_x,happiness_x)
subject_x<-1:100
chocolate_x<-round(1:100*runif(100,.5,1))
happiness_x<-round(100:1*runif(100,.5,1))
df_negative<-data.frame(subject_x,chocolate_x,happiness_x)
subject_x<-1:100
chocolate_x<-round(runif(100,0,100))
happiness_x<-round(runif(100,0,100))
df_random<-data.frame(subject_x,chocolate_x,happiness_x)
all_data<-rbind(df_positive,df_negative,df_random)
all_data<-cbind(all_data,correlation=rep(c("positive","negative","random"),each=100))
ggplot(all_data,aes(x=chocolate_x,y=happiness_x))+
  geom_point()+
  geom_smooth(method=lm,se=FALSE, formula=y ~ x)+
  theme_classic()+
  facet_wrap(~correlation)+
  xlab("chocolate supply")+
  ylab("happiness")
```

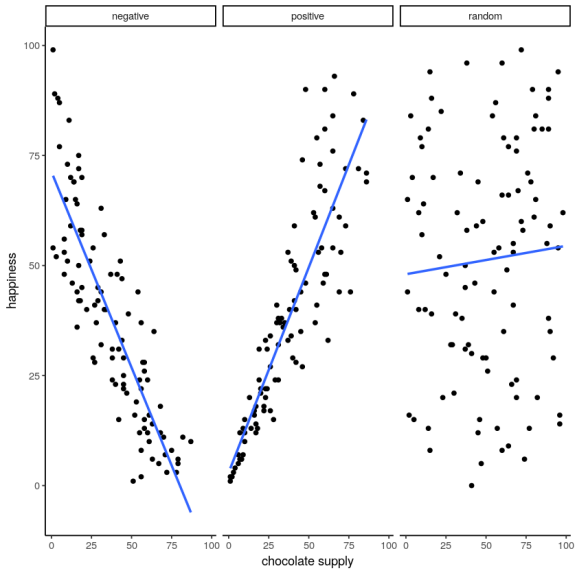


Figure 12.6.3: A reminder of what positive, negative, and zero correlation looks like.

OK, now we are ready for the movie. You are looking at the process of sampling two sets of numbers randomly, one for the X variable, and one for the Y variable. Each time we sample 10 numbers for each, plot them, then draw a line through them. Remember, these numbers are all completely random, so we should expect, on average that there should be no correlation between the numbers. However, this is not what happens. You can the line going all over the place. Sometimes we find a negative correlation (line goes down), sometimes we see a positive correlation (line goes up), and sometimes it looks like zero correlation (line is more flat).

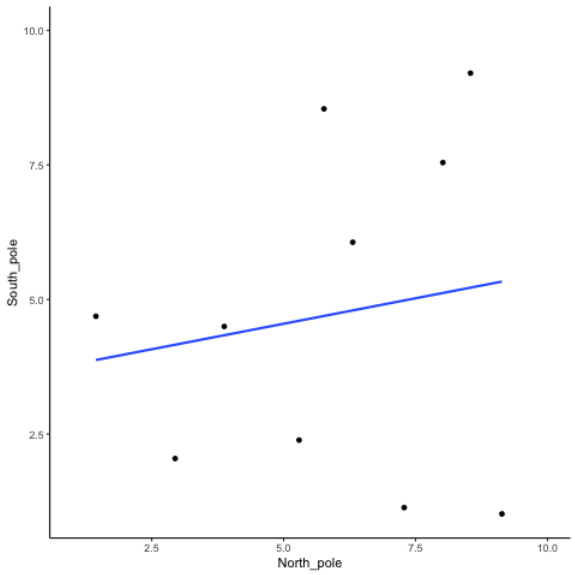


Figure 12.6.4: Completely random data points drawn from a uniform distribution with a small sampl-size of 10. The blue line twirls around sometimes showing large correlations that are produced by chance.

You might be thinking this is kind of disturbing. If we know that there should be no correlation between two random variables, how come we are finding correlations? This is a big problem right? I mean, if someone showed me a correlation between two things, and then claimed one thing was related to another, how could know I if it was true. After all, it could be chance! Chance can do that too.

Fortunately, all is not lost. We can look at our simulated data in another way, using a histogram. Remember, just before the movie, we simulated 1000 different correlations using random numbers. By, putting all of those r values into a histogram, we can get a better sense of how chance behaves. We can see what kind of correlations chance is likely or unlikely to produce. Here is a

histogram of the simulated r values.

```
simulated_correlations <- length(0)
for(sim in 1:1000){
  North_pole <- runif(10,1,10)
  South_pole <- runif(10,1,10)
  simulated_correlations[sim] <- cor(North_pole,South_pole)
}
hist(simulated_correlations,breaks=seq(-1,1,.1))
```

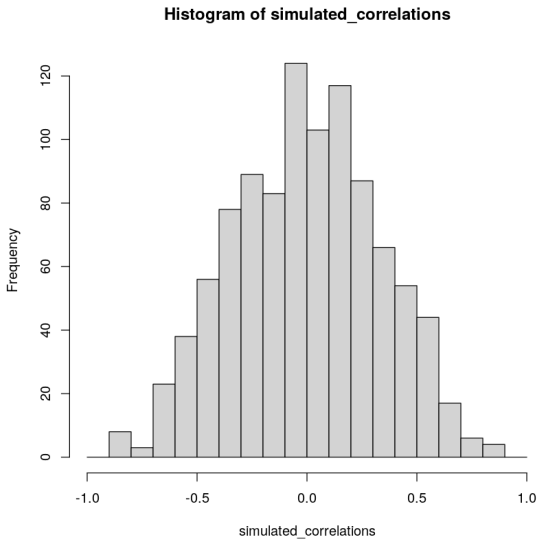


Figure 12.6.5: A histogram showing the frequency distribution of r-values for completely random values between an X and Y variable (sample-size=10). A rull range of r-values can be obtained by chance alone. Larger r-values are less common than smaller r-values.

Notice that this histogram is not flat. Most of the simulated r values are close to zero. Notice, also that the bars get smaller as you move away from zero in the positive or negative direction. The general take home here is that chance can produce a wide range of correlations. However, not all correlations happen very often. For example, the bars for -1 and 1 are very small. Chance does not produce nearly perfect correlations very often. The bars around -.5 and .5 are smaller than the bars around zero, as medium correlations do not occur as often as small correlations by chance alone.

You can think of this histogram as the window of chance. It shows what chance often does, and what it often does not do. If you found a correlation under these very same circumstances (e.g., measured the correlation between two sets of 10 random numbers), then you could consult this window. What should you ask the window? How about, could my observed correlation (the one that you found in your data) have come from this window. Let's say you found a correlation of $r = .1$. Could a .1 have come from the histogram? Well, look at the histogram around where the .1 mark on the x-axis is. Is there a big bar there? If so, this means that chance produces this value fairly often. You might be comfortable with the inference: Yes, this .1 could have been produced by chance, because it is well inside the window of chance. How about $r = .5$? The bar is much smaller here, you might think, "well, I can see that chance does produce .5 some times, so chance could have produced my .5. Did it? Maybe, maybe not, not sure". Here, your confidence in a strong inference about the role of chance might start getting a bit shakier.

How about an $r = .95$?. You might see that the bar for .95 is very very small, perhaps too small to see. What does this tell you? It tells you that chance does not produce .95 very often, hardly if at all, pretty much never. So, if you found a .95 in your data, what would you infer? Perhaps you would be comfortable inferring that chance did not produce your .95, after .95 is mostly outside the window of chance.

Increasing sample-size decreases opportunity for spurious correlation

Before moving on, let's do one more thing with correlations. In our pretend lottery game, each participant only sampled 10 balls each. We found that this could lead to a range of correlations between the numbers randomly drawn from either sides of the pole. Indeed, we even found some correlations that were medium to large in size. If you were a researcher who found such correlations,

you might be tempted to believe there was a relationship between your measurements. However, we know in our little game, that those correlations would be spurious, just a product of random sampling.

The good news is that, as a researcher, you get to make the rules of the game. You get to determine how chance can play. This is all a little bit metaphorical, so let's make it concrete.

We will see what happens in four different scenarios. First, we will repeat what we already did. Each participant will draw 10 balls, then we compute the correlation, and do this over 1000 times and look at a histogram. Second, we will change the game so each participant draws 50 balls each, and then repeat our simulation. Third, and fourth, we will change the game so each participant draws 100 balls each, and then 1000 balls each, and repeat etc.

The graph below shows four different histograms of the Pearson r values in each of the different scenarios. Each scenario involves a different sample-size, from, 10, 50, 100 to 1000.

```
library(ggplot2)
all_df<-data.frame()
for(s_size in  c(10,50,100,1000)){
  simulated_correlations <- length(0)
  for(sim in 1:1000){
    North_pole <- runif(s_size,1,10)
    South_pole <- runif(s_size,1,10)
    simulated_correlations[sim] <- cor(North_pole,South_pole)
  }
sim_df <- data.frame(sample_size=rep(s_size,1000),sims=1:1000,simulated_correlations)
all_df<-rbind(all_df,sim_df)
}
ggplot(all_df,aes(x=simulated_correlations))+
  geom_histogram(bins=30)+
  facet_wrap(~sample_size)+
  theme_classic()
```

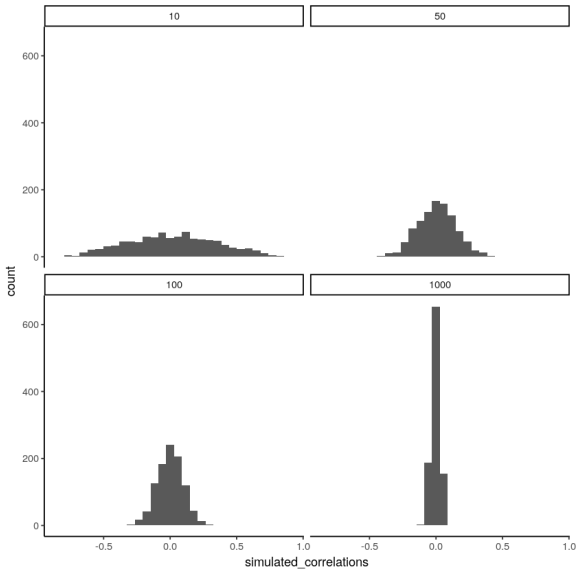


Figure 12.6.6: Four histograms showing the frequency distributions of r-values between completely random X and Y variables as a function of sample-size. The width of the distributions shrink as sample-size increases. Smaller sample-sizes are more likely to produce a wider range of r-values by chance. Larger sample-sizes always produce a narrow range of small r-values.

By inspecting the four histograms you should notice a clear pattern. The width or range of each histogram shrinks as the sample-size increases. What is going on here? Well, we already know that we can think of these histograms as windows of chance. They tell us which r values occur fairly often, which do not. When our sample-size is 10, lots of different r values happen. That histogram is very flat and spread out. However, as the sample-size increases, we see that the window of chance gets pulled in. For example, by the time we get to 1000 balls each, almost all of the Pearson r values are very close to 0.

One take home here, is that increasing sample-size narrows the window of chance. So, for example, if you ran a study involving 1000 samples of two measures, and you found a correlation of .5, then you can clearly see in the bottom right histogram that .5 does not occur very often by chance alone. In fact, there is no bar, because it didn't happen even once in the simulation. As a result, when you have a large sample size like n = 1000, you might be more confident that your observed correlation (say of .5) was not a spurious correlation. If chance is not producing your result, then something else is.

Finally, notice how your confidence about whether or not chance is mucking about with your results depends on your sample size. If you only obtained 10 samples per measurement, and found $r = .5$, you should not be as confident that your correlation reflects a real relationship. Instead, you can see that r's of .5 happen fairly often by chance alone.

> Pro tip: when you run an experiment you get to decide how many samples you will collect, which means you can choose to narrow the window of chance. Then, if you find a relationship in the data you can be more confident that your finding is real, and not just something that happened by chance.

Some more movies

Let's ingrain these idea with some more movies. When our sample-size is small (N is small), sampling error can cause all sort "patterns" in the data. This makes it possible, and indeed common, for "correlations" to occur between two sets of numbers. When we increase the sample-size, sampling error is reduced, making it less possible for "correlations" to occur just by chance alone. When N is large, chance has less of an opportunity to operate.

Watching how correlation behaves when there is no correlation

Below we randomly sample numbers for two variables, plot them, and show the correlation using a line. There are four panels, each showing the number of observations in the samples, from 10, 50, 100, to 1000 in each sample.

Remember, because we are randomly sampling numbers, there should be no relationship between the X and Y variables. But, as we have been discussing, because of chance, we can sometimes observe a correlation (due to chance). The important thing to watch is how the line behaves across the four panels. The line twirls around in all directions when the sample size is 10. It is also moves around quite a bit when the sample size is 50 or 100. It still moves a bit when the sample size is 1000, but much less. In all cases we expect that the line should be flat, but every time we take new samples, sometimes the line shows us pseudo patterns.

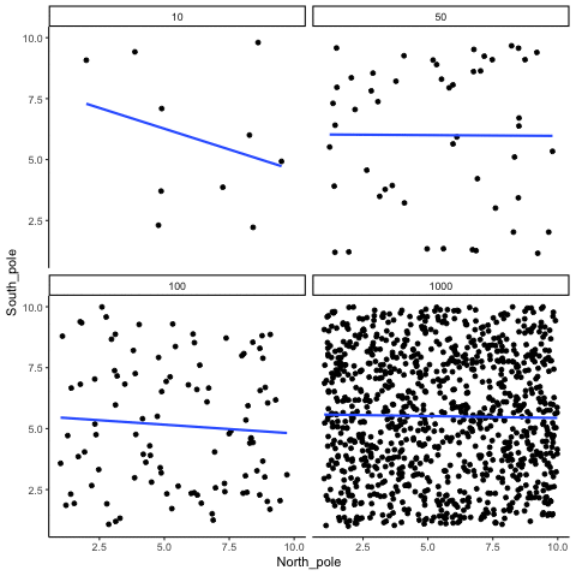


Figure 12.6.7: Animation of how correlation behaves for completely random X and Y variables as a function of sample size. The best fit line is not very stable for small sample-sizes, but becomes more reliably flat as sample-size increases.

Which line should you trust? Well, hopefully you can see that the line for 1000 samples is the most stable. It tends to be very flat every time, and it does not depend so much on the particular sample. The line with 10 observations per sample goes all

over the place. The take home here, is that if someone told you that they found a correlation, you should want to know how many observations they hand in their sample. If they only had 10 observations, how could you trust the claim that there was a correlation? You can't!!! Not now that you know samples that are that small can do all sorts of things by chance alone. If instead, you found out the sample was very large, then you might trust that finding a little bit more. For example, in the above movie you can see that when there are 1000 samples, we never see a strong or weak correlation; the line is always flat. This is because chance almost never produces strong correlations when the sample size is very large.

In the above example, we sampled numbers random numbers from a uniform distribution. Many examples of real-world data will come from a normal or approximately normal distribution. We can repeat the above, but sample random numbers from the same normal distribution. There will still be zero actual correlation between the X and Y variables, because everything is sampled randomly. But, we still see the same behavior as above. The computed correlation for small sample-sizes fluctuate wildly, and large sample sizes do not.

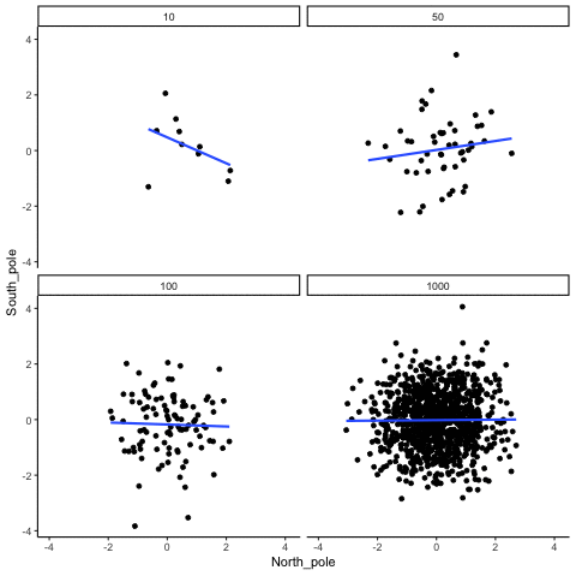


Figure 12.6.8: Animation of correlation for random values sampled from a normal distribution, rather than a uniform distribution.

OK, so what do things look like when there actually is a correlation between variables?

Watching correlations behave when there really is a correlation

Sometimes there really are correlations between two variables that are not caused by chance. Below, we get to watch a movie of four scatter plots. Each shows the correlation between two variables. Again, we change the sample-size in steps of 10, 50 100, and 1000. The data have been programmed to contain a real positive correlation. So, we should expect that the line will be going up from the bottom left to the top right. However, there is still variability in the data. So this time, sampling error due to chance will fuzz the correlation. We know it is there, but sometimes chance will cause the correlation to be eliminated.

Notice that in the top left panel (sample-size 10), the line is twirling around much more than the other panels. Every new set of samples produces different correlations. Sometimes, the line even goes flat or downward. However, as we increase sample-size, we can see that the line doesn't change very much, it is always going up showing a positive correlation.

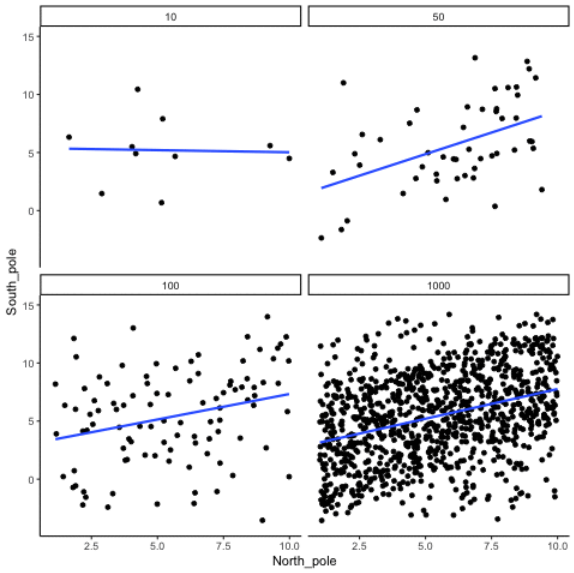


Figure 12.6.9: How correlation behaves as a function of sample-size when there is a true correlation between X and Y variables.

The main takeaway here is that even when there is a positive correlation between two things, you might not be able to see it if your sample size is small. For example, you might get unlucky with the one sample that you measured. Your sample could show a negative correlation, even when the actual correlation is positive! Unfortunately, in the real world we usually only have the sample that we collected, so we always have to wonder if we got lucky or unlucky. Fortunately, if you want to remove luck, all you need to do is collect larger samples. Then you will be much more likely to observe the real pattern, rather the pattern that can be introduced by chance.

This page titled 12.6: Interpreting Correlations is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

12.7: Summary

In this section we have talked about correlation, and started to build some intuitions about **inferential statistics**, which is the major topic of the remaining chapters. For now, the main ideas are:

1. We can measure relationships in data using things like correlation
2. The correlations we measure can be produced by numerous things, so they are hard to to interpret
3. Correlations can be produced by chance, so have the potential to be completely meaningless.
4. However, we can create a model of exactly what chance can do. The model tells us whether chance is more or less likely to produce correlations of different sizes
5. We can use the chance model to help us make decisions about our own data. We can compare the correlation we found in our data to the model, then ask whether or not chance could have or was likely to have produced our results.

This page titled 12.7: Summary is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

12.8: References

Salsburg, David. 2001. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Macmillan.

This page titled 12.8: References is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

- 1.5: Legal and Political Systems of the World is licensed CC BY-NC-SA 4.0.

13: PROBABILITY, SAMPLING, AND ESTIMATION

13.1: Prelude to Probability, Sampling, and Estimation

13.2: How are Probability and Statistics Different?

13.3: What Does Probability Mean?

13.4: Basic Probability Theory

13.5: The Binomial Distribution

13.6: The normal distribution

13.7: Other useful distributions

13.8: Summary of Probability

13.9: Samples, populations and sampling

13.10: The Law of Large Numbers

13.11: Sampling distributions and the central limit theorem

13.12: The Central Limit Theorem

13.13: z-scores

13.14: Estimating population parameters

13.15: Estimating a confidence interval

13.16: Summary

13.17: Videos

13.18: References

This page titled 13: Probability, Sampling, and Estimation is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

13.1: Prelude to Probability, Sampling, and Estimation

> I have studied many languages-French, Spanish and a little Italian, but no one told me that Statistics was a foreign language.
>
> —Charmaine J. Forde

Note

Sections 4.1 & 4.9 - *Adapted text by Danielle Navarro.*

Section 4.10 - 4.11 & 4.13 - *Mix of Matthew Crump & Danielle Navarro.*

Section 4.12-4.13 - *Adapted text by Danielle Navarro.*

Up to this point in the book, we've discussed some of the key ideas in experimental design, and we've talked a little about how you can summarize a data set. To a lot of people, this is all there is to statistics: it's about calculating averages, collecting all the numbers, drawing pictures, and putting them all in a report somewhere. Kind of like stamp collecting, but with numbers. However, statistics covers much more than that. In fact, descriptive statistics is one of the smallest parts of statistics, and one of the least powerful. The bigger and more useful part of statistics is that it provides tools **that let you make inferences about data**.

Once you start thinking about statistics in these terms – that statistics is there to help us draw inferences from data – you start seeing examples of it everywhere. For instance, here's a tiny extract from a newspaper article in the Sydney Morning Herald (30 Oct 2010):

> "I have a tough job," the Premier said in response to a poll which found her government is now the most unpopular Labor administration in polling history, with a primary vote of just 23 per cent.

This kind of remark is entirely unremarkable in the papers or in everyday life, but let's have a think about what it entails. A polling company has conducted a survey, usually a pretty big one because they can afford it. I'm too lazy to track down the original survey, so let's just imagine that they called 1000 voters at random, and 230 (23%) of those claimed that they intended to vote for the party. For the 2010 Federal election, the Australian Electoral Commission reported 4,610,795 enrolled voters in New South Whales; so the opinions of the remaining 4,609,795 voters (about 99.98% of voters) remain unknown to us. Even assuming that no-one lied to the polling company the only thing we can say with 100% confidence is that the true primary vote is somewhere between 230/4610795 (about 0.005%) and 4610025/4610795 (about 99.83%). So, on what basis is it legitimate for the polling company, the newspaper, and the readership to conclude that the ALP primary vote is only about 23%?

The answer to the question is pretty obvious: if I call 1000 people at random, and 230 of them say they intend to vote for the ALP, then it seems very unlikely that these are the **only** 230 people out of the entire voting public who actually intend to do so. In other words, we assume that the data collected by the polling company is pretty representative of the population at large. But how representative? Would we be surprised to discover that the true ALP primary vote is actually 24%? 29%? 37%? At this point everyday intuition starts to break down a bit. No-one would be surprised by 24%, and everybody would be surprised by 37%, but it's a bit hard to say whether 29% is plausible. We need some more powerful tools than just looking at the numbers and guessing.

Inferential statistics provides the tools that we need to answer these sorts of questions, and since these kinds of questions lie at the heart of the scientific enterprise, they take up the lions share of every introductory course on statistics and research methods. However, our tools for making statistical inferences are 1) built on top of **probability theory**, and 2) require an understanding of how samples behave when you take them from distributions (defined by probability theory...). So, this chapter has two main parts. A brief introduction to probability theory, and an introduction to sampling from distributions.

This page titled 13.1: Prelude to Probability, Sampling, and Estimation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

- 10: Operations Management by Stephen Skripak et al. is licensed CC BY-NC-SA 4.0. Original source: https://otn.pressbooks.pub/fundamentalsofbusiness/.

13.2: How are Probability and Statistics Different?

Before we start talking about probability theory, it's helpful to spend a moment thinking about the relationship between probability and statistics. The two disciplines are closely related but they're not identical. Probability theory is "the doctrine of chances". It's a branch of mathematics that tells you how often different kinds of events will happen. For example, all of these questions are things you can answer using probability theory:

- What are the chances of a fair coin coming up heads 10 times in a row?
- If I roll two six sided dice, how likely is it that I'll roll two sixes?
- How likely is it that five cards drawn from a perfectly shuffled deck will all be hearts?
- What are the chances that I'll win the lottery?

Notice that all of these questions have something in common. In each case the "truth of the world" is known, and my question relates to the "what kind of events" will happen. In the first question I **know** that the coin is fair, so there's a 50% chance that any individual coin flip will come up heads. In the second question, I **know** that the chance of rolling a 6 on a single die is 1 in 6. In the third question I **know** that the deck is shuffled properly. And in the fourth question, I **know** that the lottery follows specific rules. You get the idea. The critical point is that probabilistic questions start with a known *model* of the world, and we use that model to do some calculations.

The underlying model can be quite simple. For instance, in the coin flipping example, we can write down the model like this: $P(heads) = 0.5$ which you can read as "the probability of heads is 0.5".

As we'll see later, in the same way that percentages are numbers that range from 0% to 100%, probabilities are just numbers that range from 0 to 1. When using this probability model to answer the first question, I don't actually know exactly what's going to happen. Maybe I'll get 10 heads, like the question says. But maybe I'll get three heads. That's the key thing: in probability theory, the **model** is known, but the **data** are not.

So that's probability. What about statistics? Statistical questions work the other way around. In statistics, we know the truth about the world. All we have is the data, and it is from the data that we want to **learn** the truth about the world. Statistical questions tend to look more like these:

- If my friend flips a coin 10 times and gets 10 heads, are they playing a trick on me?
- If five cards off the top of the deck are all hearts, how likely is it that the deck was shuffled?
- If the lottery commissioner's spouse wins the lottery, how likely is it that the lottery was rigged?

This time around, the only thing we have are data. What I **know** is that I saw my friend flip the coin 10 times and it came up heads every time. And what I want to *infer* is whether or not I should conclude that what I just saw was actually a fair coin being flipped 10 times in a row, or whether I should suspect that my friend is playing a trick on me. The data I have look like this:

```
H H H H H H H H H H
```

and what I'm trying to do is work out which "model of the world" I should put my trust in. If the coin is fair, then the model I should adopt is one that says that the probability of heads is 0.5; that is, $P(heads) = 0.5$. If the coin is not fair, then I should conclude that the probability of heads is **not** 0.5, which we would write as $P(heads) \neq 0.5$. In other words, the statistical inference problem is to figure out which of these probability models is right. Clearly, the statistical question isn't the same as the probability question, but they're deeply connected to one another. Because of this, a good introduction to statistical theory will start with a discussion of what probability is and how it works.

This page titled 13.2: How are Probability and Statistics Different? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

13.3: What Does Probability Mean?

Let's start with the first of these questions. What is "probability"? It might seem surprising to you, but while statisticians and mathematicians (mostly) agree on what the **rules** of probability are, there's much less of a consensus on what the word really **means**. It seems weird because we're all very comfortable using words like "chance", "likely", "possible" and "probable", and it doesn't seem like it should be a very difficult question to answer. If you had to explain "probability" to a five year old, you could do a pretty good job. But if you've ever had that experience in real life, you might walk away from the conversation feeling like you didn't quite get it right, and that (like many everyday concepts) it turns out that you don't **really** know what it's all about.

So I'll have a go at it. Let's suppose I want to bet on a soccer game between two teams of robots, **Arduino Arsenal** and **C Milan**. After thinking about it, I decide that there is an 80% probability that **Arduino Arsenal** winning. What do I mean by that? Here are three possibilities...

- They're robot teams, so I can make them play over and over again, and if I did that, **Arduino Arsenal** would win 8 out of every 10 games on average.
- For any given game, I would only agree that betting on this game is only "fair" if a $1 bet on **C Milan** gives a $5 payoff (i.e. I get my $1 back plus a $4 reward for being correct), as would a $4 bet on **Arduino Arsenal** (i.e., my $4 bet plus a $1 reward).
- My subjective "belief" or "confidence" in an **Arduino Arsenal** victory is four times as strong as my belief in a **C Milan** victory.

Each of these seems sensible. However they're not identical, and not every statistician would endorse all of them. The reason is that there are different statistical ideologies (yes, really!) and depending on which one you subscribe to, you might say that some of those statements are meaningless or irrelevant. In this section, I give a brief introduction the two main approaches that exist in the literature. These are by no means the only approaches, but they're the two big ones.

The Frequentist View

The first of the two major approaches to probability, and the more dominant one in statistics, is referred to as the *frequentist view*, and it defines probability as a *long-run frequency*. Suppose we were to try flipping a fair coin, over and over again. By definition, this is a coin that has $P(H) = 0.5$. What might we observe? One possibility is that the first 20 flips might look like this:

T,H,H,H,H,T,T,H,H,H,H,T,H,H,T,T,T,T,T,H

In this case 11 of these 20 coin flips (55%) came up heads. Now suppose that I'd been keeping a running tally of the number of heads (which I'll call N_H) that I've seen, across the first N flips, and calculate the proportion of heads N_H/N every time. Here's what I'd get (I did literally flip coins to produce this!):

number of flips	1	2	3	4	5	6	7	8	9	10
number of heads	0	1	2	3	4	4	4	5	6	7
proportion	.00	.50	.67	.75	.80	.67	.57	.63	.67	.70

number of flips	11	12	13	14	15	16	17	18	19	20
number of heads	8	8	9	10	10	10	10	10	10	11
proportion	.73	.67	.69	.71	.67	.63	.59	.56	.53	.55

Notice that at the start of the sequence, the **proportion** of heads fluctuates wildly, starting at .00 and rising as high as .80. Later on, one gets the impression that it dampens out a bit, with more and more of the values actually being pretty close to the "right" answer of .50. This is the frequentist definition of probability in a nutshell: flip a fair coin over and over again, and as N grows large (approaches infinity, denoted $N \rightarrow \infty$), the proportion of heads will converge to 50%. There are some subtle technicalities that the mathematicians care about, but qualitatively speaking, that's how the frequentists define probability. Unfortunately, I don't have an infinite number of coins, or the infinite patience required to flip a coin an infinite number of times. However, I do have a computer, and computers excel at mindless repetitive tasks. So I asked my computer to simulate flipping a coin 1000 times, and then drew a picture of what happens to the proportion N_H/N as N increases. Actually, I did it four times, just to make

sure it wasn't a fluke. The results are shown in Figure 13.3.1. As you can see, the **proportion of observed heads** eventually stops fluctuating, and settles down; when it does, the number at which it finally settles is the true probability of heads.

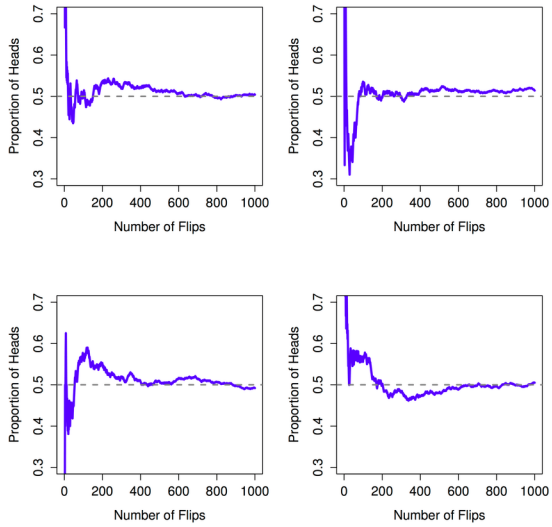


Figure 13.3.1: An illustration of how frequentist probability works. If you flip a fair coin over and over again, the proportion of heads that you've seen eventually settles down, and converges to the true probability of 0.5. Each panel shows four different simulated experiments: in each case, we pretend we flipped a coin 1000 times, and kept track of the proportion of flips that were heads as we went along. Although none of these sequences actually ended up with an exact value of .5, if we'd extended the experiment for an infinite number of coin flips they would have.

The frequentist definition of probability has some desirable characteristics. First, it is objective: the probability of an event is **necessarily** grounded in the world. The only way that probability statements can make sense is if they refer to (a sequence of) events that occur in the physical universe. Second, it is unambiguous: any two people watching the same sequence of events unfold, trying to calculate the probability of an event, must inevitably come up with the same answer.

However, it also has undesirable characteristics. Infinite sequences don't exist in the physical world. Suppose you picked up a coin from your pocket and started to flip it. Every time it lands, it impacts on the ground. Each impact wears the coin down a bit; eventually, the coin will be destroyed. So, one might ask whether it really makes sense to pretend that an "infinite" sequence of coin flips is even a meaningful concept, or an objective one. We can't say that an "infinite sequence" of events is a real thing in the physical universe, because the physical universe doesn't allow infinite anything.

More seriously, the frequentist definition has a narrow scope. There are lots of things out there that human beings are happy to assign probability to in everyday language, but cannot (even in theory) be mapped onto a hypothetical sequence of events. For instance, if a meteorologist comes on TV and says, "the probability of rain in Adelaide on 2 November 2048 is 60%" we humans are happy to accept this. But it's not clear how to define this in frequentist terms. There's only one city of Adelaide, and only 2 November 2048. There's no infinite sequence of events here, just a once-off thing. Frequentist probability genuinely **forbids** us from making probability statements about a single event. From the frequentist perspective, it will either rain tomorrow or it will not; there is no "probability" that attaches to a single non-repeatable event. Now, it should be said that there are some very clever tricks that frequentists can use to get around this. One possibility is that what the meteorologist means is something like this: "There is a category of days for which I predict a 60% chance of rain; if we look only across those days for which I make this prediction, then on 60% of those days it will actually rain". It's very weird and counterintuitive to think of it this way, but you do see frequentists do this sometimes.

The Bayesian View

The **Bayesian view** of probability is often called the subjectivist view, and it is a minority view among statisticians, but one that has been steadily gaining traction for the last several decades. There are many flavours of Bayesianism, making hard to say exactly what "the" Bayesian view is. The most common way of thinking about subjective probability is to define the probability of an event as the **degree of belief** that an intelligent and rational agent assigns to that truth of that event. From that perspective, probabilities don't exist in the world, but rather in the thoughts and assumptions of people and other intelligent beings. However, in order for this approach to work, we need some way of operationalising "degree of belief". One way that you can do this is

to formalise it in terms of "rational gambling", though there are many other ways. Suppose that I believe that there's a 60% probability of rain tomorrow. If someone offers me a bet: if it rains tomorrow, then I win $5, but if it doesn't rain then I lose $5. Clearly, from my perspective, this is a pretty good bet. On the other hand, if I think that the probability of rain is only 40%, then it's a bad bet to take. Thus, we can operationalise the notion of a "subjective probability" in terms of what bets I'm willing to accept.

What are the advantages and disadvantages to the Bayesian approach? The main advantage is that it allows you to assign probabilities to any event you want to. You don't need to be limited to those events that are repeatable. The main disadvantage (to many people) is that we can't be purely objective – specifying a probability requires us to specify an entity that has the relevant degree of belief. This entity might be a human, an alien, a robot, or even a statistician, but there has to be an intelligent agent out there that believes in things. To many people this is uncomfortable: it seems to make probability arbitrary. While the Bayesian approach does require that the agent in question be rational (i.e., obey the rules of probability), it does allow everyone to have their own beliefs; I can believe the coin is fair and you don't have to, even though we're both rational. The frequentist view doesn't allow any two observers to attribute different probabilities to the same event: when that happens, then at least one of them must be wrong. The Bayesian view does not prevent this from occurring. Two observers with different background knowledge can legitimately hold different beliefs about the same event. In short, where the frequentist view is sometimes considered to be too narrow (forbids lots of things that that we want to assign probabilities to), the Bayesian view is sometimes thought to be too broad (allows too many differences between observers).

What's the difference? And who is right?

Now that you've seen each of these two views independently, it's useful to make sure you can compare the two. Go back to the hypothetical robot soccer game at the start of the section. What do you think a frequentist and a Bayesian would say about these three statements? Which statement would a frequentist say is the correct definition of probability? Which one would a Bayesian do? Would some of these statements be meaningless to a frequentist or a Bayesian? If you've understood the two perspectives, you should have some sense of how to answer those questions.

Okay, assuming you understand the different, you might be wondering which of them is **right**? Honestly, I don't know that there is a right answer. As far as I can tell there's nothing mathematically incorrect about the way frequentists think about sequences of events, and there's nothing mathematically incorrect about the way that Bayesians define the beliefs of a rational agent. In fact, when you dig down into the details, Bayesians and frequentists actually agree about a lot of things. Many frequentist methods lead to decisions that Bayesians agree a rational agent would make. Many Bayesian methods have very good frequentist properties.

For the most part, I'm a pragmatist so I'll use any statistical method that I trust. As it turns out, that makes me prefer Bayesian methods, for reasons I'll explain towards the end of the book, but I'm not fundamentally opposed to frequentist methods. Not everyone is quite so relaxed. For instance, consider Sir Ronald Fisher, one of the towering figures of 20th century statistics and a vehement opponent to all things Bayesian, whose paper on the mathematical foundations of statistics referred to Bayesian probability as "an impenetrable jungle [that] arrests progress towards precision of statistical concepts" Fisher (1922, 311). Or the psychologist Paul Meehl, who suggests that relying on frequentist methods could turn you into "a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring" Meehl (1967, 114). The history of statistics, as you might gather, is not devoid of entertainment.

This page titled 13.3: What Does Probability Mean? is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

13.4: Basic Probability Theory

Ideological arguments between Bayesians and frequentists notwithstanding, it turns out that people mostly agree on the rules that probabilities should obey. There are lots of different ways of arriving at these rules. The most commonly used approach is based on the work of Andrey Kolmogorov, one of the great Soviet mathematicians of the 20th century. I won't go into a lot of detail, but I'll try to give you a bit of a sense of how it works. And in order to do so, I'm going to have to talk about my pants.

Introducing Probability Distributions

One of the disturbing truths about my life is that I only own 5 pairs of pants: three pairs of jeans, the bottom half of a suit, and a pair of tracksuit pants. Even sadder, I've given them names: I call them X_1, X_2, X_3, X_4 and X_5. I really do: that's why they call me Mister Imaginative. Now, on any given day, I pick out exactly one of pair of pants to wear. Not even I'm so stupid as to try to wear two pairs of pants, and thanks to years of training I never go outside without wearing pants anymore. If I were to describe this situation using the language of probability theory, I would refer to each pair of pants (i.e., each X) as an *elementary event*. The key characteristic of elementary events is that every time we make an observation (e.g., every time I put on a pair of pants), then the outcome will be one and only one of these events. Like I said, these days I always wear exactly one pair of pants, so my pants satisfy this constraint. Similarly, the set of all possible events is called a *sample space*. Granted, some people would call it a "wardrobe", but that's because they're refusing to think about my pants in probabilistic terms. Sad.

Okay, now that we have a sample space (a wardrobe), which is built from lots of possible elementary events (pants), what we want to do is assign a *probability* of one of these elementary events. For an event X, the probability of that event $P(X)$ is a number that lies between 0 and 1. The bigger the value of $P(X)$, the more likely the event is to occur. So, for example, if $P(X) = 0$, it means the event X is impossible (i.e., I never wear those pants). On the other hand, if $P(X) = 1$ it means that event X is certain to occur (i.e., I always wear those pants). For probability values in the middle, it means that I sometimes wear those pants. For instance, if $P(X) = 0.5$ it means that I wear those pants half of the time.

At this point, we're almost done. The last thing we need to recognise is that "something always happens". Every time I put on pants, I really do end up wearing pants (crazy, right?). What this somewhat trite statement means, in probabilistic terms, is that the probabilities of the elementary events need to add up to 1. This is known as the *law of total probability*, not that any of us really care. More importantly, if these requirements are satisfied, then what we have is a *probability distribution*. For example, this is an example of a probability distribution

Which pants?	Label	Probability
Blue jeans	X_1	$P(X_1) = .5$
Grey jeans	X_2	$P(X_2) = .3$
Black jeans	X_3	$P(X_3) = .1$
Black suit	X_4	$P(X_4) = 0$
Blue tracksuit	X_5	$P(X_5) = .1$

Each of the events has a probability that lies between 0 and 1, and if we add up the probability of all events, they sum to 1. Awesome. We can even draw a nice bar graph to visualise this distribution, as shown in Figure 13.4.1. And at this point, we've all achieved something. You've learned what a probability distribution is, and I've finally managed to find a way to create a graph that focuses entirely on my pants. Everyone wins!

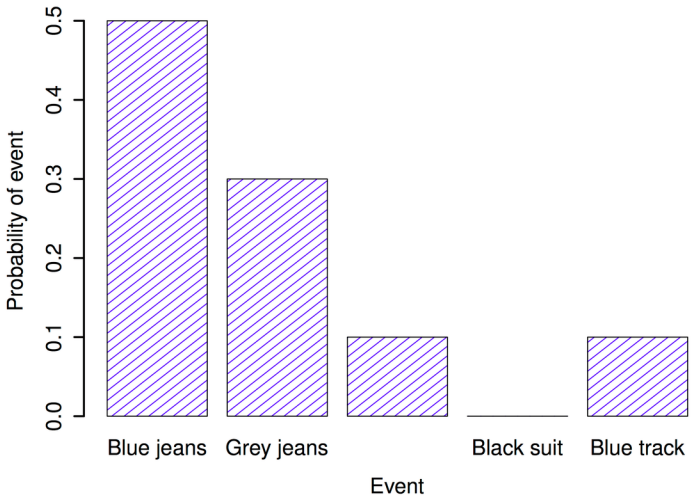


Figure 13.4.1: A visual depiction of the pants probability distribution. There are five elementary events, corresponding to the five pairs of pants that I own. Each event has some probability of occurring: this probability is a number between 0 to 1. The sum of these probabilities is 1.

The only other thing that I need to point out is that probability theory allows you to talk about *non elementary events* as well as elementary ones. The easiest way to illustrate the concept is with an example. In the pants example, it's perfectly legitimate to refer to the probability that I wear jeans. In this scenario, the "Dan wears jeans" event said to have happened as long as the elementary event that actually did occur is one of the appropriate ones; in this case "blue jeans", "black jeans" or "grey jeans". In mathematical terms, we defined the "jeans" event E to correspond to the set of elementary events (X_1, X_2, X_3). If any of these elementary events occurs, then E is also said to have occurred. Having decided to write down the definition of the E this way, it's pretty straightforward to state what the probability $P(E)$ is: we just add everything up. In this particular case

$$P(E) = P(X_1) + P(X_2) + P(X_3)$$

and, since the probabilities of blue, grey and black jeans respectively are .5, .3 and .1, the probability that I wear jeans is equal to .9.

At this point you might be thinking that this is all terribly obvious and simple and you'd be right. All we've really done is wrap some basic mathematics around a few common sense intuitions. However, from these simple beginnings it's possible to construct some extremely powerful mathematical tools. I'm definitely not going to go into the details in this book, but what I will do is list some of the other rules that probabilities satisfy. These rules can be derived from the simple assumptions that I've outlined above, but since we don't actually use these rules for anything in this book, I won't do so here.

Table 13.4.1: *Some basic rules that probabilities must satisfy. You don't really need to know these rules in order to understand the analyses that we'll talk about later in the book, but they are important if you want to understand probability theory a bit more deeply.*

English	Notation		Formula
not A	$P(\neg A)$	$=$	$1 - P(A)$
A or B	$P(A \cup B)$	$=$	$P(A) + P(B) - P(A \cap B)$
A and B	$P(A \cap B)$	$=$	$P(A\|B)P(B)$

Now that we have the ability to "define" non-elementary events in terms of elementary ones, we can actually use this to construct (or, if you want to be all mathematicallish, "derive") some of the other rules of probability. These rules are listed above, and while I'm pretty confident that very few of my readers actually care about how these rules are constructed, I'm going to show you anyway: even though it's boring and you'll probably never have a lot of use for these derivations, if you read through it once or twice and try to see how it works, you'll find that probability starts to feel a bit less mysterious, and with any luck a lot less daunting. So here goes. Firstly, in order to construct the rules I'm going to need a sample space X that consists of a bunch of elementary events x, and two non-elementary events, which I'll call A and B. Let's say:

$$\begin{aligned} X &= (x_1, x_2, x_3, x_4, x_5) \\ A &= (x_1, x_2, x_3) \\ B &= (x_3, x_4) \end{aligned}$$

To make this a bit more concrete, let's suppose that we're still talking about the pants distribution. If so, A corresponds to the event "jeans", and B corresponds to the event "black":

$$\begin{aligned} \text{"jeans"} &= (\text{"blue jeans"}, \text{"grey jeans"}, \text{"black jeans"}) \\ \text{"black"} &= (\text{"black jeans"}, \text{"black suit"}) \end{aligned}$$

So now let's start checking the rules that I've listed in the table.

In the first line, the table says that

$$P(\neg A) = 1 - P(A)$$

and what it **means** is that the probability of "not A" is equal to 1 minus the probability of A. A moment's thought (and a tedious example) make it obvious why this must be true. If A coresponds to the even that I wear jeans (i.e., one of x_1 or x_2 or x_3 happens), then the only meaningful definition of "not A" (which is mathematically denoted as $\neg A$) is to say that $\neg A$ consists of **all** elementary events that don't belong to A. In the case of the pants distribution it means that $\neg A = (x_4, x_5)$, or, to say it in English: "not jeans" consists of all pairs of pants that aren't jeans (i.e., the black suit and the blue tracksuit). Consequently, every single elementary event belongs to either A or $\neg A$, but not both. Okay, so now let's rearrange our statement above:

$$P(\neg A) + P(A) = 1$$

which is a trite way of saying either I do wear jeans or I don't wear jeans: the probability of "not jeans" plus the probability of "jeans" is 1. Mathematically:

$$\begin{aligned} P(\neg A) &= P(x_4) + P(x_5) \\ P(A) &= P(x_1) + P(x_2) + P(x_3) \end{aligned}$$

so therefore

$$\begin{aligned} P(\neg A) + P(A) &= P(x_1) + P(x_2) + P(x_3) + P(x_4) + P(x_5) \\ &= \sum_{x \in X} P(x) \\ &= 1 \end{aligned}$$

Excellent. It all seems to work.

Wow, I can hear you saying. That's a lot of xs to tell me the freaking obvious. And you're right: this **is** freaking obvious. The whole **point** of probability theory to to formalise and mathematise a few very basic common sense intuitions. So let's carry this line of thought forward a bit further. In the last section I defined an event corresponding to **not** A, which I denoted $\neg A$. Let's now define two new events that correspond to important everyday concepts: A **and** B, and A **or** B. To be precise:

English statement:	Mathematical notation:
"A and B" both happen	$A \cap B$
at least one of "A or B" happens	$A \cup B$

Since A and B are both defined in terms of our elementary events (the xs) we're going to need to try to describe $A \cap B$ and $A \cup B$ in terms of our elementary events too. Can we do this? Yes we can The only way that both A and B can occur is if the elementary event that we observe turns out to belong to both A and B. Thus "$A \cap B$" includes only those elementary events that belong to both A and B...

$$\begin{aligned} A &= (x_1, x_2, x_3) \\ B &= (x_3, x_4) \\ A \cap B &= (x_3) \end{aligned}$$

So, um, the only way that I can wear "jeans" (x_1, x_2, x_3) and "black pants" (x_3, x_4) is if I wear "black jeans" (x_3). Another victory

for the bloody obvious.

At this point, you're not going to be at all shocked by the definition of $A \cup B$, though you're probably going to be extremely bored by it. The only way that I can wear "jeans" or "black pants" is if the elementary pants that I actually do wear belongs to A or to B, or to both. So...

$$\begin{aligned} A &= (x_1, x_2, x_3) \\ B &= (x_3, x_4) \\ A \cup B &= (x_1, x_2, x_3, x_4) \end{aligned}$$

Oh yeah baby. Mathematics at its finest.

So, we've defined what we mean by $A \cap B$ and $A \cup B$. Now let's assign probabilities to these events. More specifically, let's start by verifying the rule that claims that:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Using our definitions earlier, we know that $A \cup B = (x_1, x_2, x_3, x_4)$, so

$$P(A \cup B) = P(x_1) + P(x_2) + P(x_3) + P(x_4)$$

and making similar use of the fact that we know what elementary events belong to A, B and $A \cap B$....

$$\begin{aligned} P(A) &= P(x_1) + P(x_2) + P(x_3) \\ P(B) &= P(x_3) + P(x_4) \\ P(A \cap B) &= P(x_3) \end{aligned}$$

and therefore

$$\begin{aligned} P(A) + P(B) - P(A \cap B) &= P(x_1) + P(x_2) + P(x_3) + P(x_3) + P(x_4) - P(x_3) \\ &= P(x_1) + P(x_2) + P(x_3) + P(x_4) \\ &= P(A \cup B) \end{aligned}$$

Done.

The next concept we need to define is the notion of "B given A", which is typically written $B|A$. Here's what I mean: suppose that I get up one morning, and put on a pair of pants. An elementary event x has occurred. Suppose further I yell out to my wife (who is in the other room, and so cannot see my pants) "I'm wearing jeans today!". Assuming that she believes that I'm telling the truth, she knows that A is true. **Given** that she knows that A has happened, what is the **conditional probability** that B is also true? Well, let's think about what she knows. Here are the facts:

- **The non-jeans events are impossible**. If A is true, then we know that the only possible elementary events that could have occurred are x_1, x_2 and x_3 (i.e.,the jeans). The non-jeans events x_4 and x_5 are now impossible, and must be assigned probability zero. In other words, our **sample space** has been restricted to the jeans events. But it's still the case that the probabilities of these these events **must** sum to 1: we know for sure that I'm wearing jeans.
- **She's learned nothing about which jeans I'm wearing**. Before I made my announcement that I was wearing jeans, she already knew that I was five times as likely to be wearing blue jeans ($P(x_1) = 0.5$) than to be wearing black jeans ($P(x_3) = 0.1$). My announcement doesn't change this... I said **nothing** about what colour my jeans were, so it must remain the case that $P(x_1)/P(x_3)$ stays the same, at a value of 5.

There's only one way to satisfy these constraints: set the impossible events to have zero probability (i.e., $P(x|A) = 0$ if x is not in A), and then divide the probabilities of all the others by $P(A)$. In this case, since $P(A) = 0.9$, we divide by 0.9. This gives:

which pants?	elementary event	old prob, $P(x)$	new prob, $P(x\|A)$
blue jeans	x_1	0.5	0.556
grey jeans	x_2	0.3	0.333
black jeans	x_3	0.1	0.111

which pants?	elementary event	old prob, $P(x)$	new prob, $P(x\|A)$
black suit	x_4	0	0
blue tracksuit	x_5	0.1	0

In mathematical terms, we say that

$$P(x|A) = \frac{P(x)}{P(A)}$$

if $x \in A$, and $P(x|A) = 0$ otherwise. And therefore...

$$\begin{aligned} P(B|A) &= P(x_3|A) + P(x_4|A) \\ &= \frac{P(x_3)}{P(A)} + 0 \\ &= \frac{P(x_3)}{P(A)} \end{aligned}$$

Now, recalling that $A \cap B = (x_3)$, we can write this as

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

and if we multiply both sides by $P(A)$ we obtain:

$$P(A \cap B) = P(B|A)P(A)$$

which is the third rule that we had listed in the table.

This page titled 13.4: Basic Probability Theory is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

13.5: The Binomial Distribution

As you might imagine, probability distributions vary enormously, and there's an enormous range of distributions out there. However, they aren't all equally important. In fact, the vast majority of the content in this book relies on one of five distributions: the binomial distribution, the normal distribution, the t distribution, the χ^2 ("chi-square") distribution and the F distribution. Given this, what I'll do over the next few sections is provide a brief introduction to all five of these, paying special attention to the binomial and the normal. I'll start with the binomial distribution, since it's the simplest of the five.

Introducing the Binomial

The theory of probability originated in the attempt to describe how games of chance work, so it seems fitting that our discussion of the *binomial distribution* should involve a discussion of rolling dice and flipping coins. Let's imagine a simple "experiment": in my hot little hand I'm holding 20 identical six-sided dice. On one face of each die there's a picture of a skull; the other five faces are all blank. If I proceed to roll all 20 dice, what's the probability that I'll get exactly 4 skulls? Assuming that the dice are fair, we know that the chance of any one die coming up skulls is 1 in 6; to say this another way, the skull probability for a single die is approximately .167. This is enough information to answer our question, so let's have a look at how it's done.

As usual, we'll want to introduce some names and some notation. We'll let N denote the number of dice rolls in our experiment; which is often referred to as the *size parameter* of our binomial distribution. We'll also use θ to refer to the the probability that a single die comes up skulls, a quantity that is usually called the *success probability* of the binomial. Finally, we'll use X to refer to the results of our experiment, namely the number of skulls I get when I roll the dice. Since the actual value of X is due to chance, we refer to it as a *random variable*. In any case, now that we have all this terminology and notation, we can use it to state the problem a little more precisely. The quantity that we want to calculate is the probability that $X = 4$ given that we know that $\theta = 0.167$ and $N = 20$. The general "form" of the thing I'm interested in calculating could be written as

$$P(X \mid \theta, N)$$

and we're interested in the special case where $X = 4$, $\theta = .167$ and $N = 20$. There's only one more piece of notation I want to refer to before moving on to discuss the solution to the problem. If I want to say that X is generated randomly from a binomial distribution with parameters θ and N, the notation I would use is as follows:

$$X \sim \text{Binomial}(\theta, N)$$

Yeah, yeah. I know what you're thinking: notation, notation, notation. Really, who cares? Very few readers of this book are here for the notation, so I should probably move on and talk about how to use the binomial distribution. I've included the formula for the binomial distribution in Table 4.3.1, since some readers may want to play with it themselves, but since most people probably don't care that much and because we don't need the formula in this book, I won't talk about it in any detail. Instead, I just want to show you what the binomial distribution looks like. To that end, Figure 13.5.1 plots the binomial probabilities for all possible values of X for our dice rolling experiment, from $X = 0$ (no skulls) all the way up to $X = 20$ (all skulls). Note that this is basically a bar chart, and is no different to the "pants probability" plot I drew in Figure 4.3.1. On the horizontal axis we have all the possible events, and on the vertical axis we can read off the probability of each of those events. So, the probability of rolling 4 skulls out of 20 times is about 0.20 (the actual answer is 0.20220036, as we'll see in a moment). In other words, you'd expect that to happen about 20% of the times you repeated this experiment.

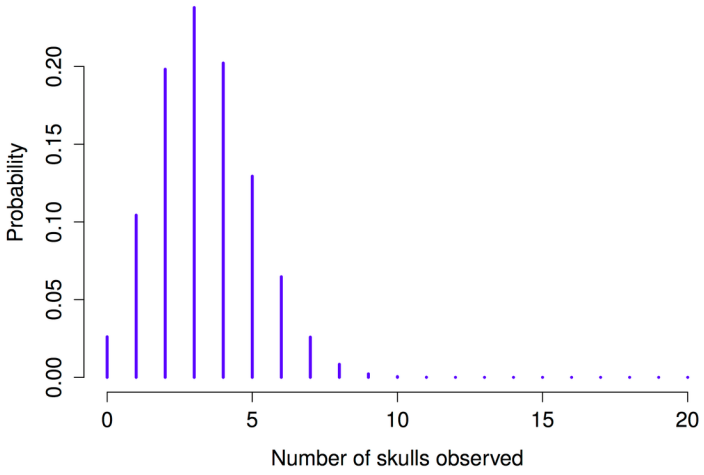


Figure 13.5.1: The binomial distribution with size parameter of N =20 and an underlying success probability of 1/6. Each vertical bar depicts the probability of one specific outcome (i.e., one possible value of X). Because this is a probability distribution, each of the probabilities must be a number between 0 and 1, and the heights of the bars must sum to 1 as well.

Working with the binomial distribution in R

R has a function called `dbinom` that calculates binomial probabilities for us. The main arguments to the function are

- `x` This is a number, or vector of numbers, specifying the outcomes whose probability you're trying to calculate.
- `size` This is a number telling R the size of the experiment.
- `prob` This is the success probability for any one trial in the experiment.

So, in order to calculate the probability of getting skulls, from an experiment of trials, in which the probability of getting a skull on any one trial is ... well, the command I would use is simply this:

```
dbinom( x = 4, size = 20, prob = 1/6 )
```

```
0.202203581217173
```

To give you a feel for how the binomial distribution changes when we alter the values of θ and N, let's suppose that instead of rolling dice, I'm actually flipping coins. This time around, my experiment involves flipping a fair coin repeatedly, and the outcome that I'm interested in is the number of heads that I observe. In this scenario, the success probability is now $\theta = 1/2$. Suppose I were to flip the coin $N = 20$ times. In this example, I've changed the success probability, but kept the size of the experiment the same. What does this do to our binomial distribution?

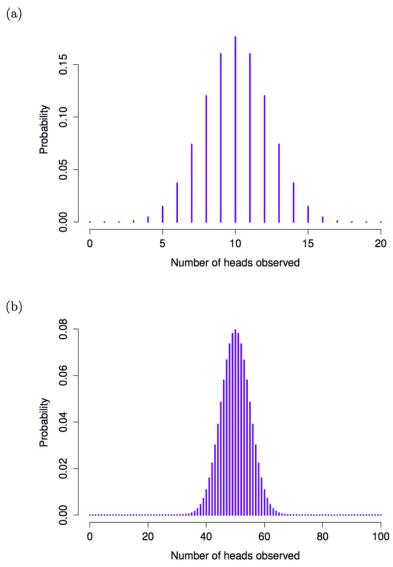


Figure 13.5.2: Two binomial distributions, involving a scenario in which I'm flipping a fair coin, so the underlying success probability is 1/2. In panel (a), we assume I'm flipping the coin N = 20 times. In panel (b) we assume that the coin is flipped N = 100 times.

Well, as Figure $13.5.2(a)$ shows, the main effect of this is to shift the whole distribution, as you'd expect. Okay, what if we flipped a coin $N = 100$ times? Well, in that case, we get Figure $13.5.2(b)$. The distribution stays roughly in the middle, but there's a bit more variability in the possible outcomes.

At this point, I should probably explain the name of the `dbinom` function. Obviously, the "binom" part comes from the fact that we're working with the binomial distribution, but the "d" prefix is probably a bit of a mystery. In this section I'll give a partial explanation: specifically, I'll explain why there is a prefix. As for why it's a "d" specifically, you'll have to wait until the next section. What's going on here is that R actually provides **four** functions in relation to the binomial distribution. These four functions are `dbinom`, `pbinom`, `rbinom` and `qbinom`, and each one calculates a different quantity of interest. Not only that, R does the same thing for **every** probability distribution that it implements. No matter what distribution you're talking about, there's a `d` function, a `p` function, `r` a function and a `q` function.

Let's have a look at what all four functions do. Firstly, all four versions of the function require you to specify the `size` and `prob` arguments: no matter what you're trying to get R to calculate, it needs to know what the parameters are. However, they differ in terms of what the other argument is, and what the output is. So let's look at them one at a time.

- The `d` form we've already seen: you specify a particular outcome `x`, and the output is the probability of obtaining exactly that outcome. (the "d" is short for *density*, but ignore that for now).
- The `p` form calculates the *cumulative probability*. You specify a particular quantile `q` , and it tells you the probability of obtaining an outcome **smaller than or equal to** `q`.
- The `q` form calculates the *quantiles* of the distribution. You specify a probability value `p`, and it gives you the corresponding percentile. That is, the value of the variable for which there's a probability `p` of obtaining an outcome lower than that value.
- The `r` form is a *random number generator*: specifically, it generates `n` random outcomes from the distribution.

This is a little abstract, so let's look at some concrete examples. Again, we've already covered `dbinom` so let's focus on the other three versions. We'll start with `pbinom`, and we'll go back to the skull-dice example. Again, I'm rolling 20 dice, and each die has a 1 in 6 chance of coming up skulls. Suppose, however, that I want to know the probability of rolling 4 **or fewer** skulls. If I wanted to, I could use the `dbinom` function to calculate the exact probability of rolling 0 skulls, 1 skull, 2 skulls, 3 skulls and 4 skulls and then add these up, but there's a faster way. Instead, I can calculate this using the `pbinom` function. Here's the command:

```
pbinom( q= 4, size = 20, prob = 1/6)
```

```
0.768749218992842
```

In other words, there is a 76.9% chance that I will roll 4 or fewer skulls. Or, to put it another way, R is telling us that a value of 4 is actually the 76.9th percentile of this binomial distribution.

Next, let's consider the `qbinom` function. Let's say I want to calculate the 75th percentile of the binomial distribution. If we're sticking with our skulls example, I would use the following command to do this:

```
qbinom( p = 0.75, size = 20, prob = 1/6 )
```

```
4
```

Hm. There's something odd going on here. Let's think this through. What the `qbinom` function appears to be telling us is that the 75th percentile of the binomial distribution is 4, even though we saw from the function that 4 is **actually** the 76.9th percentile. And it's definitely the `pbinom` function that is correct. I promise. The weirdness here comes from the fact that our binomial distribution doesn't really **have** a 75th percentile. Not really. Why not? Well, there's a 56.7% chance of rolling 3 or fewer skulls (you can type `pbinom(3, 20, 1/6)` to confirm this if you want), and a 76.9% chance of rolling 4 or fewer skulls. So there's a sense in which the 75th percentile should lie "in between" 3 and 4 skulls. But that makes no sense at all! You can't roll 20 dice and get 3.9 of them come up skulls. This issue can be handled in different ways: you could report an in between value (or **interpolated** value, to use the technical name) like 3.9, you could round down (to 3) or you could round up (to 4).

The `qbinom` function rounds upwards: if you ask for a percentile that doesn't actually exist (like the 75th in this example), R finds the smallest value for which the the percentile rank is **at least** what you asked for. In this case, since the "true" 75th percentile (whatever that would mean) lies somewhere between 3 and 4 skulls, R rounds up and gives you an answer of 4. This subtlety is tedious, I admit, but thankfully it's only an issue for discrete distributions like the binomial. The other distributions that I'll talk about (normal, t, χ^2 and F) are all continuous, and so R can always return an exact quantile whenever you ask for it.

Finally, we have the random number generator. To use the `rbinom` function, you specify how many times R should "simulate" the experiment using the `n` argument, and it will generate random outcomes from the binomial distribution. So, for instance, suppose I were to repeat my die rolling experiment 100 times. I could get R to simulate the results of these experiments by using the following command:

```
rbinom( n = 100, size = 20, prob = 1/6 )
```

```
2 4 3 5 5 1 2 7 2 1 4 3 4 1 1 4 5 2 1 2 7 6 1 1 1 2 4 4 3 6 4 3 3 3 4
2 4 4 2 1 3 2 5 4 2 4 1 1 6 2 3 1 4 1 1 3 4 4 3 6 5 5 5 3 2 4 5 5 4
4 2 1 1 5 3 6 1 6 3 4 3 5 2 7 1 5 3 2 2 0 3 4 3 6 6 4 5 1 4 3
```

As you can see, these numbers are pretty much what you'd expect given the distribution shown in Figure 13.5.1. Most of the time I roll somewhere between 1 to 5 skulls. There are a lot of subtleties associated with random number generation using a computer, but for the purposes of this book we don't need to worry too much about them.

This page titled 13.5: The Binomial Distribution is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

13.6: The normal distribution

While the binomial distribution is conceptually the simplest distribution to understand, it's not the most important one. That particular honour goes to the *normal distribution*, which is also referred to as "the bell curve" or a "Gaussian distribution".

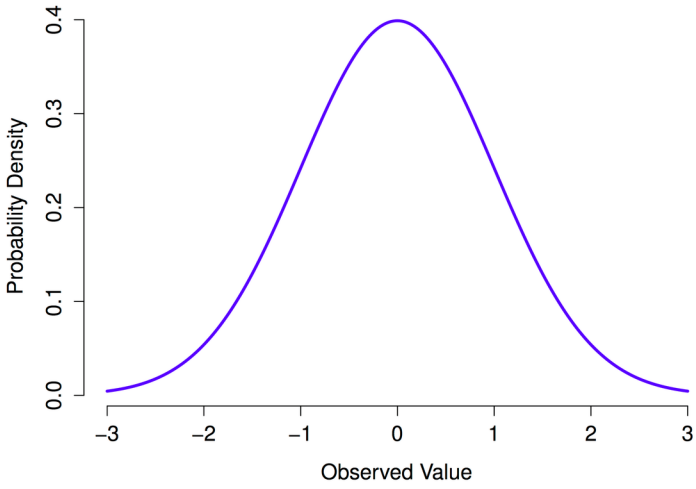


Figure 13.6.1: The normal distribution with mean = 0 and standard deviation = 1. The x-axis corresponds to the value of some variable, and the y-axis tells us something about how likely we are to observe that value. However, notice that the y-axis is labelled Probability Density and not Probability. There is a subtle and somewhat frustrating characteristic of continuous distributions that makes the y axis behave a bit oddly: the height of the curve here isn't actually the probability of observing a particular x value. On the other hand, it is true that the heights of the curve tells you which x values are more likely (the higher ones!).

A normal distribution is described using two parameters, the mean of the distribution μ and the standard deviation of the distribution σ. The notation that we sometimes use to say that a variable X is normally distributed is as follows:

$$X \sim \text{Normal}(\mu, \sigma) \tag{13.6.1}$$

Of course, that's just notation. It doesn't tell us anything interesting about the normal distribution itself. The mathematical formula for the normal distribution is:

$$p(X \mid \mu, \sigma) = \underline{\underline{\text{Normal}}} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)$$

Figure 13.6.2: Formula for the normal distribution.

The formula is important enough that everyone who learns statistics should at least look at it, but since this is an introductory text I don't want to focus on it to much. Instead, we look at how R can be used to work with normal distributions. The R functions for the normal distribution are *dnorm()*, *pnorm()*, *qnorm()* and *rnorm()*. However, they behave in pretty much exactly the same way as the corresponding functions for the binomial distribution, so there's not a lot that you need to know. The only thing that I should point out is that the argument names for the parameters are *mean* and *sd*. In pretty much every other respect, there's nothing else to add.

Instead of focusing on the maths, let's try to get a sense for what it means for a variable to be normally distributed. To that end, have a look at Figure 13.6.1, which plots a normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. You can see where the name "bell curve" comes from: it looks a bit like a bell. Notice that, unlike the plots that I drew to illustrate the binomial distribution, the picture of the normal distribution in Figure 13.6.1 shows a smooth curve instead of "histogram-like" bars. This isn't an arbitrary choice: the normal distribution is continuous, whereas the binomial is discrete. For instance, in the die rolling example from the last section, it was possible to get 3 skulls or 4 skulls, but impossible to get 3.9 skulls.

With this in mind, let's see if we can't get an intuition for how the normal distribution works. Firstly, let's have a look at what happens when we play around with the parameters of the distribution. One parameter we can change is the mean. This will shift the distribution to the right or left. The animation below shows a normal distribution with mean = 0, moving up and down from mean = 0 to mean = 5. Note, when you change the mean the whole shape of the distribution does not change, it just shifts from left to right. In the animation the normal distribution bounces up and down a little, but that's just a quirk of the animation (plus it looks fund that way).

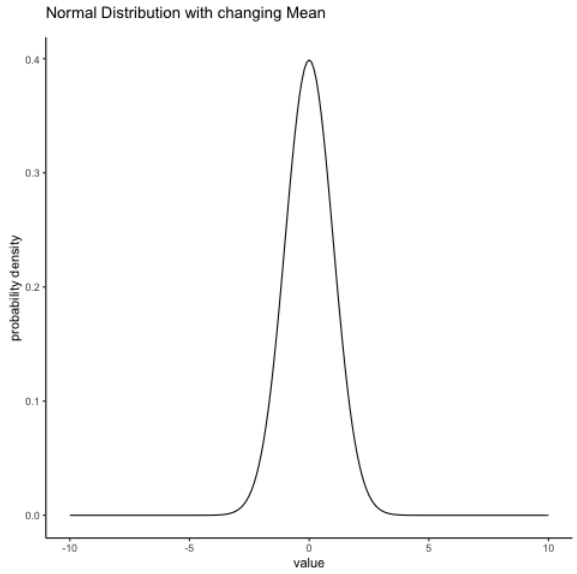


Figure 13.6.3: A normal distribution with a moving mean.

In contrast, if we increase the standard deviation while keeping the mean constant, the peak of the distribution stays in the same place, but the distribution gets wider. The next animation shows what happens when you start with a small standard deviation (sd=0.5), and move to larger and larger standard deviation (up to sd =5). As you can see, the distribution spreads out and becomes wider as the standard deviation increases.

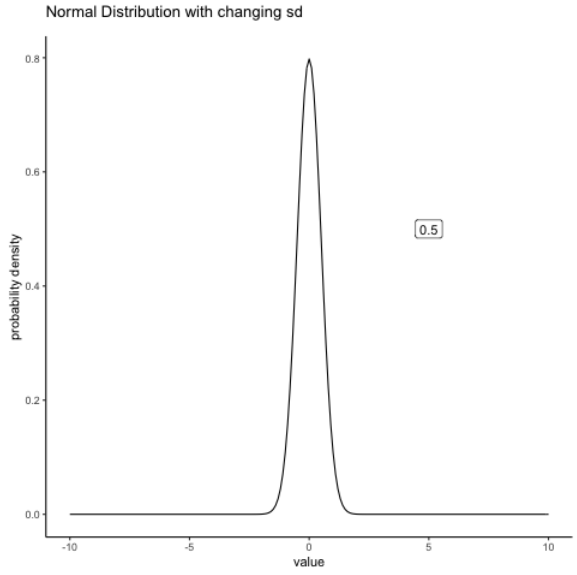


Figure 13.6.4: A normal distribution with a shifting sd.

Notice, though, that when we widen the distribution, the height of the peak shrinks. This has to happen: in the same way that the heights of the bars that we used to draw a discrete binomial distribution have to *sum* to 1, the total *area under the curve* for the normal distribution must equal 1. Before moving on, I want to point out one important characteristic of the normal distribution. Irrespective of what the actual mean and standard deviation are, 68.3% of the area falls within 1 standard deviation of the mean. Similarly, 95.4% of the distribution falls within 2 standard deviations of the mean, and 99.7% of the distribution is within 3 standard deviations.

Probability density

There's something I've been trying to hide throughout my discussion of the normal distribution, something that some introductory textbooks omit completely. They might be right to do so: this "thing" that I'm hiding is weird and counterintuitive even by the admittedly distorted standards that apply in statistics. Fortunately, it's not something that you need to understand at a deep level in order to do basic statistics: rather, it's something that starts to become important later on when you move beyond the basics. So, if it doesn't make complete sense, don't worry: try to make sure that you follow the gist of it.

Throughout my discussion of the normal distribution, there's been one or two things that don't quite make sense. Perhaps you noticed that the y-axis in these figures is labelled "Probability Density" rather than density. Maybe you noticed that I used $p(X)$ instead of $P(X)$ when giving the formula for the normal distribution. Maybe you're wondering why R uses the "d" prefix for functions like *dnorm()*. And maybe, just maybe, you've been playing around with the *dnorm()* function, and you accidentally typed in a command like this:

```
dnorm( x = 1, mean = 1, sd = 0.1 )
```

```
3.98942280401433
```

And if you've done the last part, you're probably very confused. I've asked R to calculate the probability that *x = 1*, for a normally distributed variable with *mean = 1* and standard deviation *sd = 0.1*; and it tells me that the probability is 3.99. But, as we discussed earlier, probabilities *can't* be larger than 1. So either I've made a mistake, or that's not a probability.

As it turns out, the second answer is correct. What we've calculated here isn't actually a probability: it's something else. To understand what that something is, you have to spend a little time thinking about what it really *means* to say that X is a continuous variable. Let's say we're talking about the temperature outside. The thermometer tells me it's 23 degrees, but I know that's not really true. It's not *exactly* 23 degrees. Maybe it's 23.1 degrees, I think to myself. But I know that that's not really true either, because it might actually be 23.09 degrees. But, I know that... well, you get the idea. The tricky thing with genuinely continuous quantities is that you never really know exactly what they are.

Now think about what this implies when we talk about probabilities. Suppose that tomorrow's maximum temperature is sampled from a normal distribution with mean 23 and standard deviation 1. What's the probability that the temperature will be *exactly* 23 degrees? The answer is "zero", or possibly, "a number so close to zero that it might as well be zero". Why is this?

It's like trying to throw a dart at an infinitely small dart board: no matter how good your aim, you'll never hit it. In real life you'll never get a value of exactly 23. It'll always be something like 23.1 or 22.99998 or something. In other words, it's completely meaningless to talk about the probability that the temperature is exactly 23 degrees. However, in everyday language, if I told you that it was 23 degrees outside and it turned out to be 22.9998 degrees, you probably wouldn't call me a liar. Because in everyday language, "23 degrees" usually means something like "somewhere between 22.5 and 23.5 degrees". And while it doesn't feel very meaningful to ask about the probability that the temperature is exactly 23 degrees, it does seem sensible to ask about the probability that the temperature lies between 22.5 and 23.5, or between 20 and 30, or any other range of temperatures.

The point of this discussion is to make clear that, when we're talking about continuous distributions, it's not meaningful to talk about the probability of a specific value. However, what we *can* talk about is the **probability that the value lies within a particular range of values**. To find out the probability associated with a particular range, what you need to do is calculate the "area under the curve".

Okay, so that explains part of the story. I've explained a little bit about how continuous probability distributions should be interpreted (i.e., area under the curve is the key thing), but I haven't actually explained what the *dnorm()* function actually calculates. Equivalently, what does the formula for $p(x)$ that I described earlier actually mean? Obviously, $p(x)$ doesn't describe a probability, but what is it? The name for this quantity $p(x)$ is a *probability density*, and in terms of the plots we've been drawing, it corresponds to the *height* of the curve. The densities themselves aren't meaningful in and of themselves: but they're "rigged" to

ensure that the *area* under the curve is always interpretable as genuine probabilities. To be honest, that's about as much as you really need to know for now.

This page titled 13.6: The normal distribution is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

13.7: Other useful distributions

There are many other useful distributions, these include the `t` distribution, the `F` distribution, and the chi squared distribution. We will soon discover more about the `t` and `F` distributions when we discuss t-tests and ANOVAs in later chapters.

This page titled 13.7: Other useful distributions is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

13.8: Summary of Probability

We've talked what probability means, and why statisticians can't agree on what it means. We talked about the rules that probabilities have to obey. And we introduced the idea of a probability distribution, and spent a good chunk talking about some of the more important probability distributions that statisticians work with. We talked about things like this:

- Probability theory versus statistics
- Frequentist versus Bayesian views of probability
- Basics of probability theory
- Binomial distribution, normal distribution

As you'd expect, this coverage is by no means exhaustive. Probability theory is a large branch of mathematics in its own right, entirely separate from its application to statistics and data analysis. As such, there are thousands of books written on the subject and universities generally offer multiple classes devoted entirely to probability theory. Even the "simpler" task of documenting standard probability distributions is a big topic.Fortunately for you, very little of this is necessary. You're unlikely to need to know dozens of statistical distributions when you go out and do real world data analysis, and you definitely won't need them for this book, but it never hurts to know that there's other possibilities out there.

Picking up on that last point, there's a sense in which this whole chapter is something of a digression. Many undergraduate psychology classes on statistics skim over this content very quickly (I know mine did), and even the more advanced classes will often "forget" to revisit the basic foundations of the field. Most academic psychologists would not know the difference between probability and density, and until recently very few would have been aware of the difference between Bayesian and frequentist probability. However, I think it's important to understand these things before moving onto the applications. For example, there are a lot of rules about what you're "allowed" to say when doing statistical inference, and many of these can seem arbitrary and weird. However, they start to make sense if you understand that there is this Bayesian/frequentist distinction.

This page titled 13.8: Summary of Probability is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

13.9: Samples, populations and sampling

Remember, the role of descriptive statistics is to concisely summarize what we **do** know. In contrast, the purpose of inferential statistics is to "learn what we do not know from what we do". What kinds of things would we like to learn about? And how do we learn them? These are the questions that lie at the heart of inferential statistics, and they are traditionally divided into two "big ideas": estimation and hypothesis testing. The goal in this chapter is to introduce the first of these big ideas, estimation theory, but we'll talk about sampling theory first because estimation theory doesn't make sense until you understand sampling. So, this chapter divides into sampling theory, and how to make use of sampling theory to discuss how statisticians think about estimation. We have already done lots of sampling, so you are already familiar with some of the big ideas.

Sampling theory plays a huge role in specifying the assumptions upon which your statistical inferences rely. And in order to talk about "making inferences" the way statisticians think about it, we need to be a bit more explicit about what it is that we're drawing inferences **from** (the sample) and what it is that we're drawing inferences **about** (the population).

In almost every situation of interest, what we have available to us as researchers is a **sample** of data. We might have run experiment with some number of participants; a polling company might have phoned some number of people to ask questions about voting intentions; etc. Regardless: the data set available to us is finite, and incomplete. We can't possibly get every person in the world to do our experiment; a polling company doesn't have the time or the money to ring up every voter in the country etc. In our earlier discussion of descriptive statistics, this sample was the only thing we were interested in. Our only goal was to find ways of describing, summarizing and graphing that sample. This is about to change.

Defining a population

A sample is a concrete thing. You can open up a data file, and there's the data from your sample. A **population**, on the other hand, is a more abstract idea. It refers to the set of all possible people, or all possible observations, that you want to draw conclusions about, and is generally **much** bigger than the sample. In an ideal world, the researcher would begin the study with a clear idea of what the population of interest is, since the process of designing a study and testing hypotheses about the data that it produces does depend on the population about which you want to make statements. However, that doesn't always happen in practice: usually the researcher has a fairly vague idea of what the population is and designs the study as best he/she can on that basis.

Sometimes it's easy to state the population of interest. For instance, in the "polling company" example, the population consisted of all voters enrolled at the a time of the study – millions of people. The sample was a set of 1000 people who all belong to that population. In most situations the situation is much less simple. In a typical a psychological experiment, determining the population of interest is a bit more complicated. Suppose I run an experiment using 100 undergraduate students as my participants. My goal, as a cognitive scientist, is to try to learn something about how the mind works. So, which of the following would count as "the population":

- All of the undergraduate psychology students at the University of Adelaide?
- Undergraduate psychology students in general, anywhere in the world?
- Australians currently living?
- Australians of similar ages to my sample?
- Anyone currently alive?
- Any human being, past, present or future?
- Any biological organism with a sufficient degree of intelligence operating in a terrestrial environment?
- Any intelligent being?

Each of these defines a real group of mind-possessing entities, all of which might be of interest to me as a cognitive scientist, and it's not at all clear which one ought to be the true population of interest.

Simple random samples

Irrespective of how we define the population, the critical point is that the sample is a subset of the population, and our goal is to use our knowledge of the sample to draw inferences about the properties of the population. The relationship between the two depends on the **procedure** by which the sample was selected. This procedure is referred to as a **sampling method**, and it is

important to understand why it matters.

To keep things simple, imagine we have a bag containing 10 chips. Each chip has a unique letter printed on it, so we can distinguish between the 10 chips. The chips come in two colors, black and white.

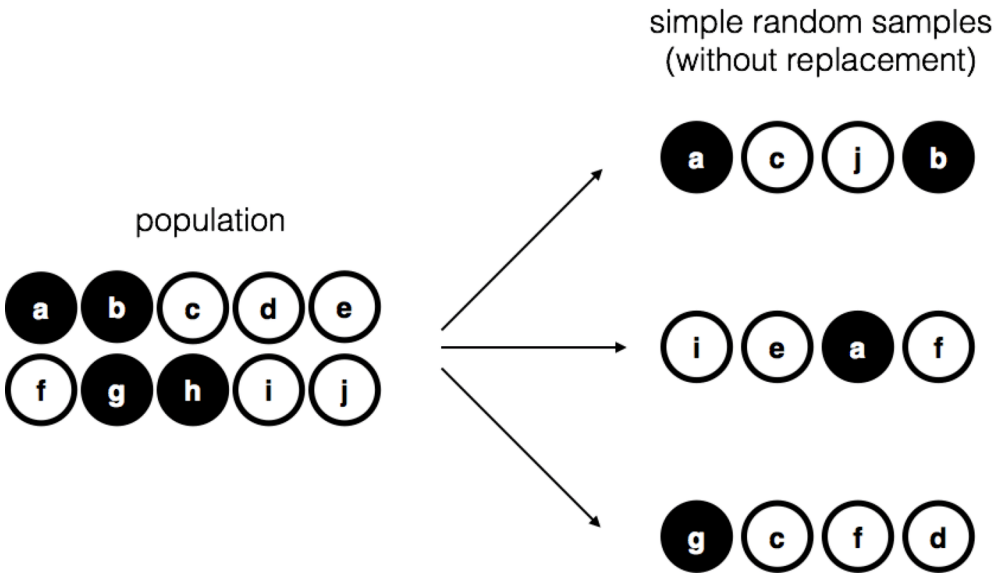


Figure 13.9.1: Simple random sampling without replacement from a finite population.

This set of chips is the population of interest, and it is depicted graphically on the left of Figure 13.9.1.

As you can see from looking at the picture, there are 4 black chips and 6 white chips, but of course in real life we wouldn't know that unless we looked in the bag. Now imagine you run the following "experiment": you shake up the bag, close your eyes, and pull out 4 chips without putting any of them back into the bag. First out comes the a chip (black), then the c chip (white), then j (white) and then finally b (black). If you wanted, you could then put all the chips back in the bag and repeat the experiment, as depicted on the right hand side of Figure 13.9.1. Each time you get different results, but the procedure is identical in each case. The fact that the same procedure can lead to different results each time, we refer to it as a **random** process. However, because we shook the bag before pulling any chips out, it seems reasonable to think that every chip has the same chance of being selected. A procedure in which every member of the population has the same chance of being selected is called a **simple random sample**. The fact that we did **not** put the chips back in the bag after pulling them out means that you can't observe the same thing twice, and in such cases the observations are said to have been sampled **without replacement**.

To help make sure you understand the importance of the sampling procedure, consider an alternative way in which the experiment could have been run. Suppose that my 5-year old son had opened the bag, and decided to pull out four black chips without putting any of them back in the bag. This **biased** sampling scheme is depicted in Figure 13.9.2.

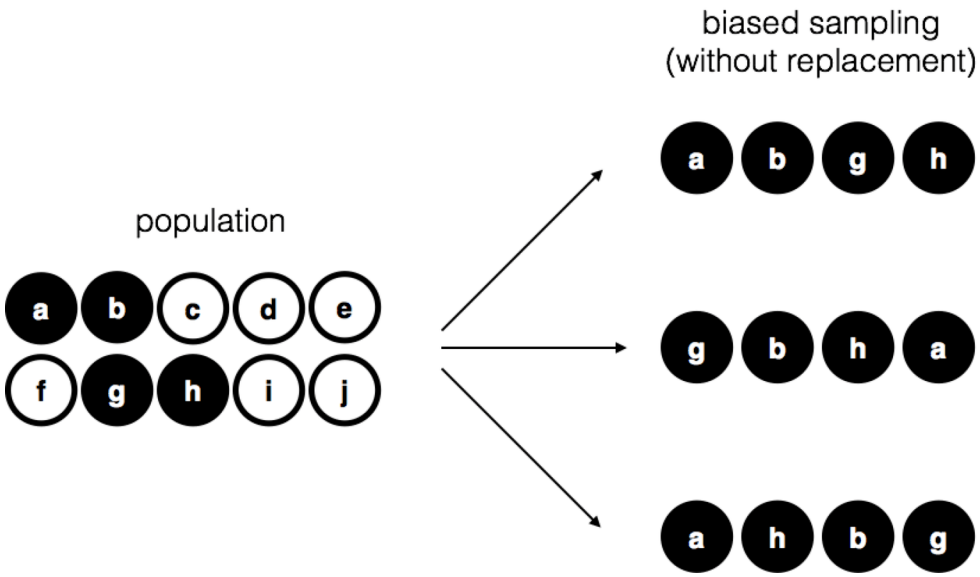


Figure 13.9.2: Biased sampling without replacement from a finite population.

Now consider the evidentiary value of seeing 4 black chips and 0 white chips. Clearly, it depends on the sampling scheme, does it not? If you know that the sampling scheme is biased to select only black chips, then a sample that consists of only black chips doesn't tell you very much about the population! For this reason, statisticians really like it when a data set can be considered a simple random sample, because it makes the data analysis **much** easier.

A third procedure is worth mentioning. This time around we close our eyes, shake the bag, and pull out a chip. This time, however, we record the observation and then put the chip back in the bag. Again we close our eyes, shake the bag, and pull out a chip. We then repeat this procedure until we have 4 chips. Data sets generated in this way are still simple random samples, but because we put the chips back in the bag immediately after drawing them it is referred to as a sample **with replacement**. The difference between this situation and the first one is that it is possible to observe the same population member multiple times, as illustrated in Figure 13.9.3.

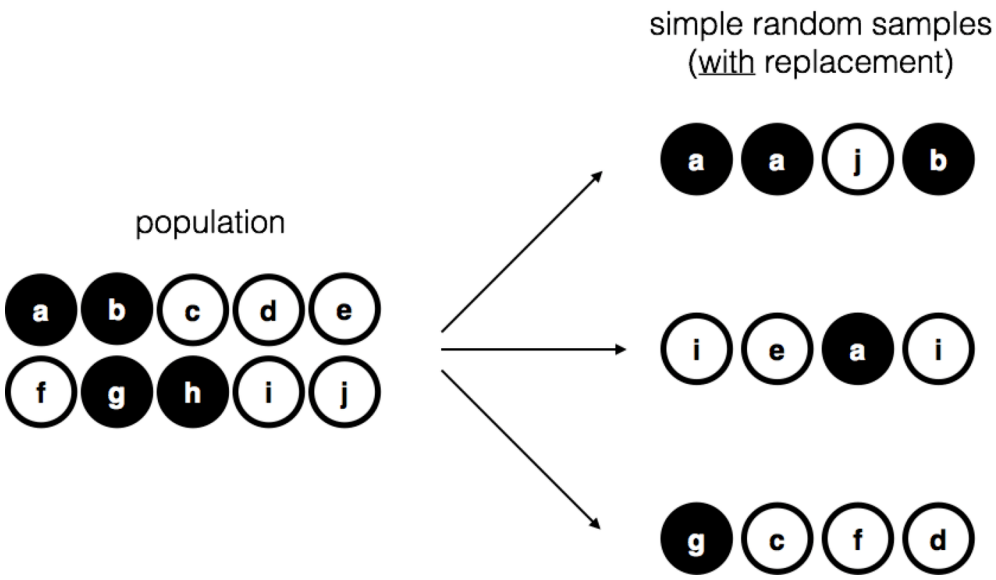


Figure 13.9.3: Simple random sampling with replacement from a finite population.

Most psychology experiments tend to be sampling without replacement, because the same person is not allowed to participate in the experiment twice. However, most statistical theory is based on the assumption that the data arise from a simple random sample **with** replacement. In real life, this very rarely matters. If the population of interest is large (e.g., has more than 10 entities!) the difference between sampling with- and without- replacement is too small to be concerned with. The difference between simple

random samples and biased samples, on the other hand, is not such an easy thing to dismiss.

Most samples are not simple random samples

As you can see from looking at the list of possible populations that I showed above, it is almost impossible to obtain a simple random sample from most populations of interest. When I run experiments, I'd consider it a minor miracle if my participants turned out to be a random sampling of the undergraduate psychology students at Adelaide university, even though this is by far the narrowest population that I might want to generalize to. A thorough discussion of other types of sampling schemes is beyond the scope of this book, but to give you a sense of what's out there I'll list a few of the more important ones:

- **Stratified sampling**. Suppose your population is (or can be) divided into several different sub-populations, or **strata**. Perhaps you're running a study at several different sites, for example. Instead of trying to sample randomly from the population as a whole, you instead try to collect a separate random sample from each of the strata. Stratified sampling is sometimes easier to do than simple random sampling, especially when the population is already divided into the distinct strata. It can also be more efficient that simple random sampling, especially when some of the sub-populations are rare. For instance, when studying schizophrenia it would be much better to divide the population into two strata (schizophrenic and not-schizophrenic), and then sample an equal number of people from each group. If you selected people randomly, you would get so few schizophrenic people in the sample that your study would be useless. This specific kind of of stratified sampling is referred to as **oversampling** because it makes a deliberate attempt to over-represent rare groups.
- **Snowball sampling** is a technique that is especially useful when sampling from a "hidden" or hard to access population, and is especially common in social sciences. For instance, suppose the researchers want to conduct an opinion poll among transgender people. The research team might only have contact details for a few trans folks, so the survey starts by asking them to participate (stage 1). At the end of the survey, the participants are asked to provide contact details for other people who might want to participate. In stage 2, those new contacts are surveyed. The process continues until the researchers have sufficient data. The big advantage to snowball sampling is that it gets you data in situations that might otherwise be impossible to get any. On the statistical side, the main disadvantage is that the sample is highly non-random, and non-random in ways that are difficult to address. On the real life side, the disadvantage is that the procedure can be unethical if not handled well, because hidden populations are often hidden for a reason. I chose transgender people as an example here to highlight this: if you weren't careful you might end up outing people who don't want to be outed (very, very bad form), and even if you don't make that mistake it can still be intrusive to use people's social networks to study them. It's certainly very hard to get people's informed consent **before** contacting them, yet in many cases the simple act of contacting them and saying "hey we want to study you" can be hurtful. Social networks are complex things, and just because you can use them to get data doesn't always mean you should.
- **Convenience sampling** is more or less what it sounds like. The samples are chosen in a way that is convenient to the researcher, and not selected at random from the population of interest. Snowball sampling is one type of convenience sampling, but there are many others. A common example in psychology are studies that rely on undergraduate psychology students. These samples are generally non-random in two respects: firstly, reliance on undergraduate psychology students automatically means that your data are restricted to a single sub-population. Secondly, the students usually get to pick which studies they participate in, so the sample is a self selected subset of psychology students not a randomly selected subset. In real life, most studies are convenience samples of one form or another. This is sometimes a severe limitation, but not always.

How much does it matter if you don't have a simple random sample?

Okay, so real world data collection tends not to involve nice simple random samples. Does that matter? A little thought should make it clear to you that it **can** matter if your data are not a simple random sample: just think about the difference between Figures 13.9.1 and 13.9.2. However, it's not quite as bad as it sounds. Some types of biased samples are entirely unproblematic. For instance, when using a stratified sampling technique you actually **know** what the bias is because you created it deliberately, often to **increase** the effectiveness of your study, and there are statistical techniques that you can use to adjust for the biases you've introduced (not covered in this book!). So in those situations it's not a problem.

More generally though, it's important to remember that random sampling is a means to an end, not the end in itself. Let's assume you've relied on a convenience sample, and as such you can assume it's biased. A bias in your sampling method is only a problem if it causes you to draw the wrong conclusions. When viewed from that perspective, I'd argue that we don't need the sample to be randomly generated in **every** respect: we only need it to be random with respect to the psychologically-relevant phenomenon of interest. Suppose I'm doing a study looking at working memory capacity. In study 1, I actually have the ability to sample randomly from all human beings currently alive, with one exception: I can only sample people born on a Monday. In study 2, I am able to sample randomly from the Australian population. I want to generalize my results to the population of all

living humans. Which study is better? The answer, obviously, is study 1. Why? Because we have no reason to think that being "born on a Monday" has any interesting relationship to working memory capacity. In contrast, I can think of several reasons why "being Australian" might matter. Australia is a wealthy, industrialized country with a very well-developed education system. People growing up in that system will have had life experiences much more similar to the experiences of the people who designed the tests for working memory capacity. This shared experience might easily translate into similar beliefs about how to "take a test", a shared assumption about how psychological experimentation works, and so on. These things might actually matter. For instance, "test taking" style might have taught the Australian participants how to direct their attention exclusively on fairly abstract test materials relative to people that haven't grown up in a similar environment; leading to a misleading picture of what working memory capacity is.

There are two points hidden in this discussion. Firstly, when designing your own studies, it's important to think about what population you care about, and try hard to sample in a way that is appropriate to that population. In practice, you're usually forced to put up with a "sample of convenience" (e.g., psychology lecturers sample psychology students because that's the least expensive way to collect data, and our coffers aren't exactly overflowing with gold), but if so you should at least spend some time thinking about what the dangers of this practice might be.

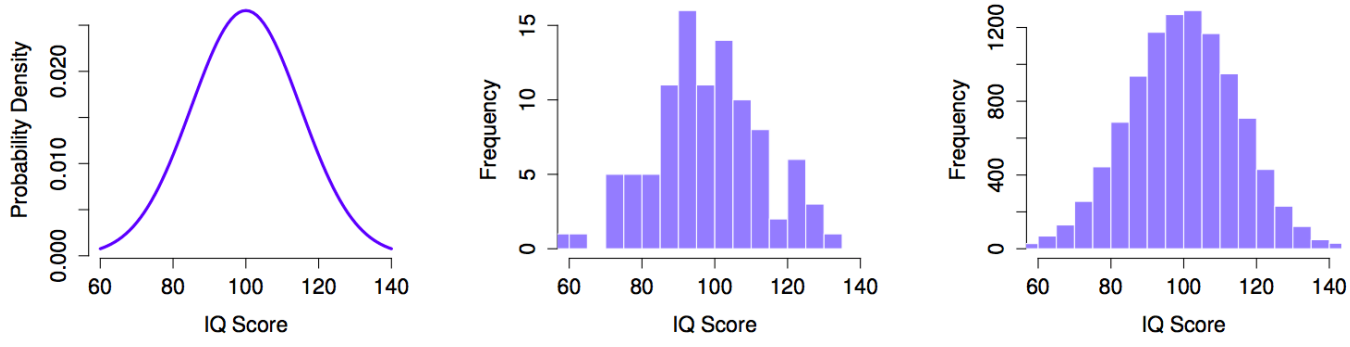
Secondly, if you're going to criticize someone else's study because they've used a sample of convenience rather than laboriously sampling randomly from the entire human population, at least have the courtesy to offer a specific theory as to **how** this might have distorted the results. Remember, everyone in science is aware of this issue, and does what they can to alleviate it. Merely pointing out that "the study only included people from group BLAH" is entirely unhelpful, and borders on being insulting to the researchers, who are aware of the issue. They just don't happen to be in possession of the infinite supply of time and money required to construct the perfect sample. In short, if you want to offer a responsible critique of the sampling process, then be **helpful**. Rehashing the blindingly obvious truisms that I've been rambling on about in this section isn't helpful.

Population parameters and sample statistics

Okay. Setting aside the thorny methodological issues associated with obtaining a random sample, let's consider a slightly different issue. Up to this point we have been talking about populations the way a scientist might. To a psychologist, a population might be a group of people. To an ecologist, a population might be a group of bears. In most cases the populations that scientists care about are concrete things that actually exist in the real world.

Statisticians, however, are a funny lot. On the one hand, they **are** interested in real world data and real science in the same way that scientists are. On the other hand, they also operate in the realm of pure abstraction in the way that mathematicians do. As a consequence, statistical theory tends to be a bit abstract in how a population is defined. In much the same way that psychological researchers operationalize our abstract theoretical ideas in terms of concrete measurements, statisticians operationalize the concept of a "population" in terms of mathematical objects that they know how to work with. You've already come across these objects they're called probability distributions (remember, the place where data comes from).

The idea is quite simple. Let's say we're talking about IQ scores. To a psychologist, the population of interest is a group of actual humans who have IQ scores. A statistician "simplifies" this by operationally defining the population as the probability distribution depicted in Figure $13.9.4a$.



(a) (b) (c)

Figure 13.9.4: The population distribution of IQ scores (panel a) and two samples drawn randomly from it. In panel b we have a sample of 100 observations, and panel c we have a sample of 10,000 observations.

IQ tests are designed so that the average IQ is 100, the standard deviation of IQ scores is 15, and the distribution of IQ scores is normal. These values are referred to as the **population parameters** because they are characteristics of the entire population. That is, we say that the population mean μ is 100, and the population standard deviation σ is 15.

Now suppose we collect some data. We select 100 people at random and administer an IQ test, giving a simple random sample from the population. The sample would consist of a collection of numbers like this:

```
106 101 98 80 74 ... 107 72 100
```

Each of these IQ scores is sampled from a normal distribution with mean 100 and standard deviation 15. So if I plot a histogram of the sample, I get something like the one shown in Figure 13.9.4*b*. As you can see, the histogram is **roughly** the right shape, but it's a very crude approximation to the true population distribution shown in Figure 13.9.4*c*. The mean of the sample is fairly close to the population mean 100 but not identical. In this case, it turns out that the people in the sample have a mean IQ of 98.5, and the standard deviation of their IQ scores is 15.9. These **sample statistics** are properties of the data set, and although they are fairly similar to the true population values, they are not the same. **In general, sample statistics are the things you can calculate from your data set, and the population parameters are the things you want to learn about.** Later on in this chapter we'll talk about how you can estimate population parameters using your sample statistics and how to work out how confident you are in your estimates but before we get to that there's a few more ideas in sampling theory that you need to know about.

This page titled 13.9: Samples, populations and sampling is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

13.10: The Law of Large Numbers

We just looked at the results of one fictitious IQ experiment with a sample size of $N = 100$. The results were somewhat encouraging: the true population mean is 100, and the sample mean of 98.5 is a pretty reasonable approximation to it. In many scientific studies that level of precision is perfectly acceptable, but in other situations you need to be a lot more precise. If we want our sample statistics to be much closer to the population parameters, what can we do about it?

The obvious answer is to collect more data. Suppose that we ran a much larger experiment, this time measuring the IQ's of 10,000 people. We can simulate the results of this experiment using R, using the **rnorm()** function, which generates random numbers sampled from a normal distribution. For an experiment with a sample size of **n = 10000**, and a population with **mean = 100** and **sd = 15**, R produces our fake IQ data using these commands:

```
IQ <- rnorm(n=10000, mean=100, sd=15) #generate IQ scores
IQ <- round(IQ) # make round numbers
```

Cool, we just generated 10,000 fake IQ scores. Where did they go? Well, they went into the variable IQ on my computer. You can do the same on your computer too by copying the above code. 10,000 numbers is too many numbers to look at. We can look at the first 100 like this:

```
IQ <- rnorm(n=10000, mean=100, sd=15)
IQ <- round(IQ)
print(IQ[1:100])
```

```
  [1]  97  98 101 114 110 105  84  95  96 103  86 118  99  93  64 101 117 104
 [19] 106  73  81  98 100 111 103 100  91 115 107  98 107  76  70 107 104  86
 [37] 120  91 103 129  92  98 105 108  96  87  94  97 102  80  98  76 131 107
 [55] 104 114  90 109 104  86 124  73 131 114 104  83  99  91  83 105 107 107
 [73] 125  74 112  87  76 103 105  88  97  86  99  90 117 121  86 109 132  89
 [91]  97 132  76 131  98 111 118  98  94  98
```

We can compute the mean IQ using the command **mean(IQ)** and the standard deviation using the command **sd(IQ)**, and draw a histogram using **hist()**. The histogram of this much larger sample is shown in Figure 4.8.4*c*. Even a moment's inspections makes clear that the larger sample is a much better approximation to the true population distribution than the smaller one. This is reflected in the sample statistics: the mean IQ for the larger sample turns out to be 99.9, and the standard deviation is 15.1. These values are now very close to the true population.s

I feel a bit silly saying this, but the thing I want you to take away from this is that large samples generally give you better information. I feel silly saying it because it's so bloody obvious that it shouldn't need to be said. In fact, it's such an obvious point that when Jacob Bernoulli – one of the founders of probability theory – formalized this idea back in 1713, he was kind of a jerk about it. Here's how he described the fact that we all share this intuition:

> For even the most stupid of men, by some instinct of nature, by himself and without any instruction (which is a remarkable thing), is convinced that the more observations have been made, the less danger there is of wandering from one's goal (see Stigler, 1986, p65).

Okay, so the passage comes across as a bit condescending (not to mention sexist), but his main point is correct: it really does feel obvious that more data will give you better answers. The question is, why is this so? Not surprisingly, this intuition that we all share turns out to be correct, and statisticians refer to it as the **law of large numbers**. The law of large numbers is a mathematical law that applies to many different sample statistics, but the simplest way to think about it is as a law about averages. The sample mean is the most obvious example of a statistic that relies on averaging (because that's what the mean is... an average), so let's look at that. **When applied to the sample mean, what the law of large numbers states is that as the sample gets larger, the sample mean tends to get closer to the true population mean.** Or, to say it a little bit more precisely, as the sample size "approaches" infinity (written as $N \rightarrow \infty$) the sample mean approaches the population mean ($\bar{X} \rightarrow \mu$).

I don't intend to subject you to a proof that the law of large numbers is true, but it's one of the most important tools for statistical theory. The law of large numbers is the thing we can use to justify our belief that collecting more and more data will eventually lead us to the truth. For any particular data set, the sample statistics that we calculate from it will be wrong, but the law of

large numbers tells us that if we keep collecting more data those sample statistics will tend to get closer and closer to the true population parameters.

This page titled 13.10: The Law of Large Numbers is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

- Front Matter has no license indicated.

13.11: Sampling distributions and the central limit theorem

The law of large numbers is a very powerful tool, but it's not going to be good enough to answer all our questions. Among other things, all it gives us is a "long run guarantee". In the long run, if we were somehow able to collect an infinite amount of data, then the law of large numbers guarantees that our sample statistics will be correct. But as John Maynard Keynes famously argued in economics, a long run guarantee is of little use in real life:

> [The] long run is a misleading guide to current affairs. In the long run we are all dead. Economists set themselves too easy, too useless a task, if in tempestuous seasons they can only tell us, that when the storm is long past, the ocean is flat again. Keynes (1923, 80)

As in economics, so too in psychology and statistics. It is not enough to know that we will **eventually** arrive at the right answer when calculating the sample mean. Knowing that an infinitely large data set will tell me the exact value of the population mean is cold comfort when my **actual** data set has a sample size of $N = 100$. In real life, then, we must know something about the behavior of the sample mean when it is calculated from a more modest data set!

Sampling distribution of the sample means

"Oh no, what is the sample distribution of the sample means? Is that even allowed in English?". Yes, unfortunately, this is allowed. The **sampling distribution of the sample means** is the next most important thing you will need to understand. IT IS SO IMPORTANT THAT IT IS NECESSARY TO USE ALL CAPS. It is only confusing at first because it's long and uses sampling and sample in the same phrase.

Don't worry, we've been prepping you for this. You know what a distribution is right? It's where numbers comes from. It makes some numbers occur more or less frequently, or the same as other numbers. You know what a sample is right? It's the numbers we take from a distribution. So, what could the sampling distribution of the sample means refer to?

First, what do you think the sample means refers to? Well, if you took a sample of numbers, you would have a bunch of numbers...then, you could compute the mean of those numbers. The sample mean is the mean of the numbers in the sample. That is all. So, what is this distribution you speak of? Well, what if you took a bunch of samples, put one here, put one there, put some other ones other places. You have a lot of different samples of numbers. You could compute the mean for each them. Then you would have a bunch of means. What do those means look like? Well, if you put them in a histogram, you could find out. If you did that, you would be looking at (roughly) a distribution, AKA **the sampling distribution of the sample means**.

"I'm following along sort of, why would I want to do this instead of watching Netflix...". Because, the sampling distribution of the sample means gives you another window into chance. A very useful one that you can control, just like your remote control, by pressing the right design buttons.

Seeing the pieces

To make a sampling distribution of the sample means, we just need the following:

1. A distribution to take numbers from
2. A bunch of different samples from the distribution
3. The means of each of the samples
4. Get all of the sample means, and plot them in a histogram

Question

Question for yourself: What do you think the sampling distribution of the sample means will look like? Will it tend to look the shape of the distribution that the samples came from? Or not? Good question, think about it.

Let's do those four things. We will sample numbers from the uniform distribution, it looks like this if we are sampling from the set

of integers from 1 to 10:

```
library(ggplot2)
df<-data.frame(a=1:10,b=seq(.1,1,.1))
df$a<-as.factor(df$a)
ggplot(df,aes(x=a,y=b))+
  geom_point(color="white")+
  geom_hline(yintercept=.1)+
  theme_classic()+
  ylab("Probability")+
  xlab("Number")+
  ggtitle("Uniform distribution for numbers 1 to 10")
```

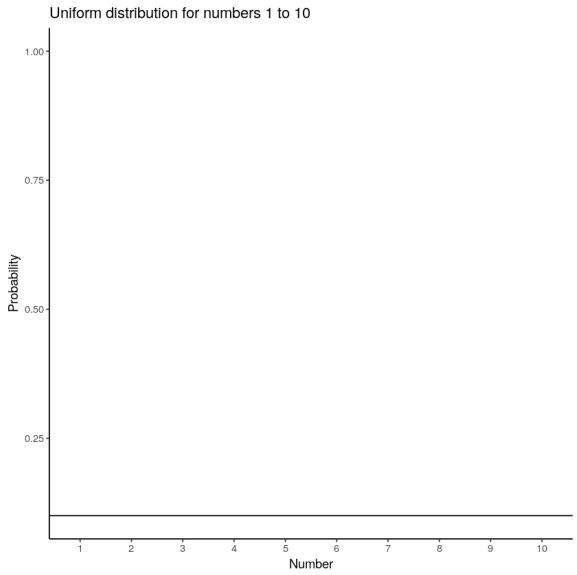


Figure 13.11.1: A uniform distribution illustrating the probabilites of sampling the numbers 1 to 10. In a uniform distribution, all numbers have an equal probability of being sampled, so the line is flat indicating all numbers have the same probability.

OK, now let's take a bunch of samples from that distribution. We will set our sample-size to 20. It's easier to see how the sample mean behaves in a movie. Each histogram shows a new sample. The red line shows where the mean of the sample is. The samples are all very different from each other, but the red line doesn't move around very much, it always stays near the middle. However, the red line does move around a little bit, and this variance is what we call the sampling distribution of the sample mean.

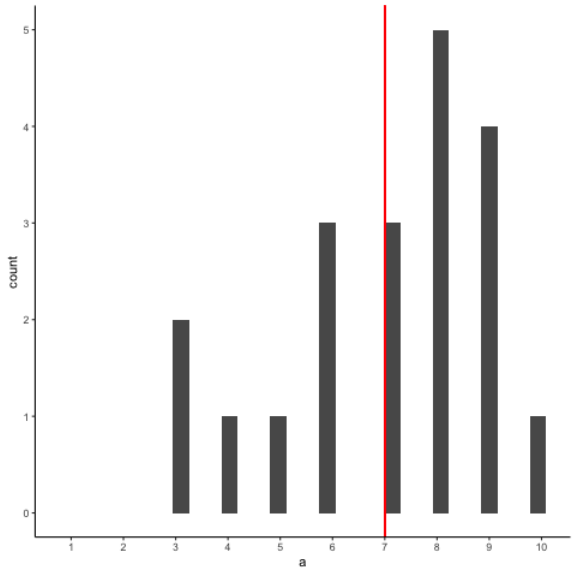


Figure 13.11.2: Animiation showing histograms for different samples of size

20 from the uniform distribution. The red line shows the mean of each sample.

OK, what have we got here? We have an animiation of 10 different samples. Each sample has 20 observations and these are summarized in each of histograms that show up in the animiation. Each histogram has a red line. The red line shows you where the mean of each sample is located. So, we have found the sample means for the 10 different samples from a uniform distribution.

First question. Are the sample means all the same? The answer is no. They are all kind of similar to each other though, they are all around five plus or minus a few numbers. This is interesting. Although all of our samples look pretty different from one another, the means of our samples look more similar than different.

Second question. What should we do with the means of our samples? Well, how about we collect them them all, and then plot a histogram of them. This would allow us to see what the distribution of the sample means looks like. The next histogram is just this. Except, rather than taking 10 samples, we will take 10,000 samples. For each of them we will compute the means. So, we will have 10,000 means. This is the histogram of the sample means:

```
library(ggplot2)
a<-round(runif(20*10000,1,10))
df<-data.frame(a,sample=rep(1:10000,each=20))
df2<-aggregate(a~sample,df,mean)
ggplot(df2, aes(x=a))+
  geom_histogram(color="white", bins=30)+
  theme_classic()+
  ggtitle("Histogram of 10,000 sample means")+
  xlab("value")
```

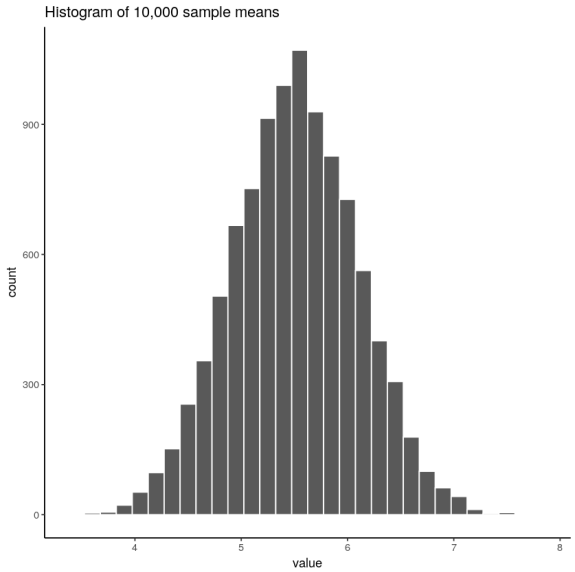


Figure 13.11.3: A histogram showing the sample means for 10,000 samples, each size 20, from the uniform distribution of numbers from 1 to 10. The expected mean is 5.5, and the histogram is centered on 5.5. The mean of each sample is not always 5.5 because of sampling error or chance.

“Wait what? This doesn’t look right. I thought we were taking samples from a uniform distribution. Uniform distributions are flat. THIS DOES NOT LOOK LIKE A FLAT DISTRIBTUION, WHAT IS GOING ON, AAAAAGGGHH.” We feel your pain.

Remember, we are looking at the distribution of sample means. It is indeed true that the distribution of sample means does not look the same as the distribution we took the samples from. Our distribution of sample means goes up and down. In fact, this will almost always be the case for distributions of sample means. This fact is called the **central limit theorem**, which we talk about later.

For now, let’s talk about about what’s happening. Remember, we have been sampling numbers between the range 1 to 10. We are supposed to get each number with roughly equal frequency, because we are sampling from a uniform distribution. So, let’s say we took a sample of 10 numbers, and happened to get one of each from 1 to 10.

```
1 2 3 4 5 6 7 8 9 10
```

What is the mean of those numbers? Well, its 1+2+3+4+5+6+7+8+9+10 = 55 / 10 = 5.5. Imagine if we took a bigger sample, say of 20 numbers, and again we got exactly 2 of each number. What would the mean be? It would be (1+2+3+4+5+6+7+8+9+10)*2 = 110 / 20 = 5.5. Still 5.5. You can see here, that the mean value of our uniform distribution is 5.5. Now that we know this, we might expect that most of our samples will have a mean near this number. We already know that every sample won't be perfect, and it won't have exactly an equal amount of every number. So, we will expect the mean of our samples to vary a little bit. The histogram that we made shows the variation. Not surprisingly, the numbers vary around the value 5.5.

Sampling distributions exist for any sample statistic!

One thing to keep in mind when thinking about sampling distributions is that **any** sample statistic you might care to calculate has a sampling distribution. For example, suppose that each time you sampled some numbers from an experiment you wrote down the largest number in the experiment. Doing this over and over again would give you a very different sampling distribution, namely the **sampling distribution of the maximum**. You could calculate the smallest number, or the mode, or the median, of the variance, or the standard deviation, or anything else from your sample. Then, you could repeat many times, and produce the sampling distribution of those statistics. Neat!

Just for fun here are some different sampling distributions for different statistics. We will take a normal distribution with mean = 100, and standard deviation =20. Then, we'll take lots of samples with n = 50 (50 observations per sample). We'll save all of the sample statistics, then plot their histograms. Let's do it:

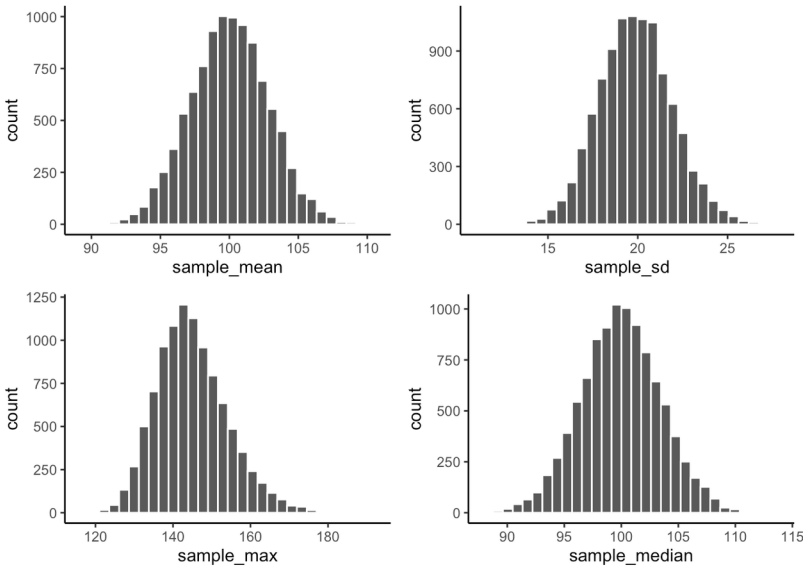


Figure 13.11.4: Each panel shows a histogram of a different sampling statistic.

We just computed 4 different sampling distributions, for the mean, standard deviation, maximum value, and the median. If you just look quickly at these histograms you might think they all basically look the same. Hold up now. It's very important to look at the x-axes. They are different. For example, the sample mean goes from about 90 to 110, whereas the standard deviation goes from 15 to 25.

These sampling distributions are super important, and worth thinking about. What should you think about? Well, here's a clue. These distributions are telling you what to expect from your sample. Critically, they are telling you what you should expect from a sample, when you take one from the specific distribution that we used (normal distribution with mean =100 and SD = 20). What have we learned. We've learned a tonne. We've learned that we can expect our sample to have a mean somewhere between 90 and 108ish. Notice, the sample means are never more extreme. We've learned that our sample will usually have some variance, and that the the standard deviation will be somewhere between 15 and 25 (never much more extreme than that). We can see that sometime we get some big numbers, say between 120 and 180, but not much bigger than that. And, we can see that the median is pretty similar to the mean. If you ever took a sample of 50 numbers, and your descriptive statistics were inside these windows, then perhaps they came from this kind of normal distribution. If your sample statistics are very different, then your sample probably did not come this distribution. By using simulation, we can find out what samples look like when they come from

distributions, and we can use this information to make inferences about whether our sample came from particular distributions.

This page titled 13.11: Sampling distributions and the central limit theorem is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

- Back Matter has no license indicated.

13.12: The Central Limit Theorem

OK, so now you've seen lots of sampling distributions, and you know what the sampling distribution of the mean is. Here, we'll focus on **how the sampling distribution of the mean changes as a function of sample size.**

Intuitively, you already know part of the answer: if you only have a few observations, the sample mean is likely to be quite inaccurate (you've already seen it bounce around): if you replicate a small experiment and recalculate the mean you'll get a very different answer. In other words, the sampling distribution is quite wide. If you replicate a large experiment and recalculate the sample mean you'll probably get the same answer you got last time, so the sampling distribution will be very narrow.

Let's give ourselves a nice movie to see everything in action. We're going to sample numbers from a normal distribution. You will see four panels, each panel represents a different sample size (n), including sample-sizes of 10, 50, 100, and 1000. The red line shows the shape of the normal distribution. The grey bars show a histogram of each of the samples that we take. The red line shows the mean of an individual sample (the middle of the grey bars). As you can see, the red line moves around a lot, especially when the sample size is small (10).

The new bits are the blue bars and the blue lines. The blue bars represent the sampling distribution of the sample mean. For example, in the panel for sample-size 10, we see a bunch of blue bars. This is a histogram of 10 sample means, taken from 10 samples of size 10. In the 50 panel, we see a histogram of 50 sample means, taken from 50 samples of size 50, and so on. The blue line in each panel is the mean of the sample means ("aaagh, it's a mean of means", yes it is).

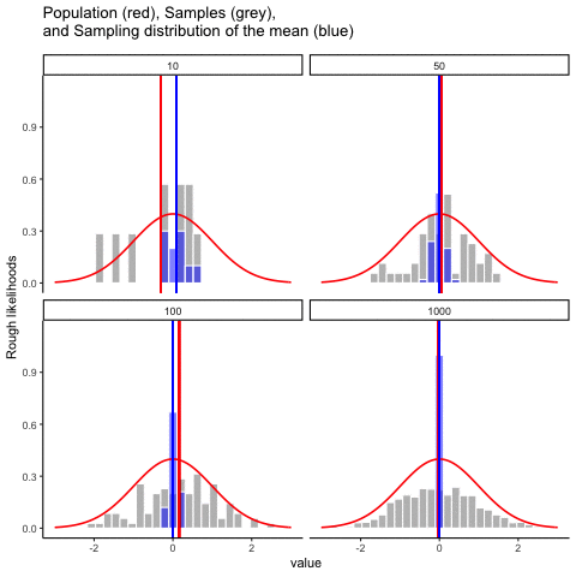


Figure 13.12.1: Animation of samples (grey histogram shows frequency counts of data in each sample), and the sampling distribution of the mean (histogram of the sampling means for many samples). Each sample is taken from the normal distribution shown in red. The moving red line is the mean of an individual sample. The blue line is the mean of the blue histogram, which represents the sampling distribution of the mean for many samples.

What should you notice? Notice that the range of the blue bars shrinks as sample size increases. The sampling distribution of the mean is quite wide when the sample-size is 10, it narrows as sample-size increases to 50 and 100, and it's just one bar, right in the middle when sample-size goes to 1000. What we are seeing is that the mean of the sampling distribution approaches the mean of the population as sample-size increases.

So, the sampling distribution of the mean is another distribution, and it has some variance. It varies more when sample-size is small, and varies less when sample-size is large. We can quantify this effect by calculating the standard deviation of the sampling distribution, which is referred to as the **standard error**. The standard error of a statistic is often denoted SE, and since we're usually interested in the standard error of the sample **mean**, we often use the acronym SEM. As you can see just by looking at the movie, as the sample size N increases, the SEM decreases.

Okay, so that's one part of the story. However, there's something we've been glossing over a little bit. We've seen it already, but

it's worth looking at it one more time. Here's the thing: **no matter what shape your population distribution is, as N increases the sampling distribution of the mean starts to look more like a normal distribution**. This is the central limit theorem.

To see the central limit theorem in action, we are going to look at some histograms of sample means different kinds of distributions. It is very important to recognize that you are looking at distributions of sample means, not distributions of individual samples! Here we go, starting with sampling from a normal distribution. The red line is the distribution, the blue bars are the histogram for the sample means. They both look normal!

```
library(ggplot2)
options(warn=-1)
get_sampling_means<-function(m,sd,s_size,iter){
  save_means<-length(iter)
  for(i in 1:iter){
    save_means[i]<-mean(rnorm(s_size,m,sd))
  }
  return(save_means)
}
all_df<-data.frame()
sims<-1
n<-50
for(n in c(10,50)){
    sample<-rnorm(n,0,1)
    sample_means<-get_sampling_means(0,1,n,1000)
    t_df<-data.frame(sims=rep(sims,1000),
                     sample,
                     sample_means,
                     sample_size=rep(n,1000),
                     sample_mean=rep(mean(sample),1000),
                     sampling_mean=rep(mean(sample_means),1000)
                     )
    all_df<-rbind(all_df,t_df)
}
ggplot(all_df, aes(x=sample))+
  geom_histogram(aes(x=sample_means,y=(..density..)/max(..density..)),fill="blue",color="w
  stat_function(fun = dnorm,
                args = list(mean = 0, sd = 1),
                lwd = .75,
                col = 'red')+
  #geom_vline(aes(xintercept=sampling_mean,frame=sims),color="blue")+
  facet_wrap(~sample_size)+xlim(-3,3)+
  theme_classic()+ggtitle("Sampling distribution of mean \n for Normal Distribution")+ylab
  xlab("value")
```

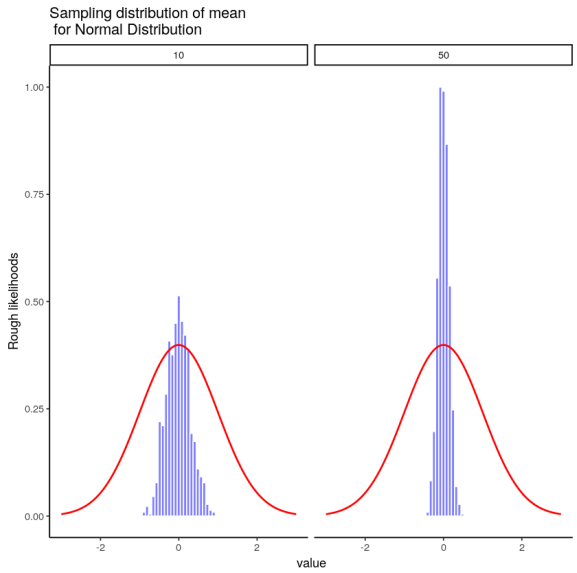


Figure 13.12.2: Comparison of two normal distributions, and histograms for the sampling distribution of the mean for different samples-sizes. The range of sampling distribution of the mean shrinks as sample-size increases.

Let's do it again. This time we sample from a flat uniform distribution. Again, we see that the distribution of the sample means is not flat, it looks like a normal distribution.

```
library(ggplot2)
options(warn=-1)
get_sampling_means<-function(mn,mx,s_size,iter){
  save_means<-length(iter)
  for(i in 1:iter){
    save_means[i]<-mean(runif(s_size,mn,mx))
  }
  return(save_means)
}
all_df<-data.frame()
sims<-1
n<-50
for(n in c(10,50)){
    sample<-rnorm(n,0,1)
    sample_means<-get_sampling_means(0,1,n,1000)
    t_df<-data.frame(sims=rep(sims,1000),
                     sample,
                     sample_means,
                     sample_size=rep(n,1000),
                     sample_mean=rep(mean(sample),1000),
                     sampling_mean=rep(mean(sample_means),1000)
                     )
    all_df<-rbind(all_df,t_df)
}
ggplot(all_df, aes(x=sample))+
  geom_histogram(aes(x=sample_means,y=(..density..)/max(..density..)),fill="blue",color="w
  geom_hline(yintercept=.1,color="red")+
  facet_wrap(~sample_size)+xlim(0,1)+
  theme_classic()+ggtitle("Sampling distribution of mean \n for samples taken from Uniform
  xlab("value")
```

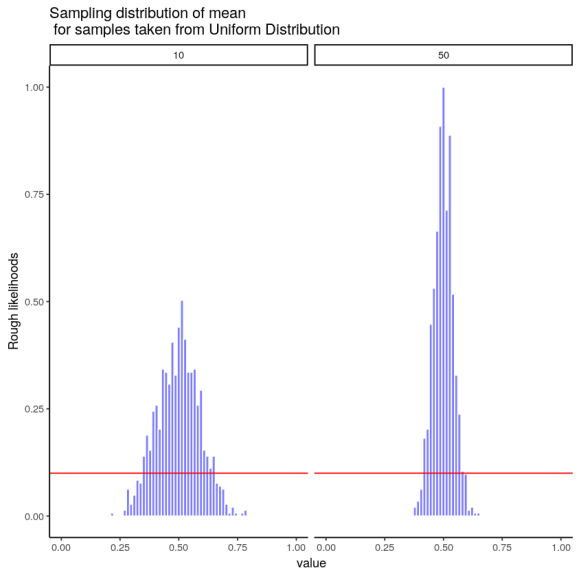


Figure 13.12.3: Illustration that the shape of the sampling distribution of the mean is normal, even when the samples come from a non-normal (uniform in this case) distribution.

One more time with an exponential distribution. Even though way more of the numbers should be smaller than bigger, then sampling distribution of the mean again does not look the red line. Instead, it looks more normal-ish. That's the central limit theorem. It just works like that.

```
library(ggplot2)
options(warn=-1)
get_sampling_means<-function(s_size,r,iter){
  save_means<-length(iter)
  for(i in 1:iter){
    save_means[i]<-mean(rexp(s_size,r))
  }
  return(save_means)
}
all_df<-data.frame()
sims<-1
n<-50
for(n in c(10,50)){
    sample<-rnorm(n,0,1)
    sample_means<-get_sampling_means(n,2,1000)
    t_df<-data.frame(sims=rep(sims,1000),
                     sample,
                     sample_means,
                     sample_size=rep(n,1000),
                     sample_mean=rep(mean(sample),1000),
                     sampling_mean=rep(mean(sample_means),1000)
                     )
    all_df<-rbind(all_df,t_df)
}
ggplot(all_df, aes(x=sample))+
  geom_histogram(aes(x=sample_means,y=(..density..)/max(..density..)),fill="blue",color="w
  stat_function(fun = dexp,
                args = list(rate=2),
                lwd = .75,
                col = 'red')+
  #geom_vline(aes(xintercept=sampling_mean,frame=sims),color="blue")+
  facet_wrap(~sample_size)+xlim(0,1)+
  theme_classic()+ggtitle("Sampling distribution of mean \n for samples from exponential D
  xlab("value")
```

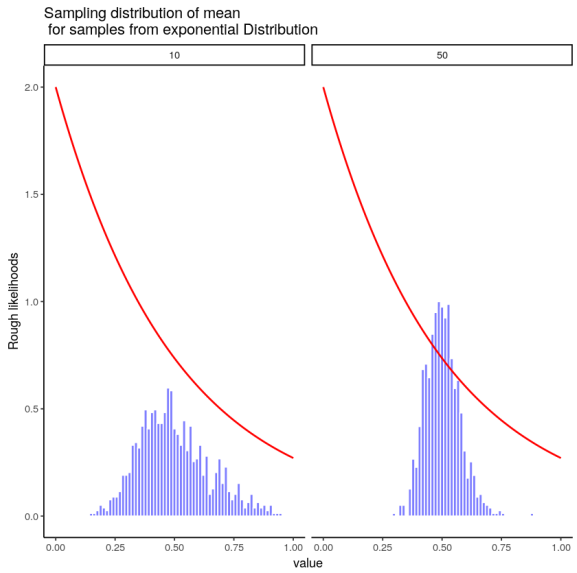


Figure 13.12.4: Illustration that the shape of the sampling distribution of the mean is normal, even when the samples come from a non-normal (exponential in this case) distribution.

On the basis of these figures, it seems like we have evidence for all of the following claims about the sampling distribution of the mean:

- The mean of the sampling distribution is the same as the mean of the population
- The standard deviation of the sampling distribution (i.e., the standard error) gets smaller as the sample size increases
- The shape of the sampling distribution becomes normal as the sample size increases

As it happens, not only are all of these statements true, there is a very famous theorem in statistics that proves all three of them, known as the **central limit theorem**. Among other things, the central limit theorem tells us that if the population distribution has mean μ and standard deviation σ, then the sampling distribution of the mean also has mean μ, and the standard error of the mean is

$$\text{SEM} = \frac{\sigma}{\sqrt{N}}$$

Because we divide the population standard deviation σ by the square root of the sample size N, the SEM gets smaller as the sample size increases. It also tells us that the shape of the sampling distribution becomes normal.

This result is useful for all sorts of things. It tells us why large experiments are more reliable than small ones, and because it gives us an explicit formula for the standard error it tells us **how much** more reliable a large experiment is. It tells us why the normal distribution is, well, **normal**. In real experiments, many of the things that we want to measure are actually averages of lots of different quantities (e.g., arguably, "general" intelligence as measured by IQ is an average of a large number of "specific" skills and abilities), and when that happens, the averaged quantity should follow a normal distribution. Because of this mathematical law, the normal distribution pops up over and over again in real data.

This page titled 13.12: The Central Limit Theorem is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

- Front Matter has no license indicated.

13.13: z-scores

We are now in a position to combine some of things we've been talking about in this chapter, and introduce you to a new tool, **z-scores**. It turns out we won't use **z-scores** very much in this textbook. However, you can't take a class on statistics and not learn about **z-scores**.

The first thing we show you seems to be something that many students remember from their statistics class. This thing is probably remembered because instructors may test this knowledge many times, so students have to learn it for the test. Let's look at this thing. We are going to look at a normal distribution, and we are going to draw lines through the distribution at 0, +/- 1, +/-2, and +/- 3 standard deviations from the mean:

```
library(ggplot2)
dnorm_vec <- dnorm(seq(-5,5,.1),mean=0,sd=1)
x_range   <- seq(-5,5,.1)
t_df<-data.frame(x_range,dnorm_vec)
ggplot(t_df, aes(x=x_range,y=dnorm_vec))+
  geom_line()+
  geom_vline(xintercept = 0)+
  geom_vline(xintercept = c(-3,-2,-1,1,2,3))+
  theme_classic()+
  ylab("Density")+
  xlab("score") +
  scale_x_continuous(breaks=seq(-5,5,1))+
  geom_label(data = data.frame(x=-.5, y=.3,
            label=round(pnorm(c(0,1),0,1)[2]-pnorm(c(0,1),0,1)[1], digits=3)),
            aes(x = x, y = y, label = label))+
  geom_label(data = data.frame(x=.5, y=.3,
            label=round(pnorm(c(0,1),0,1)[2]-pnorm(c(0,1),0,1)[1], digits=3)),
            aes(x = x, y = y, label = label))+
    geom_label(data = data.frame(x=-1.5, y=.3,
            label=round(pnorm(c(1,2),0,1)[2]-pnorm(c(1,2),0,1)[1], digits=3)),
            aes(x = x, y = y, label = label))+
  geom_label(data = data.frame(x=1.5, y=.3,
            label=round(pnorm(c(1,2),0,1)[2]-pnorm(c(1,2),0,1)[1], digits=3)),
            aes(x = x, y = y, label = label)) +
  geom_label(data = data.frame(x=-2.5, y=.3,
            label=round(pnorm(c(2,3),0,1)[2]-pnorm(c(2,3),0,1)[1], digits=3)),
            aes(x = x, y = y, label = label))+
  geom_label(data = data.frame(x=2.5, y=.3,
            label=round(pnorm(c(2,3),0,1)[2]-pnorm(c(2,3),0,1)[1], digits=3)),
            aes(x = x, y = y, label = label))+
  geom_label(data = data.frame(x=-3.5, y=.3,
            label=round(pnorm(c(3,4),0,1)[2]-pnorm(c(3,4),0,1)[1], digits=3)),
            aes(x = x, y = y, label = label))+
  geom_label(data = data.frame(x=3.5, y=.3,
            label=round(pnorm(c(3,4),0,1)[2]-pnorm(c(3,4),0,1)[1], digits=3)),
            aes(x = x, y = y, label = label))
```

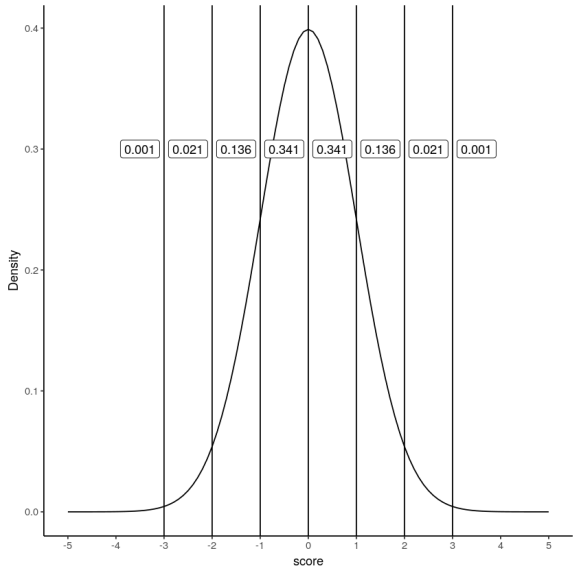


Figure 13.13.1: A normal distribution. Each line represents a standard deviation from the mean. The labels show the proportions of scores that fall between each bar.

The figure shows a normal distribution with mean = 0, and standard deviation = 1. We've drawn lines at each of the standard deviations: -3, -2, -1, 0, 1, 2, and 3. We also show some numbers in the labels, in between each line. These numbers are proportions. For example, we see the proportion is .341 for scores that fall between the range 0 and 1. Scores between 0 and 1 occur 34.1% of the time. Scores in between -1 and 1, occur 68.2% of the time, that's more than half of the scores. Scores between 1 and occur about 13.6% of the time, and scores between 2 and 3 occur even less, only 2.1% of the time.

Normal distributions always have these properties, even when they have different means and standard deviations. For example, take a look at this normal distribution, it has a mean =100, and standard deviation =25.

```
library(ggplot2)
dnorm_vec <- dnorm(seq(0,200,.1),mean=100,sd=25)
x_range   <- seq(0,200,.1)
t_df<-data.frame(x_range,dnorm_vec)
ggplot(t_df, aes(x=x_range,y=dnorm_vec))+
  geom_line()+
  geom_vline(xintercept = 100)+
  geom_vline(xintercept = c(25,50,75,125,150,175))+
  theme_classic()+
  ylab("Density")+
  xlab("score") +
  scale_x_continuous(breaks=seq(0,200,25))+
  geom_label(data = data.frame(x=87.5, y=0.01,
            label=round(pnorm(c(0,1),0,1)[2]-pnorm(c(0,1),0,1)[1], digits=3)),
            aes(x = x, y = y, label = label))+
  geom_label(data = data.frame(x=112.5, y=0.01,
            label=round(pnorm(c(0,1),0,1)[2]-pnorm(c(0,1),0,1)[1], digits=3)),
            aes(x = x, y = y, label = label))+
    geom_label(data = data.frame(x=62.5, y=0.01,
            label=round(pnorm(c(1,2),0,1)[2]-pnorm(c(1,2),0,1)[1], digits=3)),
            aes(x = x, y = y, label = label))+
  geom_label(data = data.frame(x=137.5, y=0.01,
            label=round(pnorm(c(1,2),0,1)[2]-pnorm(c(1,2),0,1)[1], digits=3)),
            aes(x = x, y = y, label = label)) +
  geom_label(data = data.frame(x=37.5, y=0.01,
            label=round(pnorm(c(2,3),0,1)[2]-pnorm(c(2,3),0,1)[1], digits=3)),
            aes(x = x, y = y, label = label))+
  geom_label(data = data.frame(x=162.5, y=0.01,
            label=round(pnorm(c(2,3),0,1)[2]-pnorm(c(2,3),0,1)[1], digits=3)),
            aes(x = x, y = y, label = label))+
```

```
  geom_label(data = data.frame(x=12.5, y=0.01,
            label=round(pnorm(c(3,4),0,1)[2]-pnorm(c(3,4),0,1)[1], digits=3)),
            aes(x = x, y = y, label = label))+
  geom_label(data = data.frame(x=187.5, y=0.01,
            label=round(pnorm(c(3,4),0,1)[2]-pnorm(c(3,4),0,1)[1], digits=3)),
            aes(x = x, y = y, label = label))
```

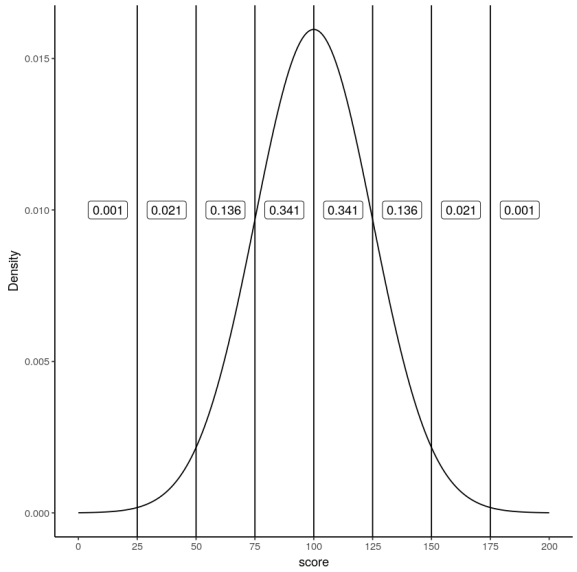


Figure 13.13.2: A normal distribution. Each line represents a standard deviation from the mean. The labels show the proportions of scores that fall between each bar.

Now we are looking at a normal distribution with mean = 100 and standard deviation = 25. Notice that the region between 100 and 125 contains 34.1% of the scores. This region is 1 standard deviation away from the mean (the standard deviation is 25, the mean is 100, so 25 is one whole standard deviation away from 100). As you can see, the very same proportions occur between each of the standard deviations, as they did when our standard deviation was set to 1 (with a mean of 0).

Idea behind z-scores

Sometimes it can be convenient to transform your original scores into different scores that are easier to work with. For example, if you have a bunch of proportions, like .3, .5, .6, .7, you might want to turn them into percentages like 30%, 50%, 60%, and 70%. To do that you multiply the proportions by a constant of 100. If you want to turn percentages back into proportions, you divide by a constant of 100. This kind of transformation just changes the scale of the numbers from between 0-1, and between 0-100. Otherwise, the pattern in the numbers stays the same.

The idea behind z-scores is a similar kind of transformation. The idea is to express each raw score in terms of it's standard deviation. For example, if I told you I got a 75% on test, you wouldn't know how well I did compared to the rest of the class. But, if I told you that I scored 2 standard deviations above the mean, you'd know I did quite well compared to the rest of the class, because you know that most scores (if they are distributed normally) fall below 2 standard deviations of the mean.

We also know, now thanks to the central limit theorem, that many of our measures, such as sample means, will be distributed normally. So, it can often be desirable to express the raw scores in terms of their standard deviations.

Let's see how this looks in a table without showing you any formulas. We will look at some scores that come from a normal distirbution with mean =100, and standard deviation = 25. We will list some raw scores, along with the z-scores

raw	z
25	-3
50	-2

raw	z
75	-1
100	0
125	1
150	2
175	3

Remember, the mean is 100, and the standard deviation is 25. How many standard deviations away from the mean is a score of 100? The answer is 0, it's right on the mean. You can see the z-score for 100, is 0. How many standard deviations is 125 away from the mean? Well the standard deviation is 25, 125 is one whole 25 away from 100, that's a total of 1 standard deviation, so the z-score for 125 is 1. The z-score for 150 is 2, because 150 is two 25s away from 100. The z-score for 50 is -2, because 50 is two 25s away from 100 in the opposite direction. All we are doing here is re-expressing the raw scores in terms of how many standard deviations they are from the mean. Remember, the mean is always right on target, so the center of the z-score distribution is always 0.

Calculating z-scores

To calculate z-scores all you have to do is figure out how many standard deviations from the mean each number is. Let's say the mean is 100, and the standard deviation is 25. You have a score of 97. How many standard deviations from the mean is 97?

First compute the difference between the score and the mean:

$$97 - 100 = -3$$

Alright, we have a total difference of -3. How many standard deviations does -3 represent if 1 standard deviation is 25? Clearly -3 is much smaller than 25, so it's going to be much less than 1. To figure it out, just divide -3 by the standard deviation.

$$\frac{-3}{25} = -.12$$

Our z-score for 97 is -.12.

Here's the general formula:

$$z = \frac{\text{raw score} - \text{mean}}{\text{standard deviation}}$$

So, for example if we had these 10 scores from a normal distribution with mean = 100, and standard deviation =25

```
72.23  73.48  96.25  91.60  56.84 105.56 128.96  91.33  70.96 120.23
```

The z-scores would be:

```
-1.1108 -1.0608 -0.1500 -0.3360 -1.7264  0.2224  1.1584 -0.3468
-1.1616  0.8092
```

Once you have the z-scores, you could use them as another way to describe your data. For example, now just by looking at a score you know if it is likely or unlikely to occur, because you know how the area under the normal curve works. z-scores between -1 and 1 happen pretty often, scores greater than 1 or -1 still happen fairly often, but not as often. And, scores bigger than 2 or -2 don't happen very often. This is a convenient thing to do if you want to look at your numbers and get a general sense of how often they happen.

Usually you do not know the mean or the standard deviation of the population that you are drawing your sample scores from. So, you could use the mean and standard deviation of your sample as an estimate, and then use those to calculate z-scores.

Finally, z-scores are also called **standardized scores**, because each raw score is described in terms of it's standard deviation. This may well be the last time we talk about z-scores in this book. You might wonder why we even bothered telling you about them.

First, it's worth knowing they are a thing. Second, they become important as your statistical prowess becomes more advanced. Third, some statistical concepts, like correlation, can be re-written in terms of z-scores, and this illuminates aspects of those statistics. Finally, they are super useful when you are dealing with a normal distribution that has a known mean and standard deviation.

This page titled 13.13: z-scores is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

- Back Matter has no license indicated.

13.14: Estimating population parameters

Let's pause for a moment to get our bearings. We're about to go into the topic of **estimation**. What is that, and why should you care? First, population parameters are things about a distribution. For example, distributions have means. The mean is a parameter of the distribution. The standard deviation of a distribution is a parameter. Anything that can describe a distribution is a potential parameter.

OK fine, who cares? This I think, is a really good question. There are some good concrete reasons to care. And there are some great abstract reasons to care. Unfortunately, most of the time in research, it's the abstract reasons that matter most, and these can be the most difficult to get your head around.

Concrete population parameters

First some concrete reasons. There are real populations out there, and sometimes you want to know the parameters of them. For example, if you are a shoe company, you would want to know about the population parameters of feet size. As a first pass, you would want to know the mean and standard deviation of the population. If your company knew this, and other companies did not, your company would do better (assuming all shoes are made equal). Why would your company do better, and how could it use the parameters? Here's one good reason. As a shoe company you want to meet demand with the right amount of supply. If you make too many big or small shoes, and there aren't enough people to buy them, then you're making extra shoes that don't sell. If you don't make enough of the most popular sizes, you'll be leaving money on the table. Right? Yes. So, what would be an optimal thing to do? Perhaps, you would make different amounts of shoes in each size, corresponding to how the demand for each shoe size. You would know something about the demand by figuring out the frequency of each size in the population. You would need to know the population parameters to do this.

Fortunately, it's pretty easy to get the population parameters without measuring the entire population. Who has time to measure every-bodies feet? Nobody, that's who. Instead, you would just need to randomly pick a bunch of people, measure their feet, and then measure the parameters of the sample. If you take a big enough sample, we have learned that the sample mean gives a very good estimate of the population mean. We will learn shortly that a version of the standard deviation of the sample also gives a good estimate of the standard deviation of the population. Perhaps shoe-sizes have a slightly different shape than a normal distribution. Here too, if you collect a big enough sample, the shape of the distribution of the sample will be a good estimate of the shape of the populations. All of these are good reasons to care about estimating population parameters. But, do you run a shoe company? Probably not.

Abstract population parameters

Even when we think we are talking about something concrete in Psychology, it often gets abstract right away. Instead of measuring the population of feet-sizes, how about the population of human happiness. We all think we know what happiness is, everyone has more or less of it, there are a bunch of people, so there must be a population of happiness right? Perhaps, but it's not very concrete. The first problem is figuring out how to measure happiness. Let's use a questionnaire. Consider these questions:

> How happy are you right now on a scale from 1 to 7? How happy are you in general on a scale from 1 to 7? How happy are you in the mornings on a scale from 1 to 7? How happy are you in the afternoons on a scale from 1 to 7?

1. = very unhappy
2. = unhappy
3. = sort of unhappy
4. = in the middle
5. = sort of happy
6. = happy
7. = very happy

Forget about asking these questions to everybody in the world. Let's just ask them to lots of people (our sample). What do you think would happen? Well, obviously people would give all sorts of answers right. We could tally up the answers and plot them in a histogram. This would show us a distribution of happiness scores from our sample. "Great, fantastic!", you say. Yes, fine and

dandy.

So, on the one hand we could say lots of things about the people in our sample. We could say exactly who says they are happy and who says they aren't, after all they just told us!

But, what can we say about the larger population? Can we use the parameters of our sample (e.g., mean, standard deviation, shape etc.) to estimate something about a larger population. Can we infer how happy everybody else is, just from our sample? HOLD THE PHONE.

Complications with inference

Before listing a bunch of complications, let me tell you what I think we can do with our sample. Provided it is big enough, our sample parameters will be a pretty good estimate of what another sample would look like. Because of the following discussion, this is often all we can say. But, that's OK, as you see throughout this book, we can work with that!

Problem 1: Multiple populations: If you looked at a large sample of questionnaire data you will find evidence of multiple distributions inside your sample. People answer questions differently. Some people are very cautious and not very extreme. Their answers will tend to be distributed about the middle of the scale, mostly 3s, 4s, and 5s. Some people are very bi-modal, they are very happy and very unhappy, depending on time of day. These people's answers will be mostly 1s and 2s, and 6s and 7s, and those numbers look like they come from a completely different distribution. Some people are entirely happy or entirely unhappy. Again, these two "populations" of people's numbers look like two different distributions, one with mostly 6s and 7s, and one with mostly 1s and 2s. Other people will be more random, and their scores will look like a uniform distribution. So, is there a single population with parameters that we can estimate from our sample? Probably not. Could be a mixture of lots of populations with different distributions.

Problem 2: What do these questions measure?: If the whole point of doing the questionnaire is to estimate the population's happiness, we really need wonder if the sample measurements actually tell us anything about happiness in the first place. Some questions: Are people accurate in saying how happy they are? Does the measure of happiness depend on the scale, for example, would the results be different if we used 0-100, or -100 to +100, or no numbers? Does the measure of happiness depend on the wording in the question? Does a measure like this one tell us everything we want to know about happiness (probably not), what is it missing (who knows? probably lots). In short, nobody knows if these kinds of questions measure what we want them to measure. We just hope that they do. Instead, we have a very good idea of the kinds of things that they actually measure. It's really quite obvious, and staring you in the face. Questionnaire measurements measure how people answer questionnaires. In other words, how people behave and answer questions when they are given a questionnaire. This might also measure something about happiness, when the question has to do about happiness. But, it turns out people are remarkably consistent in how they answer questions, even when the questions are total nonsense, or have no questions at all (just numbers to choose!) Maul (2017).

The take home complications here are that we can collect samples, but in Psychology, we often don't have a good idea of the populations that might be linked to these samples. There might be lots of populations, or the populations could be different depending on who you ask. Finally, the "population" might not be the one you want it to be.

Experiments and Population parameters

OK, so we don't own a shoe company, and we can't really identify the population of interest in Psychology, can't we just skip this section on estimation? After all, the "population" is just too weird and abstract and useless and contentious. HOLD THE PHONE AGAIN!

It turns out we can apply the things we have been learning to solve lots of important problems in research. These allow us to answer questions with the data that we collect. Parameter estimation is one of these tools. We just need to be a little bit more creative, and a little bit more abstract to use the tools.

Here is what we know already. The numbers that we measure come from somewhere, we have called this place "distributions". Distributions control how the numbers arrive. Some numbers happen more than others depending on the distribution. We assume, even if we don't know what the distribution is, or what it means, that the numbers came from one. Second, when get some numbers, we call it a sample. This entire chapter so far has taught you one thing. When your sample is big, it resembles the distribution it came from. And, when your sample is big, it will resemble very closely what another big sample of the same thing will look like. We can use this knowledge!

Very often as Psychologists what we want to know is what causes what. We want to know if X causes something to change in Y. Does eating chocolate make you happier? Does studying improve your grades? There a bazillions of these kinds of questions. And, we want answers to them.

I've been trying to be mostly concrete so far in this textbook, that's why we talk about silly things like chocolate and happiness, at least they are concrete. Let's give a go at being abstract. We can do it.

So, we want to know if X causes Y to change. What is X? What is Y? X is something you change, something you manipulate, the independent variable. Y is something you measure. So, we will be taking samples from Y. "Oh I get it, we'll take samples from Y, then we can use the sample parameters to estimate the population parameters of Y!" NO, not really, but yes sort of. We will take sample from Y, that is something we absolutely do. In fact, that is really all we ever do, which is why talking about the population of Y is kind of meaningless. We're more interested in our samples of Y, and how they behave.

So, what would happen if we removed X from the universe altogether, and then took a big sample of Y. We'll pretend Y measures something in a Psychology experiment. So, we know right away that Y is variable. When we take a big sample, it will have a distribution (because Y is variable). So, we can do things like measure the mean of Y, and measure the standard deviation of Y, and anything else we want to know about Y. Fine. What would happen if we replicated this measurement. That is, we just take another random sample of Y, just as big as the first. What should happen is that our first sample should look a lot like our second example. After all, we didn't do anything to Y, we just took two big samples twice. Both of our samples will be a little bit different (due to sampling error), but they'll be mostly the same. The bigger our samples, the more they will look the same, especially when we don't do anything to cause them to be different. In other words, we can use the parameters of one sample to estimate the parameters of a second sample, because they will tend to be the same, especially when they are large.

We are now ready for step two. You want to know if X changes Y. What do you do? You make X go up and take a big sample of Y then look at it. You make X go down, then take a second big sample of Y and look at it. Next, you compare the two samples of Y. If X does nothing then what should you find? We already discussed that in the previous paragraph. If X does nothing, then both of your big samples of Y should be pretty similar. However, if X does something to Y, then one of your big samples of Y will be different from the other. You will have changed something about Y. Maybe X makes the mean of Y change. Or maybe X makes the variation in Y change. Or, maybe X makes the whole shape of the distribution change. If we find any big changes that can't be explained by sampling error, then we can conclude that something about X caused a change in Y! We could use this approach to learn about what causes what!

The very important idea is still about estimation, just not population parameter estimation exactly. We know that when we take samples they naturally vary. So, when we estimate a parameter of a sample, like the mean, we know we are off by some amount. When we find that two samples are different, we need to find out if the size of the difference is consistent with what sampling error can produce, or if the difference is bigger than that. If the difference is bigger, then we can be confident that sampling error didn't produce the difference. So, we can confidently infer that something else (like an X) did cause the difference. This bit of abstract thinking is what most of the rest of the textbook is about. Determining whether there is a difference caused by your manipulation. There's more to the story, there always is. We can get more specific than just, is there a difference, but for introductory purposes, we will focus on the finding of differences as a foundational concept.

Interim summary

We've talked about estimation without doing any estimation, so in the next section we will do some estimating of the mean and of the standard deviation. Formally, we talk about this as using a sample to estimate a parameter of the population. Feel free to think of the "population" in different ways. It could be concrete population, like the distribution of feet-sizes. Or, it could be something more abstract, like the parameter estimate of what samples usually look like when they come from a distribution.

Estimating the population mean

Suppose we go to Brooklyn and 100 of the locals are kind enough to sit through an IQ test. The average IQ score among these people turns out to be $\bar{X} = 98.5$. So what is the true mean IQ for the entire population of Brooklyn? Obviously, we don't know the answer to that question. It could be 97.2, but if could also be 103.5. Our sampling isn't exhaustive so we cannot give a definitive answer. Nevertheless if forced to give a "best guess" I'd have to say 98.5. That's the essence of statistical estimation: giving a best guess. We're using the sample mean as the best guess of the population mean.

In this example, estimating the unknown population parameter is straightforward. I calculate the sample mean, and I use that

as my **estimate of the population mean**. It's pretty simple, and in the next section we'll explain the statistical justification for this intuitive answer. However, for the moment let's make sure you recognize that the sample statistic and the estimate of the population parameter are conceptually different things. A sample statistic is a description of your data, whereas the estimate is a guess about the population. With that in mind, statisticians often use different notation to refer to them. For instance, if true population mean is denoted μ, then we would use $\hat{\mu}$ to refer to our estimate of the population mean. In contrast, the sample mean is denoted $\bar{X}$ or sometimes m. However, in simple random samples, the estimate of the population mean is identical to the sample mean: if I observe a sample mean of $\bar{X} = 98.5$, then my estimate of the population mean is also $\hat{\mu} = 98.5$. To help keep the notation clear, here's a handy table:

Symbol	What is it?	Do we know what it is?
$\bar{X}$	Sample mean	Yes, calculated from the raw data
μ	True population mean	Almost never known for sure
$\hat{\mu}$	Estimate of the population mean	Yes, identical to the sample mean

Estimating the population standard deviation

So far, estimation seems pretty simple, and you might be wondering why I forced you to read through all that stuff about sampling theory. In the case of the mean, our estimate of the population parameter (i.e. $\hat{\mu}$) turned out to identical to the corresponding sample statistic (i.e. $\bar{X}$). However, that's not always true. To see this, let's have a think about how to construct an **estimate of the population standard deviation**, which we'll denote $\hat{\sigma}$. What shall we use as our estimate in this case? Your first thought might be that we could do the same thing we did when estimating the mean, and just use the sample statistic as our estimate. That's almost the right thing to do, but not quite.

Here's why. Suppose I have a sample that contains a single observation. For this example, it helps to consider a sample where you have no intuitions at all about what the true population values might be, so let's use something completely fictitious. Suppose the observation in question measures the **cromulence** of my shoes. It turns out that my shoes have a cromulence of 20. So here's my sample:

```
20
```

This is a perfectly legitimate sample, even if it does have a sample size of $N = 1$. It has a sample mean of 20, and because every observation in this sample is equal to the sample mean (obviously!) it has a sample standard deviation of 0. As a description of the **sample** this seems quite right: the sample contains a single observation and therefore there is no variation observed within the sample. A sample standard deviation of $s = 0$ is the right answer here. But as an estimate of the **population** standard deviation, it feels completely insane, right? Admittedly, you and I don't know anything at all about what "cromulence" is, but we know something about data: the only reason that we don't see any variability in the **sample** is that the sample is too small to display any variation! So, if you have a sample size of $N = 1$, it **feels** like the right answer is just to say "no idea at all".

Notice that you **don't** have the same intuition when it comes to the sample mean and the population mean. If forced to make a best guess about the population mean, it doesn't feel completely insane to guess that the population mean is 20. Sure, you probably wouldn't feel very confident in that guess, because you have only the one observation to work with, but it's still the best guess you can make.

Let's extend this example a little. Suppose I now make a second observation. My data set now has $N = 2$ observations of the cromulence of shoes, and the complete sample now looks like this:

```
20, 22
```

This time around, our sample is **just** large enough for us to be able to observe some variability: two observations is the bare minimum number needed for any variability to be observed! For our new data set, the sample mean is $\bar{X} = 21$, and the sample standard deviation is $s = 1$. What intuitions do we have about the population? Again, as far as the population mean goes, the best guess we can possibly make is the sample mean: if forced to guess, we'd probably guess that the population mean cromulence is 21. What about the standard deviation? This is a little more complicated. The sample standard deviation is only based on two observations, and if you're at all like me you probably have the intuition that, with only two observations, we haven't given the population "enough of a chance" to reveal its true variability to us. It's not just that we suspect that the estimate is **wrong**: after all, with only two observations we expect it to be wrong to some degree. The worry is that the error is **systematic**.

If the error is systematic, that means it is **biased**. For example, imagine if the sample mean was always smaller than the population mean. If this was true (it's not), then we couldn't use the sample mean as an estimator. It would be biased, we'd be using the wrong number.

It turns out the sample standard deviation is a **biased estimator** of the population standard deviation. We can sort of anticipate this by what we've been discussing. When the sample size is 1, the standard deviation is 0, which is obviously to small. When the sample size is 2, the standard deviation becomes a number bigger than 0, but because we only have two sample, we suspect it might still be too small. Turns out this intuition is correct.

It would be nice to demonstrate this somehow. There are in fact mathematical proofs that confirm this intuition, but unless you have the right mathematical background they don't help very much. Instead, what I'll do is use R to simulate the results of some experiments. With that in mind, let's return to our IQ studies. Suppose the true population mean IQ is 100 and the standard deviation is 15. I can use the **rnorm()** function to generate the the results of an experiment in which I measure $N = 2$ IQ scores, and calculate the sample standard deviation. If I do this over and over again, and plot a histogram of these sample standard deviations, what I have is the **sampling distribution of the standard deviation**. I've plotted this distribution in Figure 13.14.1.

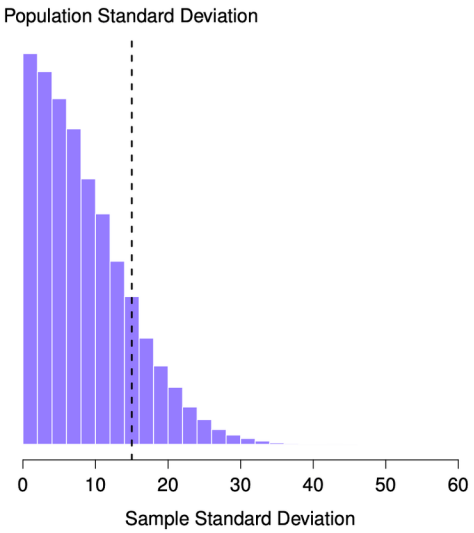


Figure 13.14.1: The sampling distribution of the sample standard deviation for a two IQ scores experiment. The true population standard deviation is 15 (dashed line), but as you can see from the histogram, the vast majority of experiments will produce a much smaller sample standard deviation than this. On average, this experiment would produce a sample standard deviation of only 8.5, well below the true value! In other words, the sample standard deviation is a biased estimate of the population standard deviation.

Even though the true population standard deviation is 15, the average of the **sample** standard deviations is only 8.5. Notice that this is a very different from when we were plotting sampling distributions of the sample mean, those were always centered around the mean of the population.

Now let's extend the simulation. Instead of restricting ourselves to the situation where we have a sample size of $N = 2$, let's repeat the exercise for sample sizes from 1 to 10. If we plot the average sample mean and average sample standard deviation as a function of sample size, you get the following results.

Figure 13.14.2 shows the sample mean as a function of sample size. Notice it's a flat line. The sample mean doesn't underestimate or overestimate the population mean. It is an unbiased estimate!

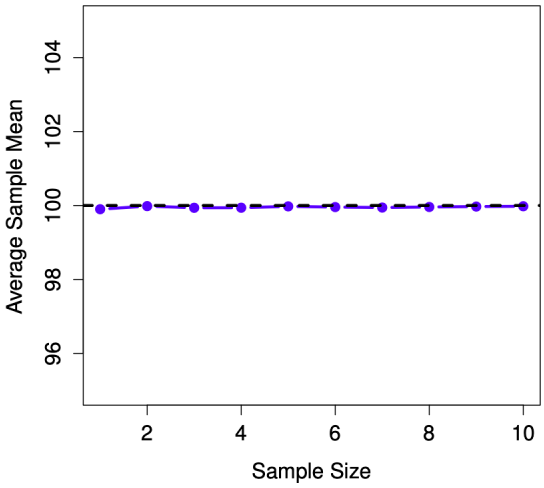


Figure 13.14.2: An illustration of the fact that the sample mean is an unbiased estimator of the population mean.

Figure 13.14.3 shows the sample standard deviation as a function of sample size. Notice it is not a flat line. The sample standard deviation systematically underestimates the population standard deviation!

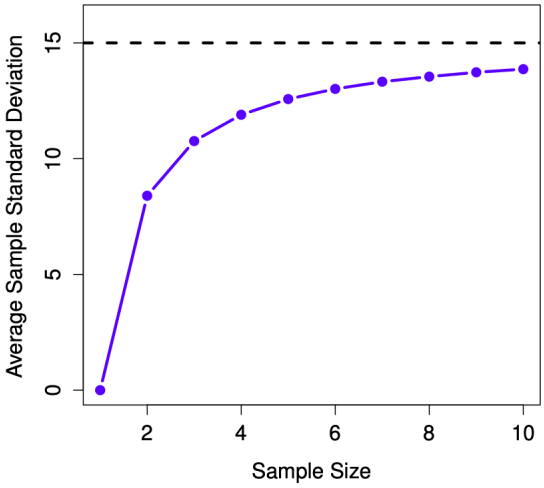


Figure 13.14.3: An illustration of the fact that the the sample standard deviation is a biased estimator of the population standard deviation.

In other words, if we want to make a "best guess" ($\hat{\sigma}$, our estimate of the population standard deviation) about the value of the population standard deviation σ, we should make sure our guess is a little bit larger than the sample standard deviation s.

The fix to this systematic bias turns out to be very simple. Here's how it works. Before tackling the standard deviation, let's look at the variance. If you recall from the second chapter, the sample variance is defined to be the average of the squared deviations from the sample mean. That is:

$$s^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

The sample variance s^2 is a biased estimator of the population variance σ^2. But as it turns out, we only need to make a tiny tweak to transform this into an unbiased estimator. All we have to do is divide by $N - 1$ rather than by N. If we do that, we obtain the following formula:

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2$$

This is an unbiased estimator of the population variance σ.

A similar story applies for the standard deviation. If we divide by $N-1$ rather than N, our estimate of the population standard deviation becomes:

$$\hat{\sigma} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2}$$

It is worth pointing out that software programs make assumptions **for you**, about which variance and standard deviation **you** are computing. Some programs automatically divide by $N-1$, some do not. You need to check to figure out what they are doing. Don't let the software tell you what to do. Software is for you telling it what to do.

One final point: in practice, a lot of people tend to refer to $\hat{\sigma}$ (i.e., the formula where we divide by $N-1$) as the **sample** standard deviation. Technically, this is incorrect: the **sample** standard deviation should be equal to s (i.e., the formula where we divide by N). These aren't the same thing, either conceptually or numerically. One is a property of the sample, the other is an estimated characteristic of the population. However, in almost every real life application, what we actually care about is the estimate of the population parameter, and so people always report $\hat{\sigma}$ rather than s.

Note

Note, whether you should divide by N or N-1 also depends on your philosophy about what you are doing. For example, if you don't think that what you are doing is estimating a population parameter, then why would you divide by N-1? Also, when N is large, it doesn't matter too much. The difference between a big N, and a big N-1, is just -1.

This is the right number to report, of course, it's that people tend to get a little bit imprecise about terminology when they write it up, because "sample standard deviation" is shorter than "estimated population standard deviation". It's no big deal, and in practice I do the same thing everyone else does. Nevertheless, I think it's important to keep the two **concepts** separate: it's never a good idea to confuse "known properties of your sample" with "guesses about the population from which it came". The moment you start thinking that s and $\hat{\sigma}$ are the same thing, you start doing exactly that.

To finish this section off, here's another couple of tables to help keep things clear:

Symbol	What is it?	Do we know what it is?
s^2	Sample variance	Yes, calculated from the raw data
σ^2	Population variance	Almost never known for sure
$\hat{\sigma}^2$	Estimate of the population variance	Yes, but not the same as the sample variance

This page titled 13.14: Estimating population parameters is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

- Table of Contents has no license indicated.

13.15: Estimating a confidence interval

> Statistics means never having to say you're certain – Unknown origin

Up to this point in this chapter, we've outlined the basics of sampling theory which statisticians rely on to make guesses about population parameters on the basis of a sample of data. As this discussion illustrates, one of the reasons we need all this sampling theory is that every data set leaves us with some of uncertainty, so our estimates are never going to be perfectly accurate. The thing that has been missing from this discussion is an attempt to **quantify** the amount of uncertainty in our estimate. It's not enough to be able guess that the mean IQ of undergraduate psychology students is 115 (yes, I just made that number up). We also want to be able to say something that expresses the degree of certainty that we have in our guess. For example, it would be nice to be able to say that there is a 95% chance that the true mean lies between 109 and 121. The name for this is a **confidence interval** for the mean.

Armed with an understanding of sampling distributions, constructing a confidence interval for the mean is actually pretty easy. Here's how it works. Suppose the true population mean is μ and the standard deviation is σ. I've just finished running my study that has N participants, and the mean IQ among those participants is $\bar{X}$. We know from our discussion of the central limit theorem that the sampling distribution of the mean is approximately normal. We also know from our discussion of the normal distribution that there is a 95% chance that a normally-distributed quantity will fall within two standard deviations of the true mean. To be more precise, we can use the **qnorm()** function to compute the 2.5th and 97.5th percentiles of the normal distribution

```
qnorm( p = c(.025, .975) ) [1] -1.959964 1.959964
```

Okay, so I lied earlier on. The more correct answer is that a 95% chance that a normally-distributed quantity will fall within 1.96 standard deviations of the true mean.

Next, recall that the standard deviation of the sampling distribution is referred to as the standard error, and the standard error of the mean is written as SEM. When we put all these pieces together, we learn that there is a 95% probability that the sample mean $\bar{X}$ that we have actually observed lies within 1.96 standard errors of the population mean. Oof, that is a lot of mathy talk there. We'll clear it up, don't worry.

Mathematically, we write this as:

$$\mu - (1.96 \times \text{SEM}) \;\leq\; \bar{X} \;\leq\; \mu + (1.96 \times \text{SEM})$$

where the SEM is equal to $\sigma/\sqrt{N}$, and we can be 95% confident that this is true.

However, that's not answering the question that we're actually interested in. The equation above tells us what we should expect about the sample mean, given that we know what the population parameters are. What we **want** is to have this work the other way around: we want to know what we should believe about the population parameters, given that we have observed a particular sample. However, it's not too difficult to do this. Using a little high school algebra, a sneaky way to rewrite our equation is like this:

$$\bar{X} - (1.96 \times \text{SEM}) \;\leq\; \mu \;\leq\; \bar{X} + (1.96 \times \text{SEM})$$

What this is telling is is that the range of values has a 95% probability of containing the population mean μ. We refer to this range as a **95% confidence interval**, denoted CI_{95}. In short, as long as N is sufficiently large – large enough for us to believe that the sampling distribution of the mean is normal – then we can write this as our formula for the 95% confidence interval:

$$\text{CI}_{95} = \bar{X} \pm \left(1.96 \times \frac{\sigma}{\sqrt{N}} \right)$$

Of course, there's nothing special about the number 1.96: it just happens to be the multiplier you need to use if you want a 95% confidence interval. If I'd wanted a 70% confidence interval, I could have used the **qnorm()** function to calculate the 15th and 85th quantiles:

```
qnorm( p = c(.15, .85) ) [1] -1.036433 1.036433
```

and so the formula for CI_{70} would be the same as the formula for CI_{95} except that we'd use 1.04 as our magic number rather than

1.96.

A slight mistake in the formula

As usual, I lied. The formula that I've given above for the 95% confidence interval is approximately correct, but I glossed over an important detail in the discussion. Notice my formula requires you to use the standard error of the mean, SEM, which in turn requires you to use the true population standard deviation σ.

Yet, before we stressed the fact that we don't actually **know** the true population parameters. Because we don't know the true value of σ, we have to use an estimate of the population standard deviation $\hat{\sigma}$ instead. This is pretty straightforward to do, but this has the consequence that we need to use the quantiles of the t-distribution rather than the normal distribution to calculate our magic number; and the answer depends on the sample size. Plus, we haven't really talked about the t distribution yet.

When we use the t distribution instead of the normal distribution, we get bigger numbers, indicating that we have more uncertainty. And why do we have that extra uncertainty? Well, because our estimate of the population standard deviation $\hat{\sigma}$ might be wrong! If it's wrong, it implies that we're a bit less sure about what our sampling distribution of the mean actually looks like... and this uncertainty ends up getting reflected in a wider confidence interval.

This page titled 13.15: Estimating a confidence interval is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

- Table of Contents has no license indicated.

13.16: Summary

In this chapter I've covered two main topics. The first half of the chapter talks about sampling theory, and the second half talks about how we can use sampling theory to construct estimates of the population parameters. The section breakdown looks like this:

- Basic ideas about samples, sampling and populations
- Statistical theory of sampling: the law of large numbers, sampling distributions and the central limit theorem.
- Estimating means and standard deviations
- confidence intervals

As always, there's a lot of topics related to sampling and estimation that aren't covered in this chapter, but for an introductory psychology class this is fairly comprehensive I think. For most applied researchers you won't need much more theory than this. One big question that I haven't touched on in this chapter is what you do when you don't have a simple random sample. There is a lot of statistical theory you can draw on to handle this situation, but it's well beyond the scope of this book.

This page titled 13.16: Summary is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

- Front Matter has no license indicated.

13.17: Videos

Introduction to Probability

Jeff has several more videos on probability that you can view on his statistics playlist.

(https://www.youtube.com/watch?v=RoalMn9VHZg)

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Chebychev's Theorem

(https://www.youtube.com/watch?v=4RtwqCFt1IU)

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Z-scores

(https://www.youtube.com/watch?v=dQVO0KAxFaU)

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Normal Distribution I

(https://www.youtube.com/watch?v=FPJDF9fGwwE)

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

Normal Distribution II

(https://www.youtube.com/watch?v=mPdnF-GVuCo)

YouTube Video

This interactive element is not available in this format. Visit the link or scan the QR code to access it online.

View this YouTube Video online

This page titled 13.17: Videos is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

- 1.6: Ethics and International Business by Anonymous is licensed CC BY-NC-SA 3.0. Original source: https://2012books.lardbucket.org/books/challenges-and-opportunities-in-international-business/.

13.18: References

Fisher, R. A. 1922. “On the Mathematical Foundation of Theoretical Statistics.” *Philosophical Transactions of the Royal Society A* 222: 309–68.

Keynes, John Maynard. 1923. *A Tract on Monetary Reform*. London: Macmillan; Company.

Maul, Andrew. 2017. “Rethinking Traditional Methods of Survey Validation.” *Measurement: Interdisciplinary Research and Perspectives* 15 (2): 51–69. https://doi.org/10.1080/15366367.2017.1348108.

Meehl, P. H. 1967. “Theory Testing in Psychology and Physics: A Methodological Paradox.” *Philosophy of Science* 34: 103–15.

This page titled 13.18: References is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

- 1.6: A Sample Case is licensed CC BY-NC-SA 4.0.

14: PUBLIC SPEAKING

Learning Objectives

- Identify key principles of effective public speaking
- Describe delivery techniques for use during a public speech
- Identify the role and importance of your audience
- Discuss tips and tricks to giving an effective speech

14.1: How You Speak

14.2: Introduction to Audience Engagement

14.3: Audience Attention and Rapport

14.4: Audience Participation

14.5: Introduction to Speech Tips and Techniques

14.6: Overcoming Fears and Anxiety

14.7: Pro Tips

14.8: Putting It Together- Public Speaking

14.9: Discussion- Public Speaking

14.10: Why It Matters- Public Speaking

14.11: Introduction to Effective Public Speaking

14.12: What is Public Speaking?

14.13: Benefits of Public Speaking

14.14: Audience Expectations

14.15: Developing an Effective Speech

14.16: Introduction to Delivery Techniques

14.17: Starting Your Speech

14.18: Body Language and Gestures

This page titled 14: Public Speaking is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 15: Recruiting and Selecting New Employees by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.1: How You Speak

Learning Objectives

- Discuss strategies to effectively use vocal variety to emphasize your message
- Identify types of language to avoid in your speech

Vocal Variety

Just as gestures and body language affect how you are perceived, vocal variety effects how you are heard. As presentation skills training consultant Gavin Meikle notes, "A carefully crafted speech can be ruined by a dull vocal delivery."[1] In a series of posts on vocal variety, Meikle identifies six key elements, common errors, and good practices to develop greater vocal impact.[2]

- Volume. Develop your range and vary your volume. To help put this in perspective, consider the saying, "A good speech needs light and shade."
- Pitch and Resonance. Research suggests a general preference for lower vocal pitch, with participants ascribing more positive personality traits to lower pitched voices. For example, Margaret Thatcher was considered to have a voice of leadership.
- Pace and Pause. Be aware of and manage your speaking speed and practice your pauses. It's been found that people who slow down their pace when speaking to groups are thought to have greater gravitas, credibility, and authority.
- Intonation.This describes changes in vocal tone within a sentence. In order to achieve the desired effect, use the three common intonation patterns appropriately.
  - Ending a spoken sentence with a rising tone indicates a question or suggestion.
  - Ending a spoken sentence with a descending tone is generally interpreted as an order.
  - A flat intonation is used to indicate a statement.

As legendary advertising creative director William Bernbach noted, "It's not just what you say that stirs people. It's the way that you say it."

Language Choices

Whether we speak to inform, persuade, or inspire, the common denominator is a desire to communicate and to arrive at a shared understanding of an idea or situation. To quote author and TED Conference curator Chris Anderson, "Your number one task as a speaker is to transfer into your listeners' minds an extraordinary gift—a strange and beautiful object that we call an idea."[3] And yet, the very expertise that makes us the right person to deliver a speech on a particular topic can make us incapable of achieving that objective. An in-depth understanding can lead us to oversimplify or over complicate the explanation of a concept foreign to our audience.

One of the most common barriers to communication is jargon, or the terminology associated with a particular profession. As the French philosopher Étienne Bonnot de Condillac observed, "Every science requires a special language because every science has its own ideas." For perspective on this challenge, and how to overcome it, watch Communications teacher Melissa Marshall's "Talk Nerdy to Me" TED Talk. Directed at scientists, but with broad applicability to communicators, Marshall describes her "Alice in Wonderland" experience teaching communication skills to engineering students. Extrapolating on her point, if we don't know about or don't understand the work of those who are trying to solve the grand challenges of our times, then we can't support it. Marshall notes that jargon in particular, is a barrier to communication. For example, "you can say 'spatial and temporal,' but why not just say "space and time," which is so much more accessible to us?"[4] A few specific recommendations:

- Eliminate bullet points (use a powerful visual instead).
  - Because you're giving a speech, rather than a business presentation as discussed in Module 6: Reports and Module 8: Developing and Delivering Business Presentations, you shouldn't need bullet points to keep you or your audience on track.
- Use stories and analogies to scaffold your important points
- Display images and diagrams to illustrate what's being described.

A related point, covered in detail in Module 13: Social Diversity in the Workplace, is to be sensitive to socio-cultural variations in language and interpretation. As the French proverb notes, “The spoken word belongs half to him who speaks and half to him who listens.”

The following video is a great talk about the mistakes and cornerstones of speech, which help you encourage your audience to listen and care about your points:

A link to an interactive elements can be found at the bottom of this page.

1. Meikle, Gavin. "Six Elements of Vocal Variety and How to Master Them." *Inter-Activ*. 18 Jun 2017. Web. 25 Jun 2018.
2. Ibid.
3. Anderson, Chris. "TED's secret to great public speaking." *TED*. Mar 2016. Web. 25 Jun 2018
4. Marshall, Melissa. "Talk Nerdy to Me." *TED*. Jun 2012. Web. 25 Jun 2018

Contributors and Attributions

CC licensed content, Original

- How You Speak. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- TED Talk: How to speak so that people want to listen. **Authored by**: Julian Treasure. **Provided by**: TED. **Located at**: https://www.youtube.com/watch?v=eIho2S0ZahI&feature=youtu.be. **License**: *CC BY-NC-ND: Attribution-NonCommercial-NoDerivatives*

This page titled 14.1: How You Speak is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 7.1: How You Speak by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.2: Introduction to Audience Engagement

What you'll learn to do: Identify the role and importance of your audience

In order to achieve the key objective identified by TED Conference curator Chris Anderson—the transfer of an idea—a speaker must effectively engage audience members. In this section, we'll discuss techniques you can use to capture and maintain audience members' attention and ways to incorporate interaction without losing control.

Contributors and Attributions

CC licensed content, Original

- Introduction to Audience Engagement. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 14.2: Introduction to Audience Engagement is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 7.2: Introduction to Audience Engagement by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.3: Audience Attention and Rapport

Learning Objectives

- Describe techniques to gain and keep an audience's attention

The key to capturing and maintaining an audience's attention is, to riff on the TED Talk tagline, having an idea worth sharing and sharing it with emotion. As expressed in one of author and cartoonist Hugh MacLeod's business card art creations: "a story without love is not worth sharing."

The following techniques are adaptations from author and communication expert Mike Parkinson's "Spark a Fire: 5 Tips to Grab and Hold Audience Attention" article on presentationexpert.com:

1. Surprise: Saying, showing, or doing something unexpected reengages the audience's brains.
2. Suspense: Use drama, slowly building your idea like a verbal puzzle.
3. Storytelling: Share a unique and compelling story to illustrate your point.
4. Senses: Engage the senses—hearing, sight, taste, touch, and smell. The greater the sensory engagement, the stronger the interest.
5. Involve: Invite participation, a point we will address in the next section.

In addition to these five techniques, consider the the six techniques mentioned on the Starting Your Speech page. These techniques, including using quotes, "what if" questions, silence, and statistics can be used to check in with and engage your audience throughout a speech. Give your audience something they can use—give them a reason to care!

Contributors and Attributions

CC licensed content, Original

- Audience Attention and Rapport. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 14.3: Audience Attention and Rapport is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 7.3: Audience Attention and Rapport by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.4: Audience Participation

Learning Objectives

- Discuss effective ways to use audience participation
- Discuss appropriate ways to respond to questions without derailing a presentation

Audience participation is not only an effective way to reinforce learning, it's associated with higher levels of attendee engagement, which may translate into higher satisfaction and understanding. Consider the following audience participation options, drawn from The Toastmasters International Guide to Successful Speaking and other sources:

1. Volunteer Exercise: Invite a member of the audience to participate in an exercise or role-play.
2. Audience Survey: Surveying the audience—for example, "by show of hands"—allows you to assess the needs and temperament of the audience and fine-tune your speech accordingly.
3. Question: Asking a leading question of the audience like, "What is your biggest hurdle when preparing for employee performance reviews?" allows a speaker to surface ideas or problems and challenges to be addressed during the speech.
4. Q&A: A variation on the question technique, time for a Q&A period at the end of the speech allows attendees to clarify open points. Make sure you anticipate likely questions and are prepared with answers. Thorough audience analysis (as discussed in previous modules) will help you anticipate both the content and the level of sophistication of the questions you might get. To overcome initial audience reluctance, you can plant questions, prompt with frequently asked questions, or draw from questions submitted in a pre-session survey.
5. Partner Exercise: Pair audience members to practice a technique learned or test learning with a think-pair-share or other collaborative learning exercise.
6. Small Group Exercise: Best used for brainstorming solutions or to generate relevant questions for deeper learning or more specific application.
7. Written exercises or note-taking: Asking attendees to take notes or complete written exercises—answering a self-assessment, identifying goals, taking a quiz, or filling out a worksheet.

You may notice that these techniques are all pretty straightforward. You should avoid using audience participation plans that are too off-the-wall—like asking your audience to sing, to mime their morning routine, or to hug the person next to them (all real examples). Such techniques are as likely to alienate your audience and lessen your credibility as they are to enhance your speech.

Responding to Questions

A key consideration when incorporating audience participation is maintaining control. To avoid having an exercise deteriorate into chaos or a question turn into an an extended digression, set clear expectations and enforce the ground rules. If you ask participants to do a partner or small group exercise, clearly communicate the process and timeframe. Let participants know how long an exercise will last and tell them when to begin and when to stop. Allow enough time for participants to get value from the exercise but not so much time that some groups become bored or distracted.



Figure 14.4.1: Questioning the Secretary of Energy. (Public Domain; via Wikipedia)

The University of Leicester's oral presentation student resources provide preparation perspective and the following 4 step approach to managing Q&A:[1]

1. Listen: Don't jump to conclusions and start framing a response before the attendee finishes stating the question. Knowing that the questioner is likely thinking on his or her feet, consider both the content and intent of the question.
2. Understand: Paraphrase the question to confirm understanding.
3. Communicate & Involve: To involve the entire audience and minimize the risk of an extended dialogue with the questioner, restate the question so all can hear and feel a part of the conversation.
4. Respond: Direct your answer to both the questioner and other audience members. Keep your response focused and confirm that you answered the question.

Their planning notes are worth considering as Q&A guidelines. For example, you may decide to limit the topics open for discussion or defer questions that are outside the scope of your talk. You might open a Q&A with a limiting phrase such as, "Are there any questions on the four techniques I've presented?" Or you could "table" a question with a response indicating that the question falls outside of the stated purpose for your speech and an softening statement such as, "I'd love discuss that with you at another time, feel free to email me." As with audience participation during the speech, it may be worth establishing and communicating a time limit for individual questions and the overall Q&A session.

Note

Q&A sessions can easily be derailed if the speaker (or moderator, depending on the set up of the event) doesn't moderate question askers well. If a speaker turns back to the questioner and says, "Did that answer your question?" it can lead to unwanted dialogue if the questioner says no or decides to elaborate. Make sure you are prepared to politely move on from the initial questioner so you don't end up in a dialogue situation.

1. www2.le.ac.uk/offices/ld/resources/presentations/questions

This page titled 14.4: Audience Participation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 7.4: Audience Participation by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.5: Introduction to Speech Tips and Techniques

What you'll learn to do: Discuss tips and tricks to giving an effective speech

We're living in a time where ideas—learning and sharing—are essential skills. According to communication coach and bestselling author Carmine Gallo, "Ideas are the true currency of the 21st century." The caveat: if you can't communicate your ideas in a way that captures attention and inspires action, it doesn't matter how good your ideas are.

This page titled 14.5: Introduction to Speech Tips and Techniques is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 7.5: Introduction to Speech Tips and Techniques by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.6: Overcoming Fears and Anxiety

Learning Objectives

- Discuss various strategies for overcoming common fears and anxiety about public speaking

In addressing fear of public speaking, author and professional speaker Michael Aun shares a quote from his grandfather who defined fear as “an absence of knowledge and a lack of information.” If you think about it, the fear of public speaking is simply a recognition of a learning gap, be it a lack of confidence in our knowledge of the material or an inability to effectively communicate our expertise. We don’t look at a bicycle and fear learning how to ride it (maybe getting hurt, but not the process of learning a new skill), so why should we fear developing skills that can improve our careers, our lives and, perhaps, our world? Logic aside, the fear of public speaking is so common that Mayo Clinic addresses it as a “specific phobia” on its website. In an article titled “Fear of Public Speaking: How Can I Overcome It?,” Dr. Craig N. Sawchuk provides ten tips for managing performance anxiety or stage fright, which are adapted below:

- **Know your topic**. In a point echoed by many professional speakers and coaches, Sawchuk notes that having a strong interest in and understanding of your material, including preparing responses to possible questions, will help you stay on point and keep your composure.
- **Get organized**. The more organized you are—regarding information, materials and logistics—the less nervous you’ll be.
- **Practice, rinse and repeat**. If possible, practice your speech and request feedback from friends, family, and colleagues. You can also record and critique your own performance.
- **Challenge worries**. Reality check your negative projections. List and then directly challenge specific worries, considering the evidence and alternative outcomes.
- **Visualize your success**. Imagining a successful speech creates a more positive frame of mind that can reduce anxiety.
- **Do some deep breathing**. To calm yourself, take a few deep, slow breaths before you get up to speak and remember to breathe during your speech.
- **Focus on your material**. People tend to focus on new information, so focus on your message rather than the messenger (you and your nerves) or the audience.
- **Don’t fear a moment of silence**. If you draw a blank or get off-topic, take a few seconds and a few deep breaths to regroup.
- **Recognize your success**. Congratulate yourself for on a completed speech. Reflect on your performance and identify what worked well and areas for improvement.
- **Get support**. Join a public speaking group that can help you develop your skills in a supportive setting.

In an article for Forbes, author Carmine Gallo cites the results of a Prezi survey of American professionals indicating that 70 percent of those who give presentations agree that presentation skills are critical to their career success. Gallo’s rejoinder: “The other 30 percent don’t know it yet!” Further, 20 percent of respondents indicated they would do almost anything to avoid giving a presentation, even if it means losing respect. Given that, developing effective speaking skills is a powerful differentiator. The good news is that you’re not in it alone. You can join a local Toastmasters group (your college or work may sponsor a group), a campus speech and debate team, a speakers bureau, or take a Dale Carnegie course. If you’re an introvert, you can start by analyzing TED Talks and reading related articles and books on public speaking skills.

This page titled 14.6: Overcoming Fears and Anxiety is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 7.6: Overcoming Fears and Anxiety by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.7: Pro Tips

Learning Objectives

- Discuss public speaking strategies of effective public speakers

In this section, we'll focus in on a few key tips and resources drawn from bestselling author and communication coach Carmine Gallo's *Talk Like TED: The 9 Public Speaking Secrets of the World's Top Minds*. For context, TED Talks started as a one-time TED (technology, entertainment, and design) event in 1985, and have since morphed into a global brand and experience. At last count, TEDx (independently organized, local-level conferences) were being produced in over 130 countries at a rate of five events per day. As Gallo notes, "the world is clearly hungry for great ideas presented in an engaging way." Consider the following four points a jump-start to further reflection and skills development from a professional public speaker.

Understand the Power of Pathos

Although emotion doesn't factor into the definition of persuasion, it is an essential ingredient. Ancient Greek philosopher Aristotle identified the three elements of persuasion as ethos (credibility), logos (logic) and pathos (emotion). When Gallo analyzed human rights lawyer Bryan Stevenson's "We Need to Talk About Injustice" talk, voted one of the most "persuasive" on TED.com, the results were surprising to many: 65% pathos, 25% logos and 10% ethos. Emotion often drives decision making and opinion formation.

Believe in Your Message

To quote law enforcement veteran Morgan Wright, "If you don't believe what you're saying, your movements will be awkward and not natural. No amount of training—unless you're a trained espionage agent or psychopath—will allow you to break that incongruence between your works and actions." You must trust in what you are sharing with your audience or it will come across as insincere.

Keep Your Speech Brief

Keep your speech succinct, about 20 minutes or less. The science behind this ideal speech length explains how too much information creates a "cognitive backlog" and state of anxiety in your audience which prevents the transfer of ideas. A shorter speech imposes a discipline that forces you to clarify your ideas and helps you communicate for effectively. Brevity is also key to this era of sharing online; if your speech is recorded and available on the internet, a shorter speech is easier to publish and access.

Make it Memorable

Invest time in distilling your big idea into a short statement that's captivating and shareable. For perspective, scan @TED Talk Quotes on Twitter or search quotes on the TED.com site. A short, memorable summary statement makes is easy for your audience to remember both you and the point of your speech when you include unforgettable phrases like:

If everything seems under control, you're not going fast enough. – Mario Andretti

The difference between successful people and really successful people is that really successful people say no to almost everything. – Warren Buffett

Design is not just what it looks like and feels like. Design is how it works. – Steve Jobs

Additional Resources

- If you're new to TED Talks, the TED in 3 Minutes playlist serves a selection of "snackable" talks.
- Toastmasters International: Find a Club
- National Speech & Debate Association
- Dale Carnegie Training: Presentation Effectiveness Course Finder

This page titled 14.7: Pro Tips is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 7.7: Pro Tips by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.8: Putting It Together- Public Speaking

Becoming a public speaker is a rite of passage, an event that marks one's emergence as a business or civic leader. Unlike a title that generally requires an external award decision, the accompanying change in status is a role you assume. Of course, being an effective oral communicator entails having something to say—that is, a compelling point of view, a truism captured in this simile: "A good speech is like a pencil; it has to have a point."

As with any leadership role, oral communication also involves responsibilities. While we all have the right to free expression, we also have the responsibility to exercise that right thoughtfully. This means applying critical thinking and conducting credible research in order to develop both our position and supporting arguments. As an orator, you also have the responsibility to use your communication skills ethically.

Public speaking isn't all upside and addressing feel-good topics. Taking a stand involves risk and requires courage. Of course, there's risk in any decision, and sometimes inaction is riskier because it allows others or current circumstances to decide for you. We also have the opportunity to use our platform to inspire change and to be the change. As Eivor Taylor put it: "There is no 'they.' We are the only ones who can make [the] change."



Figure 1. Parkland students at a political rally. Protect Kids Not Guns, Students participating in National Walkout Day at the Capitol, Washington DC. **Authored by**: Lorie Shaull. **Provided by**: flickr. **Located at**: https://www.flickr.com/photos/number7cloud/38999219860. **License**: *CC BY-SA: Attribution-ShareAlike*

For perspective, consider the #NeverAgain gun control movement started by some survivors of the mass shooting at Marjory Stoneman Douglas High School in Parkland, Florida. As Atlantic contributing editor Michelle Cottle noted in her "How Parkland Students Changed the Gun Debate" article, these teens have the training to express themselves and the confidence that they will be heard. At a rally, Stoneman Douglas senior Emma Gonzalez stated, "Every single person up here today, all these people should be home grieving. But instead we are up here standing together, because if all our government and president can do is send thoughts and prayers, then it's time for victims to be the change that we need to see." For perspective on the people behind the movement, read the *New Yorker* article "How the Survivors of Parkland Began the Never Again movement."

Becoming an effective public speaker can be transformational—for you personally, for your audience, for the cause or brand you represent, and for society. Reflect on the stories that have shaped your personal and career development as well as your understanding of the world and your place in it. What if you had a superpower that gave you that type of influence, the power to shape business practices, public policy, or "the way things work" in any number of areas of human endeavor? Public speaking can be that superpower.

This page titled 14.8: Putting It Together- Public Speaking is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 7.8: Putting It Together- Public Speaking by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.9: Discussion- Public Speaking

Part 1: Your Presentation

Your task is to create a five minute presentation supported by PowerPoint slides. Imagine you are the general manager of a grocery or clothing store and you are going to conduct a team meeting to teach your staff about a new policy or procedure. Select one of the following topics for your presentation:

- a new loss prevention policy
- a new promotional strategy
- a new cash handling policy

If you wish do present on a different topic you must first get permission from your instructor.

Your PowerPoint presentation should have a minimum of 4 slides: a title page, content slides, and conclusion.You will present your visual essay to the class. Please email the instructor a copy **before class** on the due date. You will also post your presentation in the discussion board for your classmates to provide feedback.

Part 2: Peer Feedback

Choose at least two of your classmates' presentations to review. You will review both their slides and their presentation, so make sure to take notes during your peers' presentations! You will write feedback (minimum 150 words) for each classmate based on the following questions:

- Does the speaker read each slide verbatim or include additional information than the content on the slides?
- Do they have the minimum number of required slides?
- Is there presentation the required length? (5 minutes +/– 30 seconds)?
- How is the speakers volume, speed of speech, and clarity?
- What is successful or unsuccessful about layout and look of the slides?
- Are there any grammatical or spelling errors in the slides?

Grading Rubric

Criteria	Not Evident	Developing	Exemplary	Points
Submit your slide deck	**0 pts** No slide deck submitted	**5 pts** Slide deck does not have at least 4 slides	**10 pts** Slide deck has at least 4 slides focused on a central theme	10 pts
Present within the allotted time	**0 pts** No presentation completed	**5 pts** Presentation is either 30 second under or over 5 minutes	**10 pts** Presentation is within 30 seconds of the required 5-minute length	10 pts
Speak with appropriate volume, speed, and clarity	**0 pts** No presentation completed	**5 pts** Presentation cannot be heard, is too fast or too slow, or cannot be understood	**10 pts** Presentation is spoken at an adequate volume, speed, and clarity	10 pts
Present your slides effectively	**0 pts** No presentation completed	**5 pts** Presentation is disjointed, unprofessional, or too reliant on slide deck text	**10 pts** Presentation is professional and feels cohesive	10 pts

Criteria	Not Evident	Developing	Exemplary	Points
Analyze at least two peers' presentations	**0 pts** No response to peers	**5 pts** Responded to only one peer	**10 pts** Responded to two peers	10 pts
			Total:	**50 pts**

This page titled 14.9: Discussion- Public Speaking is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy via source content that was edited to the style and standards of the LibreTexts platform.

- 7.9: Discussion- Public Speaking by Lumen Learning is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.10: Why It Matters- Public Speaking

Why learn about public speaking in business?

One of the essential business and life skills often neglected in our "schooling" broadly is the art of public speaking. It's an odd omission given that the oral tradition is as old as civilization. Indeed, public speaking is a through-line in our history as a people and as a nation. This is not only an American phenomenon but a human reality. Whether they're a signal of change or a spark that ignites change, speeches accompany the pivotal events of every time and place, rallying people around a vision of the future, be it the founding of a nation, a call to arms, or a call to action.



Figure 1. On August 28, 1963, Martin Luther King, Jr. gave his famous "I Have a Dream" speech from this spot on the Lincoln Memorial steps to an audience of a quarter million people. Image of I Have a Dream speech at the Lincoln Memorial steps. **Authored by**: Paul Ollig. **Provided by**: National Parks Service. **Located at**: www.nps.gov/thingstodo/stand-where-martin-luther-king-jr-gave-the-i-have-a-dream-speech.htm. **License**: *Public Domain: No Known Copyright*

These ideas often transcend their times, tapping into an enduring sense of possibility or responsibility. Consider, for example, this excerpt from Dr. Martin Luther King's 1963 "I have a dream" speech: "I have a dream that my four little children will one day live in a nation where they will not be judged by the color of their skin but by the content of their character."[1] Over fifty years later, Dr. King's speech remains both resonant and relevant and, as an article from the the History Channel notes, it has "endured as one of the signature moments of the civil rights movement."[2]

Hillary Rodham Clinton's 1995 "Women's Rights Are Human Rights" speech given at the United Nations Fourth World Congress on Women is another example of a defining vision and values statement. To excerpt: "As long as discrimination and inequities remain so commonplace everywhere in the world, as long as girls and women are valued less, fed less, fed last, overworked, underpaid, not schooled, subjected to violence in and outside their homes—the potential of the human family to create a peaceful, prosperous world will not be realized."[3] The title of a *New York Times* article by Amy Chozick about Clinton's positioning as a presidential candidate in 2016 sums up the impact of this speech decades after it was delivered: "Hillary Clinton's Beijing Speech on Women Resonates 20 Years Later."

Oral communication is not only a tool of preachers and politicians, it's a tool of the people. If you don't hear yourself or your world view represented, you can change that. Of course, you don't have to be working on a national or global scale to appreciate effective public speaking. To gain support for your ideas at any level, you need to be a clear and compelling oral communicator. Ultimately, public speaking matters because it can be the difference between being heard and shaping your world—however you define it—or living with the consequences of someone else's pronouncements. If you prefer the former option, let's prepare to be heard!

1. King, Martin Luther, Jr. "I Have a Dream." 1963.
2. History.com. "'I Have a Dream' Speech." History.com, 2017. Web. 26 June 2018.
3. Garau, Annie. "History's Most Powerful Speeches Given By Women." All That Is Interesting, 28 Mar 2017. Web. 26 June 2018.

This page titled 14.10: Why It Matters- Public Speaking is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 7.10: Why It Matters- Public Speaking by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.11: Introduction to Effective Public Speaking

What you'll learn to do: Identify key principles of effective public speaking

Like speaking itself, public speaking is a learned behavior. Just as no one comes out of the womb speaking eloquently, no one becomes a powerful orator without practice. In this section, we'll discuss the "why" of public speaking—the audience's expectations and the benefits that accrue to the speaker—and introduce a simple five-step process for developing an effective speech.

This page titled 14.11: Introduction to Effective Public Speaking is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 7.11: Introduction to Effective Public Speaking by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.12: What is Public Speaking?

Learning Objectives

- Discuss key characteristics of public speaking

Public speaking is, simply, an oral presentation or speech delivered to a live audience. It is generally a formal or staged event—although impromptu speeches are a common occurrence—and can be a defining career moment. For example, you may think you're attending a client meeting only to find yourself called on to explain a procedural or technical point being discussed. Or you may be sitting in a management meeting thinking you are just there to observe when you are asked to elaborate on an aspect of the supporting research and analysis or defend your recommendations.

Impromptu Speaking

Although impromptu speaking isn't the focus of this module, it is worth noting that this type of speaking is something Toastmaster members train for on an ongoing basis using a technique called "Table Topics." For more on this technique, read A Table Topics Workout: The Power Packed Exercise for Stretching Your Brain.

Executive presentation coach Peter Khoury has reverse-engineered the characteristics of great speakers for over fifteen years. Combining his findings with scientific research on leadership, he's distilled this research into the following 9 characteristics of effective public speakers:[1]

1. Confidence
2. Passion
3. Practice, don't memorize
4. Speak in a natural voice
5. Authenticity
6. Keep it Short and Sweet
7. Connect with your Audience
8. Paint a Picture through Storytelling
9. Repetition

Like computer failure and natural disasters, finding yourself in a situation requiring public speaking skills is not a matter of *whether* it will happen but *when* it will happen. Given the potential career impact, you need to prepare accordingly.

---

1. https://www.linkedin.com/pulse/9-characteristics-highly-effective-public-speakers-peter-khoury/

This page titled 14.12: What is Public Speaking? is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 7.12: What is Public Speaking? by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.13: Benefits of Public Speaking

Learning Objectives

- Discuss the importance of public speaking in a business setting

What is public speaking but a dressed up—or not, depending on your audience—version of the basic skills we've been using since we first began forming desires and shaping the words and gestures to communicate those desires? Ah, life was simple then; a baby pointing and reaching towards a bowl of grapes or a toddler repeating "more milk" until they get what they want. Then again, one fundamental dynamic hasn't changed. As Stevie Wonder put it, "If you don't ask, you don't get." This is true not only personally but especially professionally. If you want the sale, contract, funding, job, project, or promotion, you have to be willing and able to ask for it in a clear and compelling manner. Often, you'll have to do so in front of a group of deciders—those who will determine the response to your request. Welcome to public speaking!

What has changed is your potential—your potential to connect, to create or co-create and, given technology and social media/sharing, your potential reach and impact. In a statement echoed in virtually every career and leadership book and blog, the *Toastmasters International Guide to Successful Speaking* notes, "There is perhaps no greater skill [to] help you build your career or business than effective public speaking." As a testament to the tradition and enduring power of oral speech, the primary motivations for speaking are the same as they were in ancient Greece. In Aristotle's treatise on the art of persuasion titled *Rhetoric*, he identified three primary motivations: to inform, to persuade, and to inspire. Practically speaking, public speeches often include more than one element. For example, communicating a risk or potential opportunity may be done in conjunction with building support for a change in business practices or a proposed initiative.

Public speaking is also an exceptional, and cost-effective, way to build your brand and network within your organization, profession, or industry and/or to build good will for your company in the community. Whether you're pitching a product, service, idea, company or person (including yourself), public speaking differentiates you and your message from the promotional noise and general chatter. As professional speakers and authors Jeff Slutsky & Michael Aun note, public speaking "literally puts you on a pedestal." Indeed, the average audience member assumes that since you're speaking on the topic, you must be an expert. Of course, the impression they leave with depends on the quality of your speech, but the bottom line is that being a speaker gives you a level of credibility that would take a significant amount of time to cultivate otherwise. Speaking allows you to develop a reputation as a thought leader or community leader, raising your visibility and perceived market value. That's not something a cover letter and resume or pitch is likely to do—if it even makes it through the filters.

Contributors and Attributions

CC licensed content, Original

- Benefits of Public Speaking. **Authored by**: Nina Burkoas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 14.13: Benefits of Public Speaking is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 7.13: Benefits of Public Speaking by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.14: Audience Expectations

Learning Objectives

- Identify various audience needs and expectations that can be addressed by a speech



One of the finest, and rarest, gifts a person can give is their attention. When it comes to audience attention, that gift comes with an expectation. Audience expectations are simply an extension of the three speaker motivations. Specifically, audience members expect to learn from an informational speech, to be moved by a persuasive speech, or to be inspired by an inspirational speech.

Perhaps your first and most important test as a prospective speaker is to make sure you clearly communicate the purpose and benefits of attending your speech. A disconnect between what audience members thought they signed on for and what they're hearing can trigger a range of undesirable audience behaviors from zoning out to walking out. As a speaker, you also have an obligation to factor your audience into the design and development of your speech, from relevant examples to appropriate language and subject matter depth. Whatever your stated intent (benefit), the minimum audience expectation is that you fulfill it in a clear and coherent manner.

One additional point to consider is the medium. Public speeches are live events. Why would you purchase a ticket and go to see a concert or comedian or other event live rather than buying a DVD or tuning in to podcast or TV broadcast for a fraction of the price? There's a difference in the level of energy and engagement in a live "performance"—whether it's a speech, dance recital, political rally, or musical event. Keep in mind that those attending a public speech expect an experience that transcends a one-dimensional transfer of information.

Contributors and Attributions

CC licensed content, Original

- Audience Expectations. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Image of attentive audience. **Authored by**: Robert Scoble. **Located at**: https://www.flickr.com/photos/scobleizer/6293322182. **License**: *CC BY: Attribution*

This page titled 14.14: Audience Expectations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 7.14: Audience Expectations by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.15: Developing an Effective Speech

Learning Objectives

- Identify the five steps of developing an effective speech

Let's assume you see the value in developing public speaking as a skill. Where do you start? A good warm-up exercise is to watch a few TED Talks, organized by topic and popularity, among other categories. If you prefer to proceed straight to the cream of the crop, Steve Jobs' classic "How to Live Before You Die" speech delivered at Stanford University's 2005 commencement is excellent inspiration and perspective—for life as well as for speaking. If you watch a few talks, you'll notice that each presenter has a unique message and style that makes him or her compelling. This is a key point. While we all learn process and technique by copying the masters, as legions of artists have done before us, the artistry (and magic, from the audience's standpoint) is in finding your own voice and developing your personal style. In practical terms, this means that you also have to develop and curate your own material using your life experience, insights, and observations to illustrate your points.



U.S. Supreme Court Justice Sonia Sotomayor giving a speech.

Whether you're facing a blank sheet of paper or a blank screen, the start is always the hardest part of a speaking project. We're going to work through that obstacle by following this five-step jump start.

1. Choose your topic
2. Develop your benefit statement
3. Develop your positioning statement
4. Derive your title
5. Create your content

Choose Your Topic

For perspective on topics, you can scan the 194 topics used by the National Speakers Association (click on "Browse the complete topic list"). If there's a conference or Chamber of Commerce or professional association event you want to speak at, scan the associated website(s), social media posts, and publications to get a sense of what topics might be a good fit. In choosing your topic, consider your experience and expertise. That's not to say that you need to be an acknowledged expert on a particular topic—that's where research comes in—but you do need to have an interest in the topic and a base level of credibility. Although there are hundreds of potential topics, it's very likely that a particular topic has already been covered a number of times by a number of people. Given that, the essential question is what can you bring to the topic that others haven't? That is, how can you approach an exhausted topic with fresh eyes to make it feel new and engaging?

Develop Your Benefit Statement

Once you've decided on a topic, the next step is to develop a one to two sentence benefit statement that supports your credibility

as a speaker on that topic. The benefit statement should answer the question: *why you?* This is similar to the process you would go through in pitching an article to a publisher. What is the unique value—experience, expertise, point of view—that you bring to the topic? For different frames of reference on benefit statements, scan the speaker bios and bylines of writers that cover topics of interest to you.

Develop Your Positioning Statement

The positioning statement is an expansion of the last step that tailors your benefit statement to a specific audience. Working through this step helps you clarify who your audience is and what you will be presenting to them. Although the positioning statement is for internal purposes, the focus is external—what's the ROA (return on attention) for the audience? Don't skip this step; it will help you focus your thoughts, minimize interesting but off-point digressions, and help maintain a coherent structure and flow through the research, writing, editing, and ultimately, speaking phases.

Develop Your Title

In moving from your positioning statement to the speech title, think of your speech as a product or service—what would prompt someone to "buy" what you're offering? Your title is a pitch—or your bid for the audience's attention. To get to that pitch, select a few key words from your positioning statement and brainstorm a compelling headline. For additional insight and exercises, read Larry Kim's *Inc* article, "30 Ideas for Super Clickable Blog Headlines," explore the BBC News resources on writing headlines, or watch the "How to Write a Hook" YouTube video. You may also want to browse the titles of articles and blogs posted to your target audience's (i.e., industry or professional association) websites and publications. Remember that as you develop your content, your title might need some adjusting. If you don't need to submit your title far in advance (to be printed in a brochure or program), revisit it once your content is complete to make sure it still fits. If you do need to submit it before your content is fully developed, try to leave a little wiggle room and not make it too specific regarding the conclusions you might come to.

Develop Your Content

Once you have your title and framework from your positioning statement, you're ready to start developing your content. Of course, you've been building useful content all along through your life experiences. Reflect on the relevant lessons you've learned, and make note of some of the experiences—a key quote or visual, an emotion or insight, people or places. Use these events as possible connections to consider and, if applicable, work them in to your speech to illustrate your points. Research is a skill, and art, unto itself (refer to the Washington University librarians' Conducting Research pages for additional tips and resources), but a good jumping-off point is doing an internet search of your keywords. If you have lead time, you can set up a Google Alert to monitor relevant news and developments. It can also be helpful to find and follow subject-matter experts for your topic and tune in to current trends. To do this, conduct "*best of*" searches to find thought leaders. You may achieve both objectives in one search, as in this *Forbes* article: "Top Shopping Trends of 2018: Retail Experts Share What to Watch for Next Year," one of the results in a search for "*best retail marketers*."

Reminder: Remember to document your sources! Include citations in your written speech in order to give credit where credit is due and to be able to follow-up on any related audience questions.

There you have it! A simple process for sidestepping writer's or speaker's block. Next, we'll discuss another common sticking point: how to open your speech.

Contributors and Attributions

CC licensed content, Original

- Developing an Effective Speech. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Image of Sonia Sotomayor giving a speech. **Authored by**: Gage Skidmore. **Located at**: https://commons.wikimedia.org/wiki/File:Sonia_Sotomayor_(32342597022).jpg. **License**: *CC BY: Attribution*

This page titled 14.15: Developing an Effective Speech is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 7.15: Developing an Effective Speech by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.16: Introduction to Delivery Techniques

What you'll learn to do: Describe delivery techniques for use during a public speech

As alluded to in the prior section, a live speech is, in effect, a performance. In addition to what you say, audience members will be reacting—both consciously and unconsciously—to how you say it. In this section, we'll discuss fundamental considerations, including how to open your speech, how to use gestures and body language to punctuate your message, and what types of language to avoid.

Contributors and Attributions

CC licensed content, Original

- Introduction to Delivery Techniques. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 14.16: Introduction to Delivery Techniques is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 7.16: Introduction to Delivery Techniques by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.17: Starting Your Speech

Learning Objectives

- Discuss effective ways to begin your speech

Your opening comments, like the lead sentences of an article, can make or break a speech. As William Zinsser phrases it in *On Writing Well*, "The most important sentence in any article is the first one. If it doesn't induce the reader to proceed to the second sentence, your article is dead."[1] In a chapter on speaking, *Management Communication* author James O'Rourke tells the story of a plant controller who was asked to make a five-minute presentation about his value to the company. In an attempt to tap into the imagination of the audience of eighteen senior executives, the controller opened with a race car metaphor. After four sentences, he was cut off and asked to leave the room. In another instance, a speaker's opening joke, drawn from a book of speaking tips, fell flat. After the fact, the speaker reflected that a joke wasn't the best fit for a rather serious audience and noted that "when you lose something in the first two minutes of a talk, you just can't get it back."

With this type of pressure, what's a speaker to do? Often, the best option is to forget the introduction until you know what it's introducing—until you have completed a full draft of your whole speech. That is, don't force an introduction and don't become too invested in your first idea. Write a draft or "working" opening and allow additional options to emerge as you work through the research (including audience research) and content development process. The dual objectives are to capture your audience's attention and to set the stage for your speech. That is, your opening should reflect your stated intent and be an accurate indication of what will follow—the main substance of your speech.

In an article for *YPO*, an association for chief executives under the age of 45, communication strategist Matt Eventoll summarizes effective ways to open a speech and throws in one classic—and oddly common—fail. First, the effective options:

1. **Quote.** Use a relevant quote to set the tone for the speech.
2. **"What if?"** or, similarly, **"Imagine."** Asking a "what if" or "imagine" question immediately engages your audience and invites them to be a part of the creative process.
3. **Question.** Posing a question engages the brain and prompts an instinctive answer, whether internal or verbalized.
4. **Silence.** A strategic silence of two to ten seconds creates an additional level of attention and expectation. The caveat: *you had better be able to deliver!*
5. **Statistic.** A powerful, relevant statistic can convey a key idea with impact and evoke emotion.
6. **Statement.** An emphatic phrase or statement can be used to create a sense of drama and anticipation.

The epic fail, generally followed by a collective disconnect on the part of the audience, is opening with some variation of "thank you for inviting me" or "today I'm going to be talking about." If your audience isn't invested from the beginning, it's likely the point of your speech will never really be heard.

---

1. http://training.npr.org/digital/leads-are-hard-heres-how-to-write-a-good-one/

Contributors and Attributions

CC licensed content, Original

- Starting Your Speech. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 14.17: Starting Your Speech is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 7.17: Starting Your Speech by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

14.18: Body Language and Gestures

Learning Objectives

- Discuss strategies to effectively use body language and gestures to emphasize your message



One of the essential rules, and success factors, for public speaking is authenticity. This is as true for your non-verbal language as it is of the words you say and the ideas you express. Body language and gestures are a form of expression and can be either meaningful or distracting.

Toastmasters International, the global non-profit dedicated to teaching public speaking skills, believes that "gestures are probably the most evocative form of nonverbal communication a speaker can employ."[1] In their *Gestures: Your Body Speaks* publication, they identify the following seven benefits of incorporating gestures into your speech:[2]

1. Clarify and support your words
2. Dramatize your ideas
3. Lend emphasis and vitality to the spoken word
4. Help dissipate nervous tension
5. Function as visual aids
6. Stimulate audience participation
7. Are highly visible

Note: The *improper* use of gestures can have *just as powerful* an effect but will likely be *detrimental*. To avoid this, record yourself presenting and make sure your gestures are consistent with your words. When the two are telling different stories, you create confusion and lose credibility and rapport with the audience.

Body language—how you dress as well as your mannerisms—is another powerful communication element. For perspective on this point, and a powerful speaking and life hack, watch social psychologist Amy Cuddy's "Your Body Language May Shape Who You Are" TED Talk. The core idea is that we make judgments based on body language, and those judgments can predict meaningful life outcomes. In one example cited, social scientist Alex Todorov found that one-second judgments of political candidates' faces predict 70 percent of U.S. Senate and gubernatorial race outcomes. What is perhaps more important, however, is that our body language reflects how we judge, think, and feel about ourselves. The key takeaway from this is that our bodies change our minds. That is, we can change not only how we are perceived but how we perceive ourselves by managing our body language. As a speaker, you must be conscious of, and cultivate, the presence you bring to your speech.

To quote Toastmasters International, “When you present a speech, you send two kinds of messages to your audience. While your voice is transmitting a verbal message, a vast amount of information is being visually conveyed by your appearance, your manner, and your physical behavior.”[3]

Your use of gestures and body movement should reflect not only your personal communication style but should also match the audience and the environment. A good practice is to “preview” the attendees or venue by sitting in on a prior event, watching a video, or scanning the event’s social feeds. This will give you a sense for audience dynamics and the size of the room. Certainly ask the event organizers in advance about the setup of the room in which you will speak. Consider adjusting your gestures to fit the audience, room size, and acoustics. For example, you may may want to tone down your gestures in a smaller space and put more emphasis on vocal rather than physical delivery. This doesn’t mean that you should put your personality on “mute” if you’re a naturally ebullient or expressive person. The key is to manage your mannerisms so they don’t overpower either your audience or your words. If the room is a large auditorium filled with enthusiastic fans, you may want to increase your physical presence with gestures to better “fill” the space. Rehearse new elements so they become fluid and reinforce rather than detract from your message. Remember that gestures and body language are most effective when they’re used as “visual punctuation.”

---

1. Toastmasters International. *Gestures: Your Body Speaks*, p. 8. 2011. Web. 26 Jun 2018.
2. Ibid.
3. Ibid.

Contributors and Attributions

CC licensed content, Original

- Body Language and Gestures. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Man gesturing while speaking. **Authored by**: Derrick Coetzee. **Located at**: https://commons.wikimedia.org/wiki/File:Frank_Schulenberg_speaking_and_gesturing_4.jpg. **License**: *CC0: No Rights Reserved*

This page titled 14.18: Body Language and Gestures is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 7.18: Body Language and Gestures by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15: DEVELOPING AND DELIVERING BUSINESS PRESENTATIONS

Learning Objectives

- Discuss the usefulness of visual aids and identify common presentation tools
- Create a presentation using Microsoft PowerPoint
- Create a presentation using Google Slides
- Create a presentation intended for a business meeting

15.1: Text Boxes and Themes

15.2: Images, Videos, and Arranging Objects

15.3: Introduction to Making a Presentation for a Meeting

15.4: Parts of a Good Presentation

15.5: What's my Presentation About

15.6: Bad Presentations

15.7: Making a Presentation for a Meeting

15.8: Putting It Together- Developing and Delivering Business Presentations

15.9: Assignment- Developing and Delivering Business Presentations

15.10: Why It Matters- Developing and Delivering Business Presentations

15.11: Introduction to Visual Aids

15.12: Presentation Tools and Visual Aids

15.13: Introduction to Using Microsoft PowerPoint

15.14: New Presentations

15.15: Editing and Enhancing a PowerPoint Presentation

15.16: Visuals- Images, Videos, and Objects

15.17: Introduction to Using Google Slides

15.18: Creating and Organizing Presentations

This page titled 15: Developing and Delivering Business Presentations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 15: Recruiting and Selecting New Employees by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.1: Text Boxes and Themes

Learning Objectives

- Insert text boxes
- Change presentation themes

Text Boxes

As with most actions in Google slides, there are multiple was to add a text box:

- Select the Text box icon from the main menu (see Figure 1).
- Select Insert > Text Box.

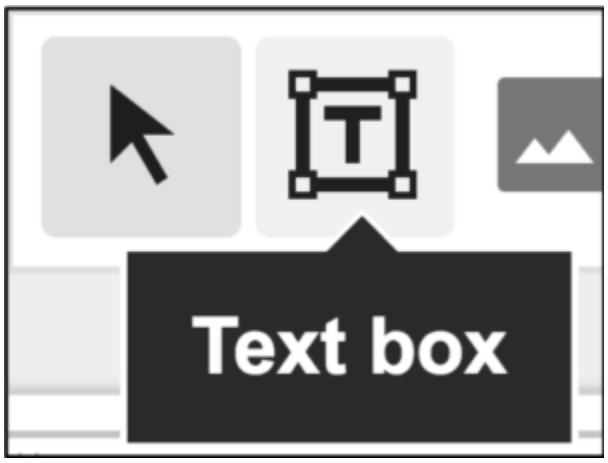


Figure 1.

Using either method, your cursor will become a crosshair; position your cursor where you want the text box to start and click to anchor. Drag the cursor the approximate width and length of your anticipated text entry and release the cursor. If you want to change the location of the text box, click within the text box to highlight it and use the arrow keys on your keyboard to reposition. To change the box dimensions, click within the text box (if it's not already highlighted) and hover over one of the anchor points (tiny boxes). A directional arrow will appear; click to stretch or reduce the size; release when complete.

To delete a text box, select to highlight and press delete. If you change your mind, click on the Undo (Ctrl+Z) arrow.

Themes

Themes are one of the formatting options within Google Slides. To view theme options, click on Slide > Change theme. A column will appear on the right of the screen, showing various color, layout and font options. To see how a particular combination would work with your content, click on your chosen theme. You can continue test-driving themes; clicking on a new theme will replace the prior choice. There's no risk of losing your original formatting; Undo (or Ctrl+Z) will return you to your original theme. You can also import a theme from outside Google Slides. Click on the Import theme button at the bottom of the themes column. Sites such as Slides Carnival allow you to copy slide formats directly into Google Slides.

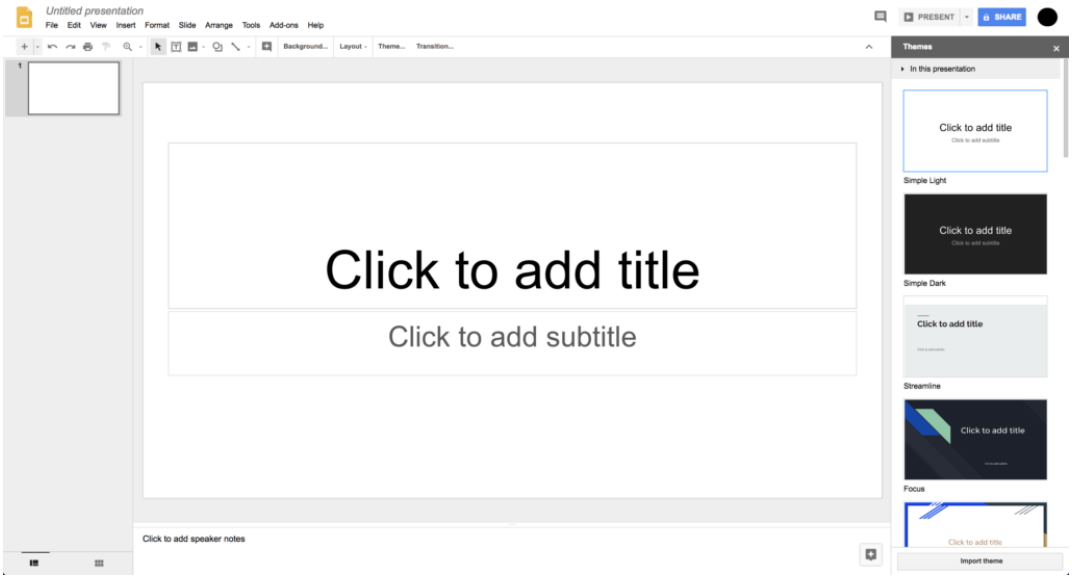


Figure 2.

Contributors and Attributions

CC licensed content, Original

- Text Boxes and Themes. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 15.1: Text Boxes and Themes is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.1: Text Boxes and Themes by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.2: Images, Videos, and Arranging Objects

Learning Objectives

- Add images and clip art
- Add videos
- Arrange objects

Images and Clip Art

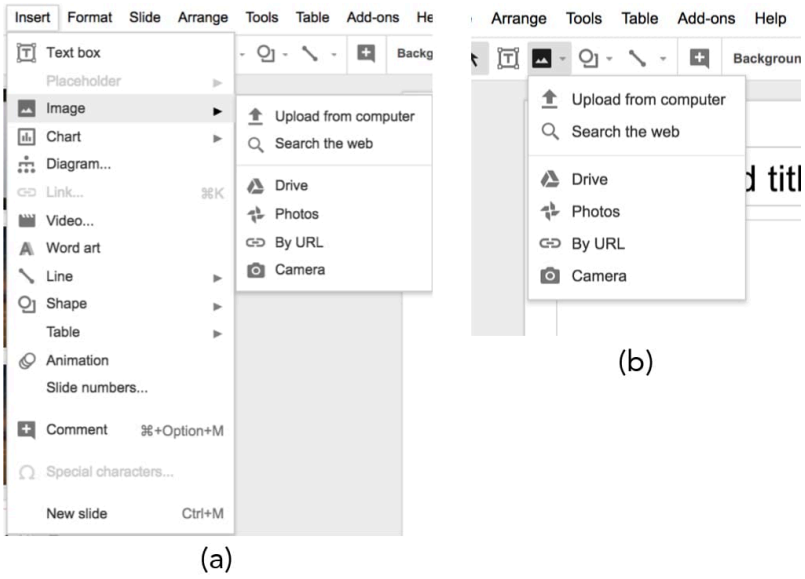


Figure 1.

Using art and images adds impact and provides welcome relief to the challenge of processing information-intensive slides. To insert an image, select Insert > image or the image icon, to the right of the Text box icon. As shown in Figure 1, there are multiple options for accessing images; you can also access these options using the drop-down menu to the right of the Image icon. You can upload images from your computer, use the camera on your device to take a photo, link to an image using a URL, or Search Google Images. Selecting Search the Web will link you to Google Images, where you can use a keyword or keywords to search for images, including quote images based on your keyword(s).

When your results appear, you'll see a small magnifying lens in the bottom right corner of the images. Clicking on the magnifying lens allows you to preview the image prior to selection/inserting. To exit preview (if you don't want to choose this image), press the left (return) arrow to return to the search options.

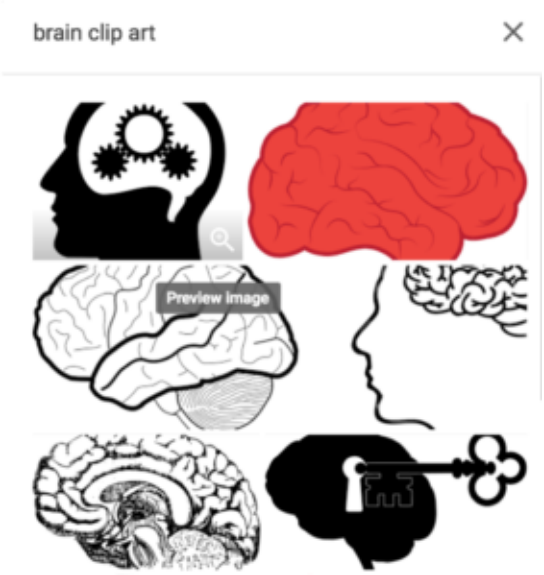


Figure 2.

To select an image, click on it and click on the INSERT button at the bottom of the window. The image will be inserted into your current (highlighted in the thumbnail) slide. If you change your mind, press the escape key (ESC), Undo, or Ctrl+Z to reverse the Insert operation.

To incorporate clip art, select Insert > Image to open the Google Image browser. In the Search area “Search for Google Images,” enter your keyword or desired image and clip art as your search terms. For example, “brain clip art” or “brain + clip art.” For more professional images, include “stock” – stock image or stock photo – in your search. For example, “brain stock image.”

Images shown are generally in the public domain or labeled for commercial use with modification, but as the disclaimer on the Google Image window notes, you are responsible for ensuring that your intended use is in accordance with the image creator’s license.

Videos

Adding video to your Google slides adds yet another level of visual interest and engagement. To incorporate video:

1. Select Insert > Video from the main menu. A window will appear where you can search for a YouTube video, or you can select URL at the top of the window, and paste a URL for a specific YouTube video.
2. Click on a video to select it and press the Select button to embed it in your current slide. Your video is now embedded in your slide.

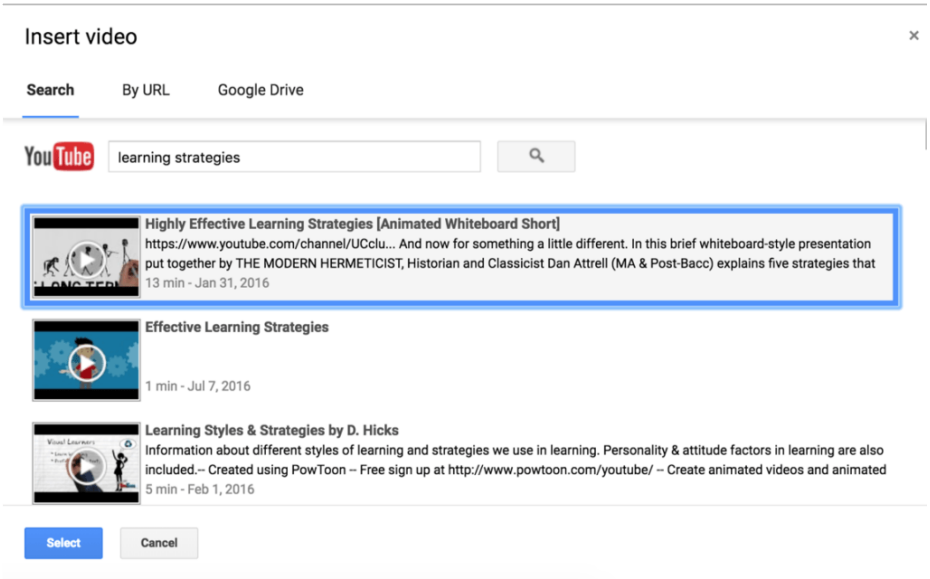


Figure 3.

Arranging Objects

When you're creating a slide with multiple elements—for example, layering text, imagery, lines, shapes or other objects—you may find a need to arrange the elements or objects.

To manage objects on a slide:

1. Click Arrange from the main menu.
2. Choose from the following options:
    - Order: Put the object behind or in front of text, other objects, or images.
    - Align horizontally or Align vertically: When you select multiple objects, you can align the edges of the objects.
    - Center on page: Center objects vertically or horizontally on a slide.
    - Rotate: Change the orientation of an object, by flipping or rotating it.
    - Distribute: When you select three or more objects, you can evenly distribute the space between them.
    - Group: Lock multiple objects together to make them easier to move around and format.

The Google Slides grid line feature, enabled by default, also provides a visual to help with alignment. When you click and drag an object, you will see colored lines appear that will help you align the object with other objects on the page. This setting (View > Snap to > Guides) is selected by default.

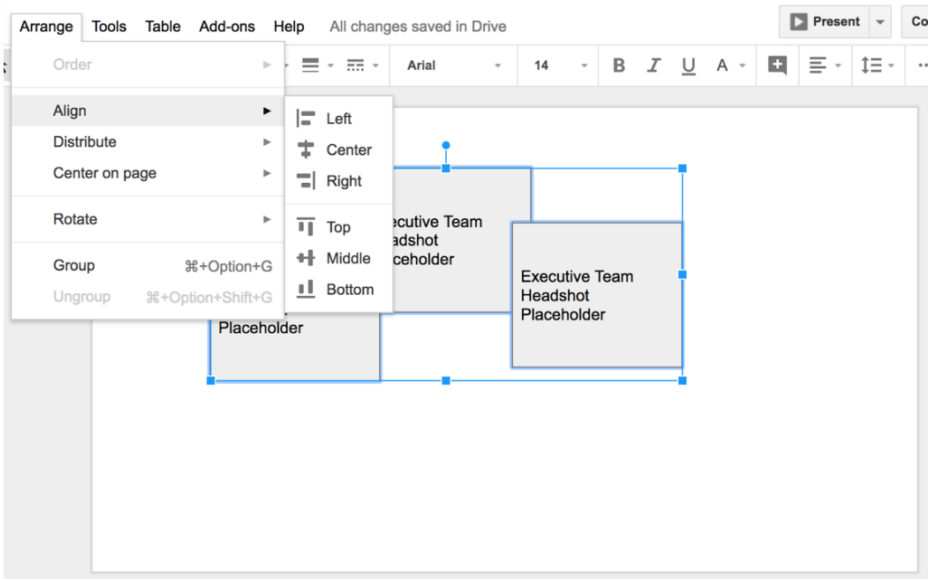


Figure 4.

Contributors and Attributions

CC licensed content, Original

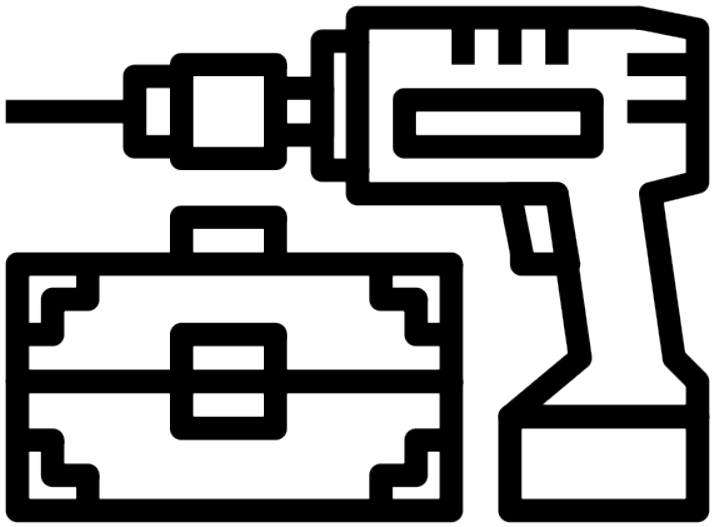
- Images, Videos, and Arranging Objects. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 15.2: Images, Videos, and Arranging Objects is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.2: Images, Videos, and Arranging Objects by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.3: Introduction to Making a Presentation for a Meeting

What you'll learn to do: Create a presentation intended for a business meeting



Tools, no matter how sophisticated, are simply tools. Moving from the right tools to a good presentation involves perspective and planning. For perspective, we'll approach the concept of a good presentation from two standpoints: identifying the key features of a good presentation and common mistakes that contribute to presentation failure. We'll also discuss what's involved in the planning process, including the three essential questions that need to be answered prior to developing content. Finally, we'll explore the classic story structure and apply that structure to a business presentation scenario.

Contributors and Attributions

CC licensed content, Original

- Introduction to Making a Presentation for a Meeting. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Toolbox. **Authored by**: Chaowalit Koetchuea. **Provided by**: Noun Project. **Located at**: https://thenounproject.com/term/toolbox/1458241/. **License**: *CC BY: Attribution*

This page titled 15.3: Introduction to Making a Presentation for a Meeting is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.3: Introduction to Making a Presentation for a Meeting by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.4: Parts of a Good Presentation

Learning Objectives

- Identify key features of a good presentation

Like reverse engineering a product, we can distill the key features of a good presentation by looking at presentation evaluation scorecards. Refer to Table 1 for a sample class presentation grading rubric.

Table 1. Presentation Grading Rubric

Criteria	Rating		Pts
Overall communication effectiveness. *Presentation reflects thoughtful planning (content) and development (structure) and engaged delivery (style)*	10 pts (Full Marks)	0 pts (No Marks)	10 pts
Presenter has a unique voice and compelling message.	10 pts (Full Marks)	0 pts (No Marks)	10 pts
Presentation reflects learning. *Content reflects knowledge and thoughtful development of subject. Speaker applies communication concepts learned and/or incorporates relevant terminology. Points are supported by credible sources/data.*	50 pts (Full Marks)	0 pts (No Marks)	50 pts
Slides reflect thoughtful design. *Slides reflect an understanding of and ability to apply design best practices (i.e., 10-20-30 rule; "bullet points kill")*	20 pts (Full Marks)	0 pts (No Marks)	20 pts
Presentation meets stated requirements. *Presentation meets subject matter, research, length & formatting (including citation) requirements.*	10 pts (Full Marks)	0 pts (No Marks)	10 pts
		Total Points:	100

At the macro level, the key elements of a good presentation are content, organization, and delivery. There are both substance and style aspects of content. Substance elements include the originality and significance of your idea, the quality of your research and analysis, clarity and potential impact of your recommendations. Style aspects of content include confidence and credibility, both of which have a significant impact on how you—and your message—are received.

Good organization starts with a strong opening and continues in a logical and well-supported manner throughout the presentation, leading to a close that serves as a resolution of the problem or a summary of the situation you've presented. The audience experiences good organization as a sense of flow—an inevitable forward movement to a satisfying close. This forward momentum also requires audiences to have a certain level of technical and information-management competency. To the latter point, good presentation requires a presenter to put thought into information design, from the structure and content of slides to the transitions between individual points, slides and topics.

Delivery entails a range of factors from body language and word choice to vocal variety. In this category, your audience is responding to your personality and professionalism. For perspective, one of the three evaluation categories on the official Toastmasters speaker evaluation form is "As I Saw You;" in parentheses: "approach, position, personal appearance, facial expression, gestures and detracting mannerisms." A good presenter has a passion for the subject and an ability to convey and perhaps elicit that emotion in the audience. Audience engagement—through eye contact, facial expression, perhaps the use of gestures or movement—also contributes to an effective presentation. However, to the point in the Toastmasters evaluation, gestures, movement other mannerisms can be distracting (see Module 7: Public Speaking for more on this). What works: natural (not staged) movement that reinforces communication of your idea.



Figure 1. The WIIFM Principle.

With those key features and presentation-evaluation criteria in mind, let's add a disclaimer. The reality is that your features won't matter if you don't deliver one essential benefit: relevance.

Whether you think in Toastmasters terminology—"What's in it for me? (WIIFM)" from the audience perspective—or put yourself in the audience's position and ask "So what?," it's a question that you need to answer early. We'll get into this more in the next section as we discuss presentation planning.

Contributors and Attributions

CC licensed content, Original

- Parts of a Good Presentation. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Modification of WIIFM. **Authored by**: Nathan Stephens. **Located at**: https://flic.kr/p/dEFKQS. **License**: *CC BY-SA: Attribution-ShareAlike*

This page titled 15.4: Parts of a Good Presentation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.4: Parts of a Good Presentation by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.5: What's my Presentation About

Learning Objectives

- Identify the purpose, audience, and message of your presentation

It may be helpful to think of your presentation as having three key moving parts or interlocking gears: purpose, audience and message. Let's walk through the presentation-development process at this planning level.

Purpose

Generally the first step in developing a presentation is identifying your purpose. Purpose is a multi-layered term, but in this context, it simply means objective or intended outcome. And why is this? To riff on the classic Yogi Berra quote, if you don't know where you're going, you might as well be somewhere else. That is, don't waste your audience's (or your own) time.

Your purpose will determine both your content and approach and suggest supplemental tools, audience materials and room layout. Perhaps your purpose is already defined for you: perhaps your manager has asked you to research three possible sites for a new store. In this case, it's likely there's an established evaluation criteria and format for presenting that information. Voila! your content and approach is defined. If you don't have a defined purpose, consider whether your objective is to inform, to educate, or to inspire a course of action. State that objective in a general sense, including what action you want your audience to take based on your presentation. Once you have that sketched in, consider your audience.

Audience

The second step in the presentation development process is audience research. Who are the members of your audience? Why are they attending this conference, meeting, or presentation? This step is similar to the demographic and psychographic research marketers conduct prior to crafting a product or service pitch—and is just as critical. Key factors to consider include your audience's age range, educational level, industry/role, subject matter knowledge, etc. These factors matter for two reasons: you need to know what they know and what they need to know.

Understanding your audience will allow you to articulate what may be the most critical aspect of your presentation: "WIIFM," or what's in it for them. Profiling your audience also allows you adapt your message so it's effective for this particular audience. That is, to present your idea (proposal, subject matter, recommendations) at a depth and in a manner (language, terminology, tools) that's appropriate. Don't expect your audience to meet you where you are; meet them where they are and then take them where you want to go together.

Returning to the site analysis example mentioned earlier, knowing your audience also means getting clear on what management expects from you. Are you serving in an analyst role—conducting research and presenting "just the facts"—to support a management decision? Or are you expected to make a specific recommendation? Be careful of power dynamics and don't overstep your role. Either way, be prepared to take a stand and defend your position. You never know when a routine stand-and-deliver could become a career-defining opportunity.

Message

The third step is honing your message. In "TED's Secret to Great Public Speaking," TED Conference curator Chris Anderson notes that there's "no single formula" for a compelling talk, but there is one common denominator: great speakers build an idea inside the minds of their audience. Take, for instance, Chimamanda Adichie's idea, which Anderson summarizes as "people are more than a single identity."[1] As Adichie expresses it: "The problem with stereotypes [of a single story or identity] is not that they are untrue, but that they are incomplete."[2] Or Sir Ken Robinson's idea that creativity is a essential building block for learning. As he expresses the idea: "My contention is that creativity now is as important in education as literacy, and we should treat it with the same status."[3] Ideas matter because they're capable of changing our perceptions, our actions and our world. As Anderson puts

it: “Ideas are the most powerful force shaping human culture.”[4]

So if ideas are that powerful, more is better, right? Perhaps a handful or a baker’s dozen? Wrong. As any seasoned sales person knows, you don’t walk into a meeting with a prospective client and launch into an overview of every item in your company’s product or service line. That’s what’s known as “throwing spaghetti on the wall to see what sticks.” And that’s an approach that will have you wearing your spaghetti—and perhaps the dust from one of your client’s shoes on your backside, as well. What audience members expect is that you’ve done your homework, that you know them and their pain, and that you have something to offer: a fresh perspective, an innovative approach or a key insight that will change things for the better. As Chris Anderson puts it: “pick one idea, and make it the through-line running through your entire talk.”[5] One message, brought vividly and relevantly to life.

So now that you have a macro view of the presentation development process, let’s review what can what can—and often does—go wrong so we can avoid the common mistakes.

---

1. Anderson, Chris. “TED’s Secret to Great Public Speaking.” *TED*, March 2016.
2. Adichie, Chimamanda Ngozi. “The danger of a single story.” *TED*, July 2009.
3. Robinson, Ken. “Do schools kill creativity?” *TED*, Feb 2006.
4. Anderson, *TED*
5. Anderson, *TED*

Contributors and Attributions

CC licensed content, Original

- What's my Presentation About. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 15.5: What’s my Presentation About is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.5: What’s my Presentation About by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.6: Bad Presentations

Learning Objectives

- Discuss common mistakes in presentations

For many, the prospect of developing and delivering a business presentation rates right up there with death and taxes. Interestingly, that same mixture of fear and loathing is often felt by audience members as well. But it doesn't have to be that way. The ability to craft a compelling story is a skill as old as the human race, and the need to communicate is as primal and potentially powerful.



Figure 1. Akhenaten as a sphinx, and was originally found in the city of Amarna.

For millions of years before the invention of modern technology, humans used the tools available to perpetuate traditions and culture and to document—and often rewrite—history. Do a few internet searches and immerse yourself in the Egyptian tombs; the caves of Chauvet; or El Castillo, the Temple of Kukulcan. What you're experiencing is a feat of both artistry and communication. Although we don't know the full significance of these early carvings and structures, there's no doubt that these early humans captured their world view in a way that is still deeply resonant. While the tools have changed, the communication challenges—and opportunity—remain the same: to communicate an engaging and inspiring point of view.

Regardless of whether you want to change the world, build your brand, or build a billion-dollar business, effective presentation skills are essential. To quote legendary investor, philanthropist and Berkshire Hathaway chairman and CEO Warren Buffet, "If you can't communicate and talk to other people and get across your ideas, you're giving up your potential."[1] As would be expected of a numbers person, Buffet has quantified his point in talks on student campuses and professional organizations. Speaking at his alma mater in 2009, Warren Buffett told Columbia Business School students that he believed learning effective communication skills could translate into 50 percent higher lifetime earnings.

Given our vibrant storytelling tradition and with so much at stake, why are there still so many bad presentations? Wouldn't you think that modern communication technology—considering the advances in graphics and communications software alone!—would lead to more compelling presentations? Interestingly, the problem is, to some extent, the technology. It's estimated that 30 million PowerPoint presentations are created every day, with (seemingly) a majority of presenters opting for default layouts and templates. The problem is, we're wired for story, not bullet points. A related failure is our use of available technology.

Seth Godin has a wonderful—and instructive—rant on these points: Really Bad PowerPoint (and how to avoid it), blaming Microsoft wizards, templates, bui -2" class="footnote">[2]

So how do we avoid causing “death by PowerPoint”—or by whatever presentation software we use? The common denominator of presentation mistakes is that they represent a failure of communication. This failure can be attributed to two errors: too much or too little. The error of too much is generally the result of trying to use slides as a teleprompter or a substitute to a report, or, it would seem, to bludgeon the audience into submission. Of course, this tends to have an alternate effect, namely, prompting audience members to walk out or tune out, turning their attention instead to doodling or their device of choice.

What bad presentations have too little of is emotion. Presentation expert and author of the classic Presentation Zen (and 4 related books) Gar Reynolds captures the crux of the problem: “a good presentation is a mix of logic, data, emotion, and inspiration. We are usually OK with the logic and data part, but fail on the emotional and inspirational end.”[3] There’s also a hybrid too little-too much mistake, where too little substance and/or no design sensibility is — in the mind of the presenter — offset by transitions and special effects. Heed Seth Godin’s advice: “No dissolves, spins or other transitions. None.”[4]

The 10/20/30 rule, generally attributed to venture capitalist Guy Kawasaki, is a good guideline to help you achieve a “just right” balance in your presentations. Geared for entrepreneurs pitching their business, his advice is a discipline that would improve the quality—and, effectiveness—of most presentations. In brief, 10/20/30 translates to a maximum of 10 slides, a maximum of 20 minutes and a minimum of 30 point font.[5]

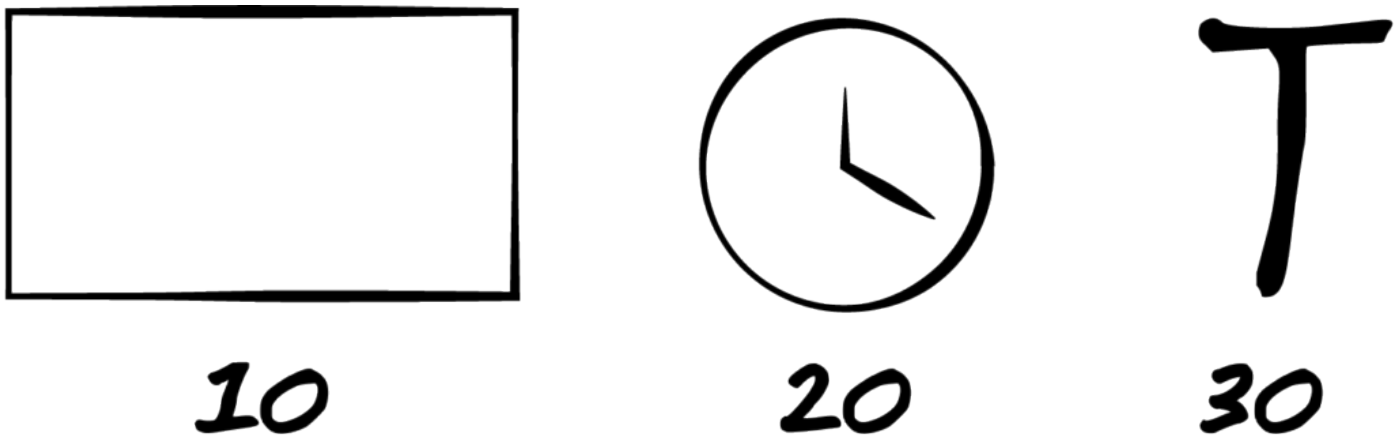


Figure 2. Your presentation should have no more than 10 slides, take no more than 20 minutes, and use type no smaller than 30 point font.

While this rule is a good starting point, it doesn’t overrule your audience analysis or understanding of your purpose. Sometimes, you may need more slides or have a more involved purpose—like training people in new software or presenting the results of a research study—that takes more than 30 minutes to address. In that case, go with what your audience needs and what will make your presentation most effective. The concept behind the 10/20/30 rule—to make new learning easy for your audience to take in, process and remember—should still be your guide even if you don’t follow the rule exactly.

How to Avoid Death By PowerPoint

For more on how to avoid causing death by PowerPoint, watch Swedish presentation expert and How to Avoid Death By PowerPoint author David Phillips TED Talk on the topic:



A YouTube element has been excluded from this version of the text. You can view it online here: pb.libretexts.org/bcsfm/?p=380

1. Gallo, Carmine. "How Warren Buffet and Joel Osteen Conquered Their Terrifying Fear of Public Speaking," *Forbes*. May 16, 2013.
2. Rouse, Margaret. "What is death by PowerPoint?" *TechTarget Network*.
3. Reynolds, Garr. "10 tips for Improving Your Presentations Today," Presentation Zen. Nov 2014.
4. Godin, Seth. *Fix Your Really Bad PowerPoint*. Ebook, sethgodin.com, 2001.
5. Kawasaki, Guy. *The 10/20/30 Rule of PowerPoint*. December 2005.

Contributors and Attributions

CC licensed content, Original

- Bad Presentations. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Echnaton (Akhenaten) as Sphinx. **Authored by**: Hans Ollermann. **Located at**: https://flic.kr/p/4nvAVm. **License**: *CC BY: Attribution*

All rights reserved content

- How to avoid death By PowerPoint. **Authored by**: David JP Phillips. **Provided by**: TED. **Located at**: https://youtu.be/Iwpi1Lm6dFo. **License**: *All Rights Reserved*

This page titled 15.6: Bad Presentations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.6: Bad Presentations by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.7: Making a Presentation for a Meeting

Learning OUtcomes

- Create a presentation intended for a business meeting

With perspective on the technical tools, communications planning and information design, let's take this learning for a test drive.

What's considered an effective (that is, persuasive) presentation structure hasn't changed fundamentally over the centuries. In his analysis of dramatic structure in the *Poetics*, Aristotle identified a play as having three parts: a beginning, middle and end. The story begins with a "complication" (problem), ends with an "unraveling" (resolution), and follows a logical sequence of events from beginning to end. Hollywood screenwriters use the same structure and dynamics. Screenwriter, producer and author Syd Field, whom CNN called "the guru of all screen writers," translated this simple three-step structure into numerous books and workbooks, including the bestsellers *Screenplay: The Foundations of Screenwriting* and *The Screenwriter's Workbook*.

In a business context, a good presentation is an effective presentation. That is, a good presentation achieves its intended outcome. Clearly, in order to achieve a specific outcome or objective, you need to know what it is. So, prior to crafting the drama (in word or slide), you need to hone in on three things:

- The purpose of your presentation
- Your audience
- Your (one) message

For a review of these elements, refer to What's my Presentation About.

Once you're clear on those points, let's proceed.

To build our presentation, we'll use presentation expert Nancy Duarte's interpretation of the classic 3-part story structure illustrated in Figure 1. For additional perspective on this structure, watch her TED Talk, "The Secret Structure of Great Talks," or read her Harvard Business Review article, "Structure Your Presentation Like a Story."

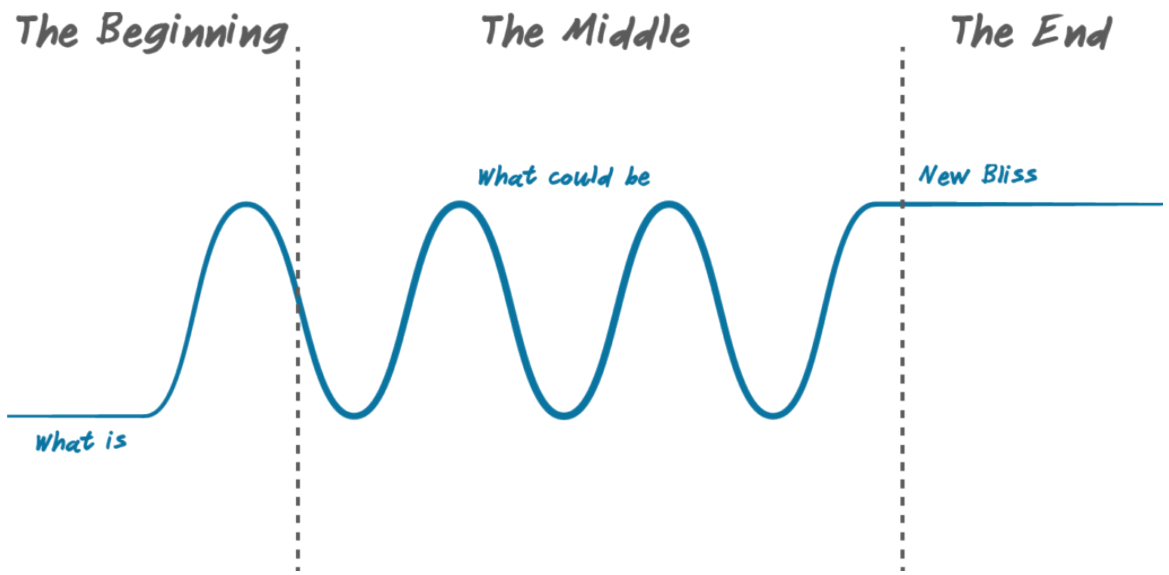


Figure 1. Persuasive story structure (Duarte, "Structure Your Presentation Like a Story," 2012).

The Beginning

The story starts with "What is"—the current state. Describe this baseline state in a way that is recognizable to the audience. This

allows you and the audience to get in sync. And with this base level of agreement, your audience will be more receptive to your proposed change.

The second step is to introduce "What could be." The gap between what is and what could be adds tension and drama to your story and largely determines the significance of your presentation. If there's no conflict, no proposed change, what's the point of the presentation?

Let's say you're an analyst on the new product development team of a retailer known for exclusive, trend-forward "house" branded products. Your company's reputation and revenue depends on consistent introduction of new consumer-product goods. Marketing and distribution are key strengths, but new-product performance is off, revenue is below expectations and the company's stock price recently fell 30 percent. Within your company, R&D (research & development) is strictly an insider's game; any ideas or innovations that weren't developed in-house are blocked. The problem is, you can't innovate fast enough—or with enough market demand accuracy—to meet financial and stock market expectations. You and the other analysts on your team have been tracking innovation trends and successes and you think the answer is opening the R&D works to outside ideas and innovations. Here's how you might lay out your presentation:

- **What Is:** We missed our quarterly earnings numbers, largely due to a failure to meet our innovation success targets over the last six months.
- **What Could Be:** Initial data suggests we could get back on track by modifying our R&D model to incorporate external innovations.

The Middle

The bulk of your the presentation is developing the contrast between what is and what could be in order to set up your proposed resolution of the conflict or challenge. The objective is also to establish the validity of your arguments, so your proposed call to action is perceived as a logical, ideally inevitable, conclusion of the conflict.

- **What Is:** We currently bear the full cost and risk of developing new products and our innovation success rate—the percentage of new products that meet financial objectives—is running 25 percent below target.
- **What Could Be:** Sourcing promising innovations from outside the company could reduce R&D costs and risk while also increasing our innovation success rate.
- **What Is:** Our R&D process is taking so long that we're missing trends and losing our market-leading brand reputation.
- **What Could Be:** We could license or buy promising innovations for a fraction of the cost it would take to develop them from scratch and leverage our marketing and distribution strengths to claim shelf and market share.
- **What Is:** Our below-plan performance and new product pipeline is costing us political capital with executive management, and we're at risk of losing budget and/or layoffs.
- **What Could Be:** Adopting an open innovation culture would allow us to create partnerships that leverage our strengths and drive revenue, regaining a position of value within the company.

The End

To craft a powerful close, heed Duarte's advice and avoid a list of bullet point to-dos. Your objective here is to achieve resolution of the conflict introduced at the beginning, to issue a call to action that inspires your audience to support your vision of what could be, a state Duarte refers to as the "new bliss."

Call to Action

To recover our position of a source of revenue and brand value, we need to start working to build a culture and networks that support open innovation and accelerate the development of new products, regardless of the source of the idea.

New Bliss

Our ability to drive value secures our position and reputations in the company, and in the marketplace, and pays off in employee stock value and profit sharing.

The new bliss articulates the proposed—and a desired future state—incorporating the WIIFM, what's in it for me, that motivates

your audience to buy into and work to support the required change.

Contributors and Attributions

CC licensed content, Original

- Making a Presentation for a Meeting. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Persuasive storytelling graphic. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 15.7: Making a Presentation for a Meeting is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.7: Making a Presentation for a Meeting by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.8: Putting It Together- Developing and Delivering Business Presentations

Let's return to your presentation for SB '18 and assume you were able to power through your doubts and accept the call. Now that you have a solid understanding of presentation tools, options and techniques, let's put it all together.

Creating a good presentation involves research and analysis, reflection and distillation. As is true of many things in life, the process will go more smoothly if you start with the end in mind. Prior to putting words on paper/slides, address the three presentation planning priorities: purpose, audience and message. These three priorities will determine (or at least inform) your content, presentation tools, and techniques. Note: Message—the idea you want to communicate—is singular, not plural. Think simple, clear and compelling.



Figure 1. We are all storytellers

In addition to keeping your end (your desired outcome or audience action) in mind, you need to consider your audience's desired outcome. As Theodore Roosevelt noted: "Nobody cares how much you know, until they know how much you care." And speaking of care, do your audience a favor and invest at least as much time in the design of your slides as the selection of your words. Remember that communication in whatever format is a fundamentally human interaction; invest your presentation with your personality: your passion, your point of view and your sense of humor (if you have one).

Finally, before going live, test drive your presentation. Ask colleagues, friends or family to listen to a dry run and rate you on the presentation evaluation criteria. In particular, identify and address any words or images that may represent a barrier to effective communication.

Let go of the memories of bad presentations—and reject the default choices and templates that contribute to ugly slides and ineffective presentations. Think of the many stories that have engaged you and that you have told over the years and channel that energy when you get up to speak.

The art of storytelling

If you're interested in learning more on about cultivating your innate storytelling skills, you can check out Pixar's The Art of Storytelling on Kahn Academy.

Contributors and Attributions

CC licensed content, Original

- Putting It Together: Developing and Delivering Business Presentations. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Hands at the Cuevas de las Manos. **Authored by**: Mariano. **Provided by**: Wikimedia. **Located at**: https://commons.wikimedia.org/wiki/File:SantaCruz-CuevaManos-P2210651b.jpg. **License**: *CC BY: Attribution*

This page titled 15.8: Putting It Together- Developing and Delivering Business Presentations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.8: Putting It Together- Developing and Delivering Business Presentations by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.9: Assignment- Developing and Delivering Business Presentations

Your task is to revisit the presentation you created for Module 7: Public Speaking and submit a revised final version.

In this version of your presentation you will incorporate the feedback from your peer group and you will use the voice over feature in PowerPoint to narrate your revised presentation.

Along with the narrated slides, you will submit a short essay (minimum 300 words) to document and reflect on any revisions you made based on the peer feedback you received. Please discuss what and why you changed or added anything, and how these changed made the presentation stronger.

Your PowerPoint presentation should have a minimum of 4 slides: a title page, content slides, and conclusion. Each slide you include in the presentation must be narrated using the voice over function. Email a final copy of your presentation and your revision essay to your instructor.

Assignment Resources

- Guide to Record a slide show with narration and slide timings
- Watch: How to create voice-over narration for your PowerPoint Presentation

Grading Rubric

Criteria	Not Evident	Developing	Exemplary	Points
Submit your revised slide deck	**0 pts** No slide deck submitted	**5 pts** Slide deck has not been revised to reflect peer review	**10 pts** Slide deck has been revised to reflect peer review	10 pts
Provided voice over functionality for each slide	**0 pts** No response to peers	**5 pts** Voice over only appears on some slides	**10 pts** Voice over is on all slides	10 pts
Speak with appropriate volume, speed, and clarity	**0 pts** No presentation completed	**5 pts** Presentation cannot be heard, is too fast or too slow, or cannot be understood	**10 pts** Presentation is spoken at an adequate volume, speed, and clarity	10 pts
Present your slides effectively	**0 pts** No presentation completed	**5 pts** Presentation is disjointed, unprofessional, or too reliant on slide deck text	**10 pts** Presentation is professional and feels cohesive	10 pts
Submit reflection essay on slide deck revisions	**0 pts** No essay submitted	**5 pts** Essay is not at least 300 words long	**10 pts** Essay is at least 300 words long, and discusses the improvements made to the slide deck	10 pts
			Total:	**50 pts**

Contributors and Attributions

CC licensed content, Original

- Assignment: Developing and Delivering Business Presentations. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- How to create voice-over narration for your PowerPoint Presentation. **Authored by**: Dawn Dubriel. **Provided by**: YouTube. **Located at**: https://www.youtube.com/watch?v=3uk4CU7uobM. **License**: *Public Domain: No Known Copyright*

This page titled 15.9: Assignment- Developing and Delivering Business Presentations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.9: Assignment- Developing and Delivering Business Presentations by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.10: Why It Matters- Developing and Delivering Business Presentations

It's mid-May; a major retail management conference is less than one month away. Your manager has been recruited as a speaker in the Redesigning Retail track at the annual Sustainable Brands conference. She's asked you to assemble background information and tune in to relevant media and social media feeds. You're scheduled to meet with her next week to discuss your findings and an approach to her topic: "To-the-Bank Trends: How Green is Driving Retail Dollars."

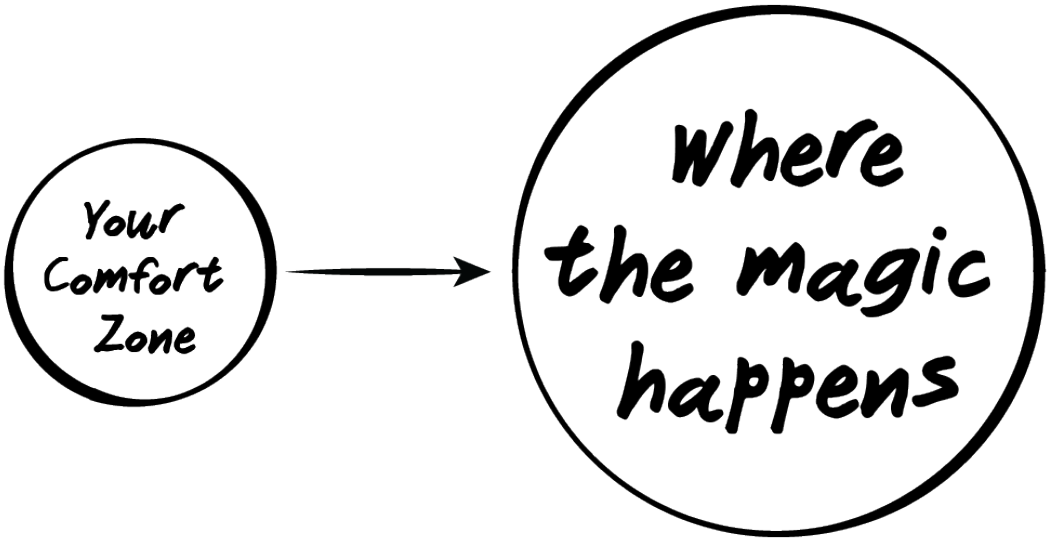


Figure 1. Career moments happen outside of our comfort zone

On Monday morning, your hear your manager shattered her leg in a rock-climbing accident over the weekend, and she will be out of commission for three months. When she calls you later that week, you're taut with anticipation—this is your opportunity to step up and demonstrate you have what it takes to be a manager. You're ready! In fact, you've been ready. But what you didn't realize is that the person who steps up will also be the person stepping out on stage at SB '18 and presenting in front of thousands of high-level attendees. This is what's known as a career moment.

In life, as in mythology, we're never really ready for the call. Regardless, ready or not, our response is what defines both our character and our career trajectory.

So let's prepare for the call.

Contributors and Attributions

CC licensed content, Original

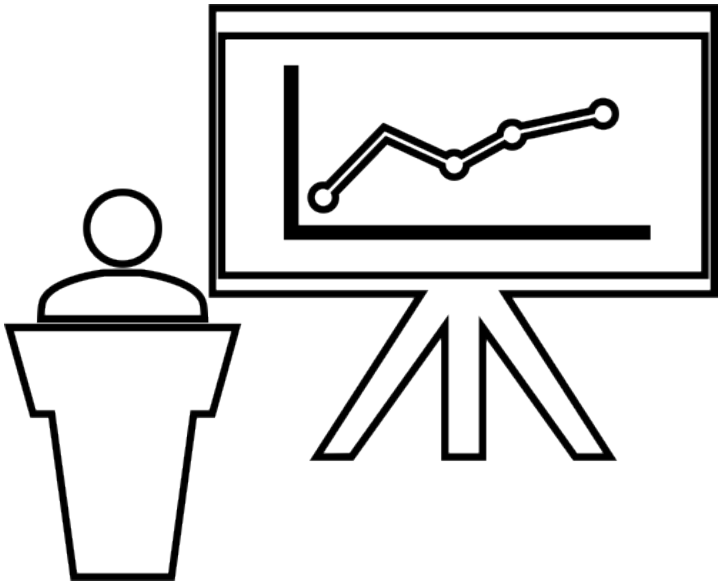
- Why It Matters: Developing and Delivering Business Presentations. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- Comfort Zone Illustrations. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 15.10: Why It Matters- Developing and Delivering Business Presentations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.10: Why It Matters- Developing and Delivering Business Presentations by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.11: Introduction to Visual Aids

What you'll learn to do: Discuss the usefulness of visual aids and identify common presentation tools



How does one prepare for the proverbial call or career moment? In this module, we'll focus specifically on business presentations. In this section, we'll explore presentation tools and factors to keep in mind when evaluating materials, from your choice of words and images to your presentation style.

Contributors and Attributions

CC licensed content, Original

- Introduction to Visual Aids. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Presentation image. **Authored by**: fahmionline. **Provided by**: The Noun Project. **Located at**: https://thenounproject.com/fahmionline/collection/presentation/?i=1368638. **License**: *CC BY: Attribution*

This page titled 15.11: Introduction to Visual Aids is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.11: Introduction to Visual Aids by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.12: Presentation Tools and Visual Aids

Learning Objectives

- Discuss key concepts to keep in mind as you create business presentations
- Discuss available presentation tools to help engage your audience

Key Considerations



Figure 1. There are four commonly accepted modalities for learning, often abbreviated as VARK.

Presentation software allows you to take an oral presentation to the next level—engaging your audience verbally and visually as well as aurally. What's particularly powerful about using presentation software and other visual aids is the ability to use imagery to bridge cultural and language gaps and arrive at a shared understanding of the issue/opportunity at hand.

A related point to keep in mind is that words have two different meanings—a literal or denotative meaning (think: Merriam-Webster or Wikipedia definition) and a more subjective or connotative meaning. The connotative meaning of a word is based on a person's cultural background and experiences and has emotional and/or judgement associations. Accomplished presenters are attuned to their audience and avoid words or references that may be misinterpreted by non-native speakers or may be perceived as emotionally "loaded" by audience members from a different subculture. In an increasingly diverse society, cultural awareness is as important for business communicators as it is for international marketers. To ensure that the message you intend to convey is what will be received, ask peers or colleagues—ideally, those with a socio-cultural profile similar to that of your audience—for feedback, with particular attention to the subtext of words and images.

Using multimedia—images, photos and video and animation—that supports your point also provides repetition and can increase retention. A memory research pioneer, German psychologist Hermann Ebbinghaus, found that we forget approximately 50 percent of new information within 18 minutes, with retention falling to 35 percent after a week. However, Ebbinghaus also discovered that repetition of the new information at key intervals can change this trajectory, a discovery known as the spacing effect. Specifically, repeating the information at a 10–20 minute, 24 hours and 7 day intervals countered the initial memory loss and reduced the subsequent rate of memory loss. The lesson for presenters: work repetition into your presentation and your follow-up. Figure 2 shows an illustration of the Forgetting Curve and Spacing Effect.

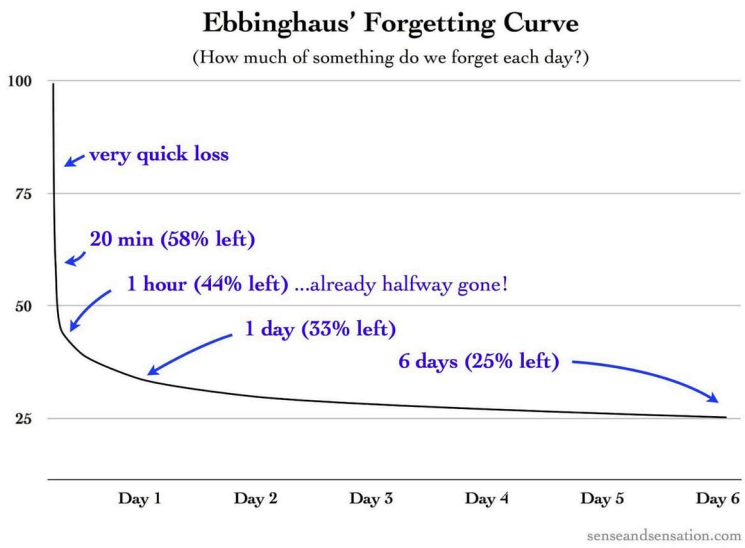


Figure 2. The Forgetting Curve

Common Presentation Tools

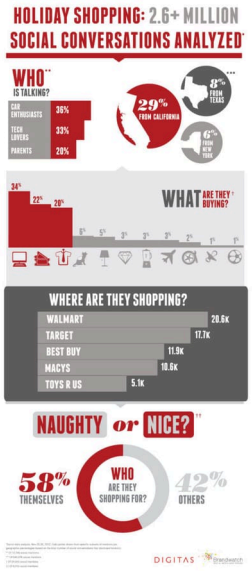


Figure 3. Infographics can be effective visual aids. Click on the image for a larger view.

The right tool for the job depends, of course, on the job. In this case, that means examining your audience and objective. If, for example, your task is simply to present "the facts," there's no need to consider interactive tools and techniques. If, however, your objective is to educate and/or inspire, you may want to consider a range of options for involving your audience, engaging them as participants or even co-presenters. For example, some workshops require participants—generally in group—to solve challenges or "stand and deliver." That is, to review and present a segment of the material to the audience or peers. Or perhaps your goal is to engage a group in a training or strategic planning exercise. In this case, you would want to incorporate tools that support participative learning and collaboration such as Post-sIt Note Pads, or packages of smaller note pads (don't forget markers, pens and highlighters) that can be arranged and rearranged as a pattern or plan emerges. Also consider easels, dry erase boards and other surfaces that lend themselves to idea sharing.

Whether you're presenting to a K-12, higher education, or business audience will also influence your choice of primary and supplemental tools: handouts, product samples, giveaways, worksheets, and snacks (yes, even for the adults). If your assignment is to develop and present a business presentation to be delivered to your Business Communications class peers, the topic, format and any supporting materials may be pre-defined. But don't stop there. If you're proposing an edible garden space on

campus, you could make or hand out seed packets. Think about how to differentiate yourself and your proposal—whatever you're proposing—in a way that's relevant and memorable.

Similarly, if you're presenting to your management, there may be a company standard template and tools that you're expected to use. Again, you can distinguish yourself by your knowledge and application of learning and design principles. Even basic facts and figures can be rendered beautifully. Instead of handing out a hard copy of your presentation or supporting charts, graphs or worksheets, consider creating an infographic that distills the insight. For inspiration, visit David McCandless's Information is Beautiful website. To understand the possibilities for presenting complex data in a compelling manner, explore the resources on Edward Tufte's website or one of his classic books on data visualization. For perspective, *The New York Times* described Tufte as the "Leonardo da Vinci of data." Not to be outdone, *Bloomberg* labeled Tufte the "Galileo of graphics."

Tools

A short-list of possible tools include the following

- Presentation software
- Add Ins: Polling
- Handouts (i.e., infographic, quick reference)—Not your presentation!
- Giveaways
- Pens/pencils/markers
- Flip Charts
- Self-Adhesive Pads
- Dry Erase Boards
- Snacks!

Also consider logistics and technical details including the room layout, lighting, temperature controls, wifi and electrical outlets and bathroom facilities.

Contributors and Attributions

CC licensed content, Original

- Presentations and Other Visual Aids. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- VARK Illustration. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Ebbinghaus Forgetting Curve. **Authored by**: Educ320. **Located at**: commons.wikimedia.org/wiki/File:Ebbinghaus%E2%80%99s_Forgetting_Curve_(Figure_1).jpg. **License**: *CC BY-SA: Attribution-ShareAlike*
- Naughty or Nice? Holiday Shopping Study. **Authored by**: Digitas Photos. **Located at**: https://flic.kr/p/dzvk1A. **License**: *CC BY: Attribution*

This page titled 15.12: Presentation Tools and Visual Aids is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.12: Presentation Tools and Visual Aids by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.13: Introduction to Using Microsoft PowerPoint

What you'll learn to do: Create a presentation using Microsoft PowerPoint

PowerPoint is one of several applications in Microsoft's Office product line, including enterprise staples Outlook (email and calendar), Excel (spreadsheet) and Word (word processing).

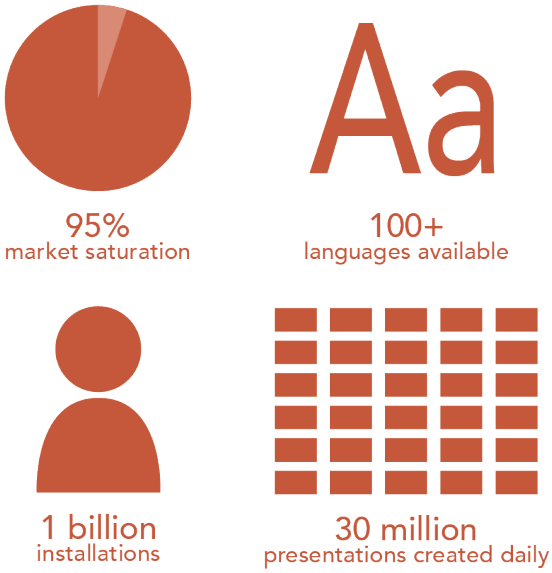
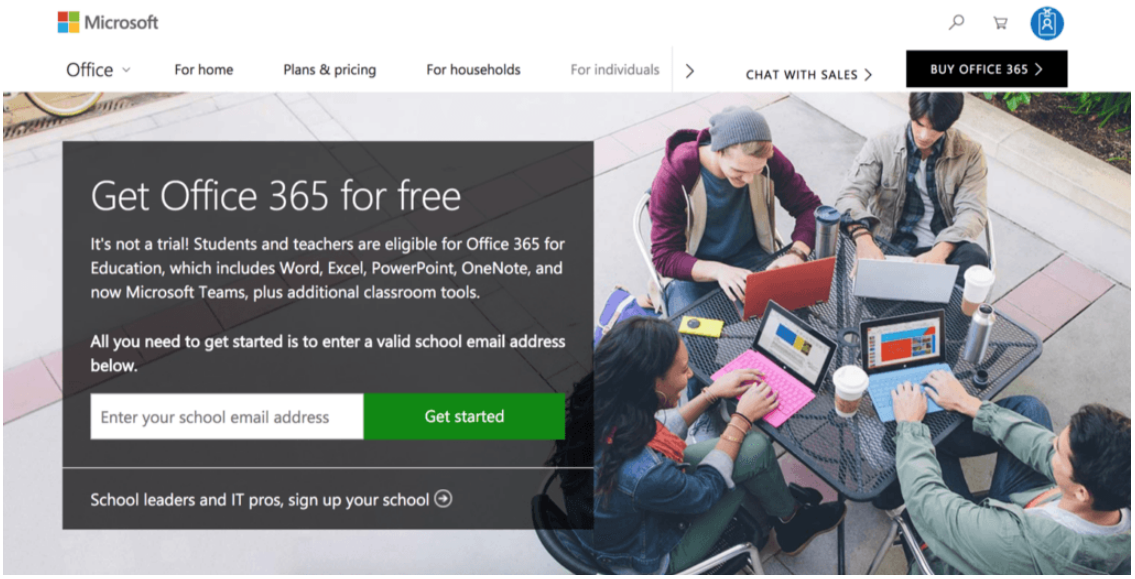


Figure 1. PowerPoint usage statistics

With a market share of approximately 95 percent, PowerPoint is considered the industry standard for both business and education. PowerPoint supports over 100 languages and can be used on both Android and Mac devices. There are over 1 billion installations of PowerPoint worldwide, with 30 million presentations created daily. Similar to Kleenex for tissue, the PowerPoint brand name has become a generic reference for all presentation software and the generally poor use of this technology has given rise to the phrase "death by PowerPoint," a scenario you will learn to avoid in subsequent sections. In this section, we will focus on the basics of creating a presentation, including how to create and save a presentation, how to change layouts and templates, how to add text, images and video and how to manage slides and objects within a slide.

As would be expected, the functional layout and logic of PowerPoint is similar to other Microsoft Office programs—in particular, Word and Excel. How you access PowerPoint varies somewhat depending on the version, both the edition (year) and whether you're using an installed package or a web-based app. The screen shots and instructions in this section are from the web-based version of PowerPoint (currently PowerPoint 2016), an app contained within Microsoft Office 365. Microsoft Office 365 is free for students and teachers. Benefits of using the web-based version include free access to the most current, fully-featured versions of Word, Excel, PowerPoint (and other apps) and automated file saving. The only requirement for a free education access is a valid school (that is, .edu) email address. Here's the link to the Get Started web page.



Contributors and Attributions

CC licensed content, Original

- Introduction to Using Microsoft PowerPoint. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*
- PowerPoint Usage Statistics. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 15.13: Introduction to Using Microsoft PowerPoint is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.13: Introduction to Using Microsoft PowerPoint by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.14: New Presentations

Learning Objectives

- Create and save a new presentation

At the Microsoft Office 365 home page, you will see a selection of apps as well as any recently-viewed documents. To open PowerPoint, click on the PowerPoint icon. The PowerPoint icon style varies depending on your operating system (PC or Mac) and the version of your software, but it's always a red-orange P, sometimes accompanied by a chart graphic, as you can see circled in Figure 1.

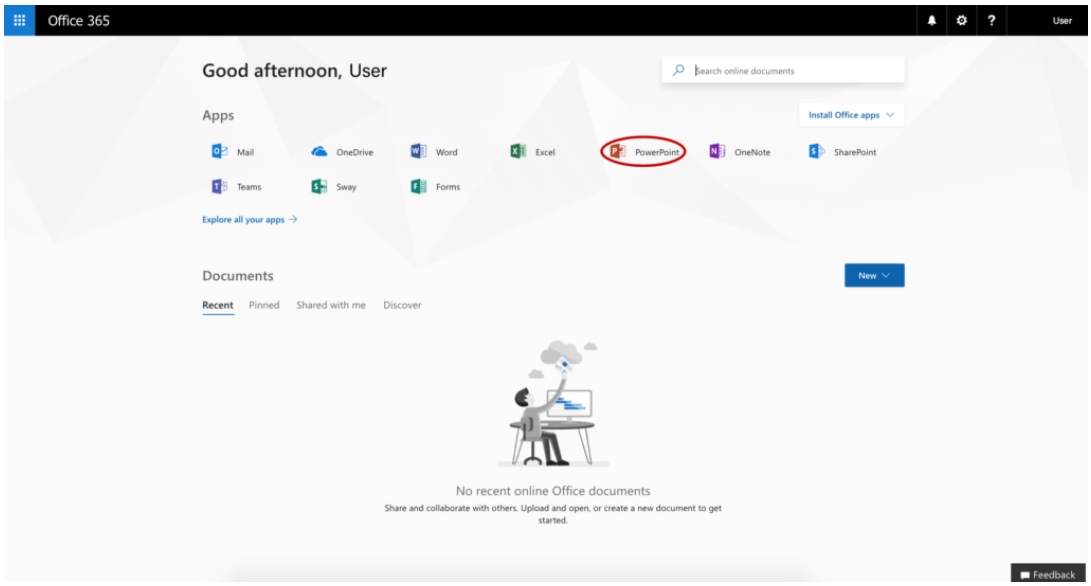


Figure 1. Accessing PowerPoint in Office 365.

When you open PowerPoint, the application will take you to a presentation gallery (see Figure 2), where you can choose to either open a new blank presentation or start a presentation using a template. To start with the proverbial blank page—no pre-set graphics, backgrounds or text—click (or double-click, depending on your version) **New blank presentation**. You'll be presented with a new blank presentation title slide containing two text boxes for a title and subtitle. These are placeholders that can be used, deleted, or simply ignored.

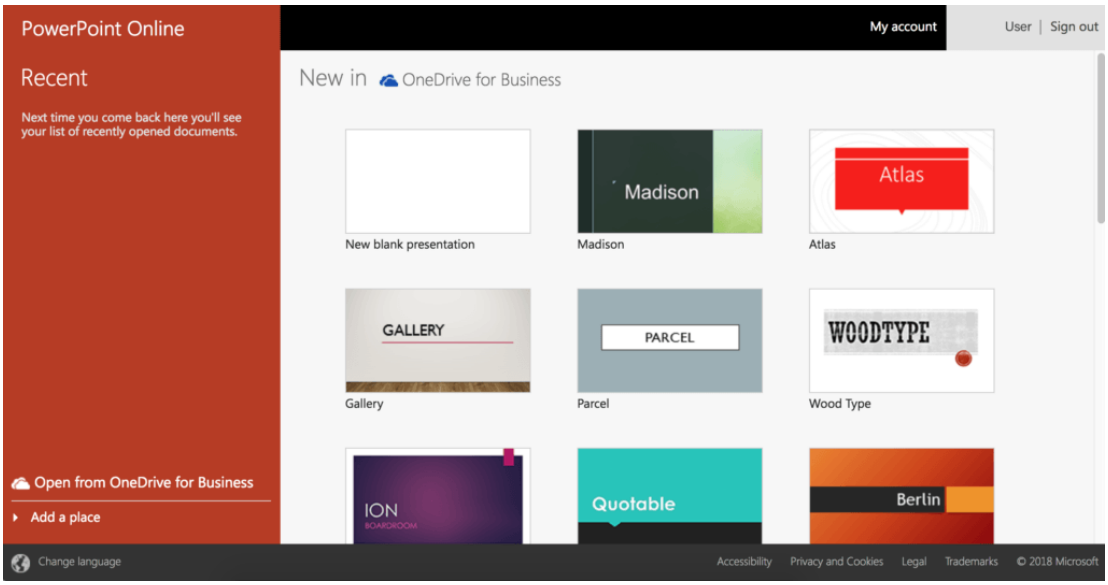


Figure 2. Template gallery.

If you want to use the default layout, click where it says **Click to add title** and type in your text. Note that when you click in the text box, the prompt text disappears and is replaced with a flashing cursor. To finish your entry, simply click outside the box. Pressing the enter or return key after typing your text will act as a return and advance the cursor to the next line. If you pressed return and don't want a two-line title, simply press the **Undo** arrow under the **File** tab at the top left of your screen (or if you're using Windows, press **Ctrl + Z** on your keyboard; if you're using a Mac, press **Cmnd + Z**). To continue adding or editing text, click the text or within another text box and continue.

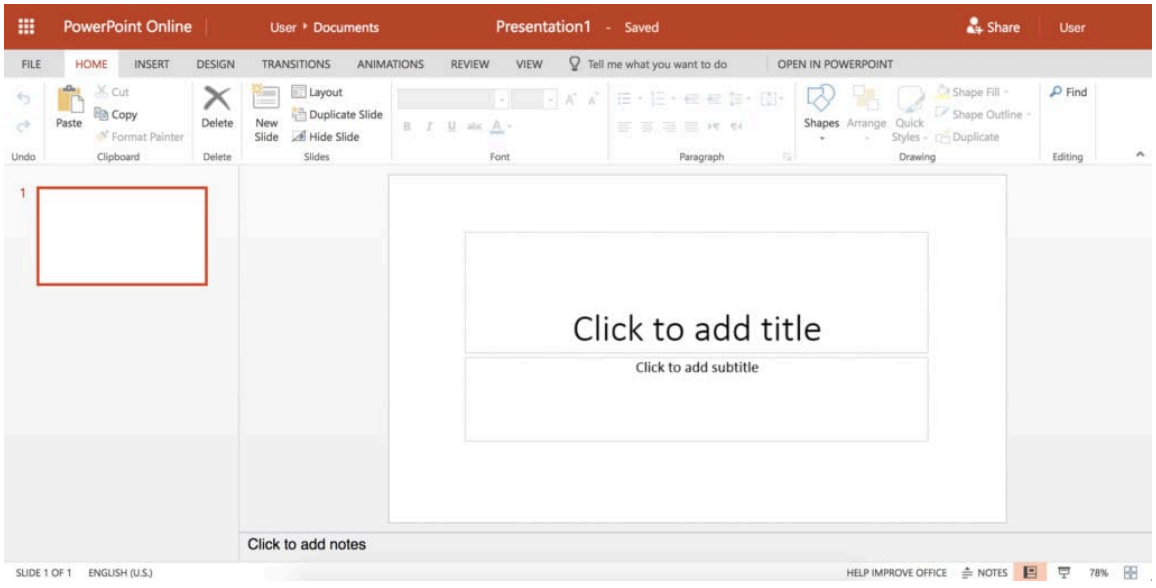


Figure 3. Your new presentation.

Placeholder boxes and prompts don't show up in Slide Show or presentation mode, so you don't necessarily need to delete those you don't use. If, however, you find them distracting, click on the border of the placeholder and press either the **delete** or **backspace** keys on your keyboard.

As mentioned above, the web-based version of PowerPoint will save your file automatically, assigning a default title of *Presentation*. If you have multiple presentations in progress, it will save subsequent files as *Presentation1*, *Presentation2*, etc. To (re)name your presentation, click the default title, type your desired title and press **enter**.

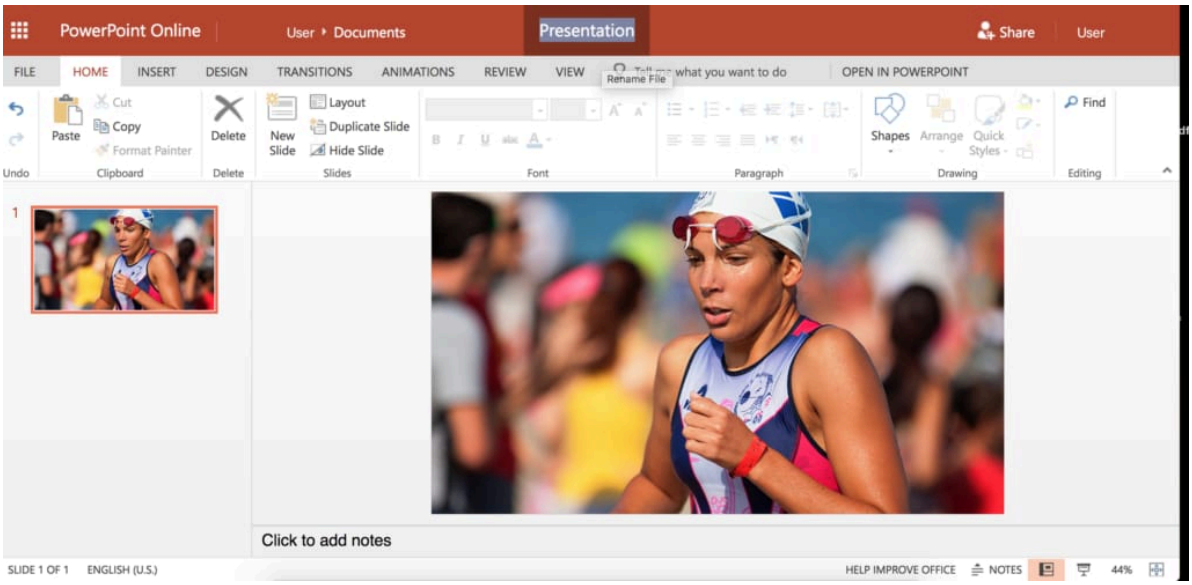


Figure 4.

Saving Documents When Working Offline

If you're not working with PowerPoint Online, you will need to save your presentation manually: click on the **File** tab.

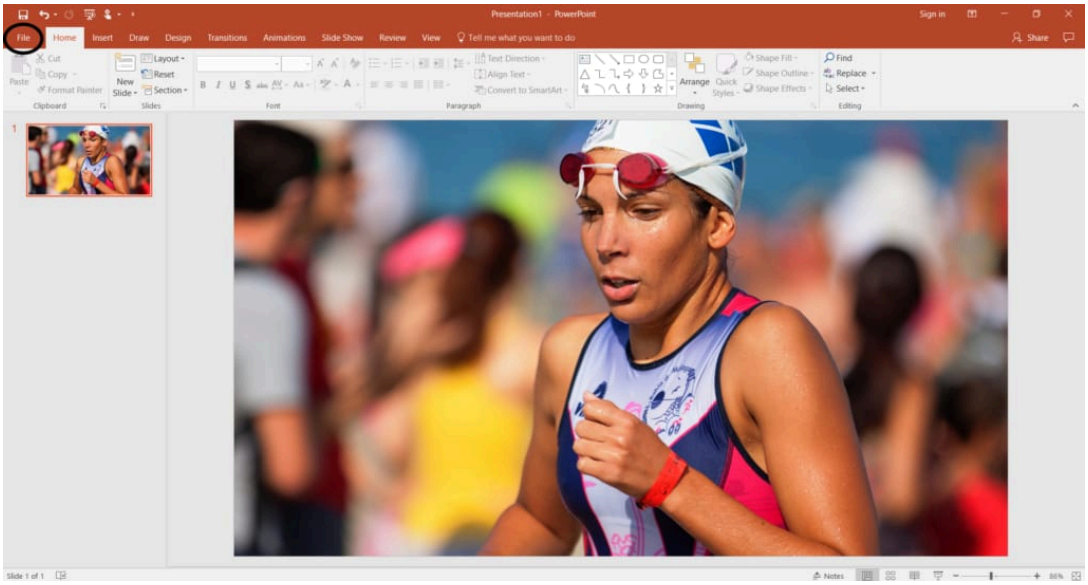


Figure 5.

You'll then click **Save** or **Save As** in the menu along the left side of the screen. Note: you can also use the common keyboard shortcuts **Ctrl + S** (Windows) or **Cmnd + S** (Mac).

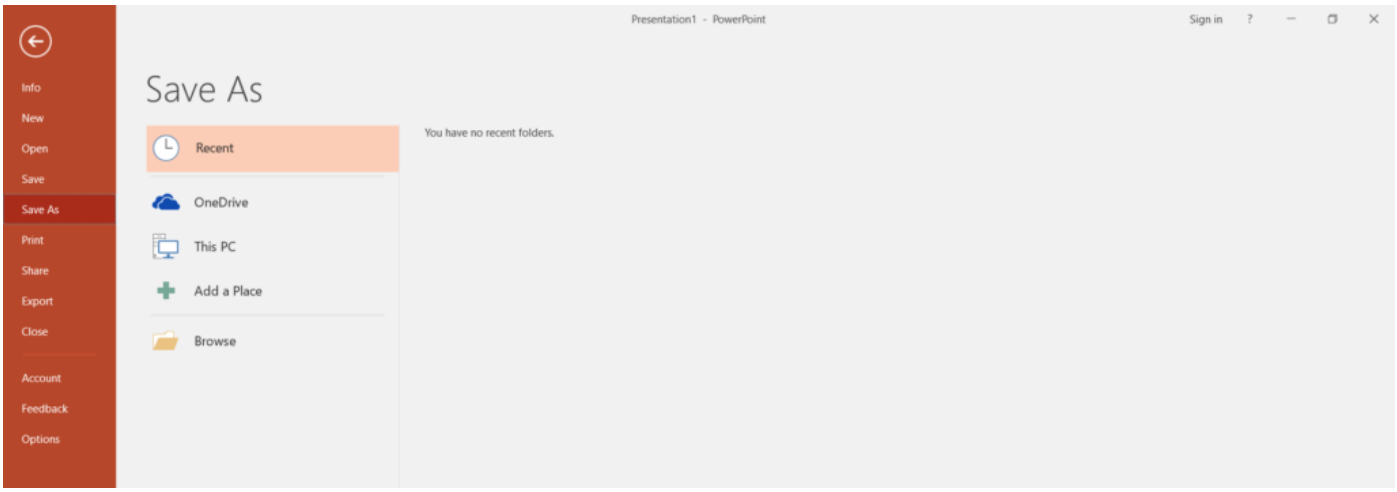


Figure 6.

To save your presentation on your desktop, choose **This PC**, select my Desktop and tile your presentation—for example, "To-the-Bank Retail Trends." At this point, you will also choose the format you want to save your presentation in **Save as type**. If you will be sharing your presentation with other students or colleagues who are using an older version of PowerPoint, select PowerPoint 97-2003 as the type from the drop-down list. Otherwise, go with the default option.

Contributors and Attributions

CC licensed content, Original

- New Presentations. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Triatleta photo. **Authored by**: Quino Al. **Provided by**: Unsplash. **Located at**: https://unsplash.com/photos/PTWyikFE3sc. **License**: *CC0: No Rights Reserved*

This page titled 15.14: New Presentations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.14: New Presentations by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.15: Editing and Enhancing a PowerPoint Presentation

Learning Objectives

- Add, delete, and move slides
- Insert text boxes
- Change theme on a presentation

Now that you've created a presentation, let's start adding to it! If you closed your presentation after saving, you can reopen it by starting PowerPoint and selecting your presentation (i.e., To-the-Bank Retail Trends) listed under Recents.

Adding Slides

Thumbnail pane.

From the Home Tab

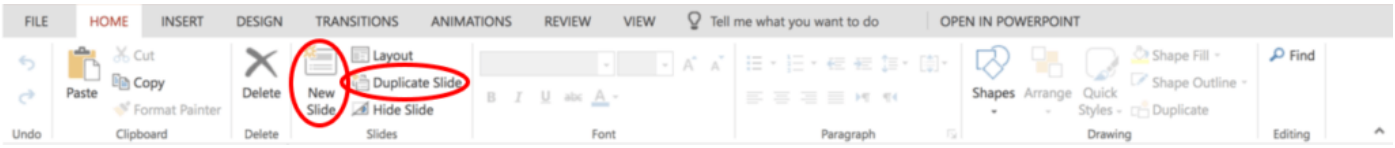
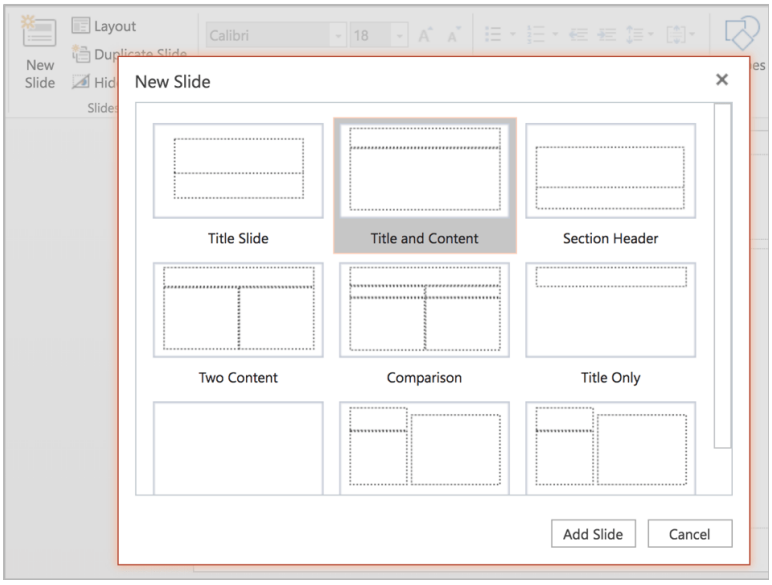
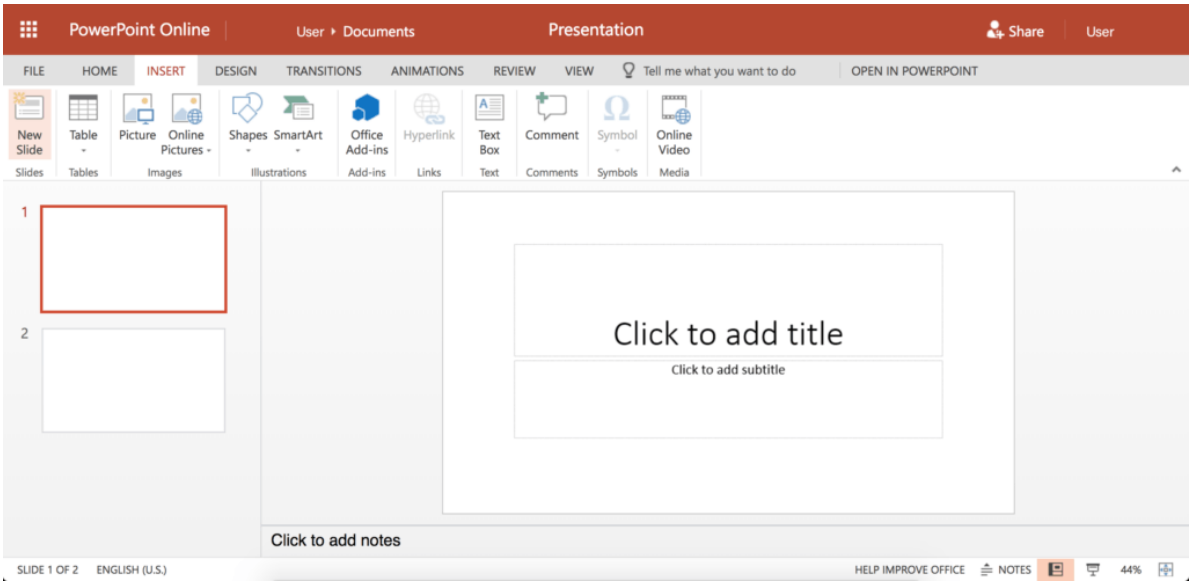


Figure 1. Home Tab.

1. Click on **New Slide** (or **Duplicate Slide**, to copy the current slide layout)
2. If you clicked **New Slide**, select a layout that best fits your content
3. Press **enter**

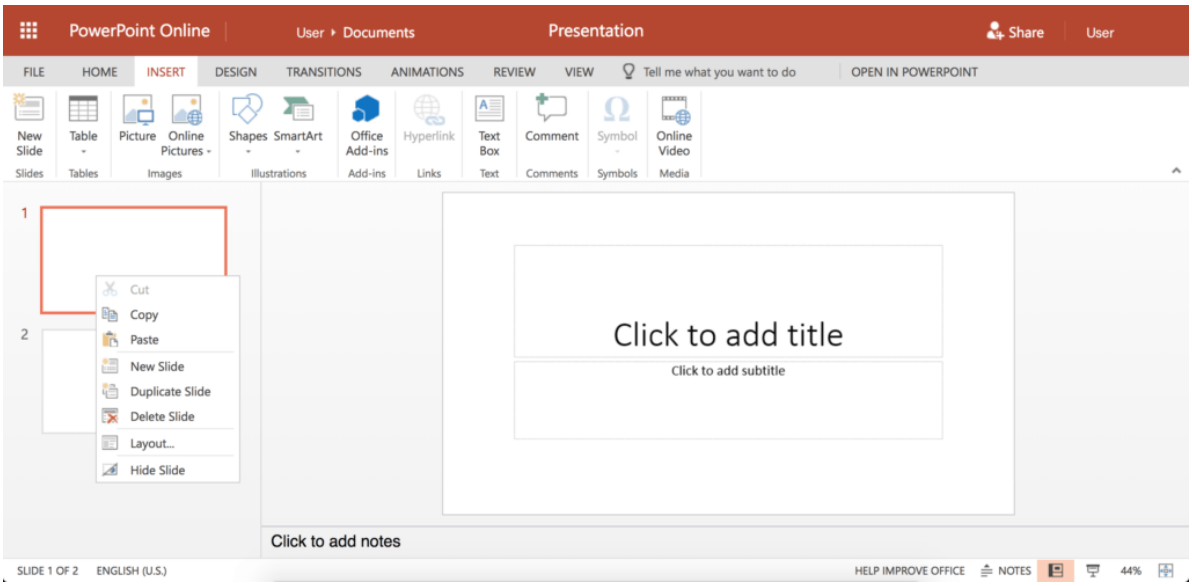


From the Insert Tab



1. Click on **New Slide**
2. Select a layout
3. Press **enter**

From the Thumbnail Pane



You can also add slides from the thumbnail pane, the lefthand column of your screen when you're in the default or Slide view.

1. Select (highlight) the slide you want the new slide to follow
2. Right-click on the slide
3. Select **New Slide** for the default content slide layout or **Duplicate Slide** to copy the slide layout and content

OR, to use a shortcut to duplicate the slide layout (not content)

1. Select (highlight) the slide you want the new slide to follow
2. Press **Ctrl + M** (PC) or **Cmnd + M** (Mac)

Note that your layout choice isn't final, so don't overthink the decision. If you're just trying to get your thoughts down on slides, select **Duplicate Slide** and fine-tune later.

Deleting a Slide

As with most operations, there are multiple ways to delete a slide:

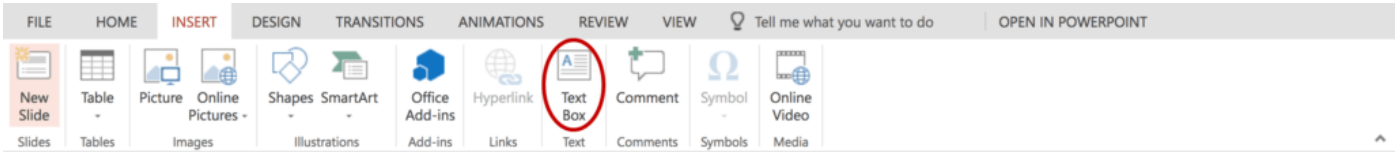
- Highlight the slide in the thumbnail pane and press **delete**
- Highlight the slide in the thumbnail pane and press the **delete** icon (Home tab)
- Highlight the slide in the thumbnail pane, right click and select **Delete Slide**

Organizing Slides

To move a slide, simply drag and drop in the thumbnail frame.

Inserting Text Boxes

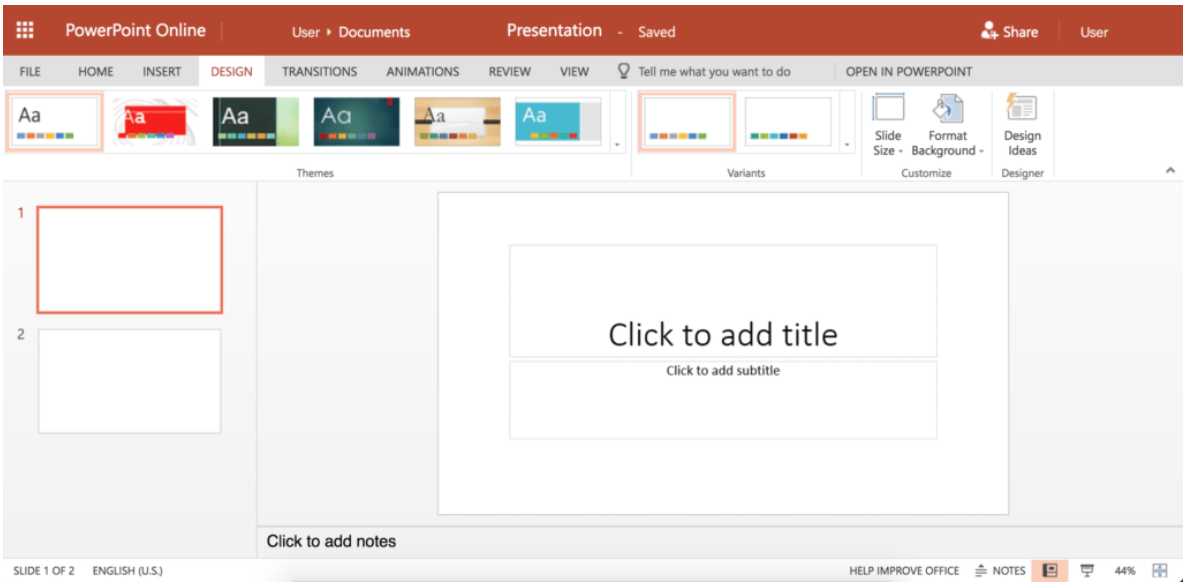
There are multiple ways to insert text into your slides. One of the ways to enter text, as discussed earlier, is to click inside a placeholder box and enter your text. However, you may want to add text to a blank slide or in a place where there is no placeholder. To do so, select the **Insert** tab on the ribbon, and choose Text Box.



A text box will be inserted in your slide with a Click to insert text message similar to the placeholder Title and Subtitle boxes discussed earlier. The text box will expand to fit your text. If you decide you want to reposition the text box, you can click and move the text box/text around the slide. If you want to modify the width of the text box, click on a corner and drag it out as desired. The circular icon at the top of the text box activates the drawing and formatting options and allows you to rotate the text box.

Changing the Theme on a Presentation

Themes are one of the formatting options within PowerPoint. A theme is a slide design that contains pre-set colors, fonts, formatting and/or images. To view formatting options, click on the design tab and scroll through the available themes.



Clicking on a theme's thumbnail applies the formatting to your slides, so you can evaluate potential fit. A second group of options in the ribbon provides color variations on that design. A third option, exclusive to the online version of PowerPoint, is Design Ideas, providing additional variations; this feature is still being built out. To revert to your original format, click Undo.

One caveat: given the limited number of templates and the number of people using PowerPoint, using a standard template doesn't do anything to distinguish you or your message. Instead of selecting a standard template, use thoughtful images and complementary fonts to add design impact.

Contributors and Attributions

CC licensed content, Original

- Editing and Enhancing a PowerPoint Presentation. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 15.15: Editing and Enhancing a PowerPoint Presentation is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.15: Editing and Enhancing a PowerPoint Presentation by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.16: Visuals- Images, Videos, and Objects

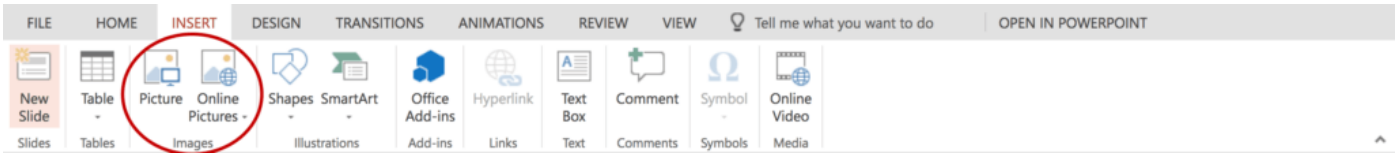
Learning Objectives

- Add images and clip art
- Add videos
- Arrange objects

Adding Images and Clip Art

Adding art, images and photographs is a way to both personalize your slides and reinforce your message.

Adding visual elements to your slides is done from the **Insert** tab. Images can also be inserted directly from a new slide using the Add Pictures, Online Pictures or Online Video icons embedded in new slide layouts.



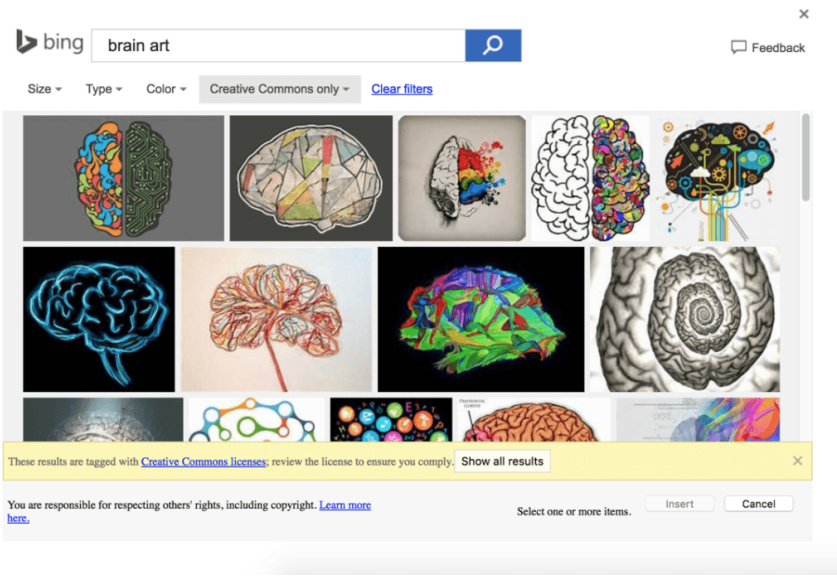
To add pictures from your computer

1. Select the **Insert tab**
2. Click on **Picture**
3. Select your desired image
4. Press **Open** to insert

The Add Pictures From the Internet option allows for two possibilities: searching via Bing and uploading images from OneDrive (Microsoft's cloud storage product). Of course, you can also do an image search using Google Images or your preferred search engine, download images to your desktop and upload them to your presentation using the steps above. Don't be put off by the initial Bing images served up. Bing's filters allow you to search by size (small, medium, large, extra large), type (all, photograph, clip art, line drawing and transparent), color (all, color, B&W and by 12 basic colors) and license (all or Creative Commons only).

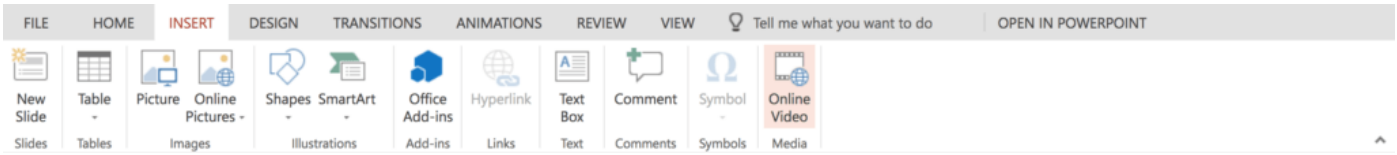
To add pictures from the Internet

1. Select the **Insert tab**
2. Click on **Online Pictures**
3. Type in your search term(s)
4. Refine your image search using Bing's size, type, color and license filters
5. Select an image
6. Press **Insert** to insert into slide
7. Use the corner image "handles" (small boxes) to size



Adding Videos

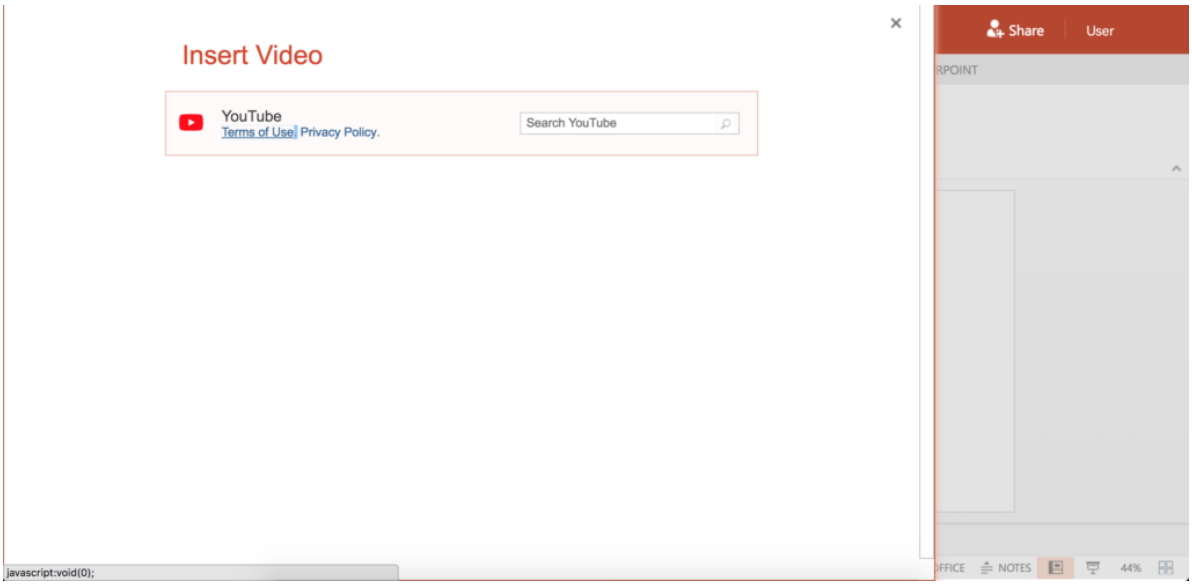
Adding video allows you to incorporate both motion and sound into your presentation and this can translate into audience energy! As with other visual elements, video is an **Insert tab** command.



The add video command launches a YouTube video search. To see video details, hover over the video. To view the video once embedded, click on the View tab, select Current Slide and press Play.

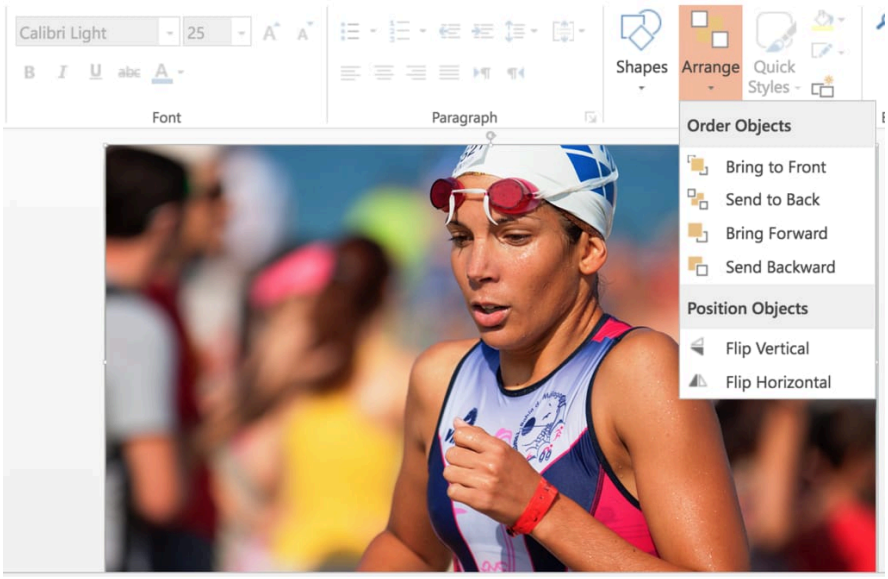
To add video

1. Select the **Insert tab**
2. Click on **Online Video**
3. Enter your search terms
4. Select your desired video
5. Press **Open** to insert



Arrange Objects

There may be times when default line spacing doesn't work or you're trying to build a slide that involves multiple blocks of text and/or imagery. That's where PowerPoint's Arrange function comes into play. Arrange allows you to order (layer) objects, bringing individual elements forward or sending them backwards. It also allows you to flip objects. Refer to Figure 4 for a screen shot.



Contributors and Attributions

CC licensed content, Original

- Visuals: Images, Videos, and Objects. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

CC licensed content, Shared previously

- Triatleta photo. **Authored by**: Quino Al. **Provided by**: Unsplash. **Located at**: https://unsplash.com/photos/PTWyikFE3sc. **License**: *CC0: No Rights Reserved*

This page titled 15.16: Visuals- Images, Videos, and Objects is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.16: Visuals- Images, Videos, and Objects by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.17: Introduction to Using Google Slides

What you'll learn to do: Create a presentation using Google Slides

Google Slides is a free program that's part of Google's suite of web-based applications, including Google Docs (word processing), Sheets (spreadsheets), Slides (presentations) and Forms (collecting & organizing information). What's particularly powerful about Google Slides is that you can create, edit, collaborate and present seamlessly across operating systems and without potential file compatibility or corruption issues, managing flash drives or figuring out connector cables. Using a web-based program also eliminates the risk of transferring a virus.

It is also fairly easy to move a presentation from PowerPoint to Google Slides and back again, though you may lose some formatting in the process. If you do move a presentation, page through to make sure the slides still look the way you want them to. However, all the content should convert just fine. If you are used to PowerPoint and find yourself needing to use Google Slides (or vice versa), fear not! The commands and procedures are pretty similar between the two of them, and if you get stuck, there are lots of good online resources to answer your questions.

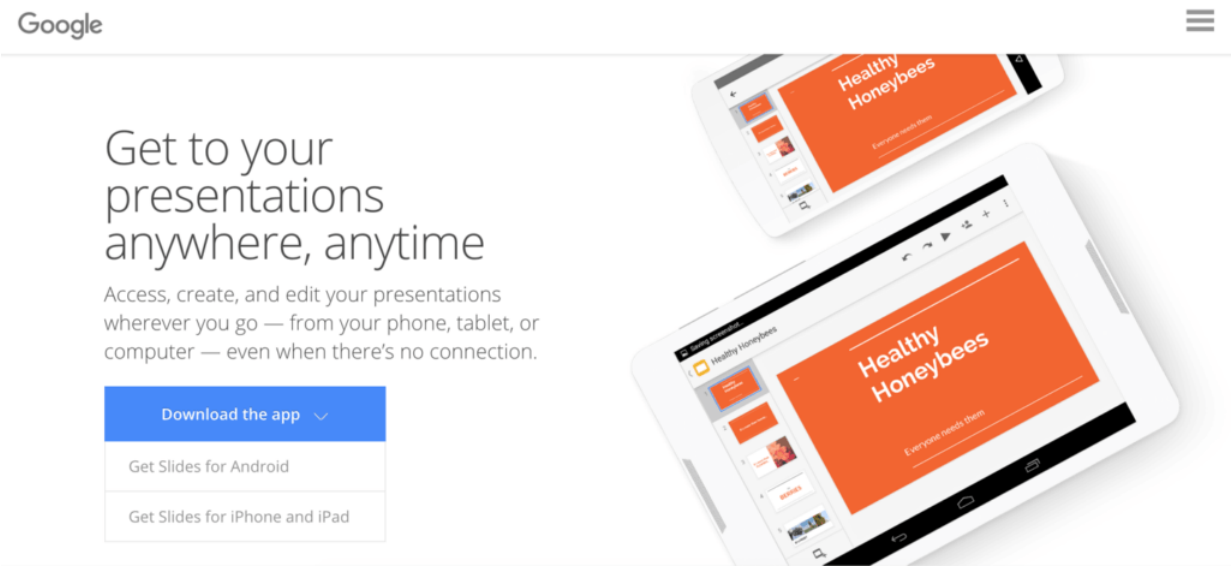


Figure 1.

For an overview of Google Slides capabilities, scroll down through the front pages on the Google Slides website.

To highlight a few key features and benefits:

- Universal access, from your phone, tablet or computer
- Support for both Android and iPhone/iPad (Google Slides apps)
- Share your presentation with options for to restrict/enable viewing, commenting and editing
- Auto-save

Embedded in Google Slides are a variety of presentation themes and templates, hundreds of fonts and support for text, images and video. You can search Google Slides template gallery for perspective or a jump start to creating your presentation. Let's jump right in!

Contributors and Attributions

CC licensed content, Original

- Introduction to Using Google Slides. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 15.17: Introduction to Using Google Slides is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.17: Introduction to Using Google Slides by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

15.18: Creating and Organizing Presentations

Learning OUtcomes

- Create and save a new presentation
- Add, delete, and move slides

Googles Slides is free to use, but to do so requires a Google account. If you don't already have a Google account (i.e., a gmail address, Google Drive or Google Calendar), you will need to create one in order to use Google Slides. To establish a Google account, search for "Google account" or go directly to the Google Account sign up page.

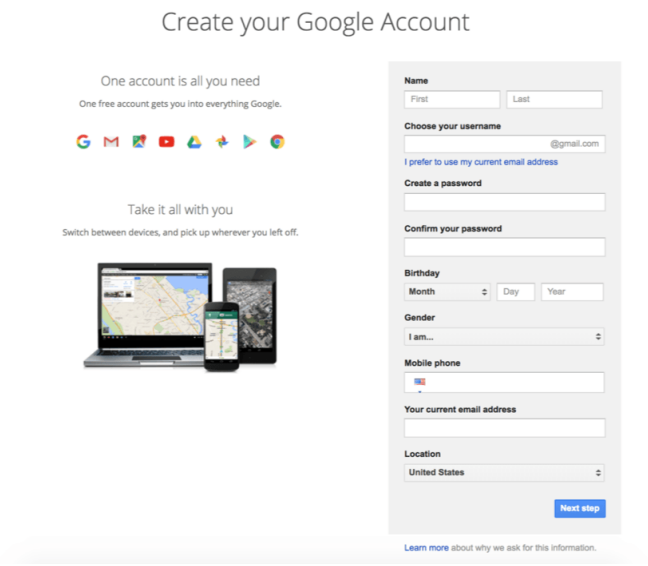


Figure 1.

New Presentations

To start a new presentation, search for "Google slides" or go to https://www.google.com/slides/about/.

See Figure 2 for a visual of the application's opening view. You can start a new blank presentation or choose from a number of existing, editable presentations or templates. If you recently viewed a shared presentation or created a presentation, you would see it listed under Recent Presentations.

To start a new presentation, click on the plus (+) sign in the Blank Presentation box.

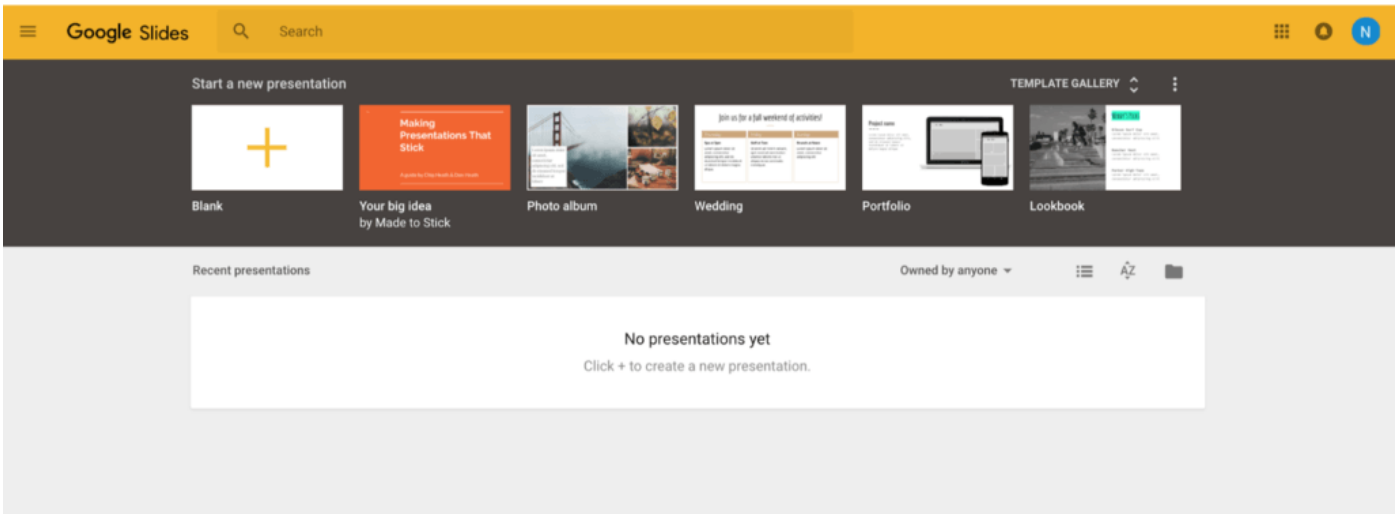


Figure 2.

Within Google Slides, as in all Google applications, your work is saved automatically. Until you name your presentation, it will save as Untitled presentation.

To title your presentation, click in the [Untitled Presentation] box and type your desired title. By default, Google slides will insert the text from the title box as the title; you can accept that or delete and enter an alternate title (see Figure 3).

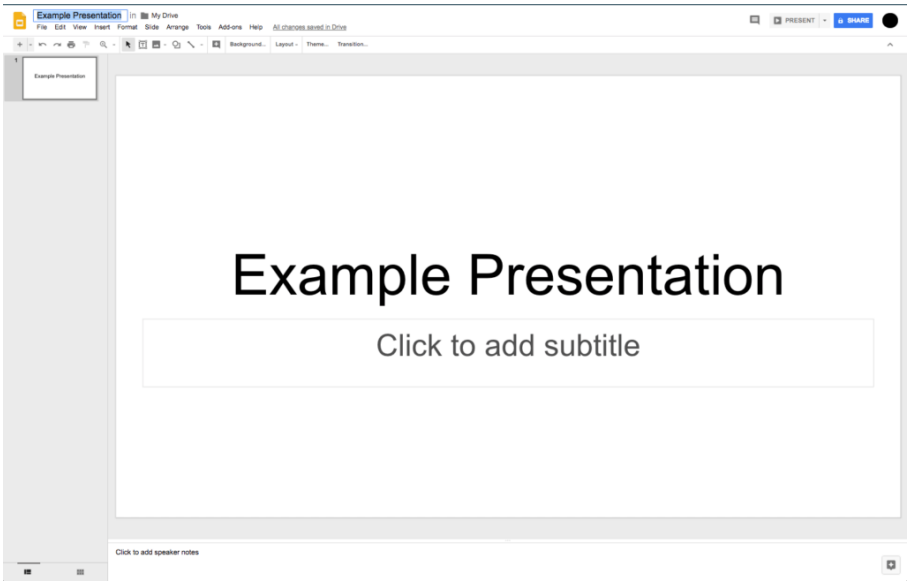


Figure 3.

Organizing Slides

There are multiple options for adding slides:

- Select “+” (New slide) from the second menu line. To select a specific slide layout, use the dropdown menu to the right of the slide to add a slide that best matches the content for that slide
- Use the New slide shortcut: Ctrl+M
- From the Insert dropdown, select New slide (Ctrl+M) Select Insert > new slide
- Select Slide > Duplicate slide, which will copy the slide that’s currently highlighted in the thumbnail section in the left column of the screen
- Right click in a slide in the thumbnail sidebar and select New slide or Duplicate slide

To delete a slide, right click on the thumbnail image and press delete OR delete from the Slide dropdown: Slide > Delete slide. If you delete a slide in error, don't panic! Click on the Undo arrow or use the Ctrl+Z shortcut.

To move a slide, click on the slide in the sidebar—this will highlight the thumbnail—and drag it to where you want it to appear. A second sorting option is to shift from the default ("Filmstrip") view to Grid view—View > Grid view—and use a select, drag and drop method. You can toggle between the default and Grid view by selecting the icons at the bottom of the thumbnail window.

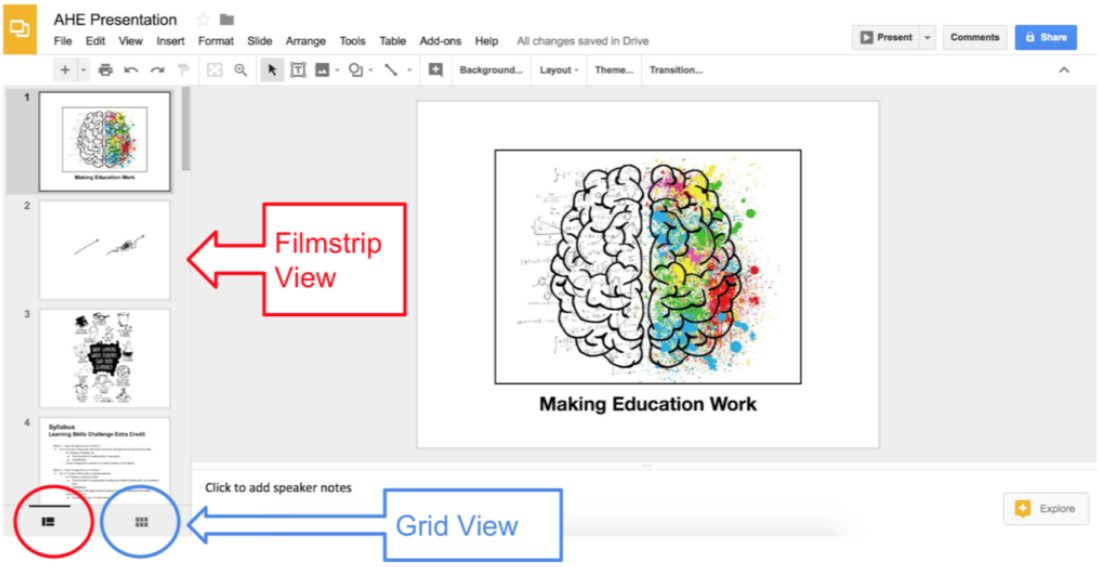


Figure 4.

Contributors and Attributions

CC licensed content, Original

- Creating and Organizing Presentations. **Authored by**: Nina Burokas. **Provided by**: Lumen Learning. **License**: *CC BY: Attribution*

This page titled 15.18: Creating and Organizing Presentations is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Nina Burokas via source content that was edited to the style and standards of the LibreTexts platform.

- 8.18: Creating and Organizing Presentations by Nina Burokas is licensed CC BY 4.0. Original source: https://courses.lumenlearning.com/wm-businesscommunicationmgrs.

INDEX

A

artifact
9.12: Confounds, Artifacts and other Threats to Validity

B

binomial distribution
13.5: The Binomial Distribution

C

central limit theorem
13.12: The Central Limit Theorem
central tendency
9.2: Measures of Center
confounding variable
9.12: Confounds, Artifacts and other Threats to Validity
consumer analysis
9.1: Introduction

D

data
9.3: Measures of Variation
data analysis
1: Welcome to Data Analysis
9.1: Introduction
data analytics
1: Welcome to Data Analysis
data science
9.1: Introduction
9.2: Measures of Center
dataset
9.2: Measures of Center
9.3: Measures of Variation
decision making
9.1: Introduction
dependent variable
9.9: The role of variables — predictors and outcomes
descriptive statistics
11.1: This is what too many numbers looks like

E

experimental research
9.10: Experimental and non-experimental research

F

financial decisions
9.1: Introduction

I

independent variable
9.9: The role of variables — predictors and outcomes

L

Law of Large Numbers
13.10: The Law of Large Numbers
learning
9.3: Measures of Variation

M

Machine Learning
9.1: Introduction
maturational effects
9.12: Confounds, Artifacts and other Threats to Validity
mean
11.4: Measures of Central Tendency (Sameness)
9.2: Measures of Center
measures of central tendency
9.3: Measures of Variation
measures of variation
9.3: Measures of Variation
median
9.2: Measures of Center
mode
11.4: Measures of Central Tendency (Sameness)
9.2: Measures of Center

N

nominal scale
9.7: Scales of measurement

O

objectives
9.3: Measures of Variation
operationalization
9.6: Introduction to Psychological Measurement
ordinal scale
9.7: Scales of measurement
outliers
9.2: Measures of Center
9.3: Measures of Variation

P

probability
9.1: Introduction
python
9.2: Measures of Center

R

randomization
9.10: Experimental and non-experimental research
regression
12.5: Regression — A mini intro

S

standard deviation
9.3: Measures of Variation
statistical measurements
9.2: Measures of Center
statistical methods
9.1: Introduction
statistics
9.1: Introduction
9.3: Measures of Variation
stock market analysis
9.1: Introduction

T

trimmed mean
9.2: Measures of Center

V

variance
9.3: Measures of Variation

This page titled Index is shared under a Public Domain license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

GLOSSARY

S

Sample Word 1 | Sample Definition 1

This page titled Glossary is shared under a CC BY 4.0 license and was authored, remixed, and/or curated by Rob Clifford, Alex Mata, and Arjen Booy.

DETAILED LICENSING

Overview

Title: Data Analysis Certificate Program

Webpages: 166

All licenses found:

- CC BY 4.0: 70.5% (117 pages)
- CC BY-SA 4.0: 27.1% (45 pages)
- Public Domain: 1.2% (2 pages)
- Undeclared: 1.2% (2 pages)

By Page

- Data Analysis Certificate Program - *CC BY-SA 4.0*
  - Front Matter - *CC BY 4.0*
    - TitlePage - *CC BY 4.0*
    - InfoPage - *Public Domain*
    - Table of Contents - *Undeclared*
    - Licensing - *CC BY 4.0*
  - 1: Welcome to Data Analysis - *CC BY 4.0*
  - 2: Introduction to Working in Microsoft Excel - *CC BY 4.0*
    - 2.1: Using Excel Cells - *CC BY 4.0*
    - 2.2: Using Excel Data - *CC BY 4.0*
    - 2.3: Using Excel Functions - *CC BY 4.0*
    - 2.4: Using Excel Charts - *CC BY 4.0*
    - 2.5: Using Conditional Formatting in Excel - *CC BY 4.0*
  - 3: Working in Google Sheets - *CC BY 4.0*
    - 3.1: Creating a New Spreadsheet - *CC BY 4.0*
    - 3.2: Using Google Sheets - *CC BY 4.0*
    - 3.3: Add-Ons - *CC BY 4.0*
  - 4: SQL Structured Query Language - *CC BY 4.0*
    - 4.1: Create Database - *CC BY 4.0*
    - 4.2: Constraints - *CC BY 4.0*
    - 4.3: User Defined Types and Altering Tables - *CC BY 4.0*
    - 4.4: Key Terms and Exercises - *CC BY 4.0*
  - 5: SQL Data Manipulation Language - *CC BY 4.0*
    - 5.1: Using SELECT Statement - *CC BY 4.0*
    - 5.2: INSERT Statement - *CC BY 4.0*
    - 5.3: UPDATE and DELETE Statements - *CC BY 4.0*
    - 5.4: Built-In Functions - *CC BY 4.0*
    - 5.5: Joining Tables - *CC BY 4.0*
    - 5.6: Key Terms and Exercises - *CC BY 4.0*
  - 6: The Relational Data Model - *CC BY 4.0*
    - 6.1: Fundamental Concepts in the Relational Data Model - *CC BY 4.0*
    - 6.2: Key Terms - *CC BY 4.0*
  - 7: The Entity Relationship Data Model - *CC BY 4.0*
    - 7.1: Entity, Entity Set and Entity Type - *CC BY 4.0*
    - 7.2: Types of Attributes - *CC BY 4.0*
    - 7.3: Keys - *CC BY 4.0*
    - 7.4: Nulls - *CC BY 4.0*
    - 7.5: Relationships - *CC BY 4.0*
    - 7.6: Key Terms and Exercises - *CC BY 4.0*
  - 8: Visual Media - *CC BY 4.0*
    - 8.1: Tables, Charts, and Infographics - *CC BY 4.0*
    - 8.2: Matching Graphics and Objectives - *CC BY 4.0*
    - 8.3: Placement, Style, and Coloring - *CC BY 4.0*
    - 8.4: Captions and Titles - *CC BY 4.0*
    - 8.5: Introduction to Contemporary Visual Aids - *CC BY 4.0*
    - 8.6: Video as a Visual Aid - *CC BY 4.0*
    - 8.7: Designing a Video for Your Needs - *CC BY 4.0*
    - 8.8: Products and Resources for Creating Videos - *CC BY 4.0*
    - 8.9: Introduction to Accessible Visual Aids - *CC BY 4.0*
    - 8.10: Accessibility Matters - *CC BY 4.0*
    - 8.11: Why It Matters- Visual Media - *CC BY 4.0*
    - 8.12: Introduction to Using Visuals - *CC BY 4.0*
    - 8.13: Visuals in a Report - *CC BY 4.0*
    - 8.14: Increasing Impact with Media - *CC BY 4.0*
    - 8.15: Revising and Enhancing Visual Media for Impact - *CC BY 4.0*
    - 8.16: Evaluating the Effectiveness of your Message - *CC BY 4.0*
    - 8.17: Putting It Together- Visual Media - *CC BY 4.0*
    - 8.18: Discussion- Visual Media - *CC BY 4.0*
    - 8.19: Assignment- Visual Media - *CC BY 4.0*
    - 8.20: Introduction to Media and Your Message - *CC BY 4.0*
    - 8.21: Using Visual Media Resources - *CC BY 4.0*

- 8.22: Visual Design Principles - *CC BY 4.0*
- 8.23: Introduction to Images - *CC BY 4.0*
- 8.24: Images Overview - *CC BY 4.0*
- 8.25: Using Images for Impact - *CC BY 4.0*
- 8.26: Using Images Legally - *CC BY 4.0*
- 8.27: Introduction to Charts, Diagrams, and Graphic Organizers - *CC BY 4.0*

◦ 9: Research and Statistics - *CC BY-SA 4.0*
- 9.1: Introduction - *CC BY 4.0*
- 9.2: Measures of Center - *CC BY 4.0*
- 9.3: Measures of Variation - *CC BY 4.0*
- 9.4: There's More to Research Methods than Statistics - *CC BY-SA 4.0*
- 9.5: A brief Introduction to Research Design - *CC BY-SA 4.0*
- 9.6: Introduction to Psychological Measurement - *CC BY-SA 4.0*
- 9.7: Scales of measurement - *CC BY-SA 4.0*
- 9.8: Assessing the Reliability of a Measurement - *CC BY-SA 4.0*
- 9.9: The role of variables — predictors and outcomes - *CC BY-SA 4.0*
- 9.10: Experimental and non-experimental research - *CC BY-SA 4.0*
- 9.11: Assessing the validity of a study - *CC BY-SA 4.0*
- 9.12: Confounds, Artifacts and other Threats to Validity - *CC BY-SA 4.0*
- 9.13: Videos - *CC BY 4.0*
- 9.14: References - *CC BY 4.0*

◦ 10: Data Validation - *CC BY-SA 4.0*
- 10.1: Importance of Statistics and Experimental Measurements - *CC BY 4.0*

◦ 11: Describing Data - *CC BY-SA 4.0*
- 11.1: This is what too many numbers looks like - *CC BY-SA 4.0*
- 11.2: Look at the data - *CC BY-SA 4.0*
- 11.3: Important Ideas - Distribution, Central Tendency, and Variance - *CC BY-SA 4.0*
- 11.4: Measures of Central Tendency (Sameness) - *CC BY-SA 4.0*
- 11.5: Measures of Variation (Differentness) - *CC BY-SA 4.0*
- 11.6: Using Descriptive Statistics with data - *CC BY-SA 4.0*
- 11.7: Rolling your own descriptive statistics - *CC BY-SA 4.0*
- 11.8: Remember to look at your data - *CC BY-SA 4.0*
- 11.9: Videos - *CC BY 4.0*
- 11.10: References - *CC BY 4.0*

◦ 12: Correlation - *CC BY-SA 4.0*
- 12.1: If something caused something else to change, what would that look like? - *CC BY-SA 4.0*
- 12.2: Pearson's r - *CC BY-SA 4.0*
- 12.3: Turning the numbers into a measure of co-variance - *CC BY-SA 4.0*
- 12.4: Examples with Data - *CC BY-SA 4.0*
- 12.5: Regression — A mini intro - *CC BY-SA 4.0*
- 12.6: Interpreting Correlations - *CC BY-SA 4.0*
- 12.7: Summary - *CC BY-SA 4.0*
- 12.8: References - *CC BY 4.0*

◦ 13: Probability, Sampling, and Estimation - *CC BY-SA 4.0*
- 13.1: Prelude to Probability, Sampling, and Estimation - *CC BY 4.0*
- 13.2: How are Probability and Statistics Different? - *CC BY-SA 4.0*
- 13.3: What Does Probability Mean? - *CC BY-SA 4.0*
- 13.4: Basic Probability Theory - *CC BY-SA 4.0*
- 13.5: The Binomial Distribution - *CC BY-SA 4.0*
- 13.6: The normal distribution - *CC BY-SA 4.0*
- 13.7: Other useful distributions - *CC BY-SA 4.0*
- 13.8: Summary of Probability - *CC BY-SA 4.0*
- 13.9: Samples, populations and sampling - *CC BY-SA 4.0*
- 13.10: The Law of Large Numbers - *CC BY-SA 4.0*
- 13.11: Sampling distributions and the central limit theorem - *CC BY-SA 4.0*
- 13.12: The Central Limit Theorem - *CC BY-SA 4.0*
- 13.13: z-scores - *CC BY-SA 4.0*
- 13.14: Estimating population parameters - *CC BY-SA 4.0*
- 13.15: Estimating a confidence interval - *CC BY-SA 4.0*
- 13.16: Summary - *CC BY-SA 4.0*
- 13.17: Videos - *CC BY 4.0*
- 13.18: References - *CC BY 4.0*

◦ 14: Public Speaking - *CC BY 4.0*
- 14.1: How You Speak - *CC BY 4.0*
- 14.2: Introduction to Audience Engagement - *CC BY 4.0*
- 14.3: Audience Attention and Rapport - *CC BY 4.0*
- 14.4: Audience Participation - *CC BY 4.0*
- 14.5: Introduction to Speech Tips and Techniques - *CC BY 4.0*
- 14.6: Overcoming Fears and Anxiety - *CC BY 4.0*
- 14.7: Pro Tips - *CC BY 4.0*
- 14.8: Putting It Together- Public Speaking - *CC BY 4.0*
- 14.9: Discussion- Public Speaking - *CC BY 4.0*
- 14.10: Why It Matters- Public Speaking - *CC BY 4.0*
- 14.11: Introduction to Effective Public Speaking - *CC BY 4.0*
- 14.12: What is Public Speaking? - *CC BY 4.0*
- 14.13: Benefits of Public Speaking - *CC BY 4.0*
- 14.14: Audience Expectations - *CC BY 4.0*
- 14.15: Developing an Effective Speech - *CC BY 4.0*
- 14.16: Introduction to Delivery Techniques - *CC*

BY 4.0
- 14.17: Starting Your Speech - *CC BY 4.0*
- 14.18: Body Language and Gestures - *CC BY 4.0*

◦ 15: Developing and Delivering Business Presentations - *CC BY 4.0*
- 15.1: Text Boxes and Themes - *CC BY 4.0*
- 15.2: Images, Videos, and Arranging Objects - *CC BY 4.0*
- 15.3: Introduction to Making a Presentation for a Meeting - *CC BY 4.0*
- 15.4: Parts of a Good Presentation - *CC BY 4.0*
- 15.5: What's my Presentation About - *CC BY 4.0*
- 15.6: Bad Presentations - *CC BY 4.0*
- 15.7: Making a Presentation for a Meeting - *CC BY 4.0*
- 15.8: Putting It Together- Developing and Delivering Business Presentations - *CC BY 4.0*
- 15.9: Assignment- Developing and Delivering Business Presentations - *CC BY 4.0*
- 15.10: Why It Matters- Developing and Delivering Business Presentations - *CC BY 4.0*
- 15.11: Introduction to Visual Aids - *CC BY 4.0*
- 15.12: Presentation Tools and Visual Aids - *CC BY 4.0*
- 15.13: Introduction to Using Microsoft PowerPoint - *CC BY 4.0*
- 15.14: New Presentations - *CC BY 4.0*
- 15.15: Editing and Enhancing a PowerPoint Presentation - *CC BY 4.0*
- 15.16: Visuals- Images, Videos, and Objects - *CC BY 4.0*
- 15.17: Introduction to Using Google Slides - *CC BY 4.0*
- 15.18: Creating and Organizing Presentations - *CC BY 4.0*

◦ Back Matter - *CC BY 4.0*
- Index - *Public Domain*
- Glossary - *CC BY 4.0*
- Detailed Licensing - *CC BY 4.0*
- Detailed Licensing - *Undeclared*

This page titled Data Analysis Certificate Program is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Alex Mata.

DETAILED LICENSING

Overview

Title: Data Analysis Certificate Program

Webpages: 166

All licenses found:

- CC BY 4.0: 70.5% (117 pages)
- CC BY-SA 4.0: 27.1% (45 pages)
- Public Domain: 1.2% (2 pages)
- Undeclared: 1.2% (2 pages)

By Page

- Data Analysis Certificate Program - *CC BY-SA 4.0*
  - Front Matter - *CC BY 4.0*
    - TitlePage - *CC BY 4.0*
    - InfoPage - *Public Domain*
    - Table of Contents - *Undeclared*
    - Licensing - *CC BY 4.0*
  - 1: Welcome to Data Analysis - *CC BY 4.0*
  - 2: Introduction to Working in Microsoft Excel - *CC BY 4.0*
    - 2.1: Using Excel Cells - *CC BY 4.0*
    - 2.2: Using Excel Data - *CC BY 4.0*
    - 2.3: Using Excel Functions - *CC BY 4.0*
    - 2.4: Using Excel Charts - *CC BY 4.0*
    - 2.5: Using Conditional Formatting in Excel - *CC BY 4.0*
  - 3: Working in Google Sheets - *CC BY 4.0*
    - 3.1: Creating a New Spreadsheet - *CC BY 4.0*
    - 3.2: Using Google Sheets - *CC BY 4.0*
    - 3.3: Add-Ons - *CC BY 4.0*
  - 4: SQL Structured Query Language - *CC BY 4.0*
    - 4.1: Create Database - *CC BY 4.0*
    - 4.2: Constraints - *CC BY 4.0*
    - 4.3: User Defined Types and Altering Tables - *CC BY 4.0*
    - 4.4: Key Terms and Exercises - *CC BY 4.0*
  - 5: SQL Data Manipulation Language - *CC BY 4.0*
    - 5.1: Using SELECT Statement - *CC BY 4.0*
    - 5.2: INSERT Statement - *CC BY 4.0*
    - 5.3: UPDATE and DELETE Statements - *CC BY 4.0*
    - 5.4: Built-In Functions - *CC BY 4.0*
    - 5.5: Joining Tables - *CC BY 4.0*
    - 5.6: Key Terms and Exercises - *CC BY 4.0*
  - 6: The Relational Data Model - *CC BY 4.0*
    - 6.1: Fundamental Concepts in the Relational Data Model - *CC BY 4.0*
    - 6.2: Key Terms - *CC BY 4.0*
  - 7: The Entity Relationship Data Model - *CC BY 4.0*
    - 7.1: Entity, Entity Set and Entity Type - *CC BY 4.0*
    - 7.2: Types of Attributes - *CC BY 4.0*
    - 7.3: Keys - *CC BY 4.0*
    - 7.4: Nulls - *CC BY 4.0*
    - 7.5: Relationships - *CC BY 4.0*
    - 7.6: Key Terms and Exercises - *CC BY 4.0*
  - 8: Visual Media - *CC BY 4.0*
    - 8.1: Tables, Charts, and Infographics - *CC BY 4.0*
    - 8.2: Matching Graphics and Objectives - *CC BY 4.0*
    - 8.3: Placement, Style, and Coloring - *CC BY 4.0*
    - 8.4: Captions and Titles - *CC BY 4.0*
    - 8.5: Introduction to Contemporary Visual Aids - *CC BY 4.0*
    - 8.6: Video as a Visual Aid - *CC BY 4.0*
    - 8.7: Designing a Video for Your Needs - *CC BY 4.0*
    - 8.8: Products and Resources for Creating Videos - *CC BY 4.0*
    - 8.9: Introduction to Accessible Visual Aids - *CC BY 4.0*
    - 8.10: Accessibility Matters - *CC BY 4.0*
    - 8.11: Why It Matters- Visual Media - *CC BY 4.0*
    - 8.12: Introduction to Using Visuals - *CC BY 4.0*
    - 8.13: Visuals in a Report - *CC BY 4.0*
    - 8.14: Increasing Impact with Media - *CC BY 4.0*
    - 8.15: Revising and Enhancing Visual Media for Impact - *CC BY 4.0*
    - 8.16: Evaluating the Effectiveness of your Message - *CC BY 4.0*
    - 8.17: Putting It Together- Visual Media - *CC BY 4.0*
    - 8.18: Discussion- Visual Media - *CC BY 4.0*
    - 8.19: Assignment- Visual Media - *CC BY 4.0*
    - 8.20: Introduction to Media and Your Message - *CC BY 4.0*
    - 8.21: Using Visual Media Resources - *CC BY 4.0*

- 8.22: Visual Design Principles - *CC BY 4.0*
- 8.23: Introduction to Images - *CC BY 4.0*
- 8.24: Images Overview - *CC BY 4.0*
- 8.25: Using Images for Impact - *CC BY 4.0*
- 8.26: Using Images Legally - *CC BY 4.0*
- 8.27: Introduction to Charts, Diagrams, and Graphic Organizers - *CC BY 4.0*

◦ 9: Research and Statistics - *CC BY-SA 4.0*
- 9.1: Introduction - *CC BY 4.0*
- 9.2: Measures of Center - *CC BY 4.0*
- 9.3: Measures of Variation - *CC BY 4.0*
- 9.4: There's More to Research Methods than Statistics - *CC BY-SA 4.0*
- 9.5: A brief Introduction to Research Design - *CC BY-SA 4.0*
- 9.6: Introduction to Psychological Measurement - *CC BY-SA 4.0*
- 9.7: Scales of measurement - *CC BY-SA 4.0*
- 9.8: Assessing the Reliability of a Measurement - *CC BY-SA 4.0*
- 9.9: The role of variables — predictors and outcomes - *CC BY-SA 4.0*
- 9.10: Experimental and non-experimental research - *CC BY-SA 4.0*
- 9.11: Assessing the validity of a study - *CC BY-SA 4.0*
- 9.12: Confounds, Artifacts and other Threats to Validity - *CC BY-SA 4.0*
- 9.13: Videos - *CC BY 4.0*
- 9.14: References - *CC BY 4.0*

◦ 10: Data Validation - *CC BY-SA 4.0*
- 10.1: Importance of Statistics and Experimental Measurements - *CC BY 4.0*

◦ 11: Describing Data - *CC BY-SA 4.0*
- 11.1: This is what too many numbers looks like - *CC BY-SA 4.0*
- 11.2: Look at the data - *CC BY-SA 4.0*
- 11.3: Important Ideas - Distribution, Central Tendency, and Variance - *CC BY-SA 4.0*
- 11.4: Measures of Central Tendency (Sameness) - *CC BY-SA 4.0*
- 11.5: Measures of Variation (Differentness) - *CC BY-SA 4.0*
- 11.6: Using Descriptive Statistics with data - *CC BY-SA 4.0*
- 11.7: Rolling your own descriptive statistics - *CC BY-SA 4.0*
- 11.8: Remember to look at your data - *CC BY-SA 4.0*
- 11.9: Videos - *CC BY 4.0*
- 11.10: References - *CC BY 4.0*

◦ 12: Correlation - *CC BY-SA 4.0*
- 12.1: If something caused something else to change, what would that look like? - *CC BY-SA 4.0*
- 12.2: Pearson's r - *CC BY-SA 4.0*
- 12.3: Turning the numbers into a measure of co-variance - *CC BY-SA 4.0*
- 12.4: Examples with Data - *CC BY-SA 4.0*
- 12.5: Regression — A mini intro - *CC BY-SA 4.0*
- 12.6: Interpreting Correlations - *CC BY-SA 4.0*
- 12.7: Summary - *CC BY-SA 4.0*
- 12.8: References - *CC BY 4.0*

◦ 13: Probability, Sampling, and Estimation - *CC BY-SA 4.0*
- 13.1: Prelude to Probability, Sampling, and Estimation - *CC BY 4.0*
- 13.2: How are Probability and Statistics Different? - *CC BY-SA 4.0*
- 13.3: What Does Probability Mean? - *CC BY-SA 4.0*
- 13.4: Basic Probability Theory - *CC BY-SA 4.0*
- 13.5: The Binomial Distribution - *CC BY-SA 4.0*
- 13.6: The normal distribution - *CC BY-SA 4.0*
- 13.7: Other useful distributions - *CC BY-SA 4.0*
- 13.8: Summary of Probability - *CC BY-SA 4.0*
- 13.9: Samples, populations and sampling - *CC BY-SA 4.0*
- 13.10: The Law of Large Numbers - *CC BY-SA 4.0*
- 13.11: Sampling distributions and the central limit theorem - *CC BY-SA 4.0*
- 13.12: The Central Limit Theorem - *CC BY-SA 4.0*
- 13.13: z-scores - *CC BY-SA 4.0*
- 13.14: Estimating population parameters - *CC BY-SA 4.0*
- 13.15: Estimating a confidence interval - *CC BY-SA 4.0*
- 13.16: Summary - *CC BY-SA 4.0*
- 13.17: Videos - *CC BY 4.0*
- 13.18: References - *CC BY 4.0*

◦ 14: Public Speaking - *CC BY 4.0*
- 14.1: How You Speak - *CC BY 4.0*
- 14.2: Introduction to Audience Engagement - *CC BY 4.0*
- 14.3: Audience Attention and Rapport - *CC BY 4.0*
- 14.4: Audience Participation - *CC BY 4.0*
- 14.5: Introduction to Speech Tips and Techniques - *CC BY 4.0*
- 14.6: Overcoming Fears and Anxiety - *CC BY 4.0*
- 14.7: Pro Tips - *CC BY 4.0*
- 14.8: Putting It Together- Public Speaking - *CC BY 4.0*
- 14.9: Discussion- Public Speaking - *CC BY 4.0*
- 14.10: Why It Matters- Public Speaking - *CC BY 4.0*
- 14.11: Introduction to Effective Public Speaking - *CC BY 4.0*
- 14.12: What is Public Speaking? - *CC BY 4.0*
- 14.13: Benefits of Public Speaking - *CC BY 4.0*
- 14.14: Audience Expectations - *CC BY 4.0*
- 14.15: Developing an Effective Speech - *CC BY 4.0*
- 14.16: Introduction to Delivery Techniques - *CC*

BY 4.0
- 14.17: Starting Your Speech - *CC BY 4.0*
- 14.18: Body Language and Gestures - *CC BY 4.0*

◦ 15: Developing and Delivering Business Presentations - *CC BY 4.0*
- 15.1: Text Boxes and Themes - *CC BY 4.0*
- 15.2: Images, Videos, and Arranging Objects - *CC BY 4.0*
- 15.3: Introduction to Making a Presentation for a Meeting - *CC BY 4.0*
- 15.4: Parts of a Good Presentation - *CC BY 4.0*
- 15.5: What's my Presentation About - *CC BY 4.0*
- 15.6: Bad Presentations - *CC BY 4.0*
- 15.7: Making a Presentation for a Meeting - *CC BY 4.0*
- 15.8: Putting It Together- Developing and Delivering Business Presentations - *CC BY 4.0*
- 15.9: Assignment- Developing and Delivering Business Presentations - *CC BY 4.0*
- 15.10: Why It Matters- Developing and Delivering Business Presentations - *CC BY 4.0*
- 15.11: Introduction to Visual Aids - *CC BY 4.0*
- 15.12: Presentation Tools and Visual Aids - *CC BY 4.0*
- 15.13: Introduction to Using Microsoft PowerPoint - *CC BY 4.0*
- 15.14: New Presentations - *CC BY 4.0*
- 15.15: Editing and Enhancing a PowerPoint Presentation - *CC BY 4.0*
- 15.16: Visuals- Images, Videos, and Objects - *CC BY 4.0*
- 15.17: Introduction to Using Google Slides - *CC BY 4.0*
- 15.18: Creating and Organizing Presentations - *CC BY 4.0*

◦ Back Matter - *CC BY 4.0*
- Index - *Public Domain*
- Glossary - *CC BY 4.0*
- Detailed Licensing - *CC BY 4.0*
- Detailed Licensing - *Undeclared*

This page titled Data Analysis Certificate Program is shared under a CC BY-SA 4.0 license and was authored, remixed, and/or curated by Alex Mata.