STAT C1000



Edward Pineda-Castro
Los Angeles City College



STAT C1000

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 and 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 born 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 was 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.

This text was compiled on 08/16/2025

TABLE OF CONTENTS

Licensing

1: Interactive Lecture Notes for Introductory Statistics

- 00: Front Matter
  - Table of Contents
- 1.1: Introduction to Statistics
  - 1.1.1: Introduction to Statistics
  - 1.1.2: Sampling Designs
  - 1.1.3: Experimental Designs
  - 1.1.4: Types of Data
- 1.2: Visual Summaries of Data
  - 1.2.1: Visual Summaries of Qualitative Data
  - 1.2.2: Visual Summaries of Quantitative Data
  - 1.2.3: Other Charts
  - 1.2.4: Distribution Shapes
  - 1.2.5: Organizing Bivariate Data
  - 1.2.6: Misleading Graphs
- 1.3: Numerical Summaries of Data
  - 1.3.1: Measures of Center
  - 1.3.2: Parameter and Statistic
  - 1.3.3: Measures of Variation
  - 1.3.4: Interpreting the Mean and Standard Deviation
  - 1.3.5: Measures of Relative Standing
  - 1.3.6: Five-Number Summaries and Boxplots
- 1.4: Probability Concepts
  - 1.4.1: Introduction to Probability
  - 1.4.2: Finding the Probability
  - 1.4.3: Probability Rules
  - 1.4.4: Probabilities from Frequency Tables
  - 1.4.5: Conditional Probability and Multiplication Rules
- 1.5: Combinatorics Concepts
  - 1.5.1: Introduction to Combinatorics
  - 1.5.2: Permutations and Combinations
  - 1.5.3: Tossing a Coin
- 1.6: Discrete Random Variables
  - 1.6.1: Discrete Random Variables
  - 1.6.2: Applications of Discrete Random Variables
  - 1.6.3: Binomial Random Variables
- 1.7: Continuous Random Variables
  - 1.7.1: Continuous Random Variables
  - 1.7.2: Standard Types of Continuous Random Variables
  - 1.7.3: Other Continuous Random Variables
- 1.8: Normal Distributions
  - 1.8.1: Normal Random Variables

- 1.8.2: Applications of Normal Random Variables
- 1.8.3: Normality Plots
- 1.8.4: Approximating Binomial Random Variables
- 1.9: The Central Limit Theorem
  - 1.9.1: The Central Limit Theorem for Sample Means
  - 1.9.2: The Central Limit Theorem for Sample Sums
  - 1.9.3: The Central Limit Theorem for Sample Proportions
- 1.10: Estimating Parameters
  - 1.10.1: Estimating the Population Mean
  - 1.10.2: The One Mean T Procedure
  - 1.10.3: The One Proportion Z Procedure
  - 1.10.4: The One Variance Chi-Squared Procedure
  - 1.10.5: The Summary of Confidence Intervals
- 1.11: Hypothesis Testing
  - 1.11.1: One Mean Z Test
  - 1.11.2: One Mean T Test
  - 1.11.3: One Proportion Z Test
  - 1.11.4: One Variance Chi-Squared Test
  - 1.11.5: The Summary of Hypothesis Testing for One Parameter
- 1.12: Inferences about Two Parameters
  - 1.12.1: Two Variances F Test
  - 1.12.2: Two Proportions Z Test
  - 1.12.3: Two Means Z Test
  - 1.12.4: Two Means T Pooled Test
  - 1.12.5: Two Means T Non-Pooled Test
  - 1.12.6: Two Paired Means T Test
  - 1.12.7: The Summary of Hypothesis Testing for Two Parameters
- 1.13: Chi-Squared Procedures
  - 1.13.1: The Goodness-of-Fit Test
  - 1.13.2: Homogeneity Test
  - 1.13.3: Test for Independence
- 1.14: ANOVA Procedures
  - 1.14.1: One-Way ANOVA
  - 1.14.2: Two-Way ANOVA
- 1.15: Regression Analysis
  - 1.15.1: Introduction to Regression Analysis
  - 1.15.2: Linear Regression Analysis

Index

Glossary

Detailed Licensing

Detailed Licensing

Licensing

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

SECTION OVERVIEW

1.1: Introduction to Statistics

1.1.1: Introduction to Statistics

1.1.2: Sampling Designs

1.1.3: Experimental Designs

1.1.4: Types of Data

This page titled 1.1: Introduction to Statistics is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.1.1: Introduction to Statistics

Section 1: The Nature of Statistics



Next, we will discuss the nature of statistics and what one should expect to learn in the introductory statistics course.

There are many quotes in literature that describe statistics, some are accurate, and some are entertaining! Here are some of my favorites:

- "There are three kinds of lies: lies, damned lies, and statistics." - Benjamin Disraeli
- "There are two kinds of statistics, the kind you look up, and the kind you make up. " - Rex Stout
- "Statistics can be used to support anything - especially statisticians." - Franklin P. Jones
- "Some use statistics as a drunken man uses lamppost - for support rather than illumination." - Andrew Lang
- "Figures don't lie, liars figured." - Mark Twain

What is the subject of interest in Statistics and what should you expect to learn in an Introductory Statistics course? The science of **statistics** deals with any aspect of the collection, analysis, interpretation, and presentation of **data**. What is data exactly? **Data** (from Latin, plural of datum) are collections of observations. Why doesn't anyone know nor use the word "datum"? Because anywhere you look you are bombarded with data! In the world we live in now, there is a great need for being able to properly handle data! Basically, anything that has to do with data is the subject of interest in a statistics course. In introductory statistics we will learn the aspects of all stages of handling data. Most importantly we will develop the critical thinking skills that will help you to process the data that you are surrounded with. But first we introduce a few fundamental terms.

The population is the collection of all individuals or items under consideration in a statistical study. A study that involves the entire population is called a **census**. Censuses were taken as long ago as Roman times, however, conducting a census may be time-consuming, costly, impractical, or even impossible. Two methods other than a census for obtaining information are sampling and experimentation. A **sample** is a part of the population from which information is obtained.

Since the sample will be used to draw conclusions about the entire population, it should be a **representative** sample – a sample that reflects as closely as possible the relevant characteristics of the population under consideration.

Did you Know? To see what can happen when a sample is not representative, consider the presidential election of 1936. Before the election, the Literary Digest magazine conducted an opinion poll of the voting population. Its survey team asked a sample of the voting population whether they would vote for Franklin D. Roosevelt, the Democratic candidate, or for Alfred Landon, the Republican candidate. Based on the results of the survey, the magazine predicted an easy win for Landon. But when the actual election results were in, Roosevelt won by the greatest landslide in the history of presidential elections! What happened? The sample was obtained from among people who owned a car or had a telephone. In 1936, that group included only the more well-to-do people, and historically such people tend to vote Republican.

Over the centuries, records of such things as births, deaths, marriages, and taxes led naturally to the development of methods for organizing and summarizing information called **descriptive statistics**. Major developments began to occur with the research of Karl Pearson (1857–1936) and Ronald Fisher (1890–1962), who published their findings in the early years of the twentieth century. The methods that were established for drawing and measuring the reliability of conclusions about a population based on information obtained from a sample of the population are now called **inferential statistics**. Historically, descriptive statistics appeared before inferential statistics. Although inferential statistics is a newer arrival, since the work of Pearson and Fisher, it has evolved rapidly and is now applied in a myriad of fields. Familiarity with statistics will help you make sense of many things you read in newspapers, magazines, and on the Internet or watch in the news.

Example 1.1.1.1

In 1948, the Washington Senators played 153 games, winning 56 and losing 97. They finished seventh in the American League and were led in hitting by Bud Stewart, whose batting average was 0.279. Baseball statisticians compiled these and many other statistics by organizing the complete records for each game of the season. Do you think this example demonstrates the methods of descriptive or inferential statistics?

Solution

The correct answer is descriptive since we have simply presented the record of Bud Stewart and his team performance without any attempt to make any predictions whatsoever.

Example 1.1.1.2

In the fall of 1948, the Gallup Poll taken just prior to the election predicted that President Truman would win only 44.5% of the vote and be defeated by the Republican nominee, Thomas E. Dewey. Do you think this example demonstrates the methods of descriptive or inferential statistics?

Solution

The correct answer is inferential since we are trying to estimate the opinion about the entire population based on a sample. In reality, the statisticians had predicted it incorrectly. Truman won more than 49% of the vote and, with it, the presidency. The Gallup Organization modified some of its procedures and has correctly predicted the winner many times since then.

The process of collecting, analyzing, interpreting, and presenting data can be referred to as a **statistical study**. We can distinguish several stages in a statistical study:

1. Prepare – at this stage, a researcher will decide the purpose and the methods of the statistical study.
2. Collect data – at this stage, a researcher will collect the data in the way decided in the previous stage.
3. Process data – at this stage, a researcher will summarize the data or will perform a certain statistical procedure with it.
4. Conclude – at this stage, a researcher will interpret the results of the study based on the purpose and the set goal.

We just discussed the nature of statistics and what one should expect to learn in an introductory statistics course. Next, we will discuss one-by-one all the stages of a statistical study.

Section 2: Preparing for a Study



When preparing for a statistical study, it is important to consider the context, the purpose, and the methods. Based on the methods we can classify a statistical study in the following way. In an **observational study**, researchers simply observe characteristics and take measurements, as in a sample survey. In a **designed experiment,** researchers impose treatments and controls first and then observe characteristics and take measurements. Note that, in an observational study, someone is observing data that already exist (i.e., the data were there and would be there regardless of whether someone was observing it or not). In a **designed experiment**, however, the data do not exist until someone does "the experiment" that produces the data.

Example 1.1.1.1.1

Several studies have been conducted to analyze the relationship between vasectomies and prostate cancer. One such study found 113 cases of prostate cancer among 22,000 men who had a vasectomy. This compares to a rate of 70 cases per 22,000 among men who didn't have a vasectomy. The study shows about a 60% elevated risk of prostate cancer for men who have had a vasectomy. Do you think this is an observational study or a designed experiment?

Solution

The correct answer is an observational study because the subjects were simply observed, and the data was recorded.

Example 1.1.1.1.2

For several years, evidence had been mounting that folic acid reduces major birth defects. In one such study, the doctors enrolled 4753 women prior to conception and divided them randomly into two groups. One group took daily multivitamins containing 0.8 mg of folic acid, whereas the other group received only trace elements. A drastic reduction in the rate of major birth defects occurred among the women who took folic acid: 13 per 1000, as compared to 23 per 1000 for those women who did not take folic acid. Do you think this is an observational study or a designed experiment?

Solution

The correct answer is designed experiment because the subjects were not just simply observed but were randomly assigned a treatment to control the experiment.

Through an observational study one may observe an association (positive or negative) between the two variables, that is the growth of one variable can be linked to the growth or decline of the other variable. Do not confuse an association with a causation when the change in one variable directly triggers the change in the other variable. Remember that association does not imply causation!!! It is important to understand that observational studies can only reveal an association, and to establish the causation between the two variables one has to design an experiment.

Two variables may be strongly associated because they are both related to another variable, called a **lurking** variable, which is a variable previously not considered that causes changes in the two variables under consideration. In the following examples, what do you think a lurking variable is assuming that a strong association between two variables was observed?

Example 1.1.1.2.1

A strong association can be observed between the average number of computers per person in a country and that country's average life expectancy, but it doesn't mean that the number of computers may increase the life expectancy nor vice versa.

Solution

Both variables can be related to the country's GDP.

Example 1.1.1.2.2

A strong association can be observed between the number of firefighters at a fire and the damage caused by the fire, but it doesn't mean that more firefighters cause more damage to the property.

Solution

Both variables can be related to the intensity of the fire.

If you want to learn more about the differences between the association and causation, I invite you to read the history of the US stop-smoking movement. The tobacco companies fought for a long time against the idea of causation as it was hard to argue the strong association between the tobacco consumption and the risk of lung cancer.

Section 3: Collecting Data



When collecting data, it is important to be aware of sampling errors and non-sampling errors. Since no two samples are alike, the actual process of sampling causes a natural variation which is what we call the **sampling error.** The only way to avoid the sampling error is to perform the census. A **non-sampling error** is an issue that affects the reliability of sampling data other than a natural variation; it includes a variety of human errors including poor study design, biased sampling methods, inaccurate information provided by study participants, data entry errors, and poor analysis. In reality, a sample will never be as good as the entire population and since the sampling error is unavoidable, we always try to minimize or eliminate the non-sampling errors.

Example 1.1.1.1

Prior to the 2016 presidential elections, Bobert surveyed ten random people of which 4 said they would vote for Trump and 6 said they will vote for Clinton. The actual popular vote was 46.1% for Trump and 48.2% for Clinton. Do you think this was a sampling error or non-sampling error?

Solution

This is an example of sampling error. Bobert did everything right, but one should not be expecting the sample to produce the same exact results as the population!

Example 1.1.1.2

Prior to the 2016 presidential elections, Bobert surveyed all his Facebook friends of which 7 (14%) said they would vote for Trump and 43 (86%) said they would vote for Clinton and 37 did not respond. The actual popular vote was 46.1% for Trump and 48.2% for Clinton. Do you think this was a sampling error or non-sampling error?

Solution

This is an example of a non-sampling error. Despite the large sample size there is no way this sample would have produced the accurate results due to coverage bias and many other implied biases due to the poor sampling method.

An alternative way to produce data is to design and perform an experiment. While studying various experimental designs is not the purpose of this course one should know the "gold standard" of experimental designs. Here is the diagram showing the sequential flow of a randomized controlled experiment called the "**gold standard**" because it is so effective:

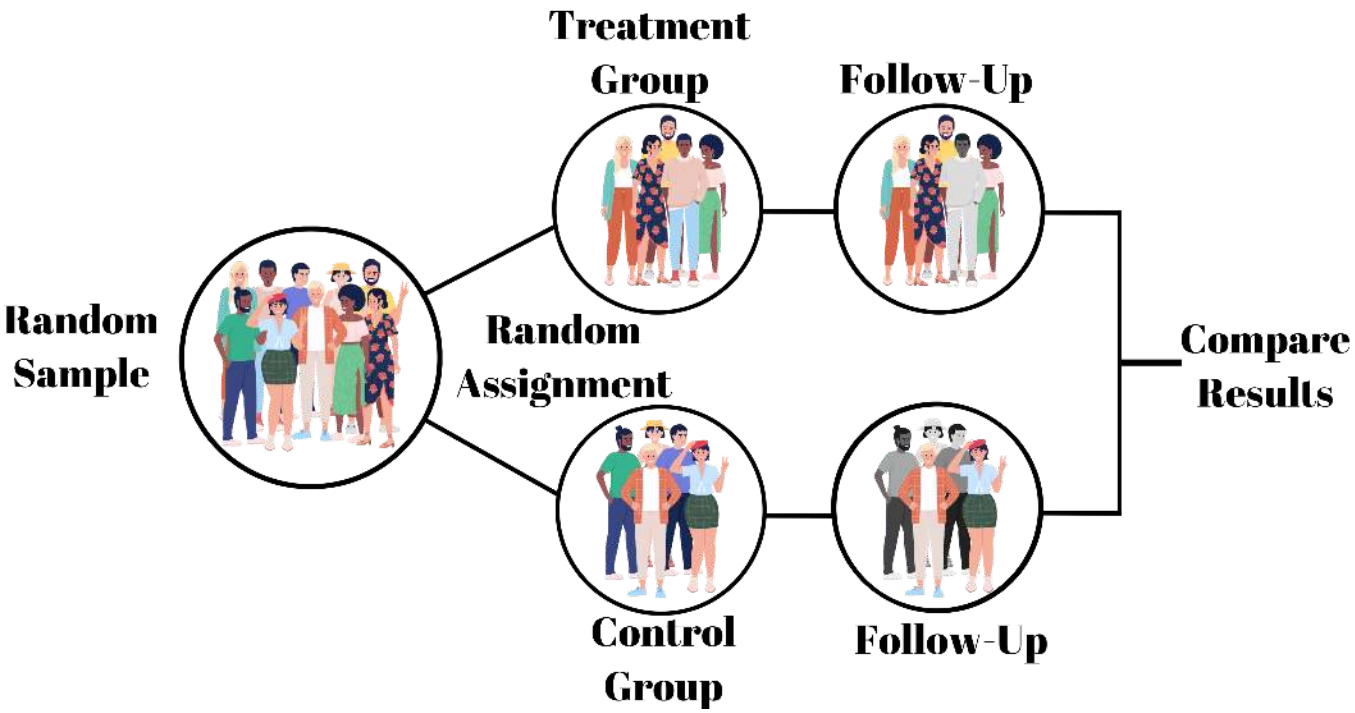


Figure 1.1.1.3.1 : A diagram showing the sequential flow of a randomized controlled experiment called the "gold standard".

A **placebo** such as a sugar pill that has no medicinal effect is sometimes used as one of the treatments. The following example describes how the gold standard was used in the largest health experiment ever conducted.

In 1954, an experiment was designed to test the effectiveness of the Salk vaccine in preventing polio, which had killed or paralyzed thousands of children.



Figure 1.1.1.3.2 : Jonas Salk on the cover of TIME magazine, March 29, 1954.

By random selection, 401,974 children were randomly assigned to two groups:

- 200,745 children were given a treatment consisting of Salk vaccine injections;
- 201,229 children were injected with a placebo that contained no drug.

Children were assigned to the treatment or placebo group through a process of random selection, equivalent to flipping a coin. Among the children given the Salk vaccine, 33 later developed paralytic polio, and among the children given a placebo, 115 later developed paralytic polio.

Section 4: Organizing Data



After the data have been collected either from sampling or from a designed experiment, comes the next stage of organizing the data. Organizing data can be a very difficult task that may involve many steps such as checking the data quality, data cleaning and many other steps. Also, when organizing the data, the storage medium must be considered and chosen. Working with big data may create its own challenges that are not in the scope of the introductory statistics course. In this course we will focus on different ways to organize the collected data **visually** and **numerically**. Now, what will be able to do with the data will depend on what type of data we are working with. Soon we will discuss different types of data and what is there to do with it. The goal of this step of the process is to organize the data so that it can be either analyzed further or explored for patterns and trends! This was a brief introduction to organizing data. We will dedicate more time to this topic later.

Section 5: Drawing Conclusions



After the data is processed and organized, comes the next stage of drawing the conclusions. Later, we will learn a variety of procedures each of which will come with a specific way of drawing the conclusions. In general, this is the stage where we reflect on what was done in a study and discuss the significance of the results. A **statistical significance** is achieved in a study when we get a result that is very unlikely to occur by chance. A common criterion is that we have statistical significance if the likelihood of an event occurring by chance is 5% or less. Results are **practically significant** when the difference is large enough to be meaningful in real life. What is meaningful may be subjective and may depend on the context. As the next example will show, it is

possible that some treatment or finding is effective, but common sense might suggest that the treatment does not make enough of a difference to justify its use or to be practically significant.

Example 1.1.1.1

In a study of a weight loss program, 31 subjects lost an average of 2 lbs. after 9 months. Methods of statistics can be used to show that it is unlikely to get these results by chance if the diet had no effect. Does the weight loss program have statistical significance?

Solution

According to the last sentence of the report the results are statistically significant! Does the weight loss program have practical significance? Would you spend any money on a program that promises that you will lose 2 oz of fat in the next 9 months? Me neither, so the results are practically insignificant!

Did you know? ProCare Industries once supplied a product named Gender Choice that supposedly increased the chance of a couple having a baby with the gender that they desired. In the absence of any evidence of its effectiveness, the product was banned by the Food and Drug Administration (FDA) as a "gross deception of the consumer." But suppose that the product was tested with 10,000 couples who wanted to have baby girls, and the results consist of 5200 baby girls born in 10,000 births. This result is statistically significant because the likelihood of it happening due to chance is only 0.003%. A chance does not seem like a feasible explanation. That 52% rate of girls is statistically significant, but it lacks practical significance because 52% is only slightly above 50%. Couples would not want to spend the time and money to increase the likelihood of a girl from 50% to 52%.

While ethics is not the main point of the course, there are typically two ways in which statistics can be used for deception:

- evil intent on the part of dishonest persons.
- unintentional errors on the part of people who don't know any better.

As responsible citizens, we should learn to distinguish between statistical conclusions that are likely to be valid and those that are seriously flawed, regardless of the source. We will look at some examples of deception after we learn how to organize and visualize data.

This page titled 1.1.1: Introduction to Statistics is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.1.2: Sampling Designs

Section 1: Sampling Designs and Biases



As previously stated, the two methods other than a census for obtaining information are sampling and experimentation. As we are already aware that the sampling error cannot be avoided, we should try to eliminate or minimize the non-sampling error. Next, we will discuss the sampling designs and different biases. One example of a non-sampling error is a sampling bias caused by the sampling method. In plain language, **sampling bias** is a measure of how "not representative" is the sample. One way to minimize the bias is to use a sampling procedure for which each possible sample of a given size is equally likely to be the one obtained. Such a procedure is called **simple random sampling**. The sample obtained from using such a procedure is called a **simple random sample**. So how exactly does one obtain a simple random sample? Let's pick a random sample of size 5! One way to arrange that is to have everyone write their names on a piece of paper and put it in a hat, and then randomly draw 5 pieces of paper from the hat.



Figure 1.1.2.1.1 : An image of a raffle.

We can classify the sampling based on whether the used piece of paper goes back in the hat or not. One is a **sampling with replacement**, whereby a member of the population can be selected more than once; the other is **sampling without replacement**, whereby a member of the population can be selected at most once. Unless we specify otherwise, a simple random sampling is done without replacement. Technically it is harder to perform the sampling without replacement as it requires additional effort of tracking and making sure the choices do not repeat. However, when a sample size is less than 5% of the population size the difference in results is almost negligible. But when population is small, or the sample size is large in relation to the size of the population there could be instances when with or without replacement matters.

Let's decide whether a sampling with or without replacement is more appropriate in the following scenario?

- A teacher picks a random sample of 10 students to answer 10 questions throughout the duration of the class. I'd say that sampling with replacement is more appropriate in this case because as a teacher you don't want a student think that he or she can't be picked again after he was already picked once.
- A teacher picks a random sample of 10 students to attend the conference with him or her.

I'd say that sampling without replacement is definitely more appropriate in this case because a teacher can't take the same student more than once on the same trip.

Obtaining a simple random sample by picking slips of paper out of a box is usually impractical, especially when the population is large. Fortunately, we can use several practical procedures to get simple random samples. One common method involves a table of random numbers, however, nowadays computers can easily replace the table.

A simple random sampling is the most natural and easily understood method of sampling - it corresponds to our intuitive notion of random selection by lot. However, simple random sampling does have its drawbacks. For instance, it may fail to provide sufficient coverage or may be impractical when the members of the population are widely scattered geographically. Next, we will examine some other commonly used sampling procedures that are often more practical than a simple random sampling.

All sampling techniques can be classified as random (probability) sampling and non-random (non-probability) sampling. **Probability sampling** is a sampling technique, in which the subjects of the population get an equal opportunity to be selected as a representative sample which is not the case with **non-probability sampling**.

Let’s assume that there are 35 desks in the classroom arranged in 5 rows with 7 desks in each row. The following diagrams illustrate the difference between random and non-random sampling of 11 students.



Figure 1.1.2.1.2: A diagram illustrating **random sampling** of 11 students from 35 in the classroom arranged in 5 rows with 7 desks in each row.





Figure 1.1.2.1.3 : A diagram illustrating **non-random sampling** of 11 students from 35 in the classroom arranged in 5 rows with 7 desks in each row.

We discussed the simple random sampling which is the ideal sampling method for producing the least biased sample. Next, we will examine some other commonly used random and non-random sampling procedures.

Section 2: Systematic Sampling



A **systematic** random sampling is a method for selecting a random sample by listing the entire population and picking every k-th object starting randomly with the object numbered between 1 and k. We select k to be as close as possible to the population size divided by the sample size.

Here is an example for picking every third person from a population of size 12 starting with number 3:

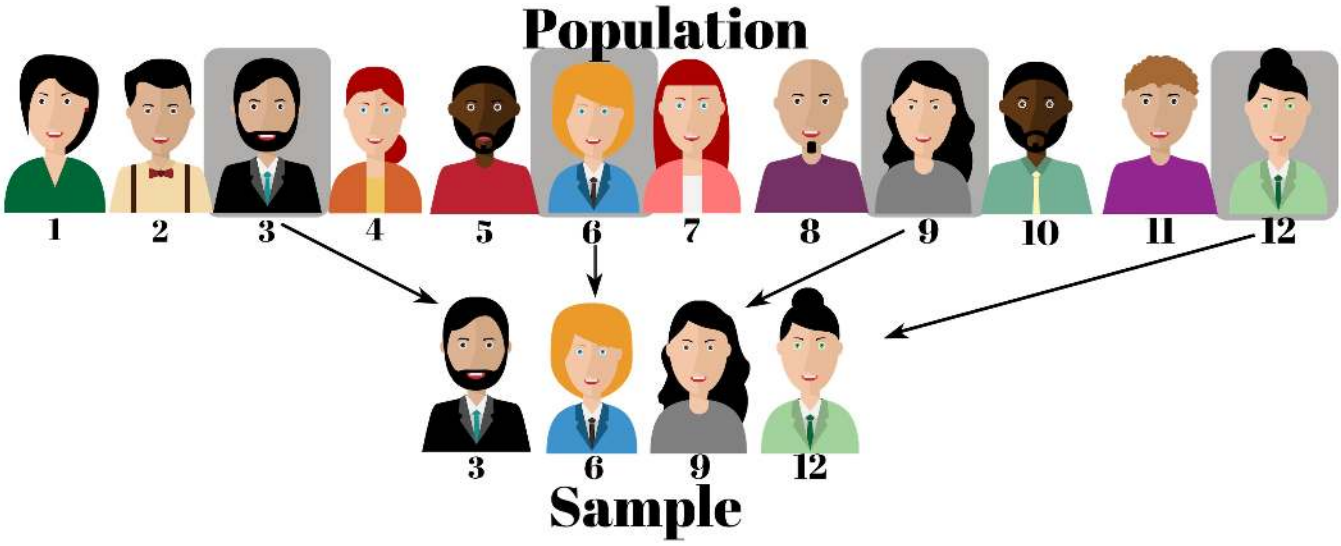


Figure 1.1.2.2.1 : A diagram showing 12 people in a row with the 3rd, 6th, 9th, and 12th selected.

Here is an example for picking every third person from a population of size 12 starting with number 2:

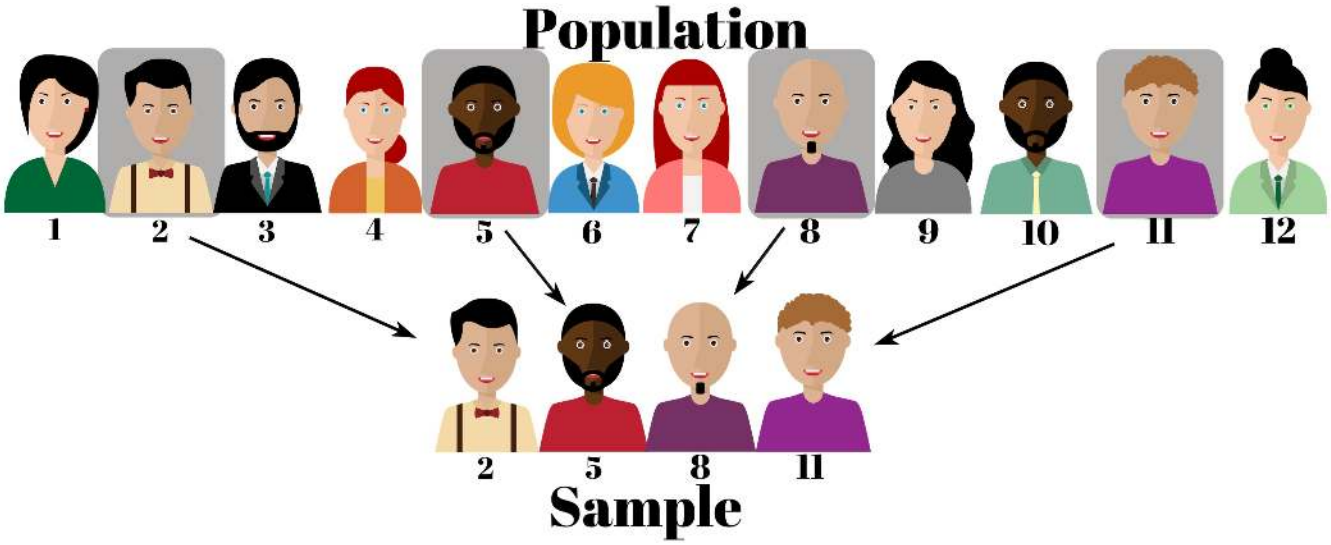


Figure 1.1.2.2.2 : A diagram showing 12 people in a row with the 2nd, 5th, 8th, and 11th selected.

Here is an example for picking every second person from a population of size 10 starting with number 2.

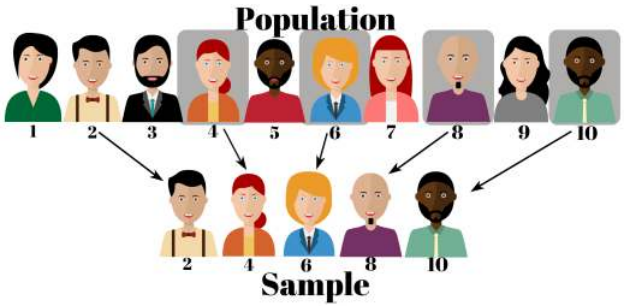


Figure 1.1.2.2.3 : A diagram showing 10 people in a row with the 2nd, 4th, 6th, 8th, and 10th selected.

The procedure can be outlined in the following steps:

1. Divide the population size by the sample size and round the result down to the nearest whole number, k.
2. Use a random-number table or a similar device to obtain a number, m, between 1 and k.
3. Select for the sample those members of the population that are numbered m, m + k, m + 2k, etc.

When is the systematic sampling useful? For example, the owner of a small convenience store in a rural area wants to survey the opinion of their customers and they expect 20 customers in the next hour. They think that a sample of size 5 will be sufficient so they follow the procedure:

1. The population size 20 divided by the sample size 5 is 4 so k=4.
2. A random number between 1 and 4 picked by a computer is 3.
3. They will survey the 3rd, 7th, 11th, 15th, and 19th customers.

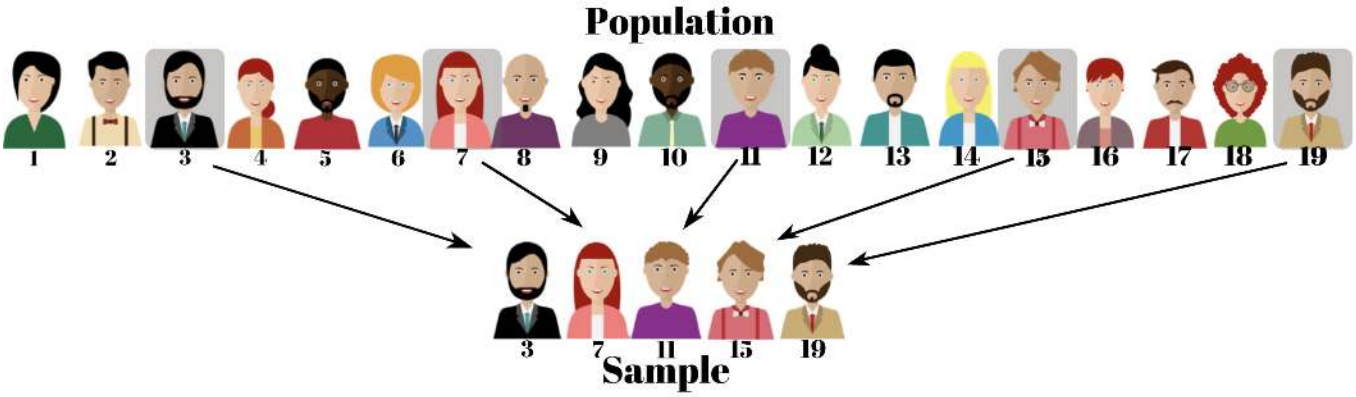


Figure 1.1.2.2.4 : A diagram showing 19 people in a row with the 3rd, 7th, 11th, 15th, and 19th selected.

There are advantages and disadvantages for every procedure. The main advantage of the systematic random sampling is that the procedure doesn't require access to the entire population, but on the other hand it does assume the knowledge of the population size.

Section 3: Stratified Sampling



A **stratified** sampling is a method for selecting a random sample used to ensure that subgroups of the population are represented adequately. Here is a diagram that summarizes the stratified sampling procedure:

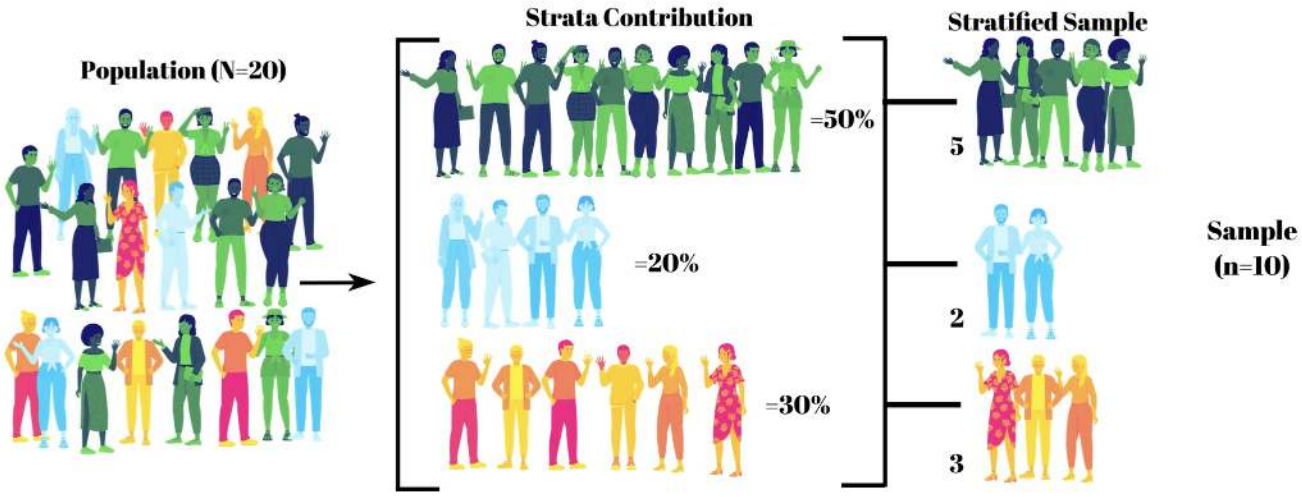


Figure 1.1.2.3.1 : A diagram that summarizes the stratified sampling procedure.

The procedure can be outlined in the following steps:

1. Divide the population into subpopulations called strata.
2. From each stratum, obtain a simple random sample of size proportional to the size of the stratum; that is, the sample size for a stratum equals the total sample size times the stratum size divided by the population size.
3. Use all the members obtained in Step 2 as the sample.

When is the stratified sampling useful? It is useful when the population is heterogeneous and contains several different groups. For example, the principal of a high school wants to survey the opinions of the students. They think that a sample of size 100 will be sufficient. There are total of 1000 students (400 freshmen, 300 sophomores, 200 juniors, 100 seniors). It won't be representative if in a sample of size 100 there were 25 students from each class. So they follow the procedure:

1. The population is already divided into classes that will be used as strata.
2. The sample size for each class will be determined by the formula k*n/N, where k is the class size, n is the total sample size, N is the population size.
3. From each class, obtain a simple random sample of size proportional to the size of the class; that is there will be:
   - Freshmen: $200 \cdot \frac{50}{500} = 20$
   - Sophomores: $150 \cdot \frac{50}{500} = 15$
   - Juniors: $100 \cdot \frac{50}{500} = 15$
   - Seniors: $50 \cdot \frac{50}{500} = 5$
4. Combine all the members obtained in Step 2 as the sample that will consist of 20 freshmen, 15 sophomores, 10 juniors, 5 seniors.

There are advantages and disadvantages for every procedure. The main advantage of the stratified random sampling is that the procedure ensures a high degree of representativeness of all the strata or layers in the population, but on the other hand it is time-consuming and tedious.

Section 4: Voluntary Response Sampling



A **voluntary response sampling** is a type of sampling that produces a sample made up of self-chosen participants. These participants volunteer to take part in different research studies to share their opinions on topics that interest them. Polling through call-in radio shows is an ideal voluntary response sample example. Only a part of the population that listens to the particular radio station (and who chooses to answer by dialing in) participate in the poll. The responses collected do not accurately reflect the feelings of the entire population as only those people who choose to call in and take part in the study will bother to respond. When is the voluntary response sampling useful? As long as the question doesn't cover a controversial topic, a voluntary response sampling can be a quick and cheap way to survey the opinion of the population.

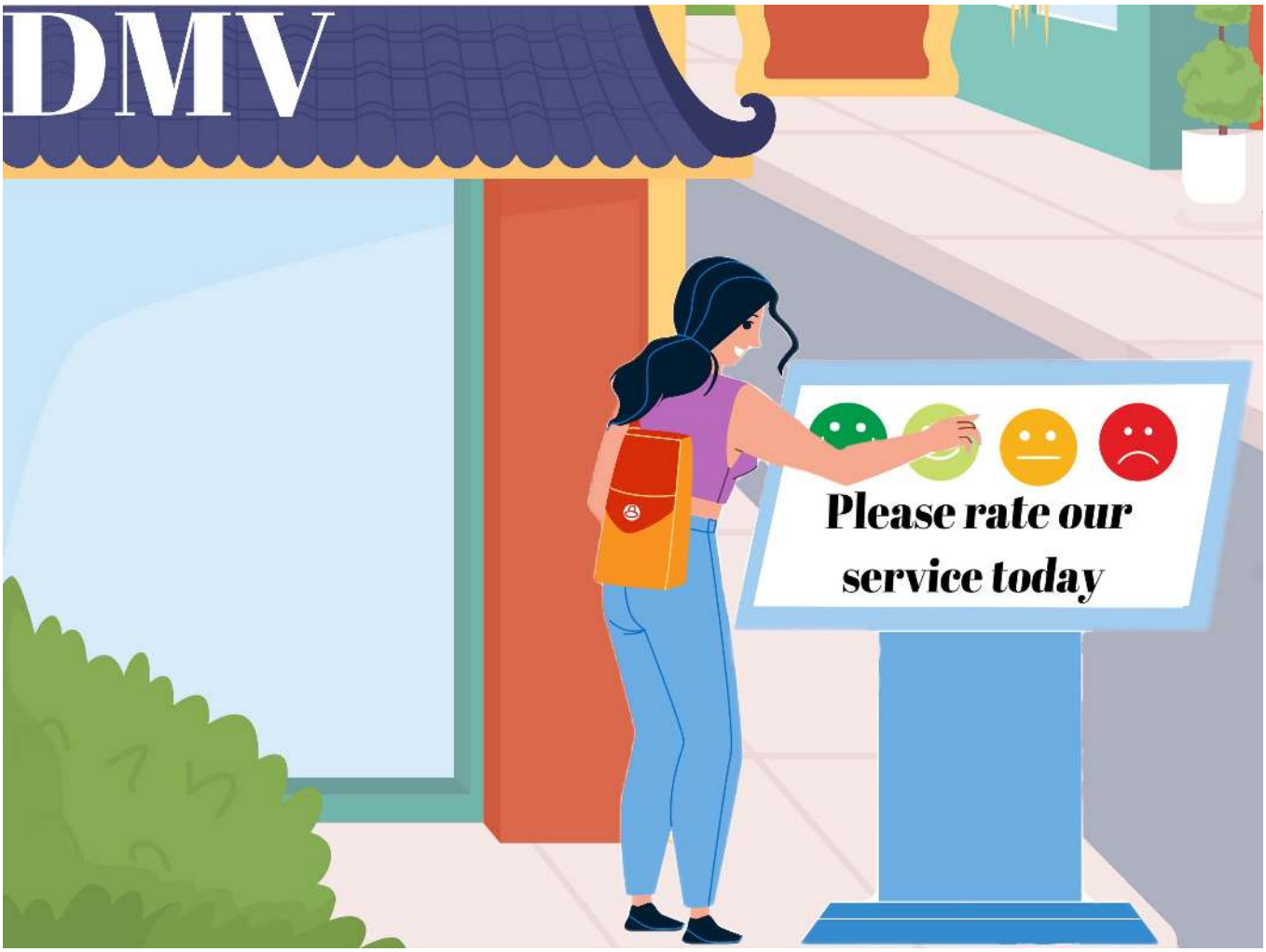


Figure 1.1.2.4.1 : An image of a person pressing the button to rate the service.

The main advantage of the voluntary response sampling is that the procedure may not accurately reflect the feelings of the entire population, but on the other hand it is inexpensive and minimal effort is required.

Section 5: Convenience Sampling



A **convenience** sampling is a nonrandom method of selecting a sample; this method selects individuals that are easily accessible. Here is a diagram that summarizes the convenience sampling procedure.

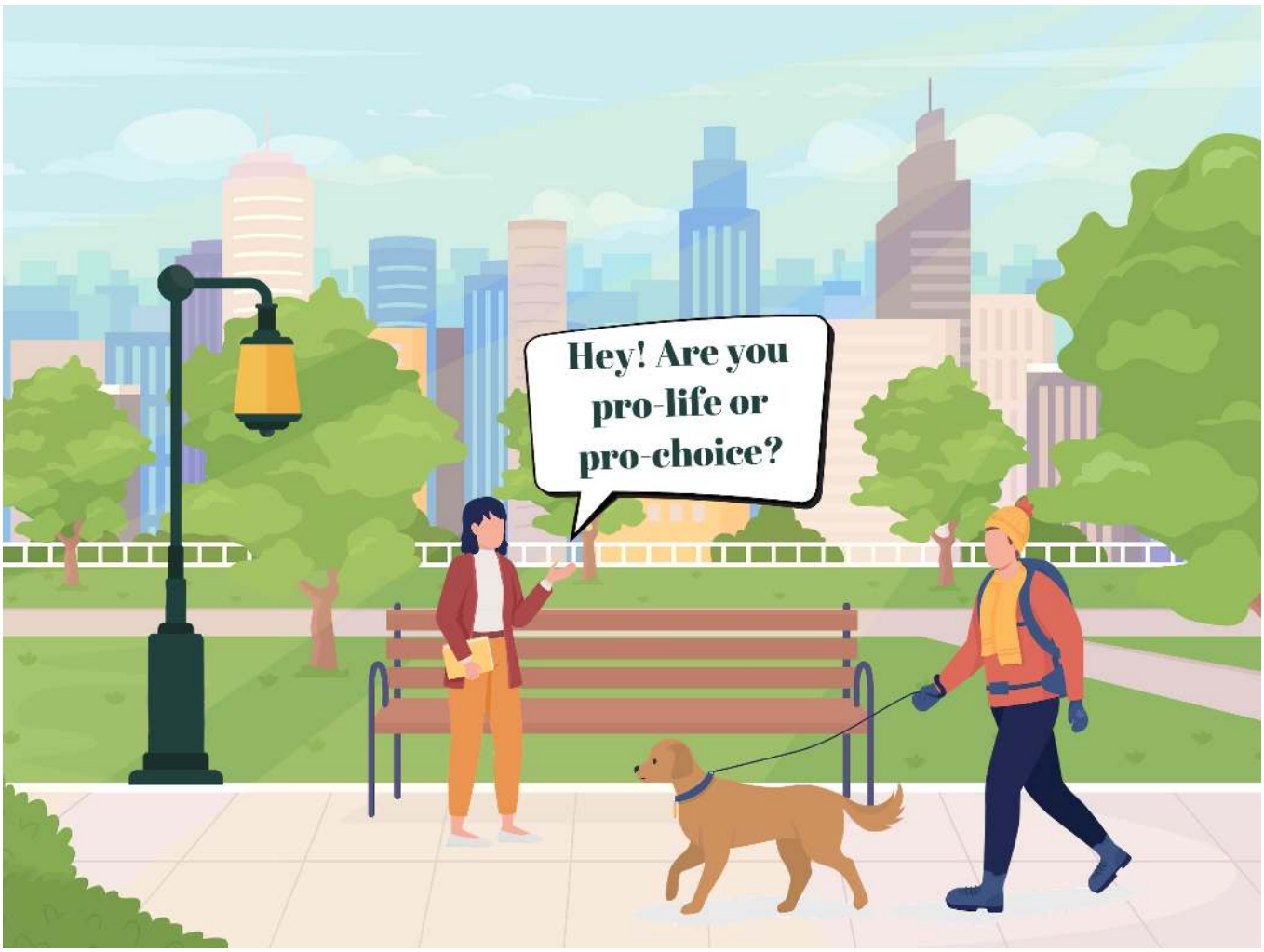


Figure 1.1.2.5.1 : An image of a person asking a passerby a question.

When is the convenience sampling useful? In pilot studies, a convenience sample is usually used because it allows the researcher to obtain basic data and trends regarding his study without the complications of using a randomized sample. The main advantage of the convenience sampling is that the procedure is convenient and inexpensive, but on the other hand it is a non-random sampling procedure, so generalization is questionable.

Section 6: Other Biases



To minimize the sampling biases, most large-scale surveys combine more than one random sampling techniques. Such multistage sampling is used frequently by pollsters and government agencies. For instance, the U.S. National Center for Health Statistics conducts surveys to obtain information on a variety of health issues. Data collection is done by a multistage probability sampling of approximately 42,000 households.

There are many other possible issues that may compromise the validity of a statistical study. Here is by no means a complete list of terms that one should be familiar with when analyzing a statistical study:

- Push polling
- Loaded questions
- The order of questions
- Social acceptability bias
- Publication bias
- Patterns in missing data
- Self-interest study
- Precise lies

Some of the terms are examples of a specific bias and some are examples of unethical practices that lead to some type of bias.

A **push poll** is an interactive marketing technique, most commonly employed during political campaigning, in which an individual or organization attempts to manipulate or alter prospective voters' views under the guise of conducting an opinion poll. Large numbers of voters are contacted with little effort made to actually collect and analyze voters' response data. Instead, the push poll is a form of telemarketing-based propaganda and rumor spreading, disguised as an opinion poll. Push polls may rely on insinuations, or information gathered from opposition research serving the political opponents.

A "**loaded question**", like a loaded gun, is a dangerous thing. A loaded question is a question with a false or questionable presupposition, and it is "loaded" with that presumption. This type of fallacious question puts the person who is being questioned in a disadvantageous and defensive position, since the assumption in the question could reflect badly on them or pressure them to answer in a way that they wouldn't otherwise.

- One especially loaded question was asked by Brian Williams during the 2011 Republican presidential candidate debate at the Reagan Library. Directed to Texas Senator Rick Perry, Williams asked: 'have you struggled to sleep at night with the idea that any one of those [234 Texas executed inmates] might have been innocent?' This loaded question threw Perry for a loop because it actually entails a number of different assumptions including the moral validity of execution and Perry's own involvement with this part of the judiciary process.
- Here is another recent example of a loaded question: "Do you regret at all, all the lying you've done?" A White House correspondent on Thursday August 13th 2020 said he's been waiting five years to ask U.S. President Donald Trump a question,

and he finally got his chance. But the blunt question was not met with an answer as this is usually the best strategy in response to loaded questions such as this.

The **order that questions are asked** in a survey or study can influence the answers that are given. That's because the human brain has a tendency to organize information into patterns. The earlier questions — in particular, the ones that come just before — may provide information that subjects use as context in formulating their subsequent answers, or affect their thoughts, feelings, and attitudes. Pew Research gave this example from a December 2008 poll: "When people were asked 'All in all, are you satisfied or dissatisfied with the way things are going in this country today?' immediately after having been asked 'Do you approve or disapprove of the way George W. Bush is handling his job as president?'; 88 percent said they were dissatisfied, compared with only 78 percent without the context of the prior question."

In social science research, **social desirability bias** is a type of response bias that describes the tendency of survey respondents to answer questions in a manner that will be viewed favorably by others.[1] It can take the form of over-reporting "good behavior" or under-reporting "bad", or undesirable behavior. The tendency poses a serious problem with conducting research with self-reports. This bias interferes with the interpretation of average tendencies as well as individual differences. When confronted with the question, "Do you use drugs/illicit substances?" the respondent may be influenced by the fact that controlled substances, including the more commonly used marijuana, are generally illegal. Respondents may feel pressured to deny any drug use or rationalize it, e.g. "I only smoke marijuana when my friends are around."

Publication bias is a type of bias that occurs in published academic research. It occurs when the outcome of an experiment or research study influences the decision whether to publish or otherwise distribute it. Publishing only results that show a significant finding disturbs the balance of findings and inserts bias in favor of positive results. Publication bias is sometimes called the **file-drawer effect**, or **file-drawer problem**. This term suggests that results not supporting the hypotheses of researchers often go no further than the researchers' file drawers, leading to a bias in published research.

Understanding the reasons why **data** are **missing** is important for handling the remaining data correctly. If values are missing completely at random, the data sample is likely still be representative of the population. But if the values are missing systematically, analysis may be biased. For example, in a study of the relation between IQ and income, if participants with an above-average IQ tend to skip the question ‘What is your salary?’, analyses that do not take into account this may falsely fail to find a positive association between IQ and salary.

Another type of bias occurs in a **self-interest study** – a study in which the researchers have an interest in the outcome of the study. Would you trust a 1930 Lucky Strike advertisement knowing that the medical study was sponsored by the tobacco company itself? Soon after e-cigarettes debuted in Europe in 2006, tobacco companies began investing heavily in vaping. Do you trust that vaping is harmless? I hope not!

Another interesting phenomenon called the **precise lies** frequently occurs in the age of mass media. The famous statistic of women earning $0.77 for every dollar men earn is exactly this sort of lie. It makes no attempt to correct for differences in experience, education, or even full versus part time status. Louis Jacobson of PolitiFact does a very nice job of walking a reader through different ways controls can be added that change that number, while also pointing out that it can be a true statement as long as it doesn’t come with an implication this difference is due to discrimination against women working equivalent jobs.

We discussed a variety of issues that may arise during the sampling process and compromise the validity of a statistical study. As a responsible citizen it is your responsibility to be aware of these issues when conducting or analyzing a statistical study.

This page titled 1.1.2: Sampling Designs is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.1.3: Experimental Designs



While studying experimental designs is not the main goal of this course, one should be familiar with the vocabulary and some basic designs including the "gold standard".

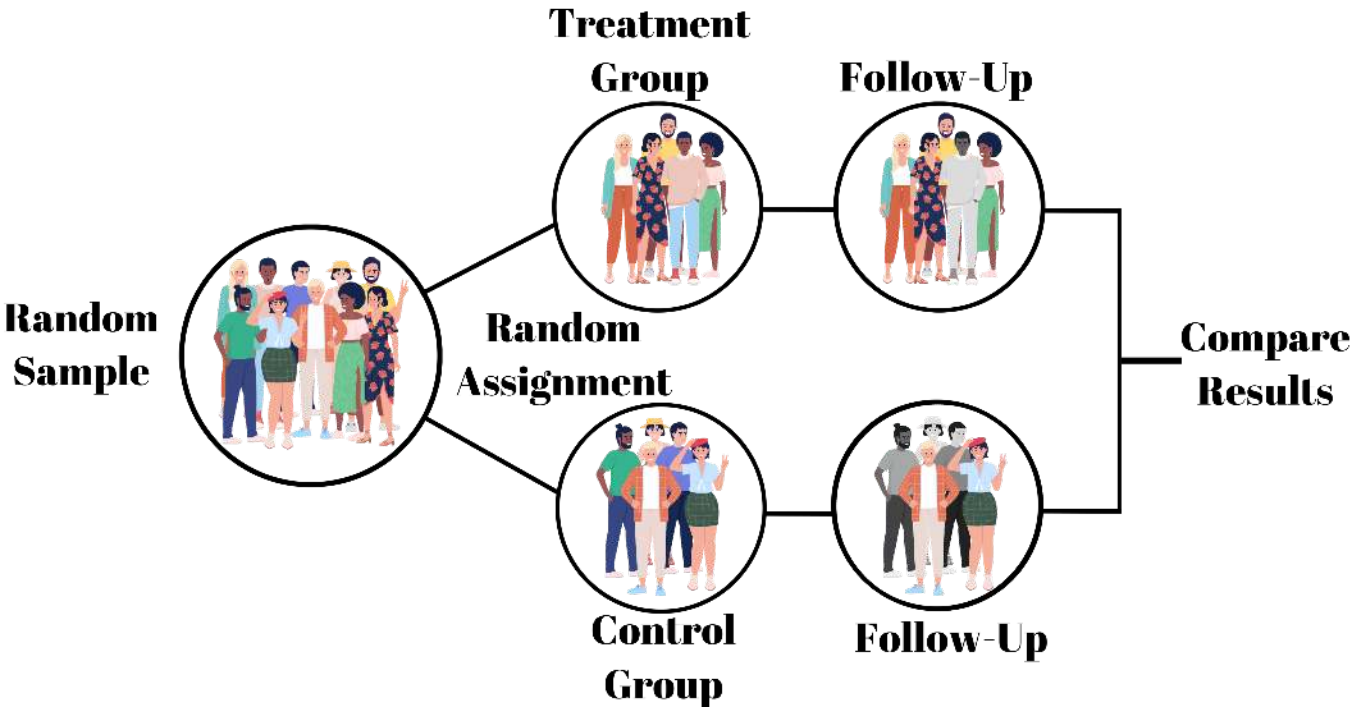


Figure 1.1.3.1.1 : A diagram showing the sequential flow of a randomized controlled experiment called the "gold standard".

Also known as **randomized controlled trial.** There are four components:

1. There is a treatment to be studied like a program, a drug, or a medical procedure.
2. There is a control condition. Sometimes, this is a group that doesn't' get any treatment at all. Often it is a group that gets some other kind of treatment, but of a different kind or smaller amount.
3. *The participants must be randomly assigned to treatment or control groups*. It is critical that nobody – not the researchers, not the people in the experiment – can participate in the decision about which group people fall into. Some kind of randomization

procedure is used to put people into groups – flipping a coin, using a computer, or some other method. This is the only way we can make sure that the people who get the intervention will be similar to those who do not.

4. There must be carefully defined outcome measures, and they must be measured before and after the treatment occurs.

In 1954, an experiment was designed to test the effectiveness of the Salk vaccine in preventing polio, which had killed or paralyzed thousands of children. By random selection, 401,974 children were randomly assigned to two groups:

- 200,745 children were given a treatment consisting of Salk vaccine injections;
- 201,229 children were injected with a placebo that contained no drug.

Children were assigned to the treatment or placebo group through a process of random selection, equivalent to flipping a coin. Among the children given the Salk vaccine, 33 later developed paralytic polio, and among the children given a placebo, 115 later developed paralytic polio.

One should be familiar with the following vocabulary.

An **experimental group** in an experiment is the group that is exposed to the treatment. A **control group** in an experiment is the group that does not receive the experimental treatment.

In the Salk vaccine experiment, the children that received the vaccine formed an experimental group, and the children that didn't receive a vaccine formed a control group.

A **treatment** is a specific condition applied to the individuals in an experiment. A **placebo** is a fake drug used in the testing of medication.

In the Salk vaccine experiment, the experimental group was treated with vaccine and the control group with placebo.

A **response variable** is a variable that measures an outcome or result of a study. An **explanatory variable** is a variable that we think explains or causes changes in the response variable.

In Salk vaccine experiment, the response variable was whether a child developed polio or not, and the explanatory variable was whether the vaccine was administered or not.

A **blinding** is a technique where the subjects do not know whether they are receiving a treatment or a placebo. A **double-blinding** is a technique where both the subject and the data recorder do not know the treatment.

One can dedicate the entire semester to study experimental designs. Since this is not the purpose of the course, we only discussed the main features of the gold standard design and the vocabulary associated with the concept.

This page titled 1.1.3: Experimental Designs is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.1.4: Types of Data



Not all data are the same. Next, we will discuss the different ways to classify statistical data.

A **statistical variable** is a characteristic that varies from one object to another. For example, age, major, GPA, and the desired grade are not the same for all students therefore each one of them is a statistical variable. The values of a variable for one or more people or things yield **data**. Here is an example of data collected from 3 students.

Table 1.1.4.1.1 : An example of data collected from 3 students showing their age, major, GPA, and desired grade.

Age	Major	GPA	Desired Grade
19	Psychology	3.5	B
24	Criminal Justice	3.8	A
21	Nursing	3.6	A

Most variables and data can be put into the following categories: qualitative and quantitative.

Qualitative data consist of names and labels describing the attributes of a population such as hair color, blood type, ethnic group, the car a person drives, and the street a person lives on. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+.

Quantitative data consist of numbers that are the result of counting or measuring attributes of a population such as the amount of money, pulse rate, weight, number of people living in your town, and number of students who take statistics.

While numbers usually mean quantities, in general, not every number represents a quantity. For example, ZIP codes do not mean quantities therefore they are qualitative rather than quantitative type. Also, phone numbers, student IDs don't mean quantities and since they are unique for each subject, they are called identifiers.

Qualitative data can be further classified as either ordinal or categorical. **Ordinal** data have a natural ordering, such as the letter grades, clothing size, and Likert scales. Qualitative data without a natural order is called **categorical**. For example, colors and college majors. There is also a special type of categorical data called **binary** when there are only two possible answer to a question such as yes/no or agree/disagree.

Quantitative data can be further classified as either discrete or continuous. All possible values of **discrete** data can be listed, while all possible values of **continuous** data form a continuous interval. For example: quantities, weights (rounded to pounds), ages (in years) are some examples of discrete type. Note that discrete data doesn't mean that the list is finite. On the other hand, the exact incomes, weights or heights, and time are some examples of continuous type.

Here is an example of different types of data collected from the same set of shows:

Table 1.1.4.1.2 : An example of different types of data collected from the same set of shows

Rank	*Show Title*	*Network*	*Viewers (millions)*
1	CSI	CBS	19.3
2	NCIS	CBS	18.0
3	Dancing with Stars	ABC	17.8
4	Desperate Housewives	ABC	15.5
5	The Mentalist	CBS	14.9

The rank is discrete type, show title is the identifier, network is categorical, and number of millions of viewers is continuous.

Which type of data is more "complex"? From the least "complex" and most “boring” on the left to the most "complex" and “interesting” on the right, we can organize the data types in the following way:

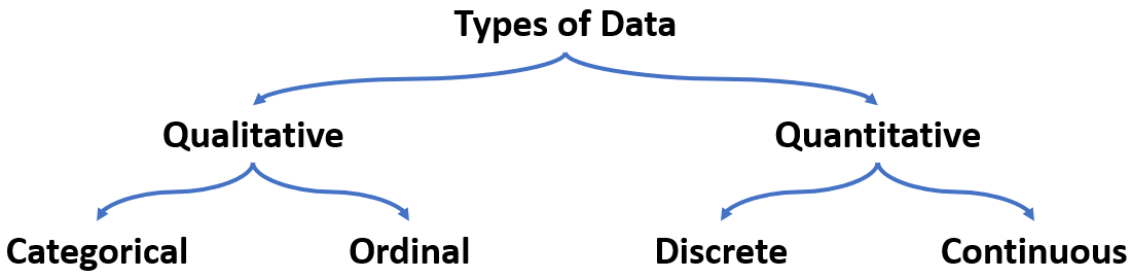


Figure 1.1.4.1.3 : The tree diagram showing the 4 types of data: categorical, ordinal, discrete, and continuous (from left to right).

Note that everything we can do with a less complex data type we would also be able to do with the more complex type of data. That is, we can do with discrete data everything that we can do with categorical data and some more!

Another way to classify the data is by the type of math operations that can be applied to the elements of the dataset. As a result, we have four **levels of measurements**: nominal, ordinal, interval, and ratio:



Figure 1.1.4.1.4 : The tree diagram showing the 4 levels of measurements: nominal, ordinal, ratio, and interval (from left to right).

The **nominal level** is when the data cannot be arranged in any ordering scheme. For example: yes/no/undecided, political party affiliation.

The **ordinal level** is when the data can be arranged in some order but differences between data values either cannot be determined or are meaningless. For example: grades, ranks.

The **interval level** is when the data can be arranged in some order, for which the differences between data values are meaningful but there is no natural zero. For example: temperature, years.

The **ratio level** is when the data can be arranged in some order, for which the differences between data values are meaningful, and there is a natural zero that represents the absence of the quantity. For example: distances, prices.

Note that the first classification addresses what the data looks like and the latter classification addresses what can be done with the data mathematically. We can match the two classifications in the following way:

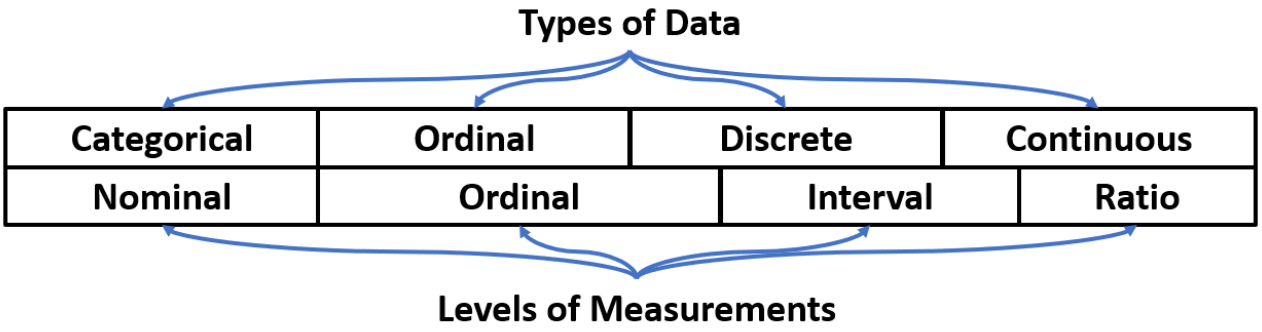


Figure 1.1.4.5: A diagram showing how the different levels of measurements align with different types of data.

Also note that the framework of distinguishing levels of measurement **originated in psychology and is widely criticized by scholars in other disciplines!**

This page titled 1.1.4: Types of Data is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

SECTION OVERVIEW

1.2: Visual Summaries of Data

1.2.1: Visual Summaries of Qualitative Data

1.2.2: Visual Summaries of Quantitative Data

1.2.3: Other Charts

1.2.4: Distribution Shapes

1.2.5: Organizing Bivariate Data

1.2.6: Misleading Graphs

This page titled 1.2: Visual Summaries of Data is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.2.1: Visual Summaries of Qualitative Data

Section 1: Categorical Data



Next, we will discuss how to visually summarize data. We will start with learning how to organize categorical data because it is the simplest type of data thus everything that we can do with categorical data can be applied to all other types of data.

Consider the following data set obtained from asking 40 students their political party affiliation:

D	D	R	R	R	R	R	R
D	R	D	D	O	D	R	D
D	R	D	D	R	R	D	R
D	R	D	D	D	R	O	D
O	D	D	D	R	R	R	D

One way of organizing categorical data is to construct a table that contains the list of all possible values. For each value we can find its **frequency** - the number of times that the value occurs in the data set. To obtain the frequencies we are going to go through the data and use tally:

Political party affiliation	Tally	Frequency
D	𝍸 𝍸 𝍸 𝍸	20
O	𝍷	3
R	𝍸 𝍸 𝍸 ||	17
Total:		20+3+17=40

A table that lists the distinct values along with their frequencies is called the **frequency distribution table**. One way to validate the table is to find the total, that is, add the entries in the frequency column and check whether the sum is equal to the sample size. Using tally can save a lot of time when constructing a frequency table because when using the tally method one would only have to browse through the data set once.

In summary, to construct a frequency distribution table for categorical data:

1. List the distinct values of the observations in the data set in the first column of a table.

2. For each observation, place a tally mark in the second column of the table in the row of the appropriate distinct value.
3. Count the tallies for each distinct value and record the totals in the third column of the table.

Let's compare the results from the same survey of a different class:

Political party affiliation	Frequency
D	40
O	
R	
Total:	

We see that the number of people that support democrats in the other class is 40. Can we conclude the democrat supporters prevail in the other class twice as much? Well, not if you see the entire frequency table!

Political party affiliation	Frequency
D	40
O	5
R	55
Total:	100

Turns out that a frequency table does not show the whole picture! We can compare apples to apples if we compare the corresponding relative frequencies. To compute a relative frequency, we divide the frequency of an observation by the total.

Class 1

Party Affiliation	Frequency	Relative Frequency
D	20	20/40=0.50=50%
O	3	3/40=0.075=7.5%
R	17	17/40=0.425=42.5%
Total:	40	40/40=1=100%

Class 2

Party Affiliation	Frequency	Relative Frequency
D	40	40/100=0.4=40%
O	5	5/100=0.05=5%
R	55	55/100=0.55=55%
Total:	100	100/100=1=100%

Relative frequencies can be expressed as fractions, decimals, or percentages. The total of a relative frequency column must add up to 1 or 100%. A listing of the distinct values and their relative frequencies is called **relative-frequency distribution table**. Now, we can compare apples with apples and see that the democrats are actually a minority in the second class!

In summary, to construct a relative-frequency distribution table for categorical data:

1. Obtain a frequency distribution of the data.
2. Divide each frequency by the total number of observations to obtain a relative frequency.

One way to visualize a frequency table is to draw a two-dimensional coordinate plane with the categories on the horizontal axis and frequencies on the vertical axis. For each category, we draw a bar above the corresponding label with the height equal the frequency.

Such a display of frequencies is called a **frequency bar chart**. The frequency of each distinct value is represented by a vertical bar whose height is equal to the frequency of that value. The bars should be positioned so that they do not touch each other.

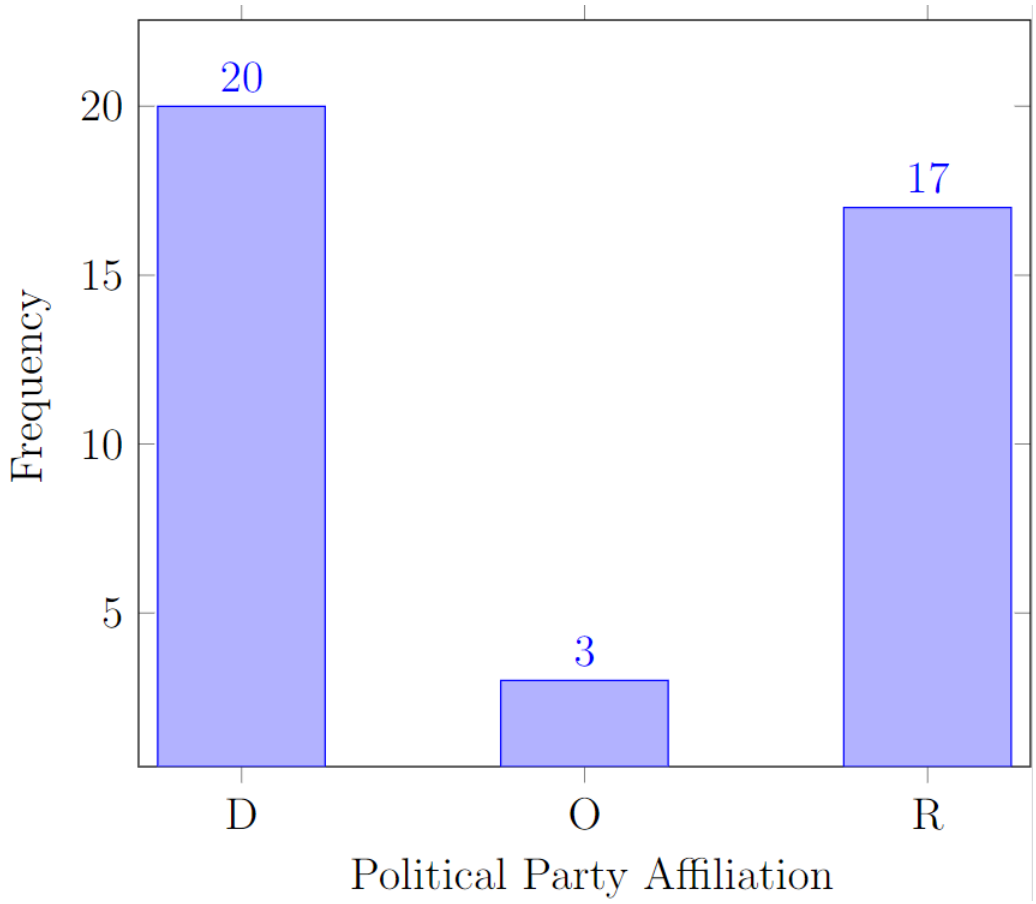


Figure 1.2.1.1.1 : A frequency bar plot showing political party affiliation of 40 students of which there are 20 democrats, 17 republicans, and 3 others. (Copyright; author via source)

In the same way we may visualize the relative frequency table.

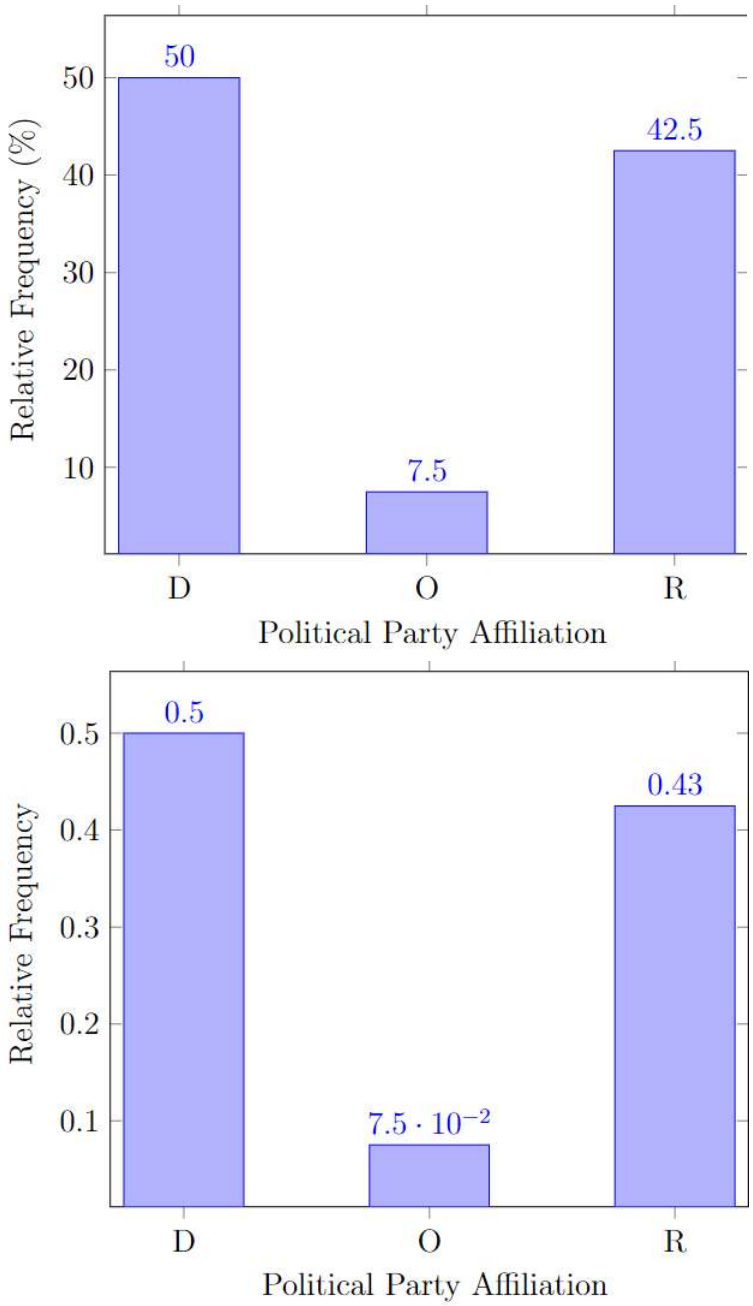


Figure 1.2.1.1.2: A relative frequency bar plot showing political party affiliation of 40 students of which there are 50% democrats, 42.5% republicans, and 7.5% others using two different ways to label the units on the vertical axis. (Copyright; author via source)

Such a summary is called the **relative frequency bar chart**. Note that the only difference between the frequency and the relative frequency bar charts is the scale and the units on the vertical axis, everything else is identical.

In summary, to construct a frequency or a relative frequency bar chart:

1. Obtain a frequency or relative-frequency distribution of the data.
2. Draw a horizontal axis on which to place the bars and a vertical axis on which to display the relative frequencies.
3. For each distinct value, construct a vertical bar whose height equals the relative frequency of that value.
4. Label the bars with distinct values, the horizontal axis with the name of the variable, and the vertical axis with “Relative frequency.”

Frequency tables and bar charts are considered the most basic ways of organizing the data! However, scientists kept inventing the new ways of organizing data throughout times. Another way to visualize the categorical data is via pie-chart. A **pie chart** is a disk divided into wedge-shaped pieces proportional to the relative frequencies of the qualitative data. In other words, the area that

corresponds to D is 50% of the total area, the area that corresponds to O is 7.5% of the total area, and the area that corresponds to R is 42.5% of the total area.

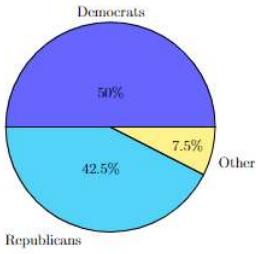


Figure 1.2.1.1.3: A relative frequency pie chart showing political party affiliation of 40 students of which there are 50% democrats, 42.5% republicans, and 7.5% others. (Copyright; author via source)

To construct this by hand one would have to use a protractor and a calculator to compute the exact values of the central angles for each sector! Luckily, every statistical software can create a pie chart with just a press of a button! For us, it is more important to be able to read the chart than to construct it because nowadays creating such chart and especially printing it out is considered a waste of ink!

In summary, to construct a pie chart:

1. Obtain a relative-frequency distribution of the data.
2. Divide a disk into wedge-shaped pieces proportional to the relative frequencies.
3. Label the slices with distinct values and their relative frequencies.

Note that due to the absence of the natural order, the categories in categorical data can be re-arranged for different purposes. For example, in a **Pareto chart**, the bars are arranged in decreasing order by the category size (from the largest to the smallest):

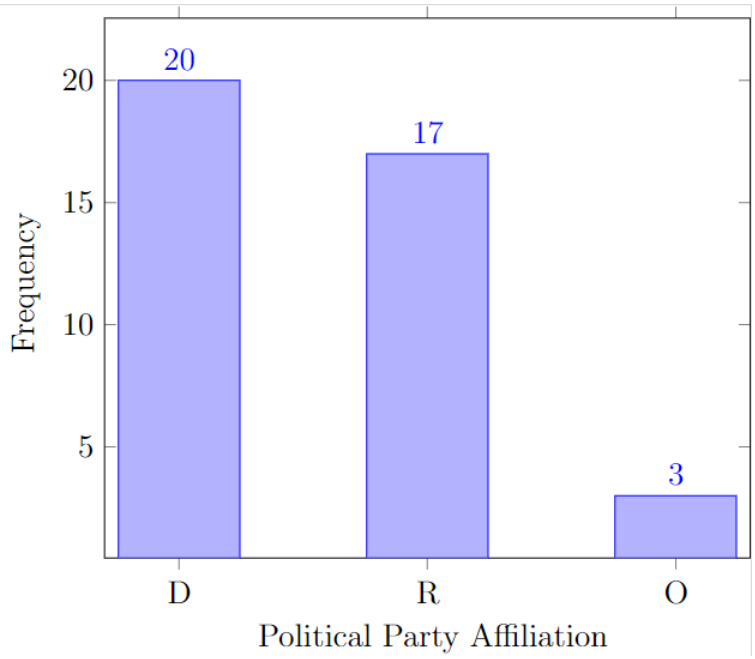


Figure 1.2.1.1.4: A Pareto chart showing political party affiliation of 40 students of which there are 20 democrats, 17 republicans, and 3 others. (Copyright; author via source)

We discussed the ways to visualize the categorical data, the two most basic ones are frequency tables and bar charts.

Section 2: Ordinal Data



Next, we will discuss how to visually summarize ordinal data. What will change when we consider the ordinal type of data? Consider the following data set obtained by recording the final grades of twenty students:

B, A, C, C, D, C, A, B, D, F, B, A, B, C, C, D, B, F, B, F

Can we treat this data set the same way we treated the categorical data? The answer is yes. We can list all possible values, do the tally, find the frequencies, and the total and then compute the relative frequencies.

Final grade	Tally	**Frequency**	**Relative Frequency**
A	|||	3	3/20=0.15=15%
B	||||| |	6	6/20=0.3=30%
C	|||||	5	5/20=0.25=25%
D	|||	3	3/20=0.15=15%
F	|||	3	3/20=0.15=15%
Total:		20	20/20=1=100%

So we can construct a relative frequency distribution table for ordinal data the same way we do it for categorical data! Can we construct the bar chart for the ordinal data the same way we did it earlier for categorical data? The answer is yes!

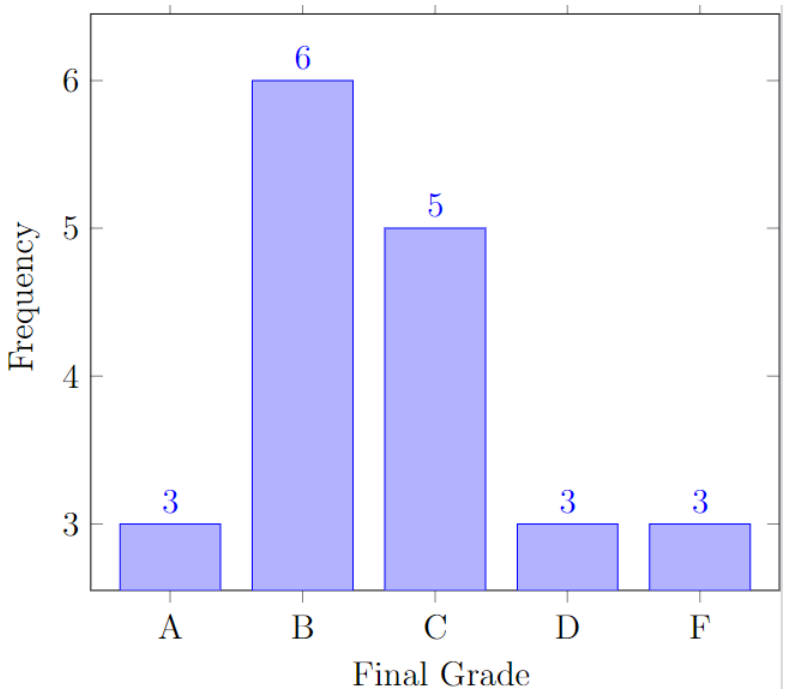


Figure 1.2.1.2.1 : A frequency bar plot showing final grades of 20 students of which there are 3 As, 6 Bs, 5 Cs, 3 Ds, and 3 Fs. (Copyright; author via source)

The same applies for the relative frequency bar chart!

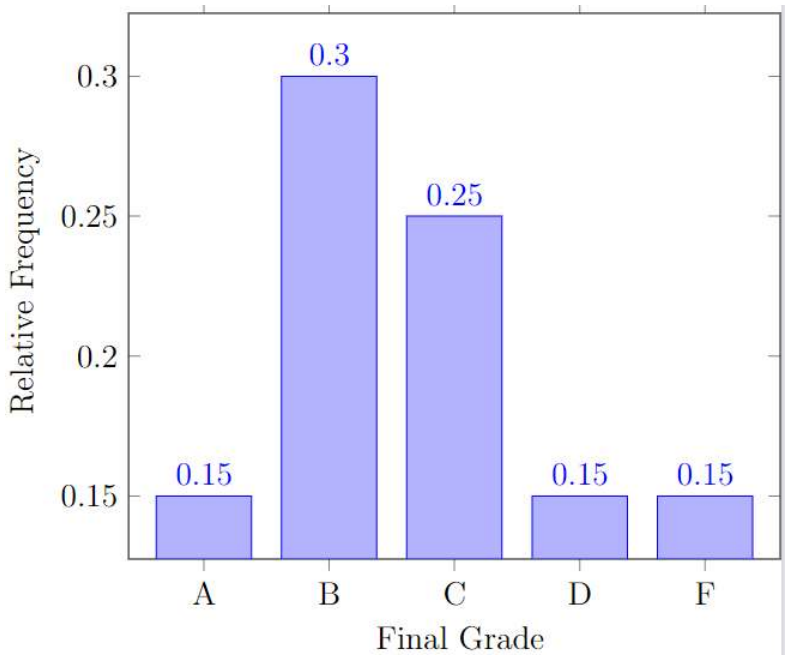


Figure 1.2.1.2.2 : A relative frequency bar plot showing final grades of 20 students of which there are 15% As, 30% Bs, 25% Cs, 15% Ds, and 15% Fs. (Copyright; author via source)

The only difference between the ways we treat ordinal and categorical data is that the ordinal data has the natural order which we would like to preserve and maintain across all visual summaries. So, while the following pie chart and Pareto chart below still make sense, we wouldn't normally use them unless we really have to.

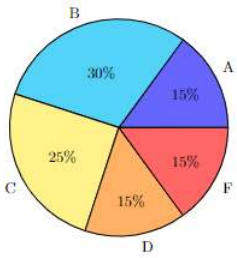


Figure 1.2.1.2.3: A relative frequency pie chart showing final grades of 20 students of which there are 15% As, 30% Bs, 25% Cs, 15% Ds, and 15% Fs.

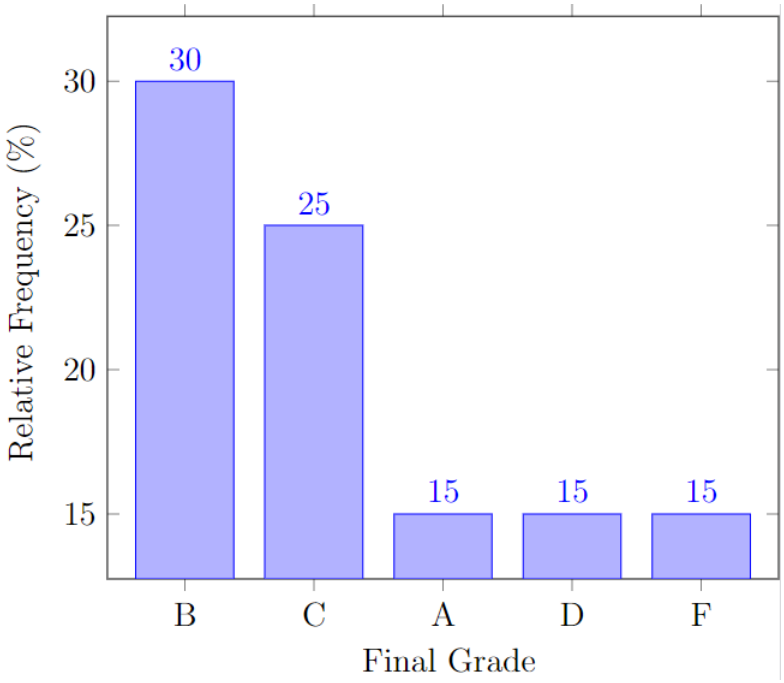


Figure 1.2.1.2.4: A Pareto chart showing final grades of 20 students of which there are 15% As, 30% Bs, 25% Cs, 15% Ds, and 15% Fs.

We discussed the most basic visual summaries of the most basic type of data.

This page titled 1.2.1: Visual Summaries of Qualitative Data is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.2.2: Visual Summaries of Quantitative Data

Section 1: Small Discrete Data



Next, we are going to learn how to organize and summarize quantitative data. What will change when we consider the small discrete type of data? Consider the following data set obtained from asking twenty students how many hours they worked per day:

5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3

Note that this discrete data is classified as "small" discrete as rarely a part-time employee works more than 8 hours per day. Can we treat this data set the same way we treated qualitative data sets previously? The answer is yes! We can list all possible values, do the tally, find the frequencies and the total, and then compute the relative frequencies.

DATA VALUE	TALLY	FREQUENCY	RELATIVE FREQUENCY
2	|||	3	3/20=0.15=15%
3	|||||	5	5/20=0.25=25%
4	|||	3	3/20=0.15=15%
5	||||| |	6	6/20=0.30=30%
6	||	2	2/20=0.10=10%
7	|	1	1/20=0.05=5%
Total:		20	20/20=1=100%

We can construct a relative frequency distribution table for small discrete data the same way we did earlier for qualitative data! Can we construct the frequency bar chart for "small' discrete data the same way we did it earlier for ordinal data? The answer is yes! And the relative frequency bar chart!

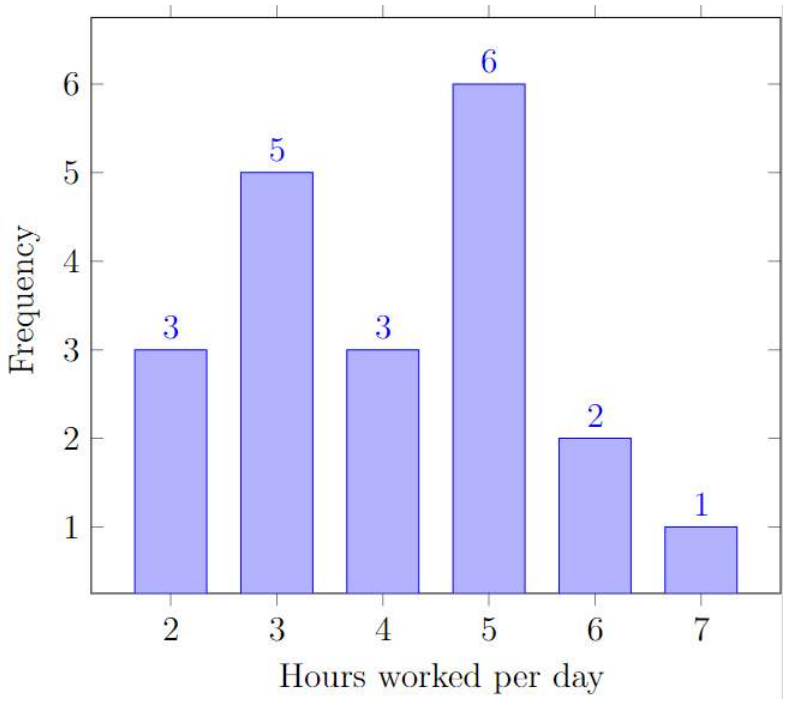


Figure 1.2.2.1.1 : A "frequency bar plot" showing hours worked per day of 20 students of which there are 3 students work 2 hours, 5 work 2, 3 work 4, 6 work 5, 2 work 6, 1 work 7.

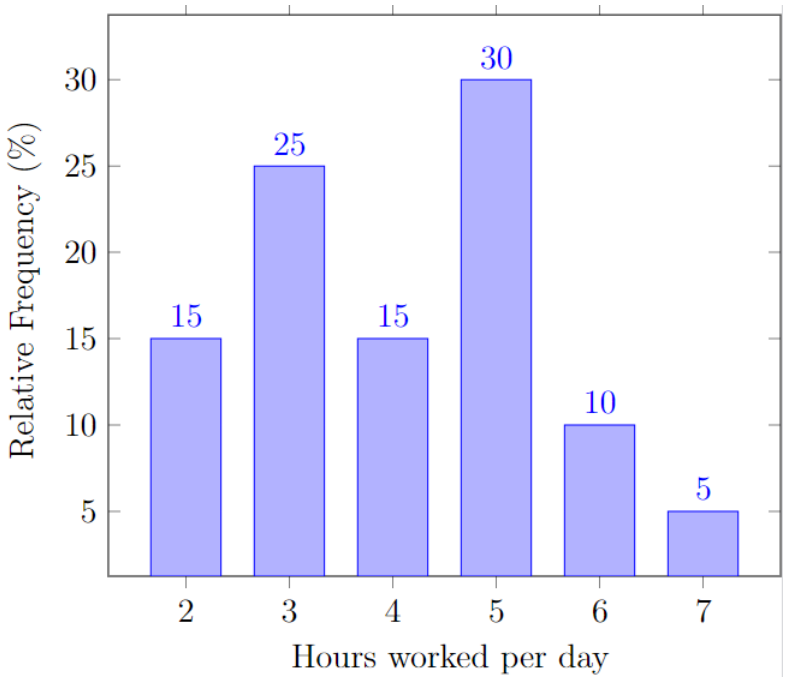


Figure 1.2.2.1.2 : A "relative frequency bar plot" showing hours worked per day of 20 students of which there are 15% students work 2 hours, 25% work 2, 15% work 4, 30% work 5, 10% work 6, 5% work 7.

However, there are a few things that we are going to do differently this time. Since quantitative (unlike qualitative) data is associated with the real number line we will draw the bars touching each other so that no number on the number line is left uncovered unless the corresponding frequency is zero. To highlight the differences between the quantitative and qualitative data, we are going to call this chart **a histogram** instead of a bar chart. As expected, a histogram that uses frequencies on the vertical axis is called a **frequency histogram**. Similarly, a histogram that uses relative frequencies on the vertical axis is called a **relative frequency histogram**.

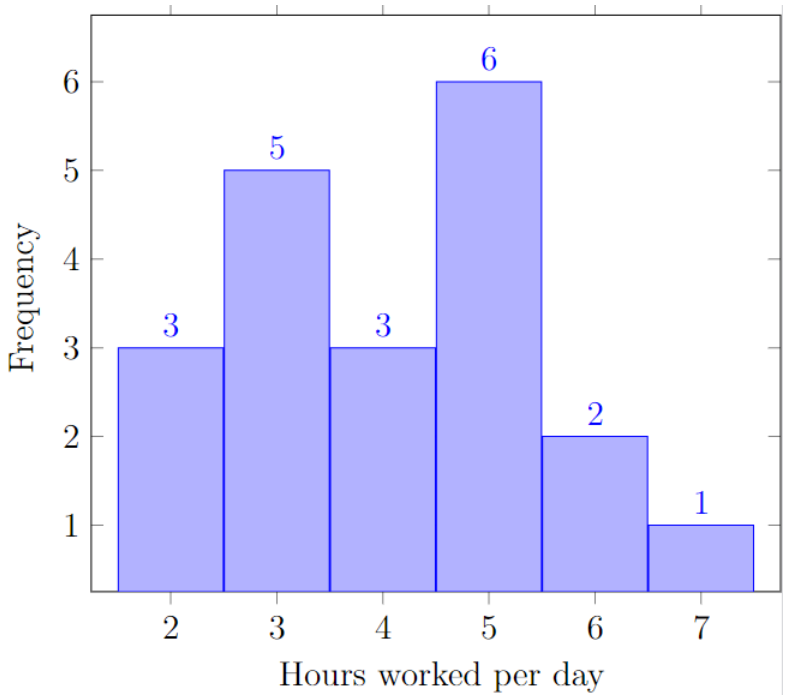


Figure 1.2.2.1.3 : A frequency histogram showing hours worked per day of 20 students of which there are 3 students work 2 hours, 5 work 2, 3 work 4, 6 work 5, 2 work 6, 1 work 7.

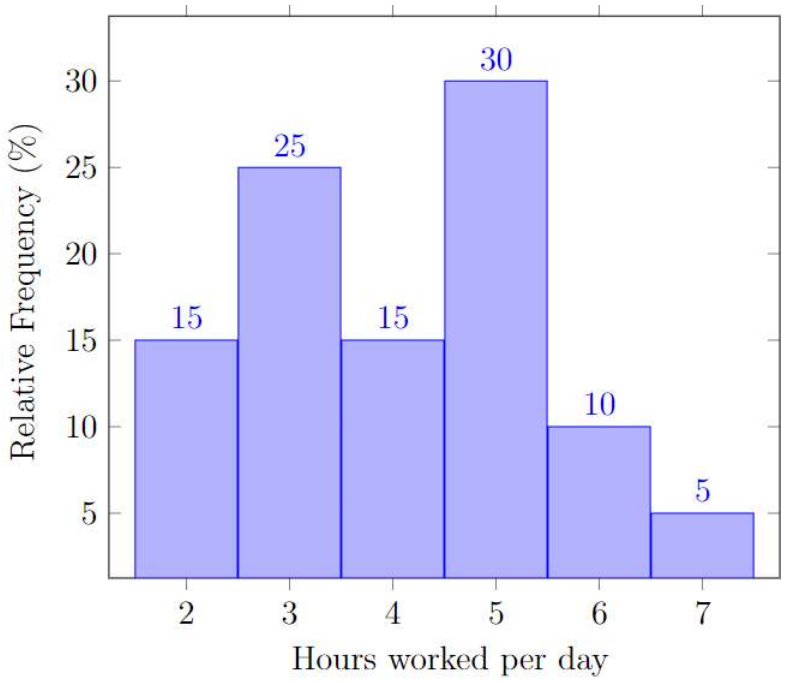


Figure 1.2.2.1.4 : A relative frequency histogram showing hours worked per day of 20 students of which there are 15% students work 2 hours, 25% work 2, 15% work 4, 30% work 5, 10% work 6, 5% work 7.

We discussed how to visually summarize small discrete data. In short, we can treat "small" discrete data the same way as qualitative data except the fact that the bar charts are now called histograms and the bars must have no gaps in between unless the corresponding frequency is zero.

Section 2: Large Discrete Data



Next, we are going to learn how to visualize large discrete data. Consider the following data set obtained by recording the presidential ages at inauguration:

President	Age	President	Age	President	Age
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	62
Monroe	58	Garfield	49	Kennedy	43
J. Q. Adams	57	Arthur	51	L. Johnson	55
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G.H.W. Bush	64
Taylor	64	Taft	51	Clinton	47
Fillmore	50	Wilson	56	G. W. Bush	54
Pierce	48	Harding	55	Obama	47
Buchanan	65	Coolidge	51	Trump	70

Will we run into any issues if you try treat "large" discrete data same way as if it was small? If we do that, then first we would end up with the following frequency table:

Frequency Table for the ages of US presidents.

President's age at inauguration	Frequency
42	1

43	1
46	1
47	3
48	1
49	2
50	1
51	5
52	2
54	5
55	4
56	3
57	4
58	1
60	1
61	3
62	1
64	2
65	1
68	1
69	1
70	1
Total:	45

Then with the following frequency histogram:

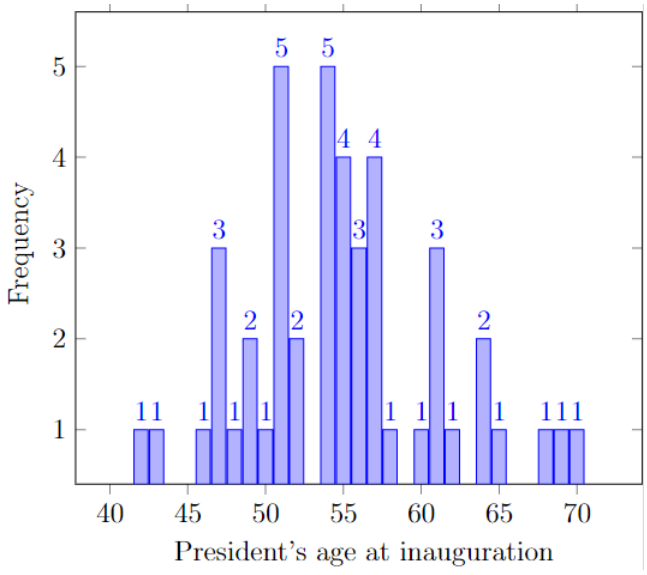


Figure 1.2.2.2.1 : A "frequency bar plot" showing ages at inauguration of 45 US presidents. (Copyright; author via source)

The problem with such a summary is that when more than half of the frequencies are 0s and 1s the summary isn't very informative! To deal with this issue, we first group the observations into classes (also known as categories or bins) and then treat each class as a distinct value. Each class is defined by a range of values from the lower-class limit up to but not including the upper-class limit. We define **the lower-class limit** (LCL) as the smallest value that could go in a class and the **upper-class limit** (UCL) as the lower-class limit of the next higher class. We define the **class midpoint** (CM) as the average of the lower-limit of a class and the upper-limit of the class. We define the **class width** (CW) is the difference between the lower limit of a class and the upper limit of the class.

Note that because of the relations between the quantities, we only need two of the four of them to define the entire class structure! For example, if LCL = 15 and CW = 10 then the classes are 15-25, 25-35, 35-45, 45-55 etc.

Also note that each value can only belong to one class! For example, 35 will belong to 35-45 and not to 25-35.

The sample size determines the number of classes in the following way:

Number of observations	Number of classes
25 or fewer	5–6
25–50	7–14
Over 50	15–20

We are going to use the following guideline for choosing the classes:

1. Decide on the (approximate) number of classes.
2. Calculate an approximate class width as $\frac{\text{Maximum observation - Minimum Observation}}{\text{Number of classes}}$ and use the result to decide on a convenient class width.
3. Choose a number for the lower limit of the first class, noting that it must be less than or equal to the minimum observation.
4. Obtain the other lower-class limits by successively adding the class width chosen in Step 2.
5. Use the results of Step 4 to specify all the classes.

Let’s take another look at our data before we decide to split it into classes. Consider the following data set obtained by recording the presidential ages at inauguration:

57 61 57 57 58 57 61 54 68 51 49 64 50 48 65

52 56 46 54 49 51 47 55 55 54 42 51 56 55 51

54 51 60 62 43 55 56 61 52 69 64 47 54 47 70

The step-by-step procedure for choosing the classes:

1. Decide on the (approximate) number of classes:

$$\sim (7-14)$$

2. Calculate an approximate class width:

$$\frac{70-42}{\sim(7-14)} = \frac{28}{7} = 4 \approx 5 \text{ (b/c 5 is a "nicer" number than 4)}$$

3. Choose a number for the lower limit of the first class as 40.
4. Obtain the other lower-class limits:

40, 45, 50, 55, 60, 65, 70

5. Use the results of Step 4 to specify all the classes:

40-45, 45-50, 50-55, 55-60, 60-65, 65-70, 70-75

Once we identified the classes, we can start the tally by assigning each value to one of the classes, count the frequency with the totals, and relative frequencies.

Classes	Tally	Frequency	Relative Frequency
40 to 45	||	2	2/45=0.044=4.4%

Classes	Tally	Frequency	Relative Frequency
45 to 50	||||| ||	7	7/45=0.156=15.6%
50 to 55	||||| ||||| |||	13	13/45=0.289=28.9%
55 to 60	||||| ||||| ||	12	12/45=0.267=26.7%
60 to 65	||||| ||	7	7/45=0.156=15.6%
65 to 70	|||	3	3/45=0.067=6.7%
70 to 75			1	1/45=0.022=2.2%
Total:	45	45	45/45=1=100%

Basically, once the classes are created, we can construct a relative frequency distribution table for large discrete data the same way we did earlier for small discrete data!

Once we have a complete frequency table, we can now construct a frequency histogram:

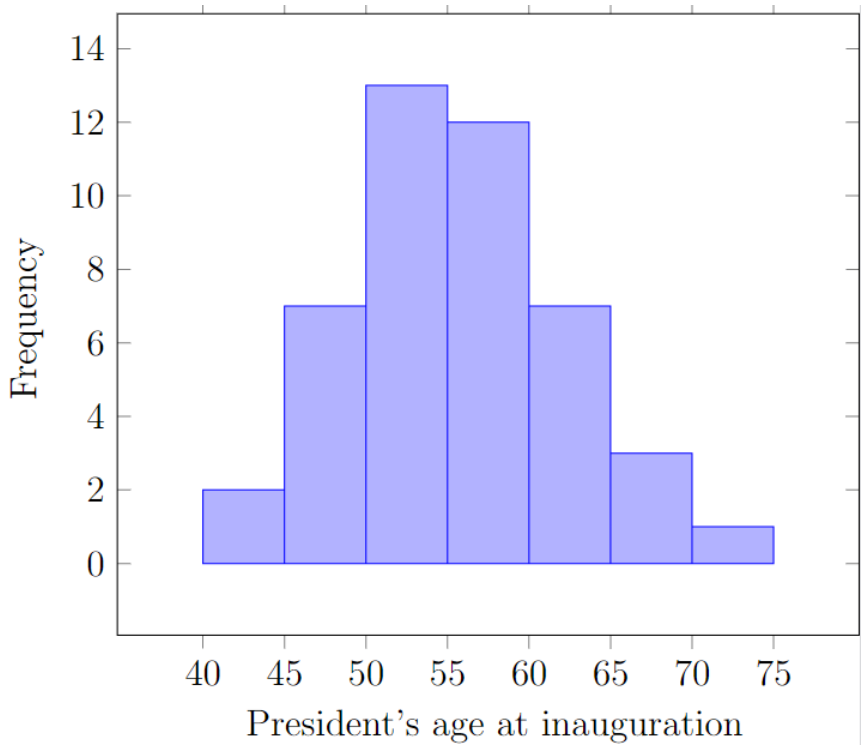


Figure 1.2.2.2.2 : A frequency histogram showing ages at inauguration of 45 US presidents of which there are 2 presidents between 40 and 45, 7 between 45-50, 13 between 50-55, 12 between 55-60, 7 between 60-65, 3 between 65-70, 1 between 70-75. (Copyright; author via source)

Similarly, we can obtain the relative frequency histogram!

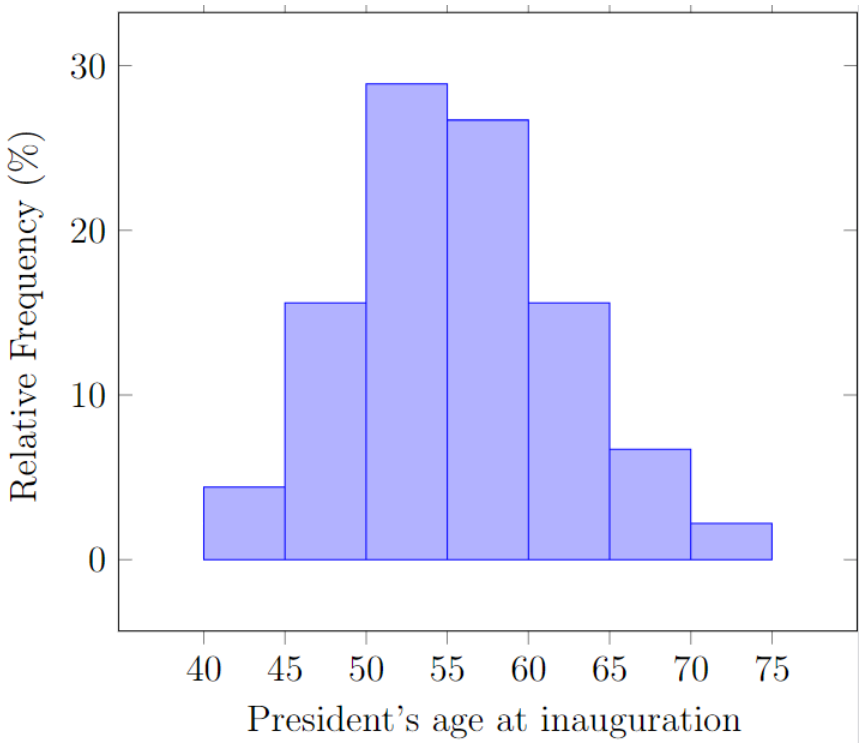


Figure 1.2.2.2.3: A relative frequency histogram showing ages at inauguration of 45 US presidents of which there are 4.4% presidents between 40 and 45, 15.6% between 45-50, 28.9% between 50-55, 26.7% between 55-60, 15.6% between 60-65, 6.7% between 65-70, 2.2% between 70-75. (Copyright; author via source)

Once the idea of grouping is understood the process of constructing the (relative) frequency table and the histogram is very intuitive!

In summary, to construct a frequency table for discrete data:

1. List the classes in the first column of a table.
2. For each observation, place a tally mark in the second column of the table in the row of the appropriate class.
3. Count the tallies for each class and record the totals in the third column of the table.

In summary, to construct a (relative) histogram for discrete data:

1. Obtain a frequency (relative frequency) distribution of the data.
2. Draw a horizontal axis on which to place the bars and a vertical axis on which to display the frequencies (relative frequencies).
3. For each class, construct a vertical bar whose height equals the frequency (relative frequency) of that class.
4. Label the bars with the classes, the horizontal axis with the name of the variable, and the vertical axis with “Frequency” (“Relative frequency” or “Percent”).
    - For single-value grouping, we use the distinct values of the observations to label the bars, with each such value centered under its bar.
    - For class grouping, we use the midpoints to label the bars. Note: Some statisticians and technology use class limits to label the bars.

While it may appear that we used different approaches to organize "small" and "large" discrete data, they both can be described by the term grouping! **Single value grouping** for "small" discrete data and **interval grouping** for "large" discrete data. The two methods have more in common than it appears at first sight because the single-value grouping can be viewed as an interval grouping with each single-value X defining a class of width 1 and X in the middle!

We discussed how to organize large discrete data and concluded that essentially it is the same way as we organize qualitative and small discrete data with addition of the step called binning or creating classes.

Section 3: Continuous Data



Next, we will discuss how to produce a visual summary of continuous data. Consider the following data set obtained from recording the heights (in inches) of 100 semiprofessional soccer players:

66.18 64.49 72.24 64.70 65.80 67.55 64.15 65.42 70.62 65.85

67.74 69.11 68.26 65.98 69.45 65.93 67.55 73.63 70.55 64.53

67.03 67.48 66.19 61.69 69.30 61.57 66.01 68.62 70.72 65.81

64.87 68.30 63.71 63.66 66.36 64.68 62.49 67.19 72.99 64.96

68.12 59.83 67.97 67.33 64.98 66.09 65.56 67.62 67.66 63.07

72.22 68.00 69.43 65.15 63.27 63.23 64.13 69.45 63.10 65.40

68.92 67.57 64.21 68.36 68.88 66.39 68.28 67.27 68.75 67.59

63.67 70.50 67.52 64.06 73.95 65.36 67.62 65.06 67.30 68.42

66.08 65.91 64.82 69.23 68.40 68.29 73.27 69.35 69.80 68.42

67.32 73.00 69.58 64.66 68.59 62.77 67.29 66.56 68.35 65.08

Does continuous dataset look more like a small discrete data or large? It certainly looks like we are going to need some interval grouping here!

The step-by-step procedure for choosing the classes:

1. Decide on the (approximate) number of classes:

$$\sim (15-20)$$

2. Calculate an approximate class width:

$$\frac{73.95-59.83}{\sim(15-20)} = \frac{\sim 15}{15} \approx 1$$

3. Choose a number for the lower limit of the first class as 59.
4. Obtain the other lower-class limits:

59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74

5. Use the results of Step 4 to specify all the classes:

59-60,60-61,61-62,62-63,63-64,64-65,65-66,66-67

67-68,68-69,69-70,70-71,71-72,72-73,73-74

Now every entry can be assigned to a single class and treat it as qualitative data!

Classes	Tally	Frequency	Relative Frequency
59 to 60	|	1	0.01
60 to 61		0	0.00
61 to 62	||	2	0.02
62 to 63	||	2	0.02
63 to 64	||||| ||	7	0.07
64 to 65	||||| ||||| |||	13	0.13
65 to 66	||||| ||||| |||	13	0.13
66 to 67	||||| |||	8	0.08
67 to 68	||||| ||||| ||||| |||	18	0.18
68 to 69	||||| ||||| ||||| |	16	0.16
69 to 70	||||| ||||	9	0.09
70 to 71	||||	4	0.04
72 to 73	|||	3	0.03
73 to 74	||||	4	0.04
Total:		100	1

Now we can turn the frequency distribution table into the frequency histogram:

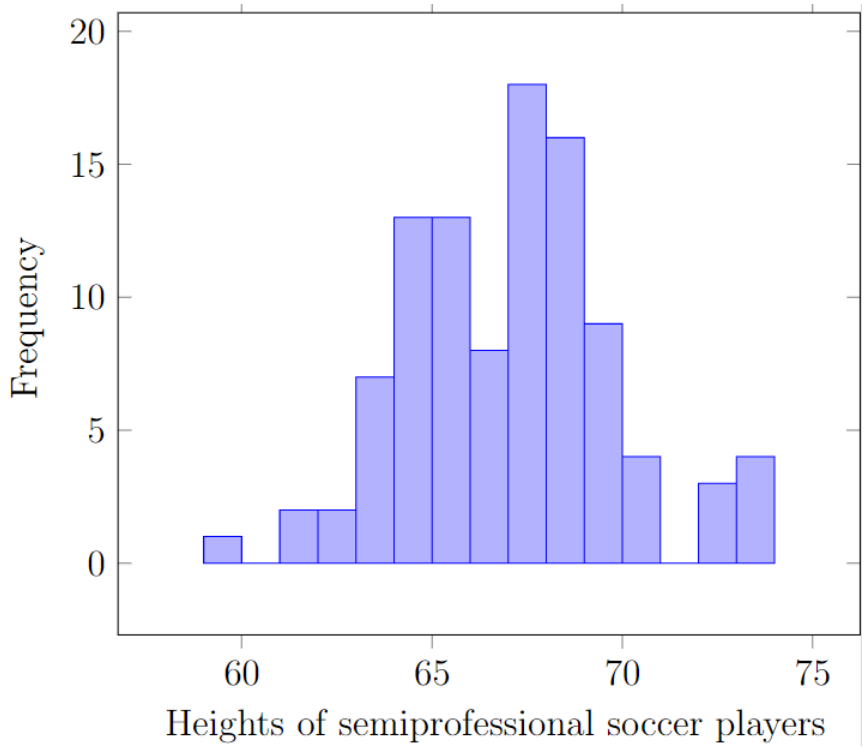


Figure 1.2.2.3.1 : A frequency histogram showing heights of 100 semiprofessional soccer players (in inches) of which there is 1 player between 59 and 60, 2 between 61-62, 2 between 62-63, 7 between 63-64, 13 between 64-65, 13 between 65-66, 8 between 66-67, 18 between 67-68, 16 between 68-69, 9 between 69-70, 4 between 70-71, 3 between 72-73, 4 between 73-74. (Copyright; author via source)

And the relative frequency histogram:

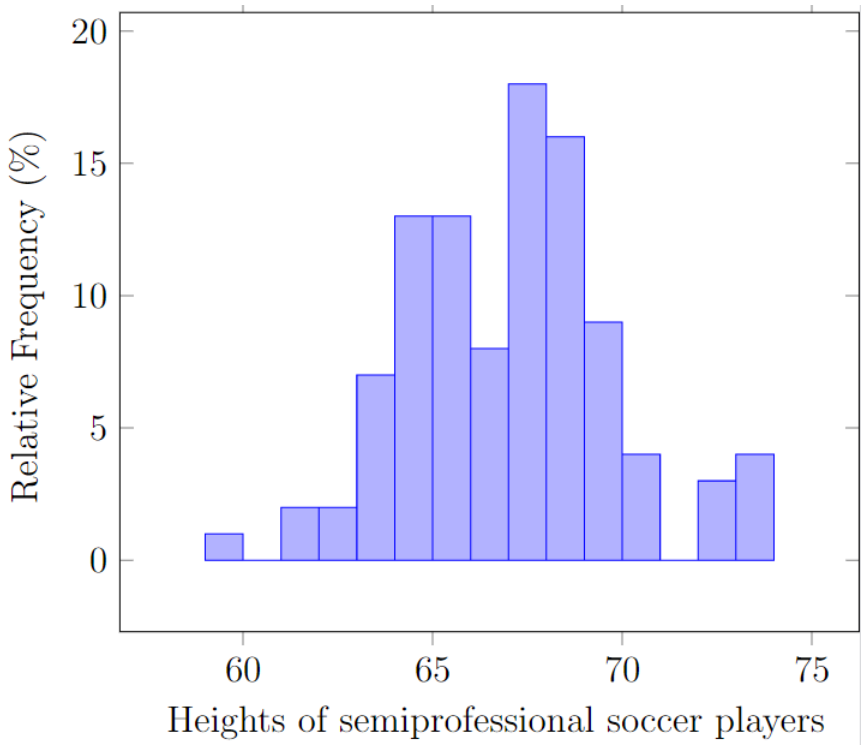


Figure 1.2.2.3.2 : A relative frequency histogram showing heights of 100 semiprofessional soccer players (in inches) of which there is 1% between 59 and 60, 2% between 61-62, 2% between 62-63, 7% between 63-64, 13% between 64-65, 13% between 65-66, 8% between 66-67, 18% between 67-68, 16% between 68-69, 9% between 69-70, 4% between 70-71, 3% between 72-73, 4% between 73-74. (Copyright; author via source)

We discussed how to summarize continuous data! in short, we use interval grouping and treat it similarly to large discrete data!

This page titled 1.2.2: Visual Summaries of Quantitative Data is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.2.3: Other Charts



Frequency and relative frequency distribution tables along with (relative) frequency bar plots and histograms are considered the basic visual summaries of data as they can be constructed for any type of data and essentially provide all the information one may need to know about the given data. However, statisticians continue to invent ways to display data. Next, we will discuss some other less common ways used to visually summarize the data.

One method, developed in the 1960s by the late Professor John Tukey of Princeton University, is called a **stem-and-leaf diagram**, or **stemplot**. Let's consider again the ages of the presidents at the time of their inauguration. We are going to construct the stem-and-leaf diagram by first constructing a **stem** - the list the first digits of all possible ages and then drawing the **leaves** one for each president's age.

Stem	Leaf (Key: 1|2 = 12)
4	9 8 6 9 7 2 3 7 7
5	7 7 7 8 7 4 1 0 2 6 4 1 5 5 4 1 6 5 1 4 1 5 6 2 4
6	1 1 8 4 5 0 2 1 9 4
7	0

Figure 1.2.3.1: Stem-and-leaf diagram for the ages of US presidents. (Copyright; author via source)

Once the diagram is complete, we can sort the values in each branch in increasing order:

Stem	Leaf (Key: 1|2 = 12)
4	2 3 6 7 7 7 8 9 9
5	0 1 1 1 1 1 2 2 4 4 4 4 4 5 5 5 5 6 6 6 7 7 7 8
6	0 1 1 1 2 4 4 5 8 9
7	0

Figure 1.2.3.2: Stem-and-leaf diagram for the ages of US presidents with leaves in ascending order. (Copyright; author via source)

When the branches are too long, we can split the branches by assigning a portion of the branch in this case values from 0-4 to one half and the values from 5-9 to the other half:

Stem	Leaf (Key: 1|2 = 12)
4	2 3
4	6 7 7 7 8 9 9
5	0 1 1 1 1 1 2 2 4 4 4 4 4
5	5 5 5 5 6 6 6 7 7 7 8
6	0 1 1 1 2 4 4
6	5 8 9
7	0

Figure 1.2.3.3: Stem-and-leaf diagram for the ages of US presidents with split stem and leaves in ascending order. (Copyright; author via source)

To construct a stem-and-leaf diagram:

1. Think of each observation as a stem—consisting of all but the rightmost digit—and a leaf, the rightmost digit.
2. Write the stems from smallest to largest in a vertical column to the left of a vertical rule.
3. Write each leaf to the right of the vertical rule in the row that contains the appropriate stem.
4. (Optional) Arrange the leaves in each row in ascending order.

5. (Optional) Split the stems.

This ingenious diagram is often easier to construct than either a frequency distribution or a histogram and generally displays more information.

Here are a few tips to consider before making a stem-and-leaf diagram also known as a stemplot:

- Stemplots do not work well for large data sets, where each stem must hold a large number of leaves.
- There is no magic number of stems to use, but five is a good minimum. Too few or too many stems will make it difficult to see the distribution's shape.
- If you split stems, be sure that each stem is assigned an equal number of possible leaf digits (two stems, each with five possible leaves; or five stems, each with two possible leaves).
- You can get more flexibility by rounding the data so that the final digit after rounding is suitable as a leaf. Do this when the data has too many digits. For example, in reporting teachers' salaries, using all five digits (for example, $42,549) would be unreasonable. It would be better to round to the nearest thousand and use 4 as a stem and 3 as a leaf.

Alternatively, we can use another type of graphical display for quantitative data called the **dot plot**.

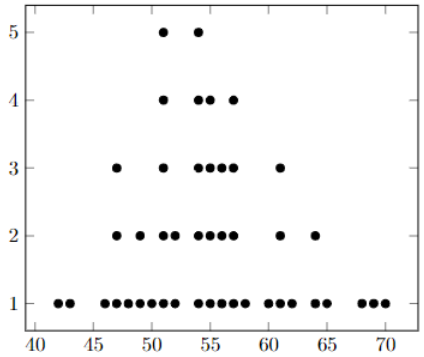


Figure 1.2.3.4: Dot plot for the ages of US presidents. (Copyright; author via source)

To Construct a Dot Plot:

1. Draw a horizontal axis that displays the possible values of the quantitative data.
2. Record each observation by placing a dot over the appropriate value on the horizontal axis.
3. Label the horizontal axis with the name of the variable.

Dot plots are particularly useful for showing the relative positions of the data in a data set or for comparing two or more data sets.

Why bother with learning different ways to summarize data?

- The advantage of dot plots and stem-and-leaf plots is that both of them can be constructed as the data being collected, that is we do not need to have access to the entire data set to start visualizing it.
- To compare two populations side-by-side we can use a back-to-back stem-and-leaf plot by drawing a stem and adding leaves for one population on one side and for the other population on the other side. We can also use the dot plot in the same way by drawing the dots for one population above the x-axis and for the other population below the x-axis.

As an alternative to drawing the vertical bars in histograms we can mark the frequencies with a point and then connect the points with lines.

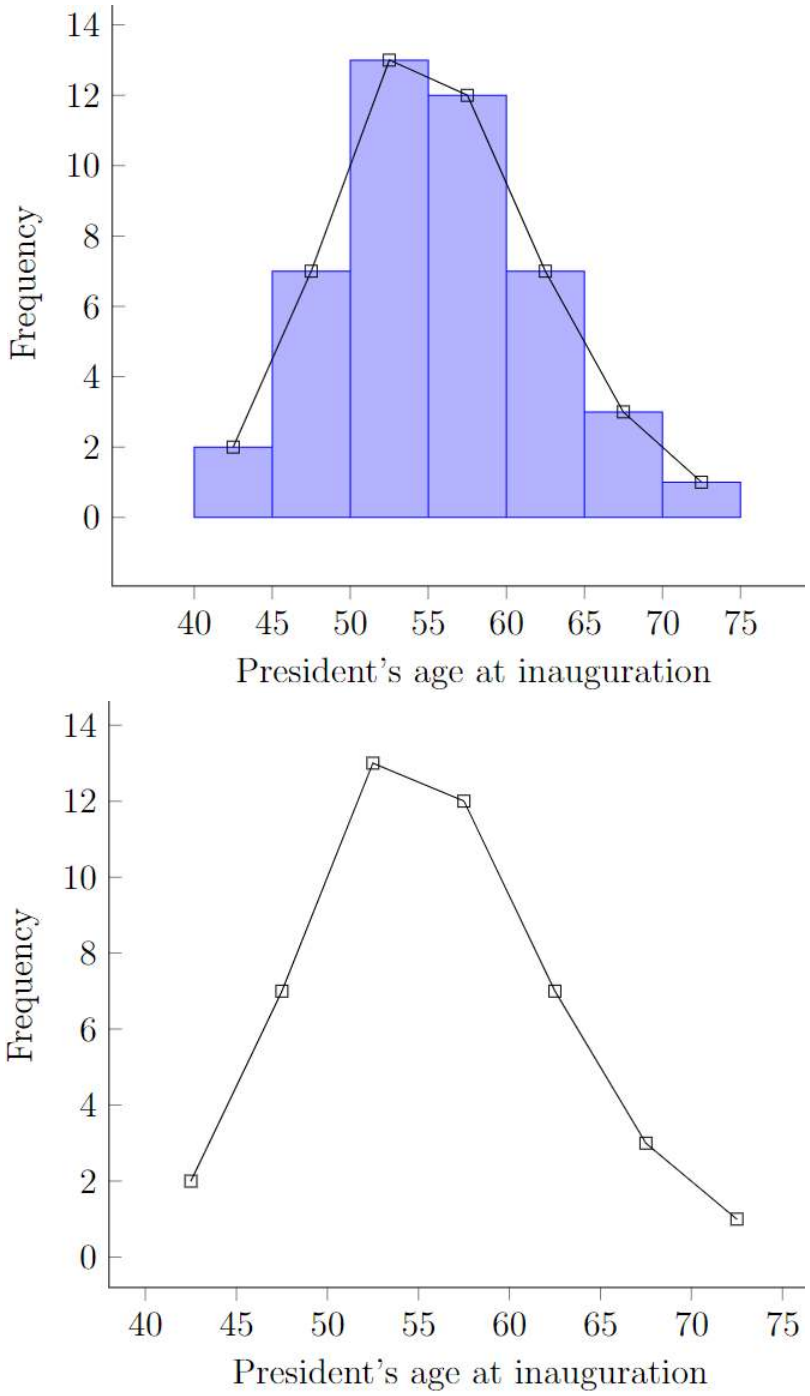


Figure 1.2.3.5: Frequency Polygon Superimposed on the Frequency Histogram (on the left) and Frequency Polygon by itself (on the right) for the ages of US presidents.

Such a summary is called **a frequency polygon**. As an alternative it can be used to save ink in your printer!

Another common way to organize the data is to use a cumulative frequency. A **cumulative frequency** of a class is obtained by summing the frequencies of all classes representing values less than the upper limit of the given class. The last entry in the cumulative frequency column must be equal to the total number of observations.

Cumulative Frequency Table for presidents' ages at inauguration

Classes	Midpoint	Frequency	Cumulative Frequency
40 to 45	42.5	2	2

Classes	Midpoint	Frequency	Cumulative Frequency
45 to 50	47.5	7	2+7=9
50 to 55	52.5	13	2+7+13=22
55 to 60	57.5	12	2+7+13+12=34
60 to 65	62.5	7	2+7+13+12+7=41
65 to 70	67.5	3	2+7+13+12+7+3=44
70 to 75	72.5	1	2+7+13+12+7+3+1=45
Total:		45	

A **cumulative frequency polygon** that is constructed using the midpoints on the horizontal axis and the cumulative frequency on the vertical axis is called **ogive**.

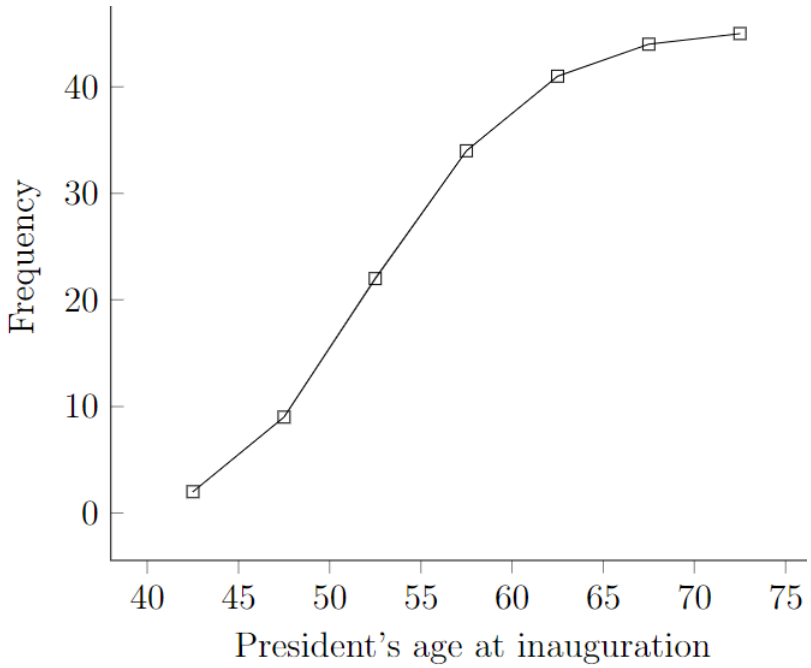


Figure 1.2.3.6: Cumulative Frequency Polygon for the ages of US presidents. (Copyright; author via source)

We can easily do the same with relative cumulative frequencies. First construct the **cumulative relative frequency table**:

Cumulative relative frequency table for the ages of US presidents.

Classes	Midpoint	Cumulative Frequency	Relative Cumulative Frequency
40 to 45	42.5	2	2/45=0.044
45 to 50	47.5	9	9/45=0.200
50 to 55	52.5	22	22/45=0.489
55 to 60	57.5	34	34/45=0.756
60 to 65	62.5	41	41/45=0.911
65 to 70	67.5	44	44/45=0.978
70 to 75	72.5	45	45/45=1.000

Then construct the cumulative relative frequency polygon, also known as **ogive**:

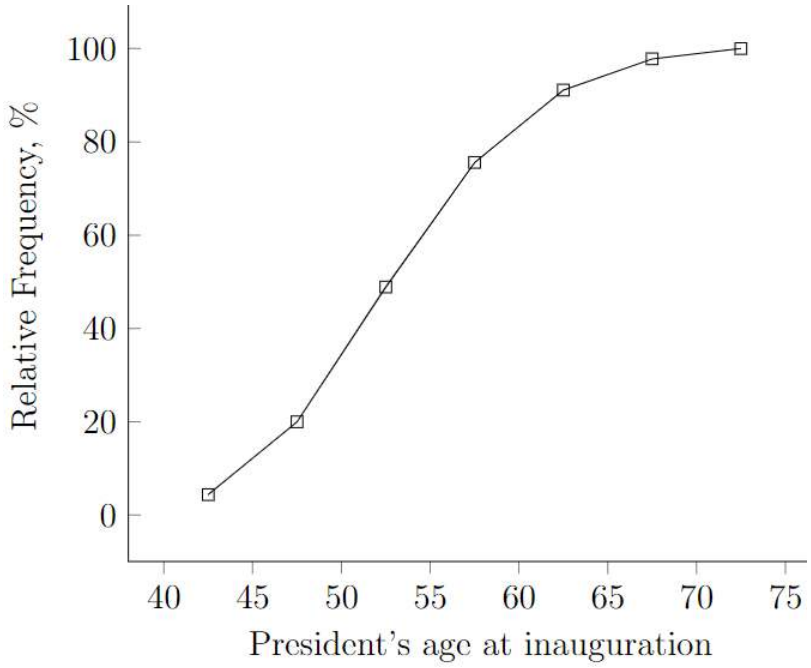


Figure 1.2.3.7: Ogive for the ages of US presidents. (Copyright; author via source)

We discussed a variety of alternatives to the most basic visual summaries – frequency tables and histograms.

This page titled 1.2.3: Other Charts is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.2.4: Distribution Shapes



Next, we will discuss the attributes of distributions that will allow us to identify the most common ones and classify the rest. Regardless of the method used to summarize the data, all visual summaries for the same data set will have one thing in common!

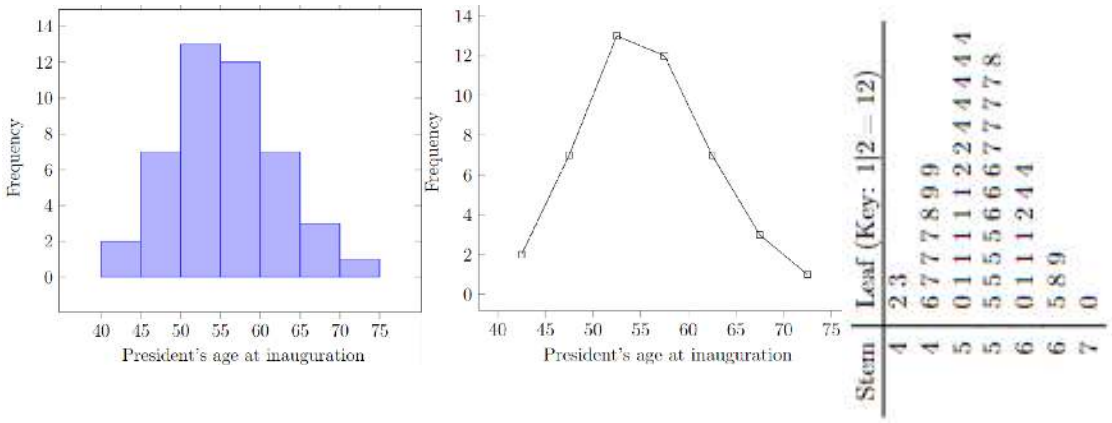
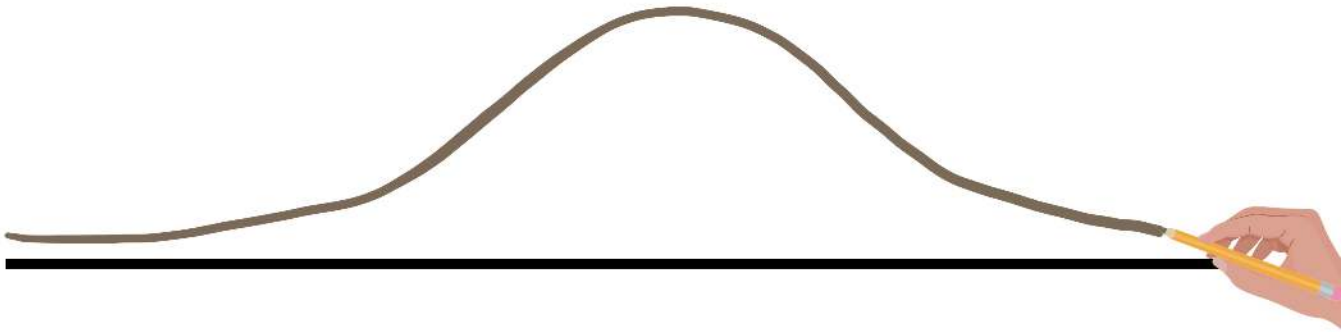
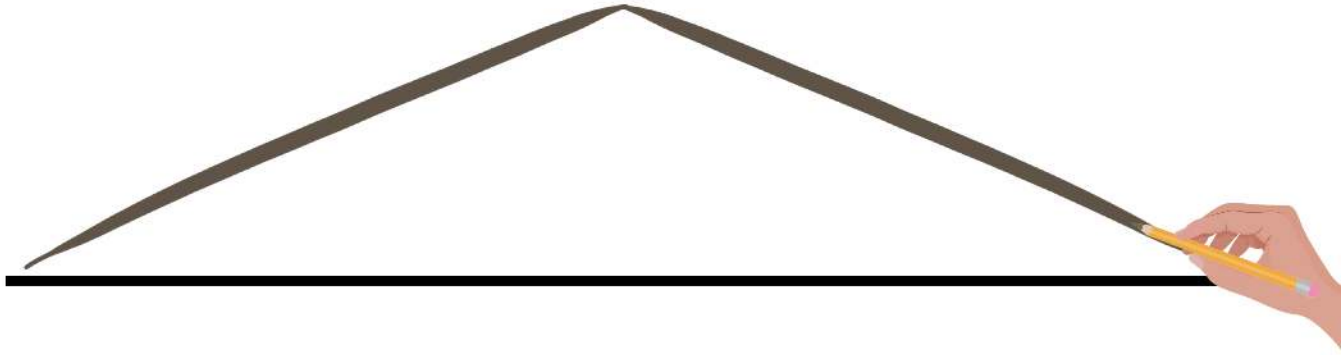


Figure 1.2.4.1: Frequency Histogram (on the left), Frequency Polygon (in the center), Stem-and-Leaf turned sideways (on the right) for the ages of US presidents.

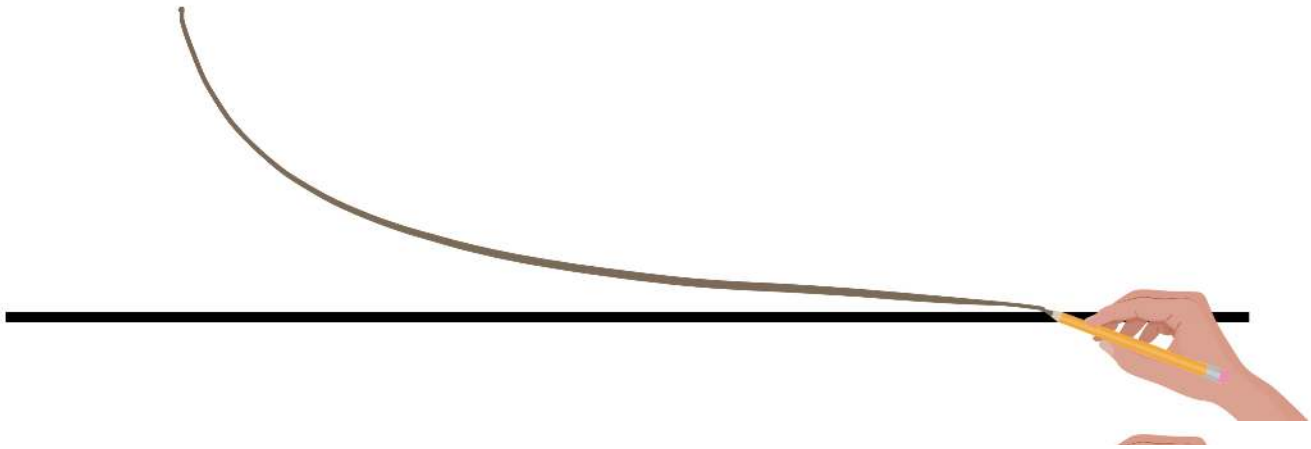
What is it? The correct answer is the shape!

One of the main characteristics of the distribution of a data set is its **shape**. Here are some examples of shapes that appear in statistics:









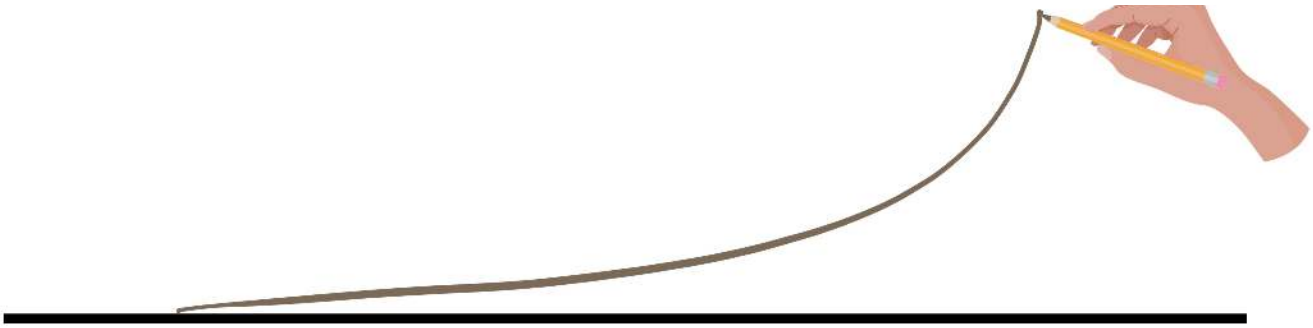


Figure 1.2.4.2: Examples of different shapes that appear in statistics. (From left to right and from top down: bell shape, triangular, uniform, reverse J-shape, J-shape)

Some shapes are more common than others. If it is not one of the common shapes, then we use the following classification.

- Based on whether there is an axis of symmetry, the shape can be classified as symmetric or asymmetric.

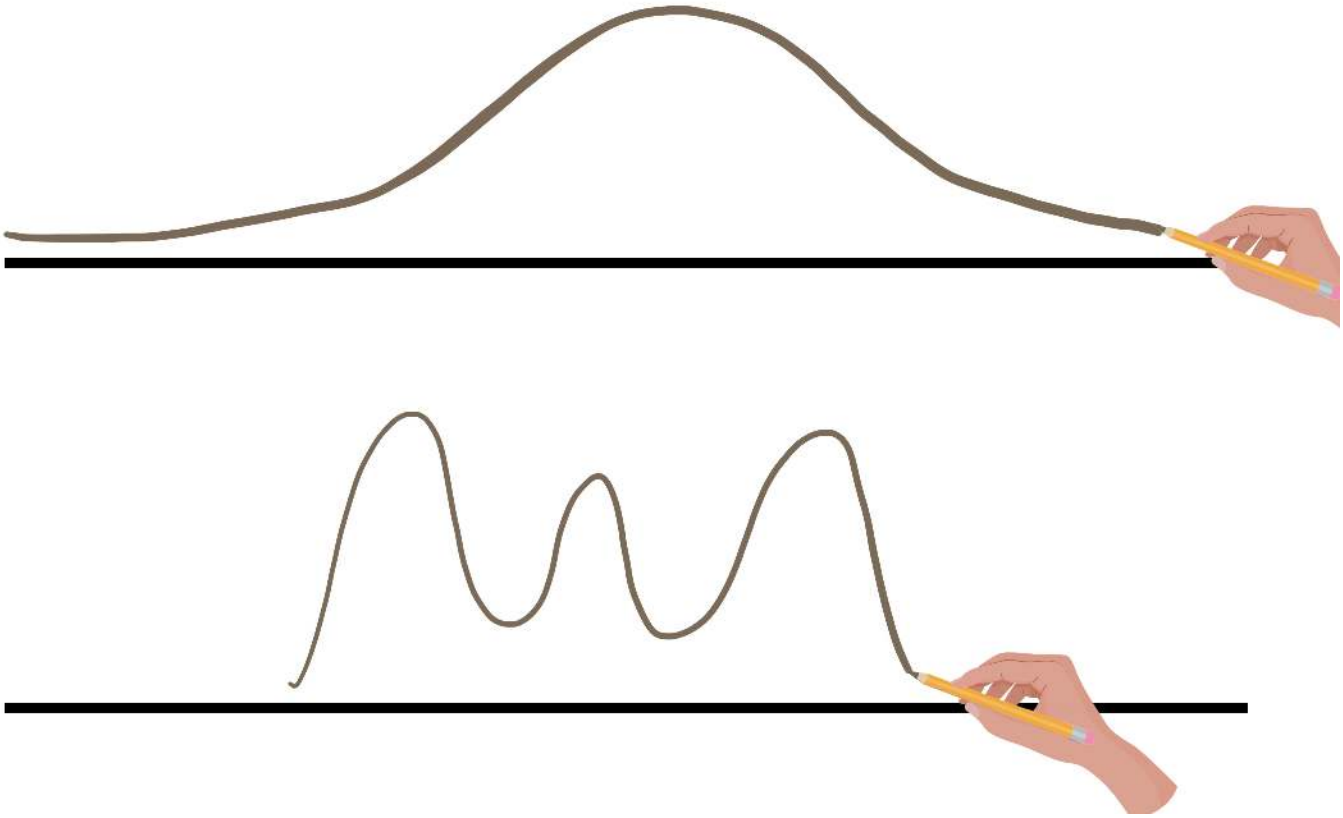


Figure 1.2.4.3: Examples of symmetric (on the left) and asymmetric (on the right) shapes.

- Some asymmetric shapes can be classified as skewed left or right.

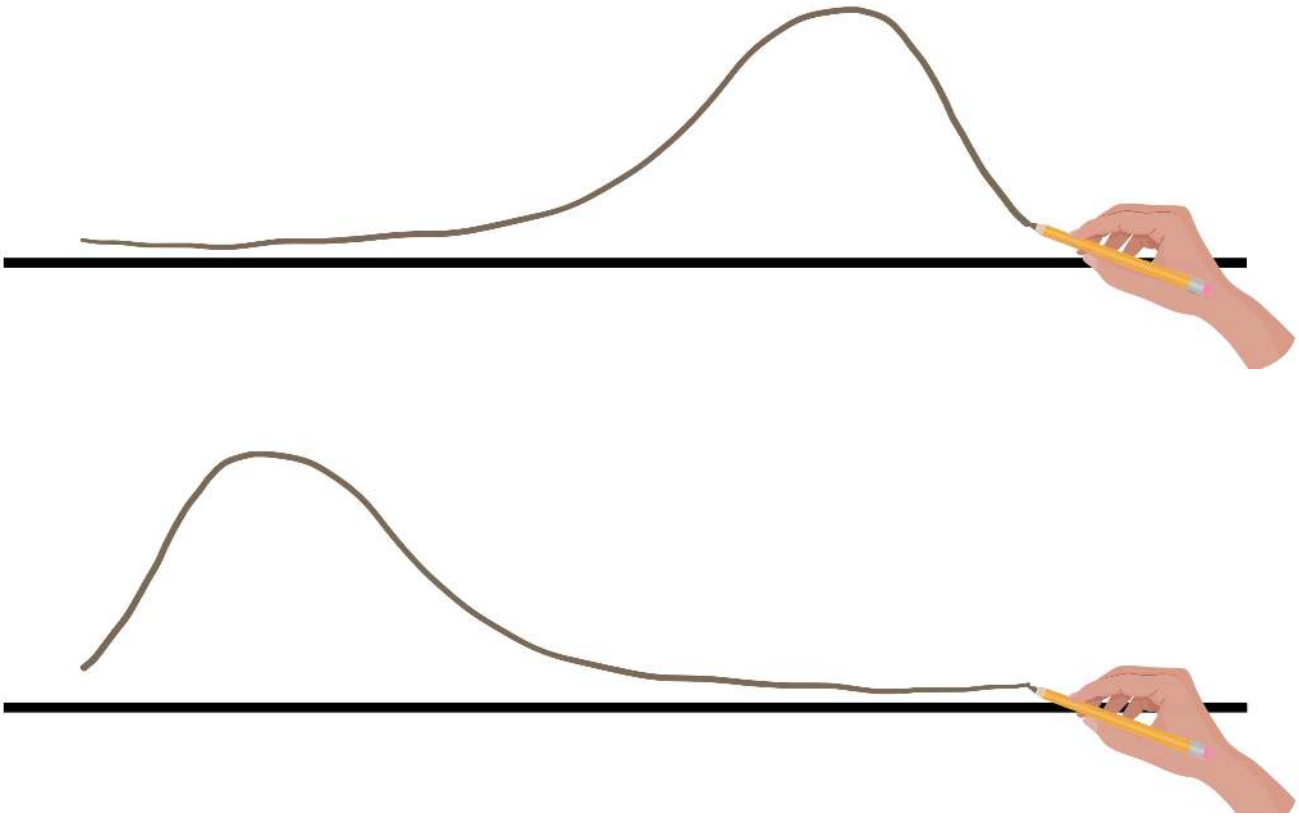


Figure 1.2.4.4: Examples of skewed left (on the left) and skewed right (on the right) shapes.

- Based on the number of peaks, the shape can be classified as unimodal, bimodal, or multimodal.

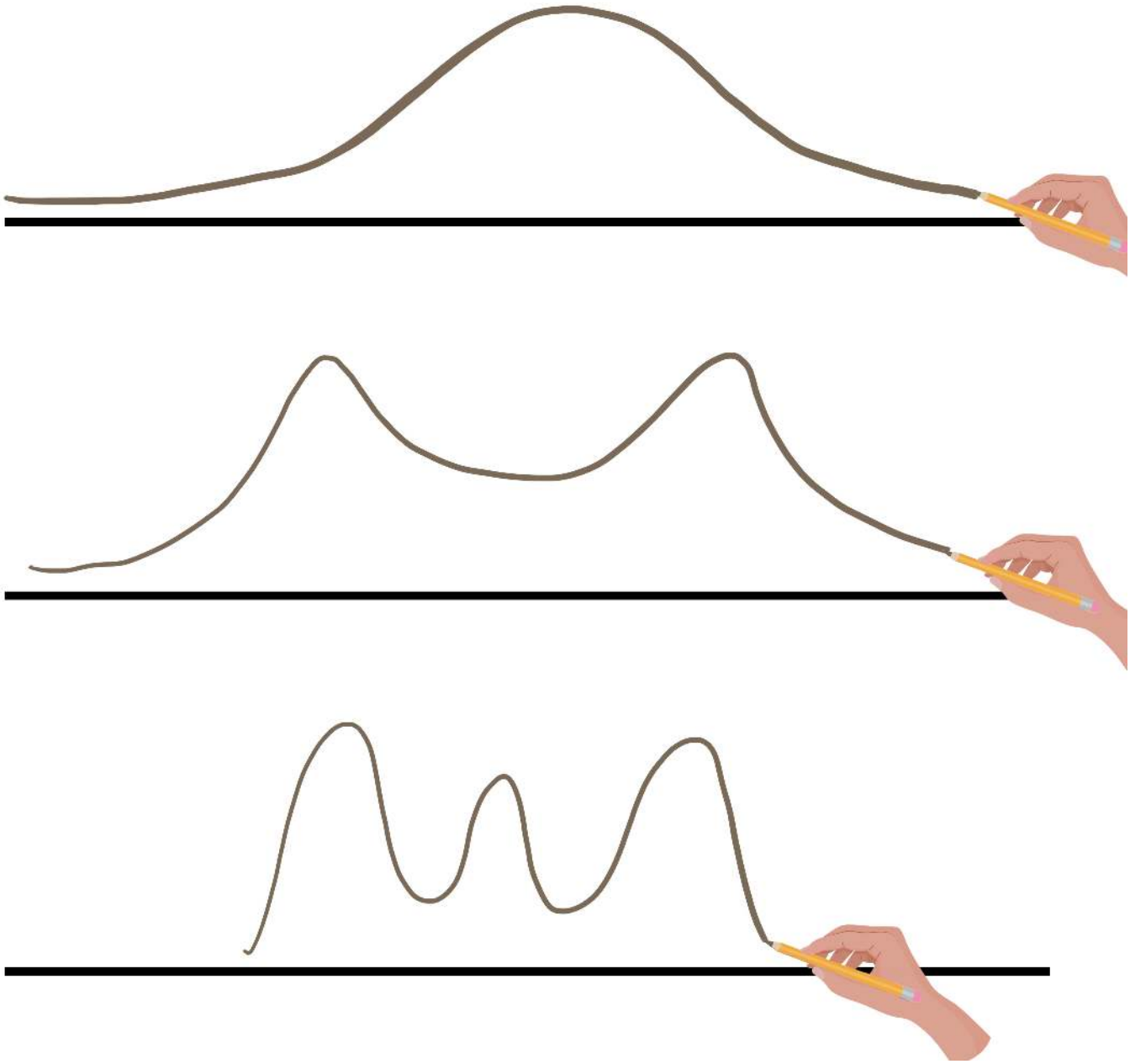


Figure 1.2.4.5: Examples of unimodal (on the left), bimodal (in the center), and multimodal shapes.

The most common distribution shape is a bell-curve, also known as a normal shape!

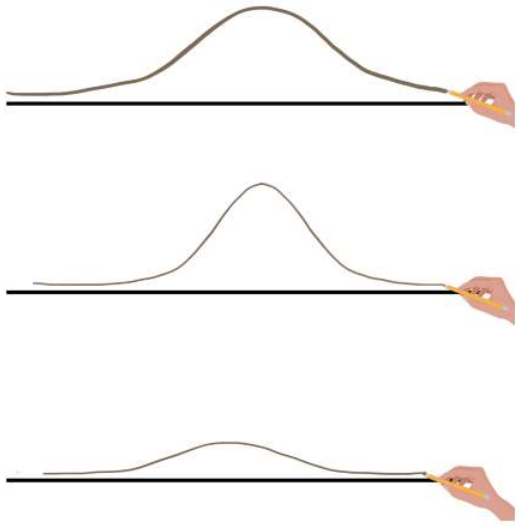


Figure 1.2.4.6: Example of a bell-curve also known as a normal shape. (Copyright; author via source)

Many things measured in nature such as heights, weights, lengths, widths have a normal distribution! Do not interpret a miniscule deviation from normal as skewed - in order to be skewed the tail must be much longer on one side than the other. For that reason, we would classify the histogram below as normal rather than skewed left or right:

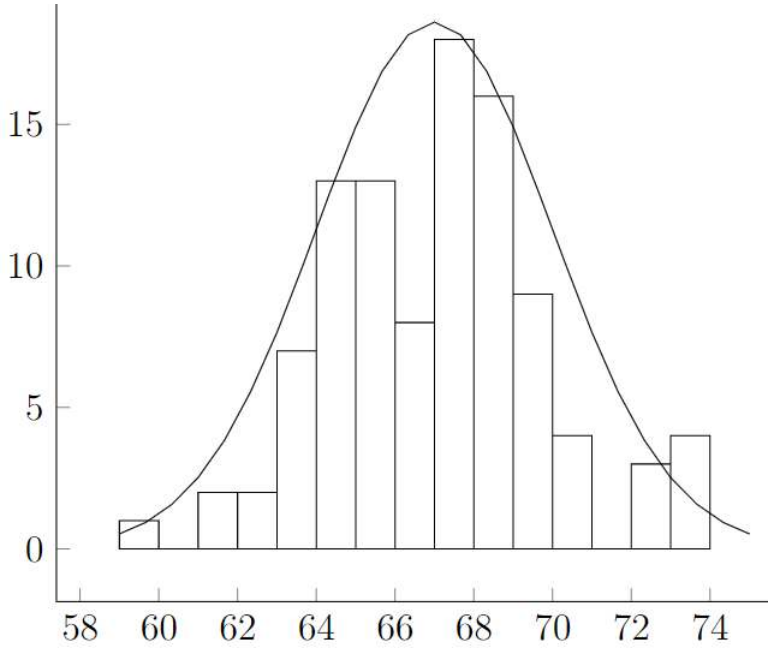


Figure 1.2.4.7: A frequency histogram showing heights of 100 semiprofessional soccer players (in inches) with the superimposed graph of a bell-curve. (Copyright; author via source)

Also note that the shape is described by a curve that roughly goes above the frequency bars without following it too closely. For instance, we wouldn't describe the histogram below as multimodal just because there appears to be more than one peak - at 51 and 54:

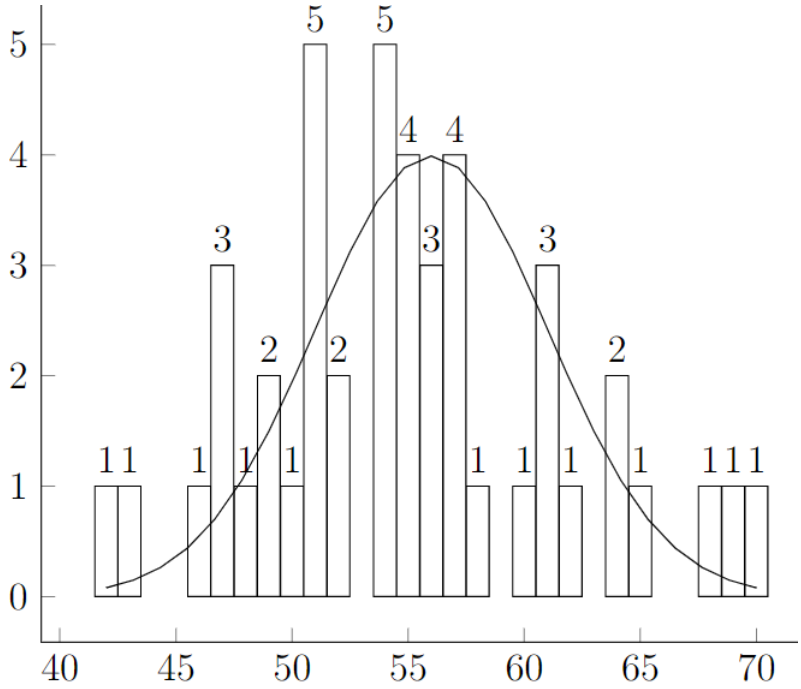


Figure 1.2.4.8: A frequency histogram showing ages at inauguration of 45 US presidents with the superimposed graph of a normal curve. (Copyright; author via source)

We just discussed a variety of different shapes of distributions and the language associated with describing them.

This page titled 1.2.4: Distribution Shapes is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.2.5: Organizing Bivariate Data

Section 1: Scatterplots



Earlier we discussed the methods of summarizing and organizing the data obtained from observing one variable. Such data are called **univariate data**.

The mileages of a sample of 5 cars.

Car	Mileage (mi)
1	78524
2	12574
3	24914
4	65813
5	39824

When we are interested in studying the relationship between a pair of variables we must collect and organize the data on two variables at the same time. Such data itself is called **bivariate data** and there are several ways to organize such data depending on the types of the variables.
The ages and mileages of a sample of 5 cars.

Car	Age (years)	Mileage (mi)
1	7	78524
2	2	12574
3	1	24914
4	5	65813
5	3	39824

When both variables are quantitative, we can construct a scatter plot to organize the data. To construct a **scatterplot**, we use a horizontal axis for the observations of one variable and a vertical axis for the observations of the other. When picking which axis to use for each variable consider whether you suspect that one variable depends on the other. The independent variable will be on the x-axis and dependent will be on the y axis. Each pair of observations is then plotted as a point. The example here demonstrates the summary of the observed relation between the age and the mileage of cars.

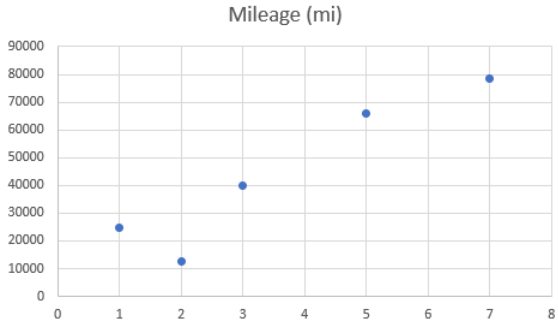


Figure 1.2.5.1: The scatterplot summarizing the mileage and conditions of a sample of 5 cars. (Copyright; author via source)

In the scatter plot above, each point is associated with a car, for example, car 1 is the point whose x-coordinate is 7 and the y-coordinate is 78.5 thousand, car 2 is the point whose x-coordinate is 2 and the y-coordinate is 12.5 thousand, etc.

Here is another summary that demonstrates the observed relation between the third exam score and the final exam score. Notice that the third exam is on the x-axis as we may suspect that it may have an effect on the final exam scores.

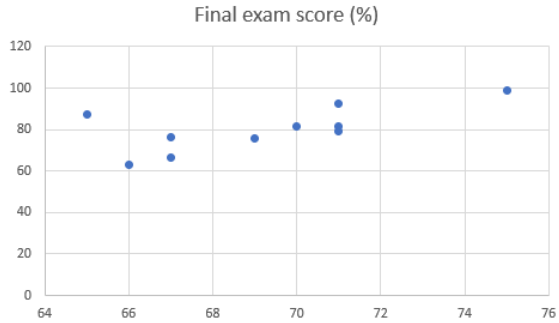


Figure 1.2.5.2: The scatterplot summarizing the third exam scores and the final exam scores of a sample of 10 students. (Copyright; author via source)

When one of the variables is time and there is only one observation for each moment of time, we can construct the scatter plot:

CO2 emissions in the United States between 2003-2009

Year	CO2 emissions in the United States
2003	5,681,664
2004	5,790,761
2005	5,826,394
2006	5,737,615
2007	5,828,697
2008	5,656,839
2009	5,299,563

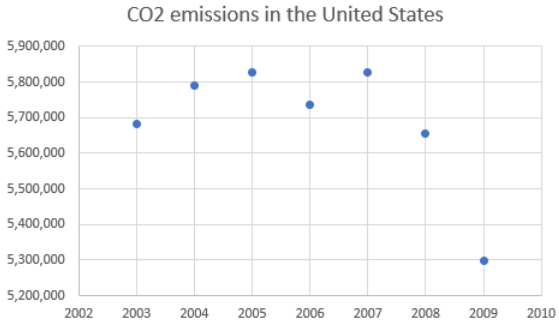


Figure 1.2.5.3: The scatterplot summarizing the CO2 emissions in the US between 2003 and 2009. (Copyright; author via source)

And then we can connect the points on the scatter plot with lines - such graph is called **time series**. The example here demonstrates the summary of the observed relation between the year and the CO_2 emissions for the United States:

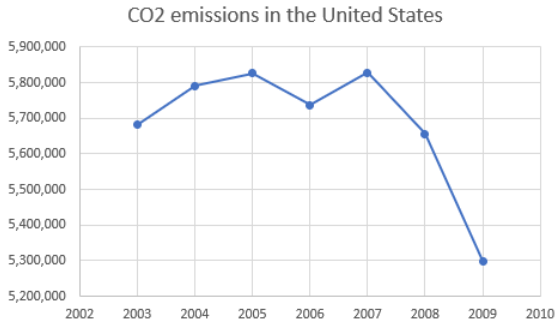


Figure 1.2.5.4: The time series summarizing the CO2 emissions in the US between 2003 and 2009. (Copyright; author via source)

Section 2: Contingency Tables



When we are interested in studying the relationship between a pair of variables we must collect and organize the data on two variables at the same time. Such data itself is called **bivariate data** and there are several ways to organize such data depending on the types of the variables.

Univariate Data: The colors of a sample of 5 cars.

Car	Color
1	Light

Car	Color
2	Dark
3	Dark
4	Light
5	Dark

Bivariate Data: The colors and conditions of a sample of 5 cars.

Car	Color	Condition
1	Light	Like new
2	Dark	Used
3	Dark	Like New
4	Light	Used
5	Dark	Used

A table called **contingency table** can be used to organize bivariate data in which one or both variables are qualitative. To construct the table, we arrange the observed frequencies into rows and columns. The intersection of a row and a column of a contingency table is called a cell. Each cell shows the frequency of observations that fit the description in the corresponding row and column.

Contingency table summarizing the colors and conditions of a sample of 5 cars.

Color\Condition	Like new	Used
Light	1	1
Dark	1	2

Many times, it makes sense to add the column and row with the totals for each column and row and for the entire table.

Contingency table summarizing the colors and conditions of a sample of 5 cars with totals.

Color\Condition	Like new	Used	Total
Light	1	1	2
Dark	1	2	3
Total	2	3	5

Alternatively, this table can be referred to as a **two-way frequency table**. Why? Cover the second and third columns and what's left can be easily seen as the frequency table for the colors of cars! Cover the second and third rows instead and what's left can be easily seen as the frequency table for whether car conditions! Also note that every contingency table can be easily broken down into two frequency tables, but the two frequency tables cannot be combined into a contingency table if the original data is lost!

Here is another example of a contingency table. For example, we can see that there are 280 people out of 755 that are cell phone users and had no speeding violation in the last year:

Contingency table with totals summarizing whether a driver is a cell phone user and whether they received a speeding violation in the last year.

	Speeding violations in the last year	No speeding violations in the last year	Total:
Cell phone user	25	280	305
Not a cell phone user	45	405	450
Total:	70	665	755

This page titled 1.2.5: Organizing Bivariate Data is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.2.6: Misleading Graphs



Next, we are going to discuss visual summaries that do not serve the purpose of telling the story but rather confuse or mislead.

Case 1.

Look at the massive decrease in unemployment from September of one year to March of the following year!



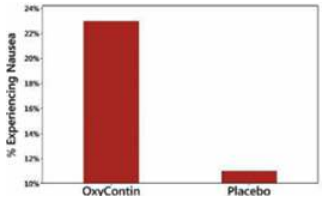
The unemployment decrease looks significant on the histogram only due to most of the histogram missing due to the cutoff on the vertical axis.



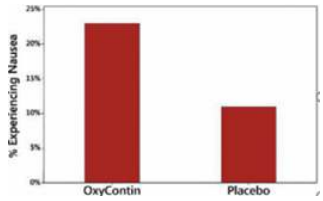
If you look at the whole picture the decrease from 5.4% to 5.0% does not seem that impressive!

Case 2.

Look at the huge difference between the effects of OxyContin and Placebo!



As in previous case, the difference between the effects looks significant on the histogram only due to most of the histogram missing due to the cutoff on the vertical axis.



And as in the previous case, if you look at the whole picture the decrease from 5.4% to 5.0% doesn't seem that impressive!

Case 3.

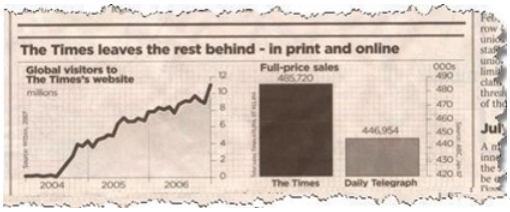
Look at the massive increase in home prices from 1998 and 1999!



Hopefully, nobody is fooled by this image and the narrative. If you look at the numbers only - the increase from $80000 to $82000 does not seem that impressive!

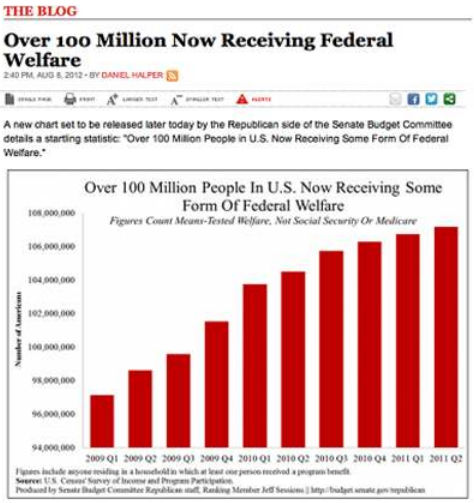
Case 4.

Please do not think that this can only happen in a textbook. This is an example of a newspaper with a misleading graph! Can you tell which graph is a misleading one?



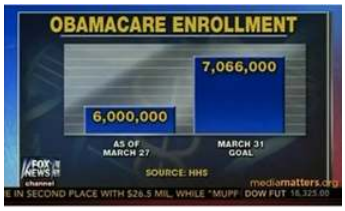
Case 5.

Of course, this can also be found on the web. Here is another example of a misleading graph!

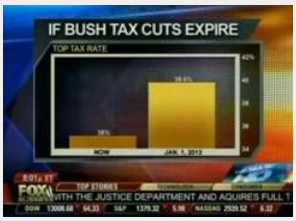


Case 6.

You can find many such examples in major news networks or at least one network – coincidentally all the examples here are from FOX NEWS.



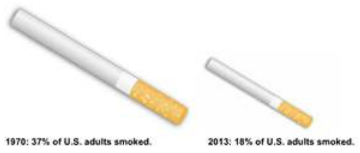






Case 7.

Look at the proportion of smoking adults decreasing from 1970 to 2013!



While the proportion of smoking adults decrease by a factor of 2, the problem with this visual demonstration is that when the dimensions are increased/decreased by a factor of 2 the area is increased/decreased by a factor of 4, so in this case, it appears that the proportion decreased by a factor of 4.

Case 8.

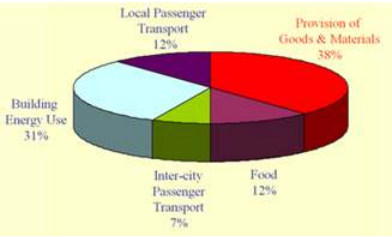
Look at how much more the basketball is more expensive to attend in comparison to the other sports!



While the average ticket prices for basketball event is the highest at $38 in comparison to $17 for the other sports, the images of a puck, a baseball, and a basketball put the focus on the area rather than the height which is rather distracting than revealing.

Case 9.

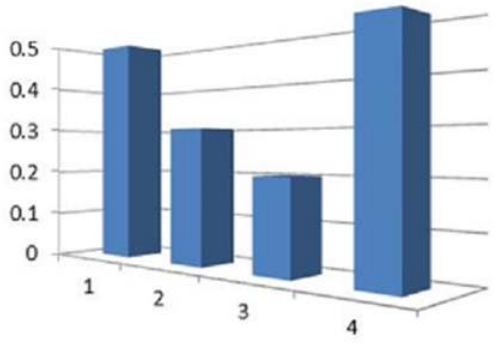
What a great 3D pie chart! But …



Which of the two categories - local passenger transport or food - appears bigger? Food? Wrong, they are both 12%! With the recent advancements in technology, it is very easy to produce a 3d image with a press of a button, however rarely the use of such graphs is justified. So be aware!

Case 10.

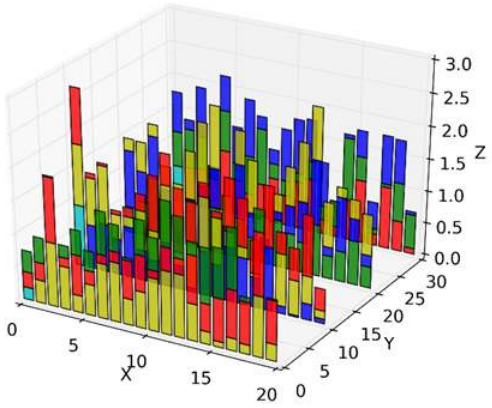
Another example of a great 3D summary! But …



Which of the two categories - 1 or 4 - appears taller? 4? Wrong, they are both 0.5!

Case 11.

Wow, look at this chart! The economy is clearly doomed!



Oh, never mind, this is the heat map of a microwave after using it for 1 minute! The problem here is that no axes are labeled, and it is hard to guess what the chart represents! Please, never hesitate to question the labels and the units!

Case 12.

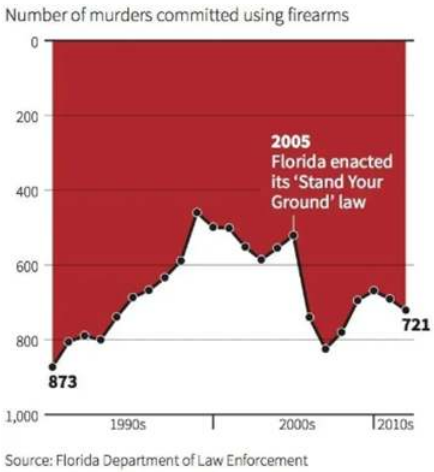
Sometimes the visual summary misrepresents the data by mistake as in the next case:



Do you think the majority of voters think that Scotland should be independent? Wrong! The other majority thinks otherwise!

Case 13.

However, sometimes the visual summary misrepresents the data with an ill purpose as in the next case:



How do you think the number of murders committed using firearms changed after Florida enacted its “stand your Ground” law? It appears that the number went down however look at the flipped vertical axis!!!

This page titled 1.2.6: Misleading Graphs is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

SECTION OVERVIEW

1.3: Numerical Summaries of Data

1.3.1: Measures of Center

1.3.2: Parameter and Statistic

1.3.3: Measures of Variation

1.3.4: Interpreting the Mean and Standard Deviation

1.3.5: Measures of Relative Standing

1.3.6: Five-Number Summaries and Boxplots

This page titled 1.3: Numerical Summaries of Data is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.3.1: Measures of Center

Introduction



In the previous section, we learned how to organize and present the collected data visually. While a picture is worth a thousand words, describing a picture may not always be an option. In this section, we will learn how to describe the data and its distribution numerically.

Consider the following example: a zoologist was sent to a distant island to study the population of lemurs. Let's say they are interested in the ages of lemurs. Of course, the goal is to produce the histogram and that histogram will represent the knowledge of the population of lemurs' ages! After collecting the data, the zoologist wants to communicate their findings, but the internet is down so the only way to send the information is via telegraph. For those of you who are not familiar with the telegraph – It is the twitter of the 20th century. The solution is to create a numerical summary which then can be communicated via telegram and then interpreted on the other end into the visual summary.

In other words, when studying data, the goal is always to have the visual summary that tells the story. When it is not possible to get the visual summary directly, we compute the numerical summary of data that later can be interpreted to describe the shape of the distribution.

There are three types of numerical measures that we will cover in this course:

- the measures of center or central tendency – describe what considered the most typical observations in the data set;
- the measures of variation or spread – describe how much data varies from the center;
- the measures of position or relative standing – describe the position of each individual observation among the entire data set.

Next, we will discuss different ways to produce the numerical summary of data.

Section 1: Measures of center



In this section, we will learn how to measure the center of the data, or in other words how to define what is a typical observation in the data set.

Definition: Measures of Center

Descriptive measures that indicate where the center or the most typical value of a data set lies are called **measures of central tendency** or, more simply, **measures of center**. The three most common measures of center: the **mean**, **median**, and **mode**.

Consider the following set of a student's 5 midterm exam scores:

86, 79, 96, 79, 100

Definition: Mean

The **mean** of a data set is the sum of the observations divided by the number of observations.

For example, the mean of {86, 79, 96, 79,100} is the sum of all five numbers divided by 5 which is 88.

Let's drop the lowest exam score so that now we have only one 79:

86, 96, 79, 100

The mean of the four remaining numbers is the sum of all four of them divided by 4 which is 90.25.

Definition: Median

The **median** of a data set is the **middle** number that divides the bottom 50% of the data from the top 50%.

For example, the median of {79, 79, 86, 96, 100} is 86 because if we arrange the data in increasing order the number 86 is such that there are two numbers below it and the other two numbers are above it!

Let's drop the lowest exam score so that now we have only one 79! The median of the four remaining numbers is the middle number by definition but since there is an even number of values in the dataset, we are going to obtain the middle value by finding the average of the middle two values. In this example, the median of the data is 91.

Definition: Mode

The **mode** of a data set is the most frequently occurring observation in the data set.

For example, the mode of {86, 79, 96, 79,100} is 79 because 79 occurs twice which more than any other observation.

Let's drop the lowest exam score so that now we have only one 79! The mode of the four remaining numbers is still the most frequent observation which in this case does not exist since every observation occurs the same number of times, so we say such data doesn't have a mode.

It is worth mentioning that the mean and median apply only to quantitative data, whereas the mode can be used with either quantitative or qualitative data. Thus, while the mode appears to be the least useful among the three measures of central tendency, it is the only way to assess the center for qualitative data! But when working with quantitative data we still have to choose between the mean and the median.

Let's consider the following example. We have four people Ann, Beth, Connor, and Diana that make the following annual salaries:

$30k, $50k, $50k, $70k

Their average salary is $50,000 and their median is also $50,000 dollars. Now, let's say Elon moves in and his salary is $300,000 a year. What effect does this have on the mean and median? The new mean is $100,000 and the new median is still $50,000. Which of the two measures produces a more representative value? In this case, the median is obviously a better choice, for example if Toyota Company was considering building a Lexus or Scion dealership and they were only considering the mean it would appear that an average person in this neighborhood can afford a Lexus when in reality only one person can afford it. So, what exactly is the difference between the mean and the median??

To understand the differences between the mean and median we introduce a new term.

Definition: Resistant measure

A **resistant measure** is not sensitive to the influence of a few extreme observations.

The median is a resistant measure of center, but the mean is not. Usually, we use the median for house prices and salaries and for everything else we would prefer the mean. One way to make the mean more resistant is to remove a percentage of the smallest and largest observations before computing the mean. Such a measure is called a **trimmed (truncated) mean**. A professor dropping the lowest midterm exam score is an example of a trimmed (truncated) mean.

Example 1.3.1.1.1

Consider a larger data set and compute its mode, mean, and median.

42	43	46	47	47	47	48	49	49
50	51	51	51	51	51	52	52	54
54	54	54	54	55	55	55	55	56
56	56	57	57	57	57	58	60	61
61	61	62	64	64	65	68	69	70

Solution

The modes are 51 and 54 because they have the highest frequency 5.

The mean is 55.02, the average of all 45 values.

The median is 55 because it is 23rd out of 45 thus it has 22 values below and 22 values above.

We discussed the three most common ways to compute the center of the data set. It should now be clear that the mean, median, and mode generally provide different information. There is no simple rule for deciding which measure of center to use in a given

situation. Even experts may disagree about the most suitable measure of center for a particular data set. But sometimes for certain variables the choice is clear.

Section 2: Measures of Center from Frequency Tables



Previously we discussed how to find the mean, median, and mode form a raw data set. Next, we will discuss how to find the center of the data when the frequency distribution is given.

Class Midpoint	Frequency
x_1	$f(x_1)$
x_2	$f(x_2)$
$\dots$	$\dots$
x_k	$f(x_k)$
Total:	n

When the data is given by the frequency distribution table the easiest measure to compute is the mode. The mode is the midpoint of the class with the largest frequency. If there are several largest frequencies, then there are many modes. If all frequencies are the same, then there are no modes.

Example 1.3.1.2.1

Consider the following frequency table produced by a single value grouping.

Class	Frequency
54	2
60	4
71	3
80	1
Total	10

Solution

The mode is 60 because 60 has the highest frequency 4.

How to find the mean of the data set that is given by the frequency distribution? The mean can be found by using the following formula in which we multiply each observation by the frequency, then add them together and divide by the total:

$$\frac{x_1 \cdot f(x_1) + x_2 \cdot f(x_2) + \cdots + x_k \cdot f(x_k)}{n}$$

Example 1.3.1.2.2

Consider the following frequency table again.

Class	Frequency
54	2
60	4
71	3
80	1
Total	10

Solution

According to the formula we compute the mean by multiplying each observation by its frequency, add them together, and divide by the total. The mean in this example is

$$\frac{54 \cdot 2 + 60 \cdot 4 + 71 \cdot 3 + 80 \cdot 1}{2 + 4 + 3 + 1} = 64.1$$

The least trivial measure of center to find from the frequency table is the median. To find the median, we are going to use the cumulative frequencies and the following process that depends on whether the number of observations is odd or even.

Class Midpoint	Class Midpoint	Frequency
x_1	$f(x_1)$	$cf(x_1)$
x_2	$f(x_2)$	$cf(x_2)$
$\cdots$	$\cdots$	$\cdots$
x_k	$f(x_k)$	$cf(x_k)$
Total:	n	

When n is odd, the median is the midpoint of the first class for which the cumulative frequency exceeds $\frac{n-1}{2}$; or when n is even, the median is the average of the midpoint of the first class for which the cumulative frequency exceeds $\frac{n}{2} - 1$ and the midpoint of the first class for which the cumulative frequency exceeds $\frac{n}{2}$. Often these are the same midpoints.

Example 1.3.1.2.3

Consider the following frequency table again along with the cumulative frequencies.

Class	Frequency	Cumulative Frequency
54	2	2
60	4	6
71	3	9
80	1	10

Class	Frequency	Cumulative Frequency
Total	10	

Solution

Since $n = 10$ is even, we find the midpoint of the first class for which the cumulative frequency exceeds $\frac{10}{2} - 1 = 4$ - it is 60, and then we find the midpoint of the first class for which the cumulative frequency exceeds $\frac{10}{2} = 5$ and it is again 60. The average of the two is 60.

Example 1.3.1.2.4

Let's consider the following frequency table produced by an interval grouping.

Class	Frequency
40 to 45	2
45 to 50	7
50 to 55	13
55 to 60	12
60 to 65	7
65 to 70	3
70 to 75	1
Total	45

Solution

The first thing we need to do is to replace each class with the class representative in this case the class's midpoint and then add the cumulative frequency column:

Class	Midpoint	Frequency	Cumulative Frequency
40 to 45	42.5	2	2
45 to 50	47.5	7	9
50 to 55	52.5	13	22
55 to 60	57.5	12	34
60 to 65	62.5	7	41
65 to 70	67.5	3	44
70 to 75	72.5	1	45
	Total	45	

Now we are going to do the same thing as before using the class midpoints as the data entries.

- The mode is 52.5 because 52.5 is the midpoint of the class that has the highest frequency 13.
- The mean is found by using the formula and it is equal to $\frac{42.5 \cdot 2 + 47.5 \cdot 7 + 52.5 \cdot 13 + 57.5 \cdot 12 + 62.5 \cdot 7 + 67.5 \cdot 3 + 72.5 \cdot 1}{2 + 7 + 13 + 12 + 7 + 3 + 1} = 55.61$
- To find the median we consider the column with the cumulative frequency values and since the cumulative frequency 34 is the first one that exceeds $\frac{45-1}{2} = 22$ we conclude that the median is 57.5.

You may not recognize this data set but previously we have found the measures of center of the same data:

42	43	46	47	47	47	48	49	49
50	51	51	51	51	51	52	52	54
54	54	54	54	55	55	55	55	56
56	56	57	57	57	57	58	60	61
61	61	62	64	64	65	68	69	70

Let’s compare the results obtained using the original data against the results that we just obtained from the frequency table:

Measure of the Center	Original Data	Frequency Table
Mean	55.02	55.61
Median	55	57.5
Mode	51, 54	52.5

Why do the answers not match? Recall that the frequency table is created by grouping the observations into classes which is the process that cannot be undone! When we use the frequency table, we essentially assume that all observations in each class are the same as the class's midpoint. As a result, the numerical summaries obtained from the frequency table are only approximations of the numerical summaries obtained from the original data set.

Now we know how to find the mean, median, and mode regardless of how the data is provided. If the formulas above, do not make sense then reconstruct the entire data set from the frequency table and find the measures using their definitions above! However, understanding the formulas simply requires an understanding of the meaning of the frequency in the frequency table and nothing more.

This page titled 1.3.1: Measures of Center is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.3.2: Parameter and Statistic



In inferential statistics, we will be trying to learn more about the population from a sample, in other words, we will be using the numerical summaries obtained from a sample to estimate the corresponding numerical summary of the population.

The following terminology will help us to distinguish between a sample and the population.

Definition: Parameter

A **parameter** is a numerical summary of the population of size N which can only be obtained by a census, and it remains constant although most of the time unknown.

Definition: Statistic

A **statistic** is a numerical summary of a sample of size n which can only be obtained by sampling, and it varies from sample to sample.

Parameter	Statistic
• Population • Size N • Census • Constant • Unknown without the census	• Sample • Size n • Sampling • Varies • Can be found by obtaining a sample

The most common use of a statistic is to estimate the value of the corresponding parameter that is usually unknown.

Definition: Estimator (estimate)

The number that is used to estimate the parameter is called **an estimator** or **an estimate** and an attempt to estimate the parameter is called **an estimating**.

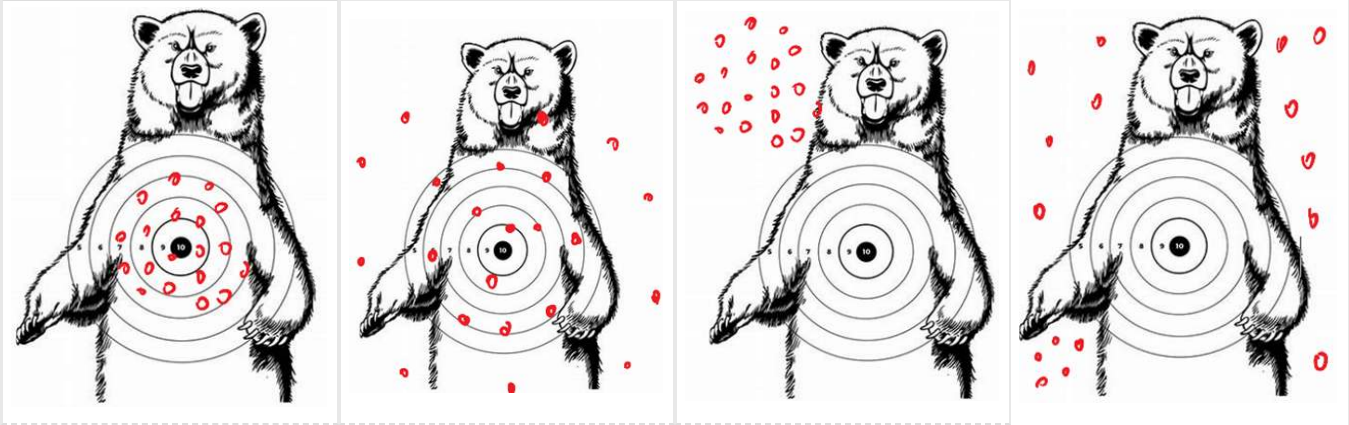
Some estimates are better than others. While a random guess may satisfy the definition of an estimate, we have an intuition that using a statistic is better than just guessing.

Definition: Biased vs unbiased estimators

An estimator is called an **unbiased estimator** of the parameter if the mean of its values, in the long run, equals the parameter; otherwise, it is said to be a **biased estimator**. In other words, an unbiased estimator yields, on average, the correct value of the parameter, whereas a biased estimator does not.

For example, the sample mean is an unbiased estimator of a population mean while a random guess is a biased estimate of the population mean.

This diagram compares the estimating process with a target shooting.



Can you guess which ones are biased and which ones are not? Clearly, any unbiased estimator is better than any biased estimator, however, some biased estimators are better than others! For example, which of the two biased shooting patterns on the right you would want your hunting partner to have? A shooting instructor can easily fix the first shooting pattern by telling the shooter to aim lower and to the right, but there is no easy fix for the other one! Please understand the difference between being correct and being correct on average - none of the unbiased shooting patterns hits the bull's eye all the time, but *on average* it appears that they do!

For popular numerical summaries, the distinguishing between being the parameter or statistic is so important that we introduce a different notation to denote the same measure, even when we use the same formula. For example, we label the population mean with a Greek letter μ, and the sample mean with $\bar{x}$ although the formulas are the same. For example, the average age of all US presidents at inauguration is μ and is equal to 55.02 and the average age of the last 5 US presidents at inauguration is $\bar{x}$ and is equal to 67.2.

Note that the same data set can be treated as the population or a sample depending on the context and depending on the context, we will use either μ or $\bar{x}$ to label the average of the data set. For example, we will treat the 5 starting players as the population when we are interested to learn the average height of the five starting players. We will treat the same group of five as a sample when we are interested to learn the average height of the team but only observed the heights of the starting five players.

In this section we started to develop the vocabulary and notation to distinguish between the parameters and statistics.

This page titled 1.3.2: Parameter and Statistic is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.3.3: Measures of Variation



Next, we will discuss the measures of variation or spread that describe how much data varies from the center. Consider the two sets of players:

Team 1	Team 2
71, 73, 76, 76, 78	67, 72, 76, 76, 84

The mean of each group is 75, but clearly the two teams are not the same. One of the differences between the two data sets that any measure of center doesn't capture is the variety of data within the set. To describe that difference quantitatively, we use a descriptive measure that indicates the amount of variation or spread, in a data set. Such descriptive measures are referred to as **measures of variation** or **measures of spread**.

Definition: Measures of variation or spread

Measures of variation and spread are numerical summaries used to describe the distribution of data points in a dataset. They provide insights into how much the data varies and how spread out the values are.

Just as there are several different measures of center, there are also several different measures of variation. In this section, we will examine the most frequently used measures of variation.

One example of a measure of variation is the range.

Definition: Range

The **range** of a data set is the difference between the maximum (largest) and minimum (smallest) observations.

Example 1.3.3.1

For example, the range of Team 1 is $78 - 72 = 6$ and the range of Team 2 is $84 - 67 = 17$.

The range only measures the total variation and doesn't capture any variation between the minimum and maximum observed values. It is not the only nor the best way to measure variation!

In contrast to the range, the **standard deviation** takes into account all the observations and is the preferred measure of variation when the mean is used as the measure of center. The formulas for finding the standard deviations of a sample and a population

differ slightly.

First, assume that we are dealing with a population. We are going to list all the values in the first column and compute the population mean:

$$\mu = 75$$

We will subtract the mean from each observation to make sure that all observations are accounted for! If we were to add these differences now then we would have gotten a zero total which is not what we want, so instead we square each difference to count each variation as positive. Now, it makes sense to add them all together to obtain the sum of squared differences:

$$\Sigma(x_i - \mu)^2 = 24$$

However, the total sum of squared differences is not a good way to measure the variation just like the grand total of a students' grades is not a good way to measure their academic achievements. We naturally divide the sum by N to find the "average":

$$\frac{\Sigma(x_i-\mu)^2}{N} = \frac{24}{5} = 4.8 = \sigma^2$$

This measure is labeled as σ^2 and is called **the population variance** and is very significant in theory, however, from a practical point of view we can make it better by taking the square root:

$$\sqrt{\frac{\Sigma(x_i-\mu)^2}{N}} = \sqrt{4}.8 = 2.19 = \sigma$$

This measure is labeled with a Greek letter σ and is called the **population standard deviation** and its advantage is that it has the same units as the original data (in this case inches) and therefore can be used as a measuring stick!

x_i	$x_i - \mu$	$(x_i - \mu)^2$
72	72-75=-3	9
73	73-75=-2	4
76	76-75=1	1
76	76-75=1	1
78	78-75=3	9
$\mu = 75$		$\Sigma(x_i - \mu)^2 = 24$
	Population Variance	$\frac{\Sigma(x_i-\mu)^2}{N} = \frac{24}{5} = 4.8 = \sigma^2$
	Population Standard Deviation	$\sqrt{\frac{\Sigma(x_i-\mu)^2}{N}} = \sqrt{4}.8 = 2.19 = \sigma$

Roughly speaking, the standard deviation measures variation by indicating how far, on average, the observations are from the mean. For a data set with a large amount of variation, the observations will, on average, be far from the mean, so the standard deviation will be large. For a data set with a small amount of variation, the observations will, on average, be close to the mean, so the standard deviation will be small.

Caution

Remember! Do not perform any rounding until the computation is complete; otherwise, substantial roundoff error can occur.

Now, let's assume that we actually are interested in the standard deviation of the heights of all players in the league, but we have only accessed the given (\5\) players, which are now became a sample. We will label the mean as $\bar{x}$ rather than μ and will perform the first few steps in the same way as before up until it is time to divide. This time, we are going to divide by 4 instead of 5. We will learn why later but for right now let's just register the fact that to compute **the sample variance** we divide by the $n-1$ rather than the sample size n:

$$\frac{\Sigma(x_i-\bar{x})^2}{n} = \frac{24}{4} = 6 = s^2$$

We finish by taking the square root:

$$\sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n}}=\sqrt{6}=s$$

This measure is called **the sample standard deviation** and it has the same units as the original data (in this case inches).

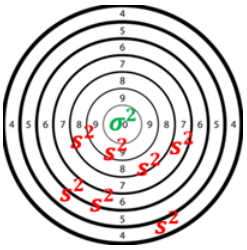
x_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$
72	72-75=-3	9
73	73-75=-2	4
76	76-75=1	1
76	76-75=1	1
78	78-75=3	9
$\bar{x}=75$		$\Sigma(x_i-\bar{x})^2=24$
	Sample Variance	$\frac{\Sigma(x_i-\bar{x})^2}{n-1}=\frac{24}{4}=6=s^2$
	Sample Standard Deviation	$\sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n-1}}=\sqrt{6}=2.45=s$

Usually, we use the same formula for the population's parameter and a sample's statistic, however as it became obvious it is not the case with the standard deviation and variance:

Population	Sample
• Variance: $\frac{\Sigma(x_i-\mu)^2}{N}=\frac{24}{5}=4.8=\sigma^2$ • Standard Deviation: $\sqrt{\frac{\Sigma(x_i-\mu)^2}{N}}=\sqrt{4.8}=2.19=\sigma$	• Variance: $\frac{\Sigma(x_i-\bar{x})^2}{n-1}=\frac{24}{4}=6=s^2$ • Standard Deviation: $\sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n-1}}=\sqrt{6}=2.45=s$

So why do we divide by $n-1$ instead of n?

Remember that the main purpose of finding a statistic is to estimate the unknown parameter (in this case the population variance, σ^2). To do that we use the statistic (in this case, the sample variance s^2). If we divide by n instead of $n-1$ we will be obtaining a value that is always underestimating the true population variance by a factor of $\frac{n}{n-1}$.



After the correction, we get a formula for the desired unbiased estimator for the population variance which we call a sample variance:

$$\frac{\Sigma(x_i-\bar{x})^2}{n}\cdot\frac{n}{n-1}=\frac{\Sigma(x_i-\bar{x})^2}{n-1}=s^2$$

We discussed the most frequently used measures of variation or spread that describe how much data varies from the center.

This page titled 1.3.3: Measures of Variation is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.3.4: Interpreting the Mean and Standard Deviation

Section 1: Chebyshev's Rule



In this section, we will learn how get a rough idea about the data distribution from the measures of the center and variation. The idea is simple - when we have data, we want to be able to construct and communicate a visual summary, but when that is impossible, we produce the numerical summaries which are very concise and yet packed with information! Next, we will learn how to interpret the numerical summaries to communicate the shape of the distribution. For example, what do you imagine, when you hear that a dataset has the average 50 and the standard deviation 10?

Naturally, there is a bond between the standard deviation and the mean, which can be used to unpack the information packed in these two values. Since the units of a standard deviation are the same as the units of data, it makes it easy to use it as a measuring stick on the data axis and create the scale by adding and subtracting 1σ, 2σ, and 3σ from and to μ. We say the values are:

- within 1 standard deviation if they are between $\mu - 1\sigma$ and $\mu + 1\sigma$.
- within 2 standard deviations if they are between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- within 3 standard deviations if they are between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Example 1.3.4.1.1

When $\mu = 50$ and $\sigma = 10$ we say the values are:

- within 1 standard deviation if they are between $\mu - 1\sigma = 50 - 10 = 40$ and $\mu + 1\sigma = 50 + 10 = 60$.
- within 2 standard deviations if they are between $\mu - 2\sigma = 50 - 20 = 30$ and $\mu + 2\sigma = 50 + 20 = 70$.
- within 3 standard deviations if they are between $\mu - 3\sigma = 50 - 30 = 20$ and $\mu + 3\sigma = 50 + 30 = 80$.

One other thing that we will keep an eye on while interpreting the numerical summaries is the outliers!

Definition: Outlier

An **outlier** is an observation that appears extreme relative to the rest of the data. The rule of thumb is the **three-standard-deviations rule**, which states that most of the observations in any dataset lie within three standard deviations to either side of the mean, so anything beyond the three standard deviations is considered to be very unlikely and therefore satisfies the definition of an outlier.

In some books, the values that are more than two standard deviations away from the mean are called **significant values that are either significantly low** or significantly **high**, depending on whether they are below or above the mean.

Example 1.3.4.1.2

When $\mu = 50$ and $\sigma = 10$ we call:

- any observation less than 20 or greater than 80 is an outlier.
- any observation less than 30 is significantly low.
- any observation more than 70 is significantly high.

It is not hard to observe that there are fewer and fewer observations the further and further away from the mean. This result was generalized by Pafnuty Chebyshev into the following theorem, named after him.

Theorem 1.3.4.1.1: Chebyshev's Rule

In any dataset, at least $(1 - \frac{1}{k^2}) \cdot 100\%$ of observations are within k standard deviations from the mean. In other words:

- At least 75% of observations are within 2 standard deviations.
- At least 89% of observations are within 3 standard deviations.
- At least 93.75% of observations are within 4 standard deviations.

Example 1.3.4.1.3

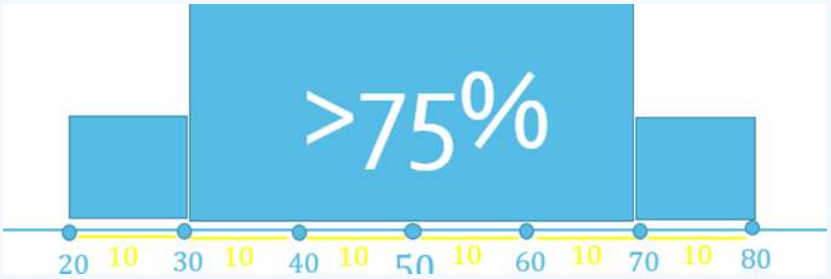
When $\mu = 50$ and $\sigma = 10$ we call:

- At least 75% of observations are between 30 and 70.
- At least 89% of observations are between 20 and 80.
- At least 93.75% of observations are between 10 and 90.

The significance of the Chebyshev's theorem is not in the exact percentages but in the fact that now we can imagine the shape of the histogram based on only two numbers.

Example 1.3.4.1.4

When $\mu = 50$ and $\sigma = 10$ we expect the histogram to look somewhat like this which of course is not accurate but just a rough description of the actual histogram.



Anyway, it is better than nothing!

Consider again the president's age at inauguration:

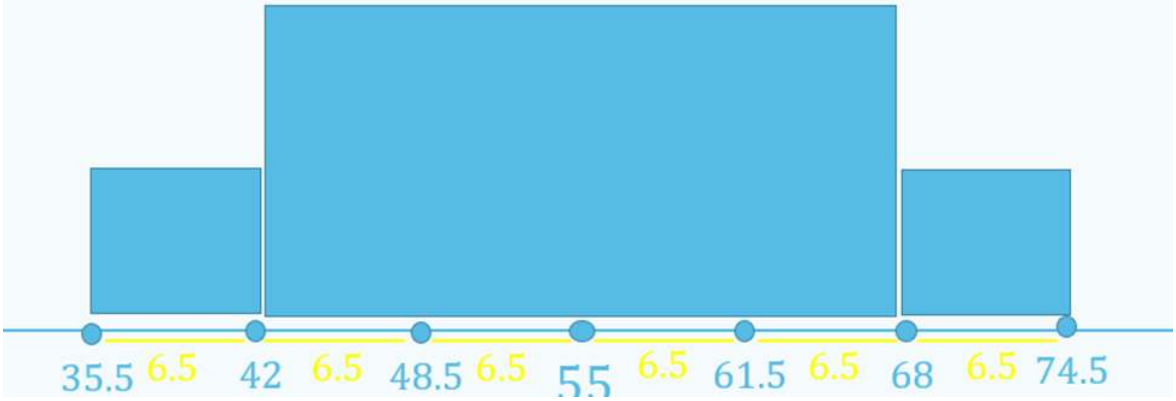
42	43	46	47	47	47	48	49	49
50	51	51	51	51	51	52	52	54
54	54	54	54	55	55	55	55	56
56	56	57	57	57	57	58	60	61
61	61	62	64	64	65	68	69	70

In the president's age at inauguration data set, the mean is approximately 55 and the standard deviation is approximately 6.5. We can compute how much of the data exactly are within 1, 2, 3, and 4 standard deviations from the mean and confirm the Chebyshev's rule.

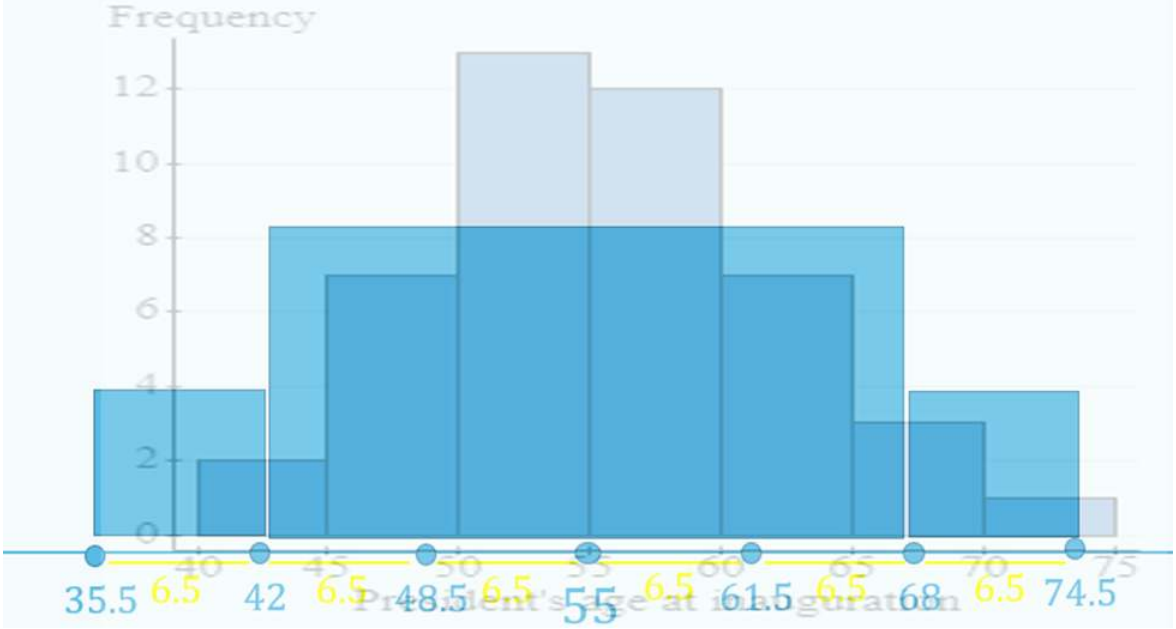
Within	1 st. dev. = 6.5	2 st. dev. = 13	3 st. dev. = 19.5	4 st. dev. = 26
Between	48.5 and 61.5	42 and 68	35.5 and 74.5	29 and 81
How many values?	31/45=68.89%	42/45=93.33%	45/45=100%	45/45=100%
Chebyshev's Rule	n/a	At least 75%	At least 89%	At least 93.75%

Chebyshev's rule is always true!

So just based on the two numbers we would draw the following scale and would imaging the following shape of the distribution.



If we superimpose the actual histogram, we will be able to see that we are not that far off!



Section 2: The Empirical Rule



Previously we learned how to get a rough shape of the distribution from a given mean and standard deviation using the Chebyshev's rule which can be applied for any dataset. Next, we will discuss the Empirical rule that allows us to better visualize the data if we can assume that it has a bell-shaped distribution.

For normal (bell-shaped) distributions the Chebyshev's estimate is too rough and can be improved.

Theorem 1.3.4.2.1: Empirical Rule

If the distribution of the data set is or can be assumed to be approximately normal (or bell-shaped), we can apply the **Empirical Rule**, which states the following:

- roughly 68% of observations are within 1 standard deviation.
- roughly 95% of observations are within 2 standard deviations.
- roughly 99.7% of observations are within 3 standard deviations.

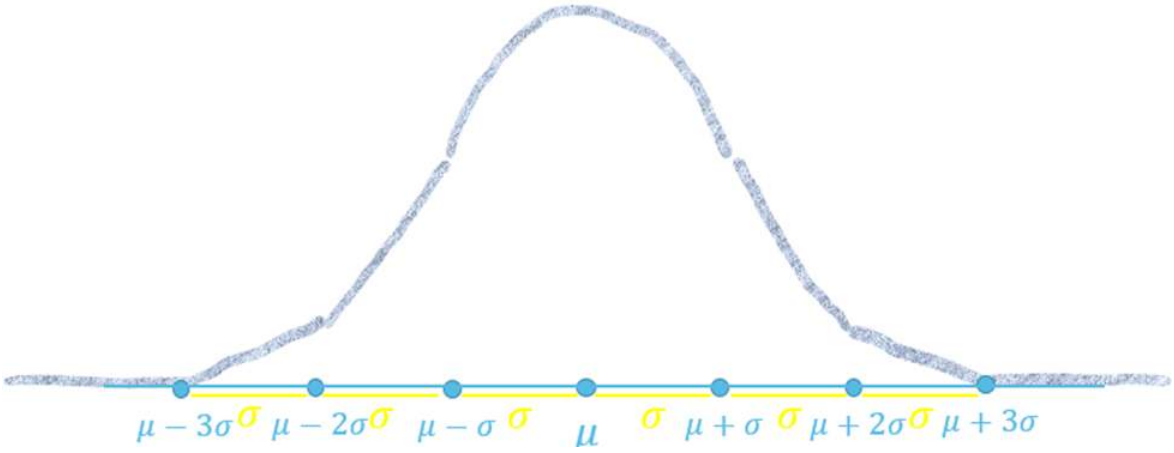
Example 1.3.4.2.1

Example: When $\mu = 50$ and $\sigma = 10$:

- Roughly 68% of observations are between 40 and 60.
- Roughly 95% of observations are between 30 and 70.
- Roughly 99.7% of observations are between 20 and 80.

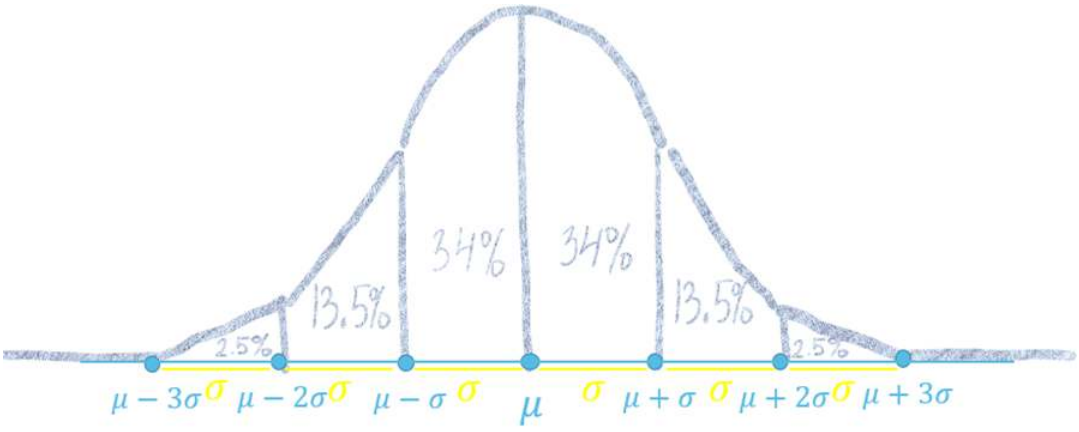
The significance of the Empirical Rule is that it enables us to determine the shape of the bell-curve from the mean and standard deviation only.

1. Draw a horizontal axis and label the values that are 1, 2, 3, standard deviations away from the mean.
2. Only 0.3% of the data are outside of 3 standard deviations – so we can draw the tails that are basically on the x-axis outside of the three standard deviations away.
3. 5% of the data are outside of 2 standard deviations - continue drawing the tails towards the middle by increasing the slope by a notch.
4. 32% of the data are outside of 1 standard deviation - continue drawing the tails towards the middle by increasing the slope by another notch.
5. 68% of the data are within 1 standard deviation - connect the tails with the bell-shaped middle part.

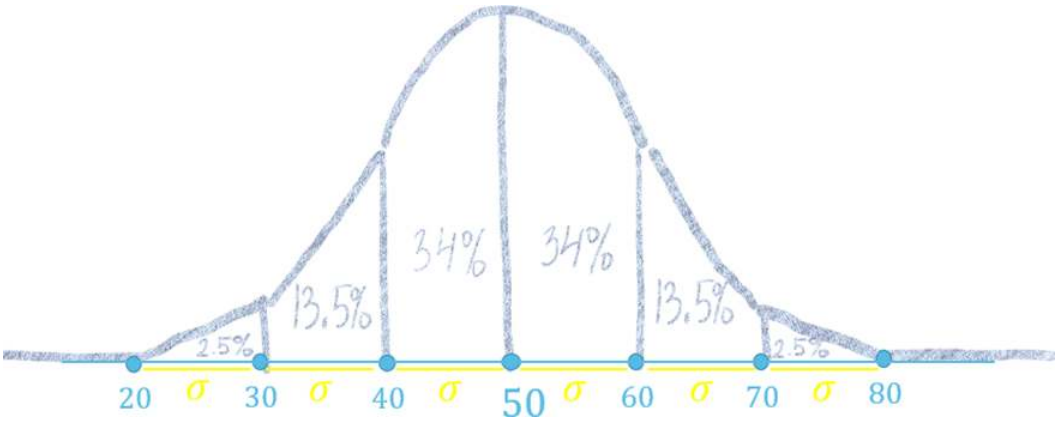


According to the Empirical rule we have the following distribution of data under the curve:

- The middle 68% are evenly split into two halves – 34% each.
- The 27% of the data is between 1 and 2 standard deviations away which means 13.5% in each half.
- The 5% of the data outside of the 2 standard deviations is mainly concentrated between 2 and 3 standard deviations which means 2.5% in each tail.



We can do the same for any μ and σ! For example, when $\mu = 50$ and $\sigma = 10$ we get the following:

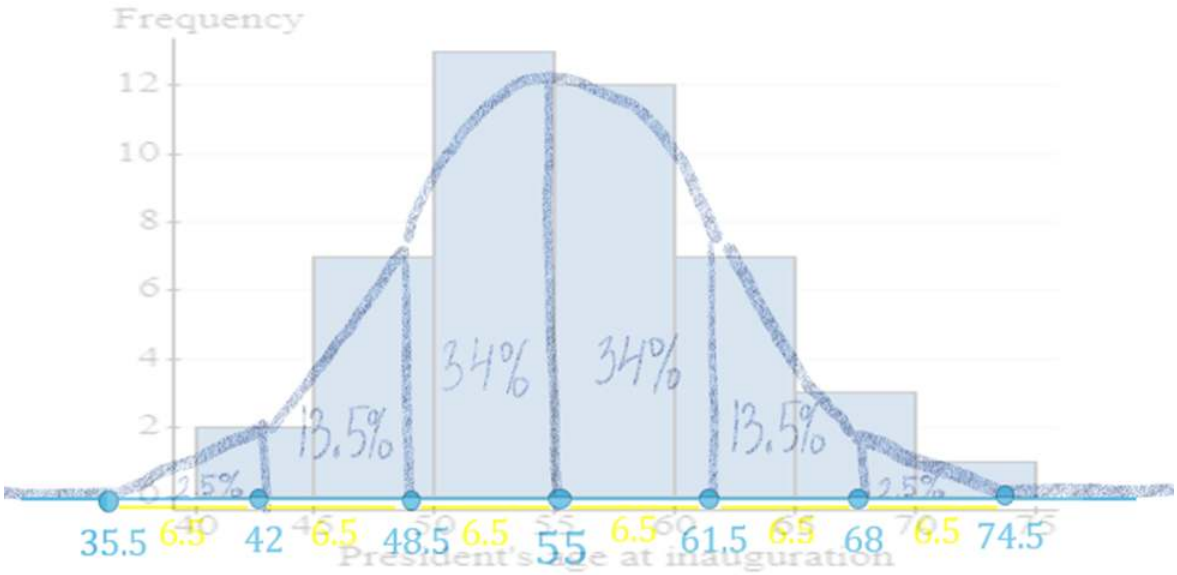


In the president's age at inauguration data set, the mean is approximately 55 and the standard deviation is approximately 6.5. We can compute how much of the data exactly are within 1, 2, 3, and 4 standard deviations from the mean and confirm the Empirical Rule.

Within	1 st. dev. = 6.5	2 st. dev. = 13	3 st. dev. = 19.5	4 st. dev. = 26

Between	48.5 and 61.5	42 and 68	35.5 and 74.5	29 and 81
How many values?	31/45=68.89%	42/45=93.33%	45/45=100%	45/45=100%
The Empirical Rule	Approx 68%	Approx 95%	Approx 99.7%	n/a

Remember that the Empirical Rule only applies when the data distribution can be assumed normal! Assuming that the presidents' ages are normally distributed, if we superimpose the actual histogram we will be able to see that we are not that far off!



This page titled 1.3.4: Interpreting the Mean and Standard Deviation is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.3.5: Measures of Relative Standing

Section 1: Z-Scores



Next, we will discuss the measures of relative standing that allow us to rank each observation in the data set.

Definition: Z-score

For an observation x_i, the value $z = \frac{x_i - \mu}{\sigma}$ is called the **z-score** of the observation x_i. The term **standard score** is often used instead of z-score.

Example 1.3.5.1.1

For example, the average age of all presidents at inauguration is $\mu = 55$ and the standard deviation is $\sigma = 6.5$. Consider the age of Lincoln, 52 and the age of Eisenhower, 62.

$$z_L = \frac{52-55}{6.5} = -0.46$$

and

$$z_L = \frac{62-55}{6.5} = 1.08$$

A negative z-score indicates that the observation is below (less than) the mean, whereas a positive z-score indicates that the observation is above (greater than) the mean. The meaning of the z-score of an observation is the number of standard deviations that the observation is away from the mean.

Example 1.3.5.1.2

We can see that Lincoln's age 52 is 0.42 standard deviations below the mean and Eisenhower's age 62 is 1.08 standard deviations above the mean.



Unlike the measures of center and variation the measures of relative standing are individual measures of each observation. The purpose of it is to identify the relative position of each observation in the entire data set in relation to other observations. The z-

score of an observation, therefore, can be used as a rough measure of its relative standing. For instance:

- a z-score of 3 or more indicates that the observation is larger than most of the other observations.
- a z-score of −3 or less indicates that the observation is smaller than most of the other observations.
- a z-score near 0 indicates that the observation is located near the mean.

If two distributions have the same shape or, more generally, if they differ only by center and variation, then z-scores can be used to compare the relative standings of two observations from those distributions. Consider the following example!

Example 1.3.5.1.3

Arthur scored 92 in the class with the mean 80 and the standard deviation 6, but Bethany scored 90 in the class with the mean 78 and the standard deviation 4. Who scored relatively better?

Solution

The Arthur's z-score is $\frac{92-80}{6}=2$ and the Bethany's z-score is $\frac{90-78}{4}=3$. Since Bethany's z-score is higher, Bethany did relatively better than Arthur despite the fact that her exam score was lower than Arthur's.

A set consisting of the z-scores of all observations is called the **standardized** dataset. For example, consider the following population:

Data, x_i
72
73
76
76
78

which has the mean 75 and standard deviation 2.2. By replacing each value with its z-score we obtain a new set called standardized:

Standardized Data, $\frac{x_i-\mu}{\sigma}$
$\frac{72-75}{2.2}=-1.36$
$\frac{73-75}{2.2}=-0.91$
$\frac{76-75}{2.2}=0.45$
$\frac{76-75}{2.2}=0.45$
$\frac{78-75}{2.2}=1.36$

Note that a standardized data set always has mean 0 and standard deviation 1.

One way to improve our interpretation skills of z-scores is to rephrase the rules from the previous section using the z-score language. For example:

- any observation with the z-score less than -3 or greater than 3 is an outlier.
- any observation with the z-score less than -2 is significantly low.
- any observation with the z-score more than 2 is significantly high.

By Chebyshev's Rule, for any data set with mean μ and standard deviation σ:

- At least 75% of observations have z-scores between -2 and 2;
- At least 89% of observations have z-scores between -3 and 3;
- At least 93.75% of observations have z-scores between -4 and 4.

By the Empirical Rule, for any bell-shaped data set with mean μ and standard deviation σ:

- Approximately 68% of observations have z-scores between -1 and 1.
- Approximately 95% of observations have z-scores between -2 and 2.
- Approximately 99.7% of observations have z-scores between -3 and 3.

We discussed the z-scores as one of the ways to measure relative standing - how to compute it and how to interpret it.

Section 2: Percentiles



Next, we will discuss another measure of relative standing that allows us to rank each observation in the data set. Another way to measure the relative standing is computing the percentiles. The p-th **percentile** is a value that separates the bottom p% of the data from the remaining values in the top of the data. To find the percentile of a particular observation x_i, we use the following formula:

$$\text{percentile of } x_i = \frac{\text{the number of values} < x_i}{\text{total number of values}} \cdot 100\%$$

Example 1.3.5.2.1

Consider the following population and before we proceed any further make sure that the data is organized in ascending order:

{72, 73, 76, 76, 78}

The percentile of x_3 is $\frac{2}{5} \cdot 100\% = 0.4 \cdot 100\% = 40$-th percentile.

To find the p-th percentile of a data set, we use the following formula:

$$x_p = x_{[i=\frac{p}{100} \cdot n]}$$

In this formula, n is the size of the data set, p is the desired percentile, and i is the index of the observation that we are trying to find!

Example 1.3.5.2.2

Consider the following population and before we proceed any further make sure that the data is organized in ascending order:

{72, 73, 76, 76, 78}

The 30-th percentile is $x_{[\frac{p}{100} \cdot n]} = x_{[\frac{30}{100} \cdot 5]} = x_{[1.5]} = x_2 = 73$.

Certain percentiles are particularly important:

- the **deciles** divide a data set into tenths (10 equal parts).

- the **quartiles** divide a data set into quarters (4 equal parts). Quartiles are the most commonly used percentiles. A data set has three quartiles, which we denote Q1, Q2, and Q3. The second quartile is also the median of the data set.

Now that we know two different ways to measure a relative standing of every observation the question is which one is better?

Percentiles usually give a more accurate and meaningful measure of relative standing than z-scores. However, to compute the percentile one needs to have access to the entire data set while to compute the z-score one needs to know only the mean and standard deviation of the data set. With very little information, z-scores provide a feasible alternative to percentiles for measuring relative standing. For example, for a student to be able to compute the percentile of their exam score the instructor must allow the access to the entire list of scores which is impossible, but to compute the z-score the instructor must share only the mean and the standard deviation which is something instructors are more comfortable doing.

This page titled 1.3.5: Measures of Relative Standing is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.3.6: Five-Number Summaries and Boxplots



Previously, we learned how to use the mean and the standard deviation of a data set to figure out the shape of the distribution and the outliers. Next, we will learn how to do the same using the other numerical summaries!

Definition: Five-number suymmary

The values Q0, Q1, Q2, Q3, and Q4 together are called **the five-number summary**.

Alternatively, we can think of the list as the 0th, 25th, 50th, 75th, 100th percentiles:

Q0 (Minimum)	Q1	Q2 (Median)	Q3	Q4 (Maximum)
0th percentile	25th percentile	50th percentile	75th percentile	100th percentile

Before finding the five-value-summary by hand, we have to make sure that the data is organized in ascending order. Then it will be easier to find the minimum and maximum. We already know how to find the median that divides the data set into two halves – top and bottom. For simplicity, we are going to define the Q1 as the median of the bottom half and Q3 as the median of the top half.

- Q0 (Minimum) – the smallest observation
- Q1 – the median of the bottom half
- Q2 (Median) – the middle observation
- Q3 – the median of the top half
- Q4 (Maximum) – the largest observation

Example 1.3.6.1

Consider the ages of the US presidents at their inauguration. For convenience it is already organized in ascending order.

42	43	46	47	47	47	48	49	49
50	51	51	51	51	51	52	52	54
54	54	54	54	55	55	55	55	56
56	56	57	57	57	57	58	60	61
61	61	62	64	64	65	68	69	70

So let's determine the five-number summary:

- The minimum is 42 and the maximum is 70.
- The number of observations is 45 which is an odd number, so the median is 55 - the 23rd observation that divides the data into upper 22 observations and lower 22 observations.
- In the bottom half, the number of observations is 22 which is an even number, so the median of the bottom half is the average between the 11th and 12th observations, which is 51.
- Similarly In the upper half, the number of observations is 22 which is an even number, so the median of the upper half is the average between the 11th and 12th observations, which is 59.

The five-number summary is provided in the table:

Min	Q1	Median	Q3	Max
42	51	55	59	70

One of the two goals that we are trying to accomplish is to learn how to identify the outliers. For that we are going to need the following vocabulary.

Definition: Interquartile Range (IQR)

The **interquartile range** (IQR) is the difference between the Q3 and Q1, or in other words, it is the width of the middle 50%.

Definition: Lower and upper limits and potential outliers

The values Q1 − 1.5IQR and Q3 + 1.5IQR are called the **lower limit** and **upper limit** of a data set. Values that are greater than upper limit or less than the lower limit are **potential outliers**.

The other goal that we are trying to accomplish is to learn how to visualize a data set. For that we are going to need the following vocabulary.

Definition: Boxplot (box-and-whiskers diagram)

A **boxplot**, also called a **box-and-whisker diagram**, is based on the five-number summary, and can be used to provide a graphical display of the center and variation of a data set. To construct a boxplot, we also need the concept of **adjacent values**. The adjacent values of a data set are the most extreme observations that still lie within the lower and upper limits; they are the most extreme observations that are not potential outliers. Note that, if a data set has no potential outliers, the adjacent values are just the minimum and maximum observations.

To Construct a Boxplot:

1. Determine the quartiles.
2. Determine potential outliers and adjacent values.
3. Draw a horizontal axis on which the numbers obtained in Steps 1 and 2 can be located. Above this axis, mark the quartiles and the adjacent values with vertical lines.
4. Connect the quartiles to make a box, and then connect the box to the adjacent values with lines.
5. Plot each potential outlier with an asterisk.

Note: One can skip steps 2 and 5 if not concerned about the outliers. In such a case the adjacent values are the minimum and maximum.

Example 1.3.6.2

For example, let's construct the boxplot for the president's ages at inauguration for which we already found the five-number summary.

IQR=Q3-Q1=59-51=8

1.5xIQR=1.5x8=12

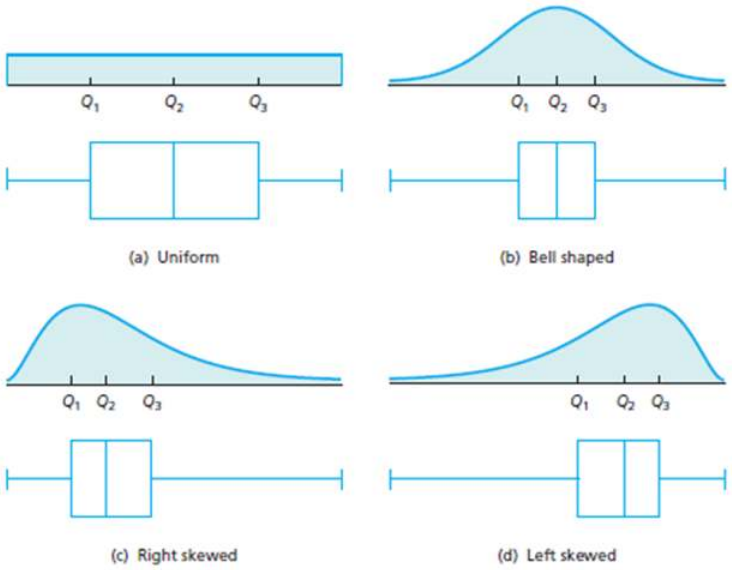
LL=Q1-1.5xIQR=51-12=39

UL=Q3+1.5xIQR=59+12=71

Since all the values are within the lower and upper limits there are no outliers. We draw the boxplot by creating the horizontal axis and drawing the vertical lines for each value in the five-number summary:



After the boxplot is constructed, we can check the following chart to identify the shape of the distribution:



Example 1.3.6.3



It appears that the shape of the distribution of the president's ages is normal!

We discussed how to use the five-number summary to identify the outliers and to visualize the data by constructing a boxplot.

This page titled 1.3.6: Five-Number Summaries and Boxplots is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

SECTION OVERVIEW

1.4: Probability Concepts

1.4.1: Introduction to Probability

1.4.2: Finding the Probability

1.4.3: Probability Rules

1.4.4: Probabilities from Frequency Tables

1.4.5: Conditional Probability and Multiplication Rules

This page titled 1.4: Probability Concepts is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.4.1: Introduction to Probability



The world is full of uncertainties. What are the chances that a student will pass a class? What are the chances that the price of a stock will go up? In this section we will introduce a framework which lays the foundation for the future study of statistics.

To quantify the uncertainties around us we use **probability** as a measure of how likely something is to happen. The more likely it is to happen the higher its probability, and the less likely something to happen the less its probability. We have to start with a few definitions.

Definition: Experiment

An **experiment** is an action whose result is not certain.

Here are some examples of actions whose outcomes are not certain:

Experiments
Flipping a coin
Flipping a coin twice (or flipping two distinct coins simultaneously)
Flipping a coin thrice (or flipping three distinct coins simultaneously)
Rolling a die
Rolling a die twice (or rolling a pair of dice simultaneously)
Rolling a die thrice (or rolling three distinct dice simultaneously)
Drawing a card from a standard 52-card deck
Spinning a roulette wheel

Definition: Simple outcome

Every experiment results in some **simple outcome**.

Here are some examples of simple outcomes for each of the experiments:

Experiments	Simple Outcome

Experiments	Simple Outcome
Flipping a coin	H
Flipping a coin twice	HT
Flipping a coin thrice	TTH
Rolling a die	3
Rolling a die twice	2-4
Rolling a die thrice	5-1-4
Drawing a card	QH
Spinning a roulette	29

Definition: Sample space

The set of all outcomes of an experiment is called the **sample space** of the experiment.

Here are the sample spaces for each of the experiments:

Experiments	Simple Outcome	Sample Space
Flipping a coin	H	{H, T}
Flipping a coin twice	HT	{HH, HT, TH, TT}
Flipping a coin thrice	TTH	{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}
Rolling a die	3	{1,2,3,4,5,6}
Rolling a die twice	2-4	{1-1,1-2,1-3, etc.}
Rolling a die thrice	5-1-4	{1-1-1,1-1-2,1-1-3, etc.}
Drawing a card	QH	{2H,3H,4H,5H, etc.}
Spinning a roulette	29	{00,0,1,2,3,4, etc.}

Not every sample space is small enough that we can list all outcomes, but when it is the case, we need to know how to efficiently write the entire sample space.

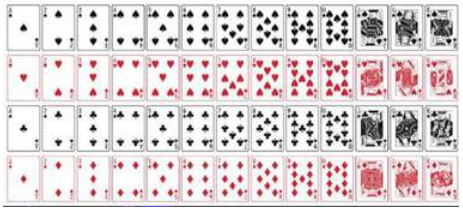
For tossing coins there is a distinct pattern.

- When tossing a coin, we can get either Heads or Tails
- When tossing two coins, we can get for the first toss either Heads or Tails but now it can be combined with either Heads or Tails for the second toss.
- When tossing three coins, we can get for the first two tosses the same pattern as we just listed for tossing two coins but now this pattern can be combined with either Heads or Tails for the third toss.

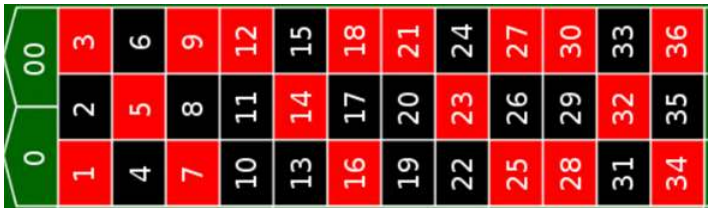
For rolling dice there is another pattern.

- When rolling a die, the outcome is the number of dots that can be any number 1 through 6.
- When rolling two dice, the outcome is a pair of numbers that represent the number of dots on die 1 and die 2. There are 36 possible outcomes and the best way to list them is to organize it in the table.
- There is not enough space here to list ALL 216 possible outcomes for rolling three dice, but we can easily imagine what each outcome looks like.

When drawing a card from a standard 52-card deck, there are 52 possible outcomes – one for each distinct card in the deck:



When spinning an American roulette, there are 38 possible outcomes – one for each of the numbers 1 through 36 and 0 and 00. The best way to visualize it is to see the roulette table:



Definition: Event

An **event** is a collection of outcomes from the sample space.

For example, consider rolling a die for which the sample space is {1,2,3,4,5,6}. Any collection of simple outcomes from the list will define an event. Here are a few examples of sets created by including some of the outcomes from the sample space:

Event	Set Description
A	{2, 4, 6}
B	{1, 3, 5}
C	{5, 6}
D	{1, 2, 3, 4, 5, 6}
E	{ }

In probability we use the capital letters from the beginning of the alphabet to denote the events - A, B, C, D, E.

Some events may have a nice verbal description in English. Which of the listed events can be verbally described as “rolling an even number”? Naturally, we would say that the list {2,4,6} matches that description. Similarly, we can label the other events in the following way:

Event	Verbal Description	Set Description
A	Roll an even number	{2, 4, 6}
B	Roll an odd number	{1, 3, 5}
C	Roll a number greater than or equal to 5	{5, 6}
D	Roll a positive number	{1, 2, 3, 4, 5, 6}
E	Roll a number greater than 7	{ }

However, the majority of events will not have such a description. There are exactly sixty-four events that can be defined from the sample space of size 6. Most of them will not have an intuitive verbal description, e.g., try to come up with a verbal description of an event $F = \{2, 3, 5\}$? This is why it is important to learn how to work with the definition of an event rather than rely on its verbal representation.

Definition: Event occurred vs didn't occur

After finishing the experiment, we say an event **occurred** if the outcome of an experiment is in the collection that defines an event. We say an event **did not occur** if the outcome of an experiment is not in the collection that defines an event.

For each event and each outcome, we can decide whether an event has occurred or not. For example, when you roll 1 we say B and D occurred and A, C, E didn't occur. Similarly, for event A we can say that it occurs when 2,4, or 6 are rolled and doesn't occur when 1,3, or 5 are rolled. The rest of the table looks like this:

Event	Result is 1	Result is 2	Result is 3	Result is 4	Result is 5	Result is 6
$A = \{2, 4, 6\}$	X	V	X	V	X	V
$B = \{1, 3, 5\}$	V	X	V	X	V	X
$C = \{5, 6\}$	X	X	X	X	V	V
$D = \{1, 2, 3, 4, 5, 6\}$	V	V	V	V	V	V
$E = \{\}$	X	X	X	X	X	X

It is important to understand why and when the event occurs and doesn't occur depending on the outcome of the experiment.

Notice that in the ongoing example event $D = \{1, 2, 3, 4, 5, 6\}$ is the entire sample space. As strange as it may look event D satisfies the definition and therefore has to be treated as an event. Since D occurs for any outcome of the experiment, we call D a **certain** event. Now, take a look at event $E = \{\}$ which is the empty set. As strange as it may look event E also satisfies the definition and therefore has to be treated as an event. Since E occurs for no outcome of the experiment, we call E an **impossible** event. Let me remind you again that this is a small sample of events from all sixty-four possible events.

Hopefully by now you can see the importance of knowing how to work with the set description. Alternatively, one can think of an event as something that someone can make a bet on happening. For example, when you throw a die, someone can walk in and say I bet the number will be less than 4, and we know exactly which outcomes will win the bet and which outcomes will lose the bet.

Now, once we have the precise definition and informal definition of an event, we can define the probability of an event.

Definition: Probability

Recall that probability is a measure of how likely an event is to occur. We denote the probability of an event E as $P(E)$. We define the probability of an **impossible** event to be equal to zero, and the probability of a **certain** event to be equal to one. Nothing can be more likely than a certain event and nothing can be less likely than an impossible event, therefore the following inequality must be true for any event E:

$$0 \leq P(E) \leq 1$$

Next, we will discuss different ways to compute the probability of an event.

This page titled 1.4.1: Introduction to Probability is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.4.2: Finding the Probability

Section 1: Probability of an Event



We already know that the probability of an impossible event is zero, and the probability of a certain event is one. As a result, the probability of any other event must be between 0 and 1. The question that we will try to answer next is about how to find the probability of any event. Before we start answering this question, we need to understand that for every event there is an experiment for which there exists the sample space which consists of all possible simple outcomes using which we can define an event as a set of simple outcomes.

Definition: Equally vs not equally likely outcomes

Consider the following two examples: tossing a coin and driving to work, both experiments have two simple outcomes – tossing a coin may result in heads or tails and driving to work may result in an accident or not. But there is one conceptual difference between these two experiments. To see the difference let us answer the following questions: Is there a reason to believe that the heads are more or less likely than tails? And is there a reason to believe that to get in an accident is as likely as not to get in an accident? Turns out that there is no reason to believe that when tossing a fair coin, the chances of getting heads is any different from getting tails. Such experiment is said to have **equally likely outcomes**. On the other hand, there is no reason to believe that the chances of getting in an accident are the same as chances of not getting in an accident. Such experiment is said to have **not equally likely outcomes**.

Example 1.4.2.1

Consider another two examples: rolling a die and counting credit cards in the wallet of a randomly selected person, both experiments have 6 simple outcomes – rolling a die may result in any number between 1 and 6 and the number of credit cards in the wallet of a randomly selected person can vary from zero to any number. Again, there is one conceptual difference between these two experiments. To see the difference let us answer the following questions: Is there a reason to believe that 1 or 3 are more or less likely than 4 or 6? And is there a reason to believe that it is as likely to have 0 cards as more than 5? It is obvious, that there is no reason to believe that when rolling a fair die the chances of getting any number are different from any other number. On the other hand, there is no reason to believe that the chances of having no credit cards are the same as chances of having more than 5.

Most classical experiments have equally likely simple outcomes but in real life most of the experiments do not have equally likely outcomes. So, it is important to be able to work with both types of experiments. When working with an experiment that has equally

likely simple outcomes, we will use an approach called **classical**, and when working with an experiment that has not equally likely simple outcomes, we will use an approach called **empirical**.

Classical Approach for Equally Likely Outcomes	Empirical Approach for Not Equally Likely Outcomes
• Flipping a coin • Flipping a coin twice • Rolling a die • Rolling a die twice • Drawing a card • Spinning a roulette wheel	• Driving (accident/no accident) • Driving (on-time/late) • Counting credit cards (0,1,2,3+) • Counting household size (1,2,3+) • Taking a class (pass/fail) • Taking a class (A/B/C/D/F)

Section 2: The Classical Approach



Next, we will discuss the classical approach which is a method of finding the probability of an event defined for an experiment with equally likely simple outcomes. We already understand that for every event there is an experiment for which there exists the sample space which consists of all possible simple outcomes. In this section, we consider only those experiments for which all simple outcomes are equally likely. In such a case, to find the probability we use the following formula:

$$P(E) = \frac{f}{N}$$

where E is an event, f is the number of ways that event E may occur or the size of the set that defines the event, and N is the total number of outcomes which is the size of the sample space.

So, to find the probability of an event we can just count the number of outcomes in the set definition and divide it by the size of the sample space.

Example 1.4.2.2.1

Find the probabilities of the following events when a fair coin is flipped three times.

a. There are two Hs
b. There are no Ts
c. There is an odd number of Ts
d. There is at least one H

Solution

To use the formula, we first need to identify the experiment and its sample space. The experiment is tossing a coin three times which has a well-known sample space of size 8:

Experiment	Sample space	Size
flipping a coin thrice	{TTT, TTH, THT, HTT, THH, HTH, HHT, HHH}	$N = 8$

For every event we will find the set description of an event that is a list all simple outcomes that define it, figure out the size and divided it by 8 - the size of the sample space.

For example, for event A there are 3 outcomes: THH, HTH, HHT, therefore the probability of event A is

$$P(A) = \frac{3}{8} = 0.375$$

Similarly, for event B there is only one outcome HHH, therefore the probability of event B is

$$P(B) = \frac{1}{8} = 0.125$$

Similarly, for event C there are 4 outcomes THH, HTH, HHT, TTT, therefore the probability of event C is

$$P(C) = \frac{4}{8} = 0.500$$

Finally, for event D there are 7 outcomes, therefore the probability of event D is

$$P(D) = \frac{7}{8} = 0.875$$

As one can see by the definition of the method it is impossible to get a number less than zero or greater than one. This result validates our choice of denoting the probability of impossible events as zero and probabilities of certain events as one.

We discussed the classical approach which allows to find the probability of an event defined for an experiment with equally likely simple outcomes.

Section 3: The Empirical Approach

As we previously said, in real life most of the experiments do not have equally likely outcomes, so what do we do? Next, we will discuss the empirical approach which is a method of finding the probability of an event defined for an experiment with not equally likely simple outcomes. We already understand that for every event there is an experiment for which there exists the sample space which consists of all possible simple outcomes.

Let our experiment consist of throwing a cone in the air and recording how it is landed.



A cone is a simple three-dimensional object, that can only land in two positions: upright (U) or sideways (S). The sample space of the experiment has two simple outcomes U and S. It is pretty clear that these outcomes are not equally likely, so the classical approach wouldn't apply here. We have a strong intuition that the probability of the cone landing on its side is more than the probability of it landing upright. The goal is to find the probabilities for each of the outcomes!

One way to compute the probability of each outcome is to look at the past experience of tossing a cone. For instance, in this case, last time I flipped a cone 100 times, I had the following frequency distribution table:

Outcome	Frequency
S	94
U	6

None of these numbers look like they can represent the probability of an event but if we construct the relative frequency table then we will start seeing that the relative frequencies look a lot like probabilities? Aren't they?

Outcome	Frequency	Relative Frequency	Probability
S	94	0.94	P(S)
U	6	0.06	P(U)
Total	100	1	

In summary, the Empirical approach consists of performing the experiment many times and constructing the relative frequency distribution table and then interpreting the relative frequencies as the approximations of the probabilities of the corresponding events. In our example, 0.94 is approximately the probability that a cone will land on its side, and 0.06 is approximately the probability that a cone will land upright. The empirical approach is also called the relative frequency approach.

To explain why the relative frequency is only an approximation of the true probability, let's consider tossing a coin which also may result in only two outcomes. This time the probability of each outcome is known, both are 0.5. Last time I flipped a fair coin 100 times I observed the following frequencies and relative frequencies:

Outcome	Frequency	Relative Frequency	Probability
H	47	0.47	0.50
T	53	0.53	0.50
Total	100	1	

Since we know that the true probability of a coin showing heads is 0.5, we conclude that the relative frequency 0.47 is only an approximation to the true value. Now if we toss a coin 1000 times, we would see the following relative frequency distribution table in which the relative frequencies are much closer to the true values of the probabilities:

Outcome	Frequency	Relative Frequency	Probability
H	487	0.487	0.50
T	513	0.513	0.50
Total	1000	1	

> This phenomenon is known as **The Law of Large Numbers** states that the larger the number of experiments the closer the relative frequencies approximate to the true value.

Example 1.4.2.3.1

Estimate the probability that a randomly selected student owns 2 cars.

Solution

To answer this question using the empirical approach I would have to check on my past experience with the students owning cars. Last time I asked students how many cars they own I received the following frequencies and relative frequencies.

Number of cars	Frequency	Relative Frequency
0	6	.24
1	16	.64
2	2	.08
3	0	.00

Number of cars	Frequency	Relative Frequency
4+	1	.04
Total	25	1

Based on the Empirical approach, we may conclude that the probability that a randomly selected student owns two cars is approximately 8%.

Note that the relative frequencies are only approximations. Knowing that we would not say that the probability of a student owning 3 cars is zero, we would say it is a very small probability. Also, it is worth mentioning that the more representative the sample is of the larger population the closer these approximations are for the entire population. For example, if we were to assume that students are representative of the entire population, we can make a statement that the probability of a random person owning two cars is approximately 8%.

So how does one interpret the probability of an event? Let us say the probability of raining tomorrow is 0.2. What does it mean? Or how was such number produced?

To interpret we must understand where the number came from. If the empirical approach was used then in the past, it rained 20 out of 100 days that were just like tomorrow, so the relative frequency and the probability are both 0.2. We interpret this as, in the future, it will also rain 20 out of 100 days or 2 out of 10 or 1 out of 5 on average. The problem with such interpretation is the implied assumption that the experiment will continue behaving the same way as in the past, which may not be necessarily the case.

Note that in general we may not know the true probability of an event, which is the case with many experiments whose outcomes are not equally likely, but we just discussed the empirical approach which is a great way to approximate it.

Section 4: Probability and Odds



Historically the probability was developed to provide the support for gambling. And historically an alternative way was used to measure the likelihood of an outcome called odds.

Definition: Odds

The odds of an event E is the ratio of the number of ways the event may occur to the number of ways that the event may not occur:

$$o(E) = |E| : |\text{no } E|$$

For an experiment with a known sample space and event E is defined by the list of x outcomes from the sample space then the remaining y outcomes are not used in the definition of E. We say the odds in favor of E are "x to y", and we write $x : y$ and the odds against E are "y to x", and we write $y : x$.

Example 1.4.2.4.1

Find the odds that there are exactly two heads out of two tosses of a fair coin.

Solution

In our experiment of tossing a coin two times the sample space is the following:

{HH, HT, TH, TT}

Let E be an event in which we have exactly two heads out of two tosses of a fair coin. Only one outcome makes up the event E and the remaining three are not E. We say the odds in favor of E are 1:3, and against E are 3:1.

Often the odds are reported as a simplified ratio, that is the ratio in which both sides are reduced by the common factor. The odds in favor of getting an odd number when rolling a die can be expresses as 1:1 instead of 3:3 and the odds in favor of rolling less than 3 when rolling a die can be expressed as 1:2 instead of 2:4.

> When there is more than one way to express the same concept, one must know how to convert from one way to another.

If the odds in favor of an event E are $x : y$, then the probability of E is $\frac{x}{x+y}$ and this table summarizes the conversion rule:

$o(E)$	$P(E)$
$x : y$	$\frac{x}{x+y}$

Example 1.4.2.4.2

If $o(E) = 3 : 1$, then $P(E) = \frac{3}{3+1} = 0.75 = 75\%$.

If the probability of an event E is $p\%$ then the odds in favor of E are $p : (100 - p)$ and this table summarizes the conversion rule:

$P(E)$	$o(E)$
$p\%$	$p : (100 - p)$

Example 1.4.2.4.3.1

If $P(E) = 20\%$, then $o(E) = 20 : (100 - 20) = 20 : 80 = 1 : 4$.

Alternatively, if the probability of an event E is given by a fraction $\frac{a}{n}$ then the odds in favor of E are $a : (n - a)$ and this table summarizes the conversion rule:

$P(E)$	$o(E)$
$\frac{a}{n}$	$a : (n - a)$

Example 1.4.2.4.3.2

If $P(E) = \frac{4}{7}$, then $o(E) = 4 : (7 - 4) = 4 : 3$.

Although it is pretty straightforward how to use odds, the only reason we covered it is to pay respect to the historical relation of probability to gambling.

This page titled 1.4.2: Finding the Probability is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.4.3: Probability Rules

Section 1: Compound Events



Next, we will introduce the ways to create new events from existing ones. First, let's discuss a way to create a new event from just a single event.

Definition: Complement of an event

Let E be an event then a new event, called the **complement of** E, can be formed from all the outcomes that are not in E. The complement of E can be labeled in many ways:

$$\text{not } E \text{ or } E^c \text{ or } E' \text{ or } \bar{E}$$

Visually we can identify the complement of E by writing out the sample space and collecting all the outcomes that are not in E.

Consider the experiment of rolling a fair die with the well-known sample space:

$$\{1, 2, 3, 4, 5, 6\}$$

Example 1.4.3.1.1

Consider the following events A, B, and C defined by the following sets:

$$A = \{1, 5\}$$

$$B = \{2, 4, 5\}$$

$$C = \{2, 3, 4, 6\}$$

Then using the definition of a complement we can find the complements of A, B, and C:

$$(\text{not } A) = \{2, 3, 4, 6\}$$

$$(\text{not } B) = \{1, 3, 6\}$$

$$(\text{not } C) = \{1, 5\}$$

Next, we will discuss ways to create a new event from a pair of events.

Definition: Intersection of two events

Let A and B be two events then a new event, called the **intersection of A and B**, can be formed from all the outcomes that are in both A and B. The intersection of A and B can be labeled in a few different ways:

$$(A \text{ and } B) \text{ or } (A \cap B)$$

Visually we can identify the intersection of A and B by writing out the sample space labeling A and B and collecting all the outcomes that are in both A and B.

Example 1.4.3.1.2

Using the definition of an intersection we can find the intersections $(A \text{ and } B)$, $(B \text{ and } C)$, and $(C \text{ and } A)$:

$$(A \text{ and } B) = \{5\}$$
$$(B \text{ and } C) = \{2, 4\}$$
$$(C \text{ and } A) = \{\ \}$$

Here is another way to create a new event from a pair of events.

Definition: Union of two events

Let A and B be two events then a new event, called the **union of A and B**, can be formed from all the outcomes that are in A, in B, and in both A and B. The union of A and B can be labeled in a few different ways:

$$(A \text{ or } B) \text{ or } (A \cup B)$$

Visually we can identify the union of A and B by writing out the sample space labeling A and B and collecting all the outcomes that are in A, B, or both.

Example 1.4.3.1.3

Using the definition of a union we can find the unions $(A \text{ or } B)$, $(B \text{ or } C)$, and $(C \text{ or } A)$:

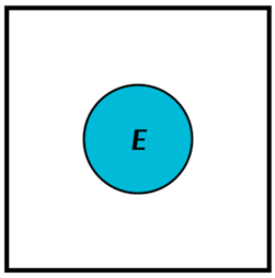
$$(A \text{ or } B) = \{1, 2, 4, 5\}$$
$$(B \text{ or } C) = \{2, 3, 4, 5, 6\}$$
$$(C \text{ or } A) = \{1, 2, 3, 4, 5, 6\}$$

We introduced and discussed the idea of creating new events from an existing one or two events.

Section 2: Venn Diagrams



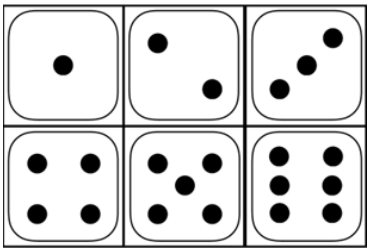
We can visualize events by first drawing all possible outcomes, and then circling the ones that are in the set description of the event. Such picture is called a **Venn diagram**. Any shape drawn on the sample space will define some event.



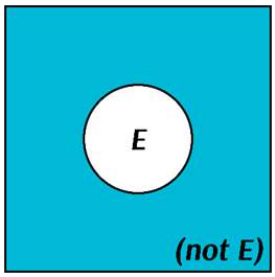
Each Venn diagram has the following properties:

- the box represents the sample space.
- the area of the region is $P(E)$.
- the area of the box is 1.

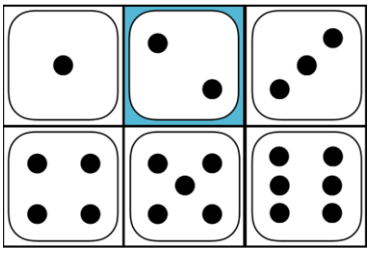
Example: This is what the Venn diagram of the sample space for rolling a die look like.



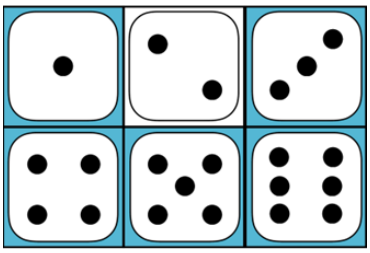
Just like the compound events are derived from events, the Venn diagrams of compound events can be derived from the Venn diagram(s) of the original event or events. For example, this is what the Venn diagram for the complement of E looks like:



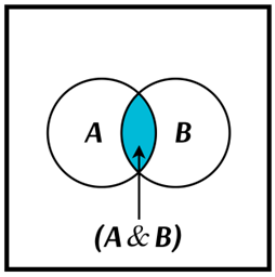
For example, let E be an event in which we roll a die and get 2 then this what the Venn diagram of the event looks like:



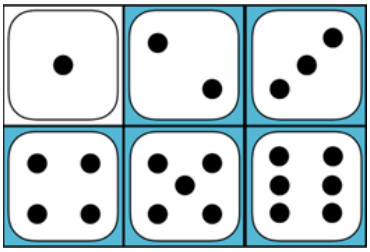
The Venn diagram for the complement of E can be easily derived from the Venn diagram of E:



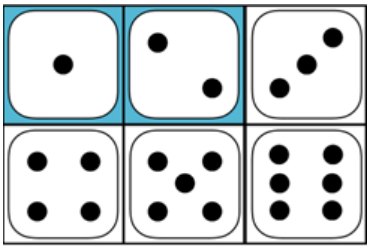
This is what the Venn diagram for the intersection of A and B looks like:



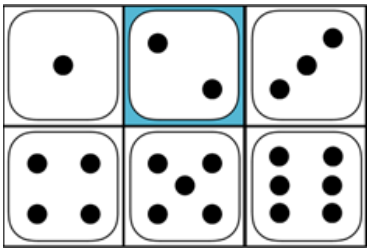
For example, let A be an event in which we roll a die and get anything but 1 with the following Venn diagram:



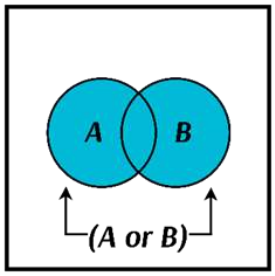
And let B be an event in which we roll a die and get a number less than 3 with the following Venn diagram:



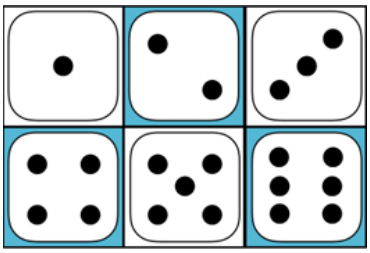
The Venn diagram for the intersection of A and B can be easily derived from the Venn diagrams of A and B:



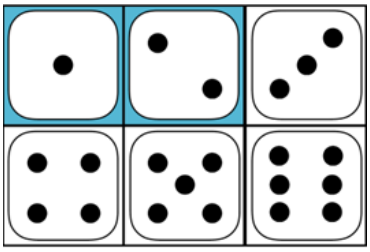
This is what the Venn diagram for the union of A and B looks like:



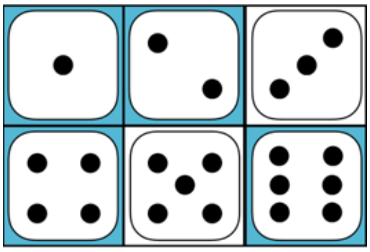
For example, let A be an event in which we roll a die and get an even number with the following Venn diagram:



And let B be an event in which we roll a die and get a number less than 3 with the following Venn diagram:



The Venn diagram for the union of A and B can be easily derived from the Venn diagrams of A and B:



We learned how to visualize the events using the concept of a Venn diagram.

Section 3: Probability Rules



Next, we will discuss how to find the probabilities of compound events if the probability of the original event or events are given.

Definition: Mutually exclusive events

Mutually exclusive (or disjoint) **events** are the events that have no common outcomes.

Example 1.4.3.3.1

- In the experiment of rolling a die with the well-known sample space, let A be an event $\{1,3\}$, B be an event $\{4,6\}$ and C be an event $\{1,3,6\}$. A and B are said to be mutually exclusive or disjoint because they have no common outcomes. The same cannot be said about the events A and C or B and C because they share at least one common outcome.
- In the experiment of tossing two coins with the well-known sample space, let A be an event $\{HH,HT\}$, B be an event $\{HH,TT\}$ and C be an event $\{TT\}$. A and C are said to be mutually exclusive or disjoint because they have no common outcomes. The same cannot be said about the events A and B or B and C because they share at least one common outcome.

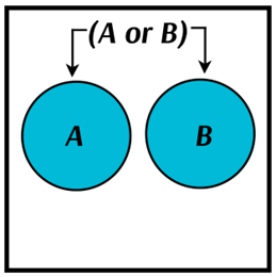
A few more remarks about mutually exclusive events:

- A pair of events are mutually exclusive if their intersection is empty, and vice versa. This is the only way to find out whether a pair of events are mutually exclusive or not.
- An event E and its complement are mutually exclusive, because $P(E and \bar{E}) = 0$.

The purpose of developing the probability rules is to be able to determine the probabilities of the complements, the union, and the intersection of events A and B from the probabilities of the event A and B, that is we want to be able to find the probabilities of compound events if the probabilities of the original events are given.

Theorem: Special Addition Rule

Consider the union of two mutually exclusive events A and B as it is shown on the Venn diagram:



Recall that the areas of the regions represent the probabilities of the corresponding events. So the area of the union A and B is equal to the sum of the areas of A and area A in other words

$$P(A \text{ or } B) = P(A) + P(B)$$

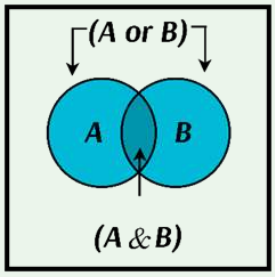
This relation is called the **special addition rule**. What makes it special is the fact that it only applies to mutually exclusive events.

Example 1.4.3.3.2

- If A, B are disjoint and $P(A) = 0.2$, $P(B) = 0.5$ then $P(A \text{ or } B) = 0.2 + 0.5 = 0.7$.
- If C, D are disjoint and $P(C) = 0.3$, $P(D) = 0.6$ then $P(C \text{ or } D) = 0.3 + 0.6 = 0.9$.

Theorem: General Addition Rule

Consider the union of two events A and B that are not necessarily mutually exclusive as it is shown on the Venn diagram:



Notice that the special addition rule doesn't work in this case because the combined area of A and B exceeds the area of the union. The good news is that we know exactly the difference between the two, so the area of the union A and B is equal to the sum of the areas of A and area B minus the area of the intersection, in other words:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

This relation is called the **general addition rule**. What makes it general is the fact that it applies regardless of whether the events are mutually exclusive or not.

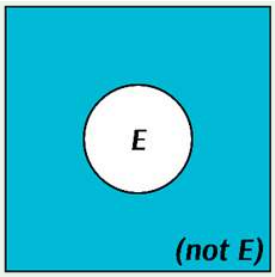
Example 1.4.3.3.3

- If $P(A \text{ and } B) = 0.1$, $P(A) = 0.2$, $P(B) = 0.5$ then $P(A \text{ or } B) = 0.2 + 0.5 - 0.1 = 0.6$.
- If $P(C \text{ and } D) = 0.2$, $P(C) = 0.3$, $P(D) = 0.6$ then $P(C \text{ or } D) = 0.3 + 0.6 - 0.2 = 0.7$.

Finally, we will discuss the complementary rule.

Theorem: Complementary Rule

Consider an event E and its complimentary event as it is shown on the Venn diagram:



Recall that the areas of the regions represent the probabilities of the corresponding events and since the total area is equal to 1 it is fair to say that

$$P(E) + P(\text{not } E) = 1$$

This relation is called the **complementary rule**. We can use this relation to find the probability of the complementary event of E by subtracting the probability of E from 1.

Example 1.4.3.3.4

- If $P(\text{Rain}) = 0.2$ then $P(\text{Not Rain}) = 1 - 0.2 = 0.8$.
- If $P(\text{Late}) = 0.01$ then $P(\text{Not Late}) = 1 - 0.01 = 0.99$.

We discussed the complementary rule, and the special and general addition rules that allow us to find the probability of the complementary event and the union of two events.

This page titled 1.4.3: Probability Rules is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.4.4: Probabilities from Frequency Tables

Section 1: Using Relative Frequency Tables



In this section, we are going to learn how to make probability statements using relative frequency distribution tables for univariate data. Let us assume that we have access to a huge database that has most of the information about the past students. Let us imagine that we requested the data and received the following frequency distribution table for a random 100 students.

Number of years	Frequency	Relative Frequency
0-2	27	0.27
2-4	42	0.42
4-6	20	0.20
6+	11	0.11
Total	100	1

We already know that the frequencies can't be interpreted as probabilities, so we turn it into the relative frequency distribution table. We have an intuition that the relative frequencies can and SHOULD be interpreted as probabilities, but we also know that in order to find the probability we must have a well-defined event which is defined for some experiment with its sample space. The question that we need to answer first is what are the experiment, the simple outcomes, the sample space, and what kind of events we can find the probabilities of when such frequency distribution table is given.

First let's define the experiment! In this case, the experiment consists of randomly selecting one of the 100 students from the sample. Based on the formatting of the table there are 4 simple outcomes 0-2, 2-4, 4-6, 6+. We label them Y1, Y2, Y3, and Y4 and these 4 outcomes form the sample space for the experiment.

$$\{Y_1, Y_2, Y_3, Y_4\}$$

We describe the outcome Y1 as an outcome in which the randomly selected student spent between 0 and 2 years in college. We describe the outcome Y2 as an outcome in which the randomly selected student spent between 2 and 4 years in college. Hopefully, it is clear now how the simple outcomes are defined.

Now that we defined the simple outcomes, we can interpret the relative frequencies as the probabilities. For example, 0.27 is the $P(Y_1)$, 0.42 is the $P(Y_2)$ and so on.

Outcome	Number of years	Frequency	Relative Frequency	Outcome

Outcome	Number of years	Frequency	Relative Frequency	Outcome
Y_1	0-2	27	0.27	$= P(Y_1)$
Y_2	2-4	42	0.42	$= P(Y_2)$
Y_3	4-6	20	0.20	$= P(Y_3)$
Y_4	6+	11	0.11	$= P(Y_4)$
Y_5	Total	100	1	$= P(Y_5)$

It is worth noting that this experiment doesn't have equally likely outcomes.

Now that we understand the experiment and its simple outcomes that form the sample space, we can start defining different events.

Let A be an event in which a randomly selected student spent 0-4 years in college. Event A is made of the two outcomes Y_1 and Y_2 so its probability is

$$P(A) = P(Y_1 \text{ or } Y_2) = P(Y_1) + P(Y_2) = 0.27 + 0.42 = 0.69$$

Let B be an event in which a randomly selected student spent 2-6 years in college. Event B is made of the two outcomes Y_2 and Y_3 so its probability is

$$P(B) = P(Y_2 \text{ or } Y_3) = P(Y_2) + P(Y_3) = 0.42 + 0.20 = 0.62$$

Let C be the complement of A. Event C is made of the two outcomes Y_3 and Y_4 so its probability is

$$P(C) = P(Y_3 \text{ or } Y_4) = P(Y_3) + P(Y_4) = 0.20 + 0.11 = 0.31$$

We can also obtain the same result by using the complementary rule.

Let D be the intersection of A and B. Event D is made of only one outcome Y_2 so its probability is

$$P(D) = P(Y_2) = 0.42$$

Let E be the union of A and B. Event E is made of three outcomes Y_1, Y_2, and Y_3 so its probability is

$$P(E) = P(Y_1 \text{ or } Y_2 \text{ or } Y_3) = P(Y_1) + P(Y_2) + P(Y_3) = 0.27 + 0.42 + 0.20 = 0.89$$

We can also obtain the same result by using the general addition rule.

We discussed how to use the frequency distribution tables to produce the probability statements about the events and compound events that are properly defined in the context of the given frequency distribution table.

Section 2: Contingency Tables



In this section, we are going to learn how to make probability statements using contingency tables for bivariate data. Let's assume that we have access to a huge database that has most of the information about the past students. Let's imagine that along with requesting the enrollment length we also requested the information about whether a student transferred, graduated or dropped. And received the following contingency table for a random 100 students.

	Dropped	Transferred	Graduated	Total
0-2	8	16	3	27
2-4	14	13	15	42
4-6	10	4	6	20
6+	8	2	1	11
Total	40	35	25	100

We already know that the frequencies can't be interpreted as probabilities, so we need to turn it into the relative frequency. We add the total for columns and rows and divide everything by the total to obtain the relative frequency contingency table.

	Dropped	Transferred	Graduated	Total
0-2	0.08	0.16	0.03	0.27
2-4	0.14	0.13	0.15	0.42
4-6	0.10	0.04	0.06	0.20
6+	0.08	0.02	0.01	0.11
Total	0.40	0.35	0.25	1.00

Now, in order to find the probability, we must have a well-defined event which is defined for some experiment with its sample space. The question that we need to answer first is what are the experiment, the simple outcomes, the sample space, and what kind of events we can find the probabilities of when such frequency distribution table is given.

First let's define the experiment! In this case, the experiment consists of randomly selecting one of the 100 students from the sample. For bivariate data, we are not going to define the simple outcomes, but we will highlight a number of special events. There are 4 special events, one for each row of the table. We label them Y_1, Y_2, Y_3, and Y_4. And there are also 3 special events, one for each column of the table. We label them R_1, R_2, R_3. We describe event Y_1 as an event in which a randomly selected student spent between 0 and 2 years in college. We describe event R_1 as an event in which a randomly selected student Dropped out of college. Hopefully, it is clear now how the other special events are defined.

	Dropped (R_1)	Transferred (R_2)	Graduated (R_3)	Total
0-2 (Y_1)	0.08	0.16	0.03	0.27
2-4 (Y_2)	0.14	0.13	0.15	0.42
4-6 (Y_3)	0.10	0.04	0.06	0.20
6+ (Y_4)	0.08	0.02	0.01	0.11
Total	0.40	0.35	0.25	1.00

The relative frequencies in the right-most column and bottom row are called **marginal probabilities** as they can be interpreted directly as the probabilities of the 7 special events. For example, the probability of Y_3 is 0.20 and the probability of R_2 is 0.35.

All the cells inside the table are the simple outcomes and form the sample space. The cell where the event Y_3 and event R2 intersect represent the intersection of the two events and the number written in the cell is the probability of the intersection of the

two events. Similarly, the probability of Y_4 and R_1 is 0.08. All the relative frequencies inside the contingency table are called **joint probabilities**. The probability of Y_1 and R_3 is 0.03 and the probability of Y_3 and R_1 is 0.10.

The cells formed by combining the events Y_3 and R_2 represent the union of the two events and to find its probability we can add all the numbers in the region one-by-one, or we can use the general addition rule the ingredients for application of which are 2 marginal and 1 joint probabilities. Of course, we are more interested in applying the rules so let's find the probability of Y_2 OR R_1 using the general addition rule:

$$P(Y_2 \text{ or } R_1) = P(Y_2) + P(R_1) - P(Y_2 \text{ and } R_1) = 0.42 + 0.40 - 0.14 = 0.68$$

We discussed how to use the contingency tables to produce the probability statements about the events and compound events that are properly defined in the context of the given contingency table.

This page titled 1.4.4: Probabilities from Frequency Tables is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.4.5: Conditional Probability and Multiplication Rules

Section 1: Introduction to Conditional Probability



Sometimes the occurrence of one event may affect the probability of the other event. For example, the chances of a student passing a class depend on whether they ace the final exam or not. This relation between two events is called dependency, and to quantify such dependency we introduce the concept of conditional probability.

Definition: Conditional Probability

Conditional probability $P(A|B)$ (probability of A given B) is the probability of event A knowing that B has occurred. In the context of conditional probability, $P(A)$ is called a **prior** probability, $P(A|B)$ is called **posterior** probability.

In general, the $P(A|B)$ is not the same as the $P(B|A)$.

Consider rolling a die. Let A be an event whose outcomes are 1,3,5. The probability of A is known and is equal to 0.5. Let's consider event B whose outcomes are $\{4,5,6\}$. Now, we are ready to roll a die and we are interested in the event A. As of right now, prior to rolling a die, the probability of A is ½ but imagine that before the outcome of the experiment is announced to us someone whispers that event B has occurred. Does the knowledge of B occurring change how we evaluate the probability of A? What exactly changed once we learned that B occurred?

To understand what changes, we must look into how the probabilities are computed! The prior probability of A is computed by dividing the number of outcomes in A by the size of the sample size.

$$P(A) = \frac{|\{1,3,5\}|}{|\{1,2,3,4,5,6\}|} = \frac{3}{6} = \frac{1}{2}$$

So what exactly changes when we know that B has occurred. We may not know what exact outcome has happened but we know that it must be one of the three that describe event B, so event B is the new sample space for event A. Since 1 and 3 are not in the list of possible outcomes anymore we remove them so event A can only occur if 5 was the outcome of the experiment. In other words, when B occurred the event A may only occur on one way out the remaining three outcomes, so the probability of A given B:

$$P(A|B) = \frac{|\{5\}|}{|\{4,5,6\}|} = \frac{1}{3}$$

In general, any conditional probability can be found by finding the restricted sample space, and then finding the number of ways the event can occur in the restricted sample space.

Let's consider for the same experiment the events $C = \{1, 2\}$ and $D = \{5, 6\}$ in addition to previously defined A and B. Let's find some prior and posterior probabilities.

The prior probability of B can be computed by dividing the total number of outcomes that describe B by the total number of outcomes in the sample space. The prior probability of D can be computed by dividing the total number of outcomes that describe D by the total number of outcomes in the sample space. The probability of B given D can be computed by dividing the total number of outcomes in D that describe B by the total number of outcomes in D. The probability of D given B can be computed by dividing the total number of outcomes in B that describe D by the total number of outcomes in B.

Prior	Posterior
$P(B) = \frac{\lvert\{4,5,6\}\rvert}{\lvert\{1,2,3,4,5,6\}\rvert} = \frac{3}{6} = \frac{1}{2}$	$P(B\lvert D) = \frac{\lvert\{5,6\}\rvert}{\lvert\{5,6\}\rvert} = \frac{2}{2} = 1$
$P(D) = \frac{\lvert\{5,6\}\rvert}{\lvert\{1,2,3,4,5,6\}\rvert} = \frac{2}{6} = \frac{1}{3}$	$P(D\lvert B) = \frac{\lvert\{5,6\}\rvert}{\lvert\{4,5,6\}\rvert} = \frac{2}{3}$

The prior probability of A can be computed by dividing the total number of outcomes that describe A by the total number of outcomes in the sample space. The prior probability of C can be computed by dividing the total number of outcomes that describe C by the total number of outcomes in the sample space. The probability of A given C can be computed by dividing the total number of outcomes in C that describe A by the total number of outcomes in C.

Prior	Posterior
$P(A) = \frac{\lvert\{1,3,5\}\rvert}{\lvert\{1,2,3,4,5,6\}\rvert} = \frac{3}{6} = \frac{1}{2}$	$P(A\lvert C) = \frac{\lvert\{1\}\rvert}{\lvert\{1,2\}\rvert} = \frac{1}{2}$
$P(C) = \frac{\lvert\{1,2\}\rvert}{\lvert\{1,2,3,4,5,6\}\rvert} = \frac{2}{6} = \frac{1}{3}$	$P(C\lvert A) = \frac{\lvert\{1\}\rvert}{\lvert\{1,3,5\}\rvert} = \frac{1}{3}$

If we write out the results of prior and conditional probabilities, we will be able to see an interesting phenomenon. The probability of A didn't change after we learned that C occurred, so as the probability of C didn't change after we learned that A has occurred. Such events are called **independent**. On the other hand, the probability of B changed after we learned that D occurred, so as the probability of D changed after we learned that B has occurred. Such events are called **dependent**.

Definition: Dependent and independent events

If A and B are such that occurrence of one doesn't affect the probability of another, then we call A and B **independent**, otherwise we call such events **dependent**.

Symbolically we can express the definition of two independent events by the following statements: $P(A|B) = P(A)$ and $P(B|A) = P(B)$. If one of them is true, the other one must be true as well. So if anyone wants to check whether two events are independent one must find the prior probability and compare it to the conditional probability to see if they are equal.

Note that frequently students confuse mutually exclusive and independent events. Mutually exclusive events have no common outcomes that is $P(A \text{ and } B) = 0$. With independent events, the occurrence of one does not affect the probability of another that is $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

Consider the birth of a child. Before the sex of the child is known, there are two possible outcomes B or G. Normally, they are considered to be mutually exclusive outcomes. The prior probability that the child is a boy is 0.5 If we ask ourselves what is the probability that the child is a boy given that she is a girl we will find the conditional probability of $P(B|G)$ which is obviously 0, so the two events are dependent because the probability of B went down to zero once we learned that event G has occurred. So the only relation between the two concepts is that the two mutually exclusive events are always dependent!

Section 2: Multiplication Rules



Previously, we set the goal to be able to find the probabilities of compound events if the probabilities of the original events are given. For finding the probability of the complementary events we developed the complementary rule, for finding the probability of the union of two events we developed the special and general addition rules. The next goal is to develop the rule for finding the probability of the intersection of two events.

First, let's formalize the approach for finding the conditional probability.

Theorem: Conditional Probability Formula

For any two events A and B, the way we found the probability of A given B was to divide the number of outcomes in B that describe A by the number of outcomes in B:

$$P(A|B) = \frac{\text{outcomes of B that describe A}}{\text{outcomes of B}}$$

This process can be easily converted into the following formula

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

which we call the **conditional probability formula.**

Contingency tables are the best playground for practicing the conditional probability formula. Let's recall the following contingency table with events defined as shown.

	Dropped (R_1)	Transferred (R_2)	Graduated (R_3)	Total
0-2 (Y_1)	0.08	0.16	0.03	0.27
2-4 (Y_2)	0.14	0.13	0.15	0.42
4-6 (Y_3)	0.10	0.04	0.06	0.20
6+ (Y_4)	0.08	0.02	0.01	0.11
Total	0.40	0.35	0.25	1.00

Example 1.4.5.2.1.1

Let's find the conditional probability of Y_3 given R_2 using the conditional probability formula where the numerator is the joint probability of Y_3 and R_2 and the denominator is the marginal probability of R_2:

$$P(Y_3|R_2) = \frac{P(Y_3 \text{ and } R_2)}{P(R_2)} = \frac{0.04}{0.35} = 0.11$$

Example 1.4.5.2.1.2

Let's find another conditional probability of R_1 given Y_4 using the conditional probability formula where the numerator is the joint probability of R_1 and Y_4 and the denominator is the marginal probability of Y_4:

$$P(R_1|Y_4) = \frac{P(R_1 \text{ and } Y_4)}{P(Y_4)} = \frac{0.08}{0.11} = 0.73$$

Theorem: General Multiplication Rule

Consider the conditional probability formula again.

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Let's multiply both sides by $P(B)$ and swap the sides of the equation.

$$P(A \text{ and } B) = P(A|B) \cdot P(B)$$

The result is known as the **general multiplication rule** that allows to find the probability of the intersection of two events.

Theorem: Special Multiplication Rule

When A and B are independent, in other words when the probability of A given B is the same as the probability of A in the general multiplication rule, we can replace $P(A|B)$ with $P(A)$ and get the following result

$$P(A \text{ and } B) = P(A) \cdot P(B)$$

which is known as the **special multiplication rule**. What makes it special is the fact that it can only be applied for two independent events.

Example 1.4.5.2.1

if A and B are independent and the probability of A is 0.4 and the probability of B is 0.5 then the probability of A and B is 0.2. In this particular example, we can also use the general addition rule and find the probability of the union of two independent events A and B, i.e., $P(AorB) = P(A) + P(B) - P(AandB) = 0.4 + 0.5 - 0.2 = 0.7$.

Using the special multiplication rule we can verify whether two events are independent or not. For example in the following contingency table consider events Y_2 and R_3 and let's check whether they are independent or not?

	Dropped (R_1)	Transferred (R_2)	Graduated (R_3)	Total
0-2 (Y_1)	0.08	0.16	0.03	0.27
2-4 (Y_2)	0.14	0.13	0.15	0.42
4-6 (Y_3)	0.10	0.04	0.06	0.20
6+ (Y_4)	0.08	0.02	0.01	0.11
Total	0.40	0.35	0.25	1.00

If they were independent then the special multiplication rule will be true that is the

$$P(Y_2 \text{ and } R_3) = P(Y_2) \cdot P(R_3)$$

On the left-hand side there is the joint probability 0.15 and on the right hand side there is the product of the marginal probabilities of Y_2 and R_3. It is easy to verify that the two sides are not equal, therefore the events are dependent!

$$0.15 = P(Y_2 \text{ and } R_3) \neq P(Y_2) \cdot P(R_3) = 0.42 \cdot 0.25$$

We finally developed a complete set of probability rules that we can use to find the probabilities of the complement of an event, and the union and intersection of any two events.

This page titled 1.4.5: Conditional Probability and Multiplication Rules is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

SECTION OVERVIEW

1.5: Combinatorics Concepts

1.5.1: Introduction to Combinatorics

1.5.2: Permutations and Combinations

1.5.3: Tossing a Coin

This page titled 1.5: Combinatorics Concepts is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.5.1: Introduction to Combinatorics

Section 1: Introduction to Combinatorics



Combinatorics is the science of counting but not the counting that you do in first grade but the type of counting that it takes to compute the size of a sample space or the number of outcomes that satisfy a certain condition.

Consider the experiment of rolling a die with the well-known sample space

$$\{1, 2, 3, 4, 5, 6\}$$

and let A be an event in which the outcome is 6. To find the probability of A we divide the number of outcomes in the event by the size of the sample size, which is equal to 1 over 6.

$$P(A) = \frac{|\{6\}|}{|\{1,2,3,4,5,6\}|} = \frac{1}{6}$$

Next, let's consider the experiment of rolling two dice with the well-known sample space

$$\{1:1, 1:2, 1:3, \cdots\}$$

and let A be an event in which the outcome is a double 6s. To find the probability of A we divide the number of outcomes in the event by the size of the sample space, which is equal to 1 over 36.

$$P(A) = \frac{|\{6:6\}|}{|\{1:1,1:2,1:3,\cdots\}|} = \frac{1}{36}$$

Finally, consider the experiment of rolling three dice with the sample space the size of which we don't know

$$\{1:1:1, 1:1:2, \cdots\}$$

and let A be an event in which the outcome is three 6s. To find the probability of A we divide the number of outcomes in the event by the size of the sample space, which is equal to 1 over a number that we don't know this time.

$$P(A) = \frac{|\{6:6:6\}|}{|\{1:1:1,1:1:2,\cdots\}|} = \frac{1}{?}$$

But can we figure it out? When there is only one die there are 6 outcomes in the sample space. When there are two dice there are 36 outcomes in the sample space. So how many outcomes are in the sample space when there are three dice. The pattern appears to be straight forward - the sample space size for rolling 1 die is 6 raised to the power 1, the sample space size for rolling 2 dice is 6 raised to the power 2, so the formula for the sample space size for rolling three dice is 6 raised to the power 3.

Roll a die	once	twice	thrice
Sample Space	6	36	216

Roll a die	once	twice	thrice
n	1	2	3
Pattern?	6^1	6^2	6^3

Therefore, the probability from the previous example is 1 over 216.

Now we can roll a die four times and without knowing exactly the size of the sample space we can let A be the event in which there are 4 sixes, and the probability of A is the number of outcomes in A divided by the size of the sample space which we can speculate to be equal to 6 raised to the power 4, so the probability is very little.

$$P(A) = \frac{|\{6:6:6:6\}|}{|\text{Sample Space}|} = \frac{1}{6^4} = 0.00077 = 0.077$$

This is a small example of counting which we used to figure out the size of the sample space and compute the probability of an event. Next, we will take a more formal approach and introduce the topic of combinatorics.

Section 2: The Basic Counting Principle



Why exactly there are 8 simple outcomes when flipping a coin three times? And 216 when rolling a die three times? Next, we will learn the basic counting principle that allows us to compute these and many other quantities.

Consider that you have the following menu with the following choices

Soup	Meat	Side	Dessert
Lentils French Onion Clam Chowder Tomato	Beef Pork Lamb	Potatoes Veggies Rice Beans Pasta	Carrot Cake Cheesecake
4	3	5	2

And the numbers of choices are summarized in the bottom of the table.

If a meal consisted of a meat and a side, how many possible meal combinations are there?

$$3 \cdot 5 = 15$$

If a dessert is added to the meal, how many possible meal combinations are there?

$$15 \cdot 2 = 30$$

If a soup is added to the meal, how many possible meal combinations are there?

$$4 \cdot 30 = 120$$

Theorem: Basic Counting Principle

The above result can be summarized as **the basic counting principle** that states that if r actions are to be performed in a definite order and that there are m_i possibilities for i-th action then there are $m_1 \cdot \cdots \cdot m_r$ possibilities altogether for the r actions. This principle can be summarized in the following way. When there is a process that consists of r actions with the known number of choices for each action, then the total number of choices is the product of the number of choices at each step.

Process	Action 1	Action 2	$\cdots$	Action r
Number of choices	m_1	m_2	$\cdots$	m_r

Total number of choices is

$$m_1 \cdot m_2 \cdot \cdots \cdot m_r$$

Example 1.5.1.2.1

If there are 2 different ways to get from point A to point B and then 3 different ways to get from point B to point C, and 7 different ways to get from point C to point D. How many different paths are there from point A to point D?

Solution

Travelling	from A to B	from B to C	from C to D
Number of choices	2	3	7

The number of paths from point A to point D is

$$2 \cdot 3 \cdot 7 = 42$$

Example 1.5.1.2.2

How many 3-letter words (meaningful or meaningless) one can create using the letters a-z?

Solution

It all depends on if the repetition is allowed or not.

- if repetition is allowed then the process of creating a 3-letter word consists of picking 3 letters and since the repetition is allowed the number of choices for each letter is 26 and the total number of words is 17,576 by the basic counting principle.
- if repetition is not allowed then the process of creating a 3-letter word still consists of picking 3 letters but since the repetition is NOT allowed the number of choices for each letter is decreasing by 1 with every step and the total number of words is 15,600 by the basic counting principle.

Example 1.5.1.2.3

How many 3-digit numbers one can create using the 10 digits 0-9?

Solution

AGAIN, it all depends on if the repetition is allowed or not.

- if repetition is allowed then the process of creating a 3-digit number consists of picking 3 digits and since the repetition is allowed the number of choices for each digits except the first one is 10 and the total number of numbers is 900 by the basic counting principle. Why the number of choices for the first digit is 9? Because in a 3-digit number the first digit cannot be 0!

- if repetition is not allowed then the process of creating a 3-digit number still consists of picking 3 digits but since the repetition is NOT allowed the number of choices for all but the first digit is decreasing by 1 with every step and the total number of numbers is 648 by the basic counting principle.

Are there enough license plates in California to cover all the registered vehicles?



To find out, let's do the following example.

Example 1.5.1.2.4.1

How many possible license plate arrangements are in California(#LLL###)? (Note: Repetition is allowed.)

Solution

The process of creating a license plate consists of picking 7 symbols in order and the number of choices for each symbol is determined by whether it is a letter or a digit.

License Plate	Symbol 1	Symbol 2	Symbol 3	Symbol 4	Symbol 5	Symbol 6	Symbol 7
# of choices	10	26	26	26	10	10	10

By the basic counting principle, the total number of license plates is

$$10 \cdot 26 \cdot 26 \cdot 26 \cdot 10 \cdot 10 \cdot 10 = 175,760,000$$

What will change if we require the symbols to be nonrepeating? In this case, the repetition is NOT allowed.

Example 1.5.1.2.4.2

How many possible license plate arrangements are in California(#LLL###) with nonrepeating symbols? (Note: Repetition is not allowed.)

Solution

The process of creating a license plate still consists of picking 7 symbols in order and the number of choices for each symbol is determined by whether it is a letter or a digit. Since there are 26 letters and 10 digits and repetition is NOT allowed the number of choices at consecutive steps is one less than in the previous, and we get the following table:

License Plate	Symbol 1	Symbol 2	Symbol 3	Symbol 4	Symbol 5	Symbol 6	Symbol 7
# of choices	10	26	25	24	9	8	7

By the basic counting principle, the total number of license plates is

$$10 \cdot 26 \cdot 25 \cdot 24 \cdot 9 \cdot 8 \cdot 7 = 78,624,000$$

What will change if we limit the number of available symbols?

Example 1.5.1.2.4.3

How many possible license plate arrangements are in California(#LLL###) with only consonants and odd digits? (Note: Repetition is allowed.)

Solution

The process of creating a license plate still consists of picking 7 symbols in order and the number of choices for each symbol is determined by whether it is a consonant or an odd digit. Since there are 21 consonants and 5 odd digits and repetition is

allowed, we get the following table:

License Plate	Symbol 1	Symbol 2	Symbol 3	Symbol 4	Symbol 5	Symbol 6	Symbol 7
# of choices	5	21	21	21	5	5	5

By the basic counting principle, the total number of license plates is

$$5 \cdot 21 \cdot 21 \cdot 21 \cdot 5 \cdot 5 \cdot 5 = 5,788,125$$

And again, what will change if we require the symbols to be nonrepeating?

Example 1.5.1.2.4.4

How many possible license plate arrangements are in California(#LLL###) with nonrepeating consonants and odd digits? (Note: Repetition is not allowed.)

Solution

The process of creating a license plate still consists of picking 7 symbols in order and the number of choices for each symbol is determined by whether it is a consonant or an odd digit. Since there are 21 consonants and 5 odd digits and repetition is NOT allowed the number of choices at consecutive steps is one less than in the previous, and we get the following table:

License Plate	Symbol 1	Symbol 2	Symbol 3	Symbol 4	Symbol 5	Symbol 6	Symbol 7
# of choices	5	21	20	19	4	3	2

By the basic counting principle, the total number of license plates is

$$5 \cdot 21 \cdot 20 \cdot 19 \cdot 4 \cdot 3 \cdot 2 = 957,600$$

The basic counting principle can tell us how many events are possible from a sample space of size n. One may think of the process of creating an event as a sequence of n choices with 2 options each - whether to include or exclude each simple outcome. Therefore, the number of choices at each step is exactly two – to include or not include, and by the basic counting principle, there are 2^n total choices to create an event. Thus, the number of events from a sample space of size n is 2^n!

Example 1.5.1.2.5

Now using the formula we can find how many events can be created from a sample space of size 6, size 10 or any other size?

Solution

- The number of events from a sample space of size 6 is $2^6 = 64$.
- The number of events from a sample space of size 10 is $2^{10} = 1,024$.

We used the basic counting principle to find the number of events that can be created from a sample space of size n, next we will discuss how to use the basic counting principle to figure out the size of the sample space for certain experiments such as tossing a coin or a die multiple time. So let's find out how to use the basic counting principle to find the number of simple outcomes are there when tossing a coin twice and thrice.

Example 1.5.1.2.6.1

- To create an outcome for tossing a coin twice we consider the outcomes for coin 1 and 2. And the number of choices for each coin toss is exactly 2. So the number of possible outcomes is 4 by the basic counting principle.
- To create an outcome for tossing a coin thrice we consider the outcomes for coin 1, 2, and 3. And the number of choices for each coin toss is exactly 2. So the number of possible outcomes is 8 by the basic counting principle.

Similarly, we can use the basic counting principle to find the number of simple outcomes are there when rolling a die twice and thrice.

Example 1.5.1.2.6.2

- To create an outcome for rolling a die twice we consider the outcomes for die 1 and 2. And the number of choices for each die roll is exactly 6. So the number of possible outcomes is 36 by the basic counting principle.
- To create an outcome for rolling a die thrice we consider the outcomes for die 1, 2, and 3. And the number of choices for each die roll is exactly 6. So the number of possible outcomes is 6 cubed by the basic counting principle.

Earlier we were able to guess the same formula by observing the pattern, however this time we used the basic counting principle which is a more formal approach.

This page titled 1.5.1: Introduction to Combinatorics is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.5.2: Permutations and Combinations

Section 1: Factorial Notation



Definition: $n!$

$n!$ denotes the product of natural numbers from 1 to n, i.e.

$$n! = n \cdot (n-1) \cdot \cdots \cdot 3 \cdot 2 \cdot 1$$

Example 1.5.2.1

$$4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$$

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120$$

$$8! = 8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 40320$$

Note that

$$0! = 1$$

In combinatorics, we would want to be able to simplify the following expressions:

$$\frac{7!}{3!}, \frac{7!}{4!}, \frac{7!}{3!4!}$$

The easiest way to simplify the factorials is to write them out using the definition, and then reduce by the greatest common factor:

$$\frac{7!}{3!} = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1} = 7 \cdot 6 \cdot 5 \cdot 4$$

$$\frac{7!}{4!} = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{4 \cdot 3 \cdot 2 \cdot 1} = 7 \cdot 6 \cdot 5$$

$$\frac{7!}{3!4!} = \frac{7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{3 \cdot 2 \cdot 1 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 7 \cdot 5$$

Frequently we will have to work with expressions like the very last one. Note that after writing out the factorials and after reducing by the common factorial we still have some common factors left which we need to reduce.

Section 2: Permutations



Definition: Permutation

An ordered arrangement of distinct objects is called a **permutation**.

Example 1.5.2.2.1

THR is a permutation of length three from letters $\{T, H, R\}$.

HRT is a permutation of length three from letters $\{T, H, R\}$.

HR is a permutation of length 2 from letters $\{T, H, R\}$.

Example 1.5.2.2.2

123 is a permutation of length three from letters $\{1, 2, 3\}$.

231 is a permutation of length three from letters $\{1, 2, 3\}$.

31 is a permutation of length 2 from letters $\{1, 2, 3\}$.

How many permutations of length 3 out of 3 letters are there?

To create a permutation, we have to consecutively choose 3 letters. Since the letters are all distinct the number of choices at each step is decreasing by one. So by the basic counting principle the total number of permutations is $3 \cdot 2 \cdot 1 = 3!$.

How many permutations of length 4 out of 4 letters are there?

To create a permutation, we have to consecutively choose 4 letters. Since the letters are all distinct the number of choices at each step is decreasing by one. So by the basic counting principle the total number of permutations is $4 \cdot 3 \cdot 2 \cdot 1 = 4!$.

How many permutations of length n out of n letters are there?

To create a permutation, we have to consecutively choose n letters. Since the letters are all distinct the number of choices at each step is decreasing by one. So by the basic counting principle the total number of permutations is $n!$.

How many permutations of length 2 out of 5 letters are there?

To create a permutation, we have to consecutively choose 2 letters. Since the letters are all distinct the number of choices at each step is decreasing by one. So by the basic counting principle the total number of permutations is $5 \cdot 4$.

How many permutations of length 3 out of 5 letters are there?

To create a permutation, we have to consecutively choose 3 letters. Since the letters are all distinct the number of choices at each step is decreasing by one. So by the basic counting principle the total number of permutations is $5 \cdot 4 \cdot 3$.

How many permutations of length k out of n letters are there?

To create a permutation, we have to consecutively choose k letters. Since the letters are all distinct the number of choices at each step is decreasing by one. So by the basic counting principle the total number of permutations is given by the following product:

$$n \cdot (n-1) \cdots (n-k+1)$$

Theorem: Permutation Rule

In general, we want to find out how many permutations of length k out of n letters are there? Let's denote this quantity as P_k^n. Based on the observed pattern, we obtain the following formula called the **permutation rule**:

$$P_k^n = \frac{n!}{(n-k)!}$$

Let's do a few applications of it.

Example 1.5.2.2.3.1

How many permutations of length 2 out of 5 letters are there?

Solution

The answer is P_2^5 which we can now compute using the formula:

$$P_2^5 = \frac{5!}{(5-2)!} = \frac{5!}{3!} = 5 \cdot 4 = 20$$

Example 1.5.2.2.3.2

How many permutations of length 4 out of 10 letters are there?

Solution

The answer is P_4^{10} which we can now compute using the formula:

$$P_4^{10} = \frac{10!}{(10-4)!} = \frac{10!}{6!} = 10 \cdot 9 \cdot 8 \cdot 7 = 5040$$

Theorem: Special Permuations Rule

If the length of a permutation is the same as the number of available letters, then we replace k with n in the permutation formula and obtain **the special permutation rule**:

$$P_n^n = \frac{n!}{(n-n)!} = \frac{n!}{0!} = n!$$

Let's do a few applications of it.

Example 1.5.2.2.4.1

How many permutations of length 4 out of 4 letters are there?

Solution

$$P_4^4 = 4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$$

Example 1.5.2.2.4.2

How many permutations of length 6 out of 6 letters are there?

Solution

$$P_6^6 = 6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$$

The following questions:

• How many ways are there to select 3 people out of 6 for a committee with distinct roles?

• How many ways are there to select 3 pieces of art out 6 for an exhibition in which the order matters?

and many other questions, all have the same answer:

$$P_3^6$$

The following questions:

• How many ways are there to select 5 people out of 9 for a committee with distinct roles?

• How many ways are there to select 5 pieces of art out 9 for an exhibition in which the order matters?

and many other questions, all have the same answer:

$$P_5^9$$

Permutations is one of the building blocks of the combinatorics.

Section 3: Combinations



Definition: Combination

A collection of distinct objects is called a **combination**.

Example 1.5.2.3.1.1

$\{T, H, R\}$ is a combination of length three from letters $\{T, H, R\}$.

$\{H, R, T\}$ is a combination of length three from letters $\{T, H, R\}$.

$\{T, H, R\}$ and $\{H, R, T\}$ are the same combinations.

Example 1.5.2.3.1.2

$\{1, 2\}$ is a combination of length two from letters $\{1, 2, 3\}$.

$\{2, 3, 1\}$ is a combination of length three from letters $\{1, 2, 3, 4\}$.

$\{1, 3\}$ is a permutation of length two from letters $\{1, 2, 3, 4\}$.

Let's discuss the difference between permutations and combinations.

Permutations	Combinations
• Definition: An ordered arrangement of distinct objects is called a **permutation**. • Order matters, i.e., $AB \neq BA$ • Example: selecting a committee of 3 people	• Definition: A collection of distinct objects is called a **combination**. • No order, i.e., $\{A, B\} = \{B, A\}$ • Example: selecting 3 people for a survey

Since every combination of length k can be further arranged in order, the number of permutations of length k is $k!$ times larger than the number of combinations of length k.

Theorem: Combinations Rule

In general, we want to find out how many combinations of length k out of n letters are there? Let's denote this quantity as C_k^n. Based on the fact that the number of permutations of length k is $k!$ times larger than the number of combinations of length k we obtain the following formula called the **combinations rule**:

$$C_k^n = \frac{n!}{(n-k)!k!}$$

Let's do a few applications of it.

Example 1.5.2.3.2.1

How many combinations of length 2 out of 5 letters are there?

Solution

The answer is C_2^5 which we can now compute using the formula:

$$C_2^5 = \frac{5!}{(5-2)!2!} = \frac{5!}{3!2!} = \frac{5 \cdot 4}{2 \cdot 1} = 10$$

Example 1.5.2.3.2.2

How many combinations of length 4 out of 10 letters are there?

Solution

The answer is C_4^10 which we can now compute using the formula:

$$C_4^10 = \frac{10!}{(10-4)!4!} = \frac{10!}{6!4!} = \frac{10 \cdot 9 \cdot 8 \cdot 7}{4 \cdot 3 \cdot 2 \cdot 1} = 210$$

Alternatively, the following notation is commonly used in the literature:

$$C_k^n = \binom{n}{k}$$

and we read it as "n choose k".

The following questions:

• How many ways are there to select 3 people out of 6 for a survey?

• How many ways are there to select 3 pieces of art out 6 to donate?

and many other questions, all have the same answer:

$$C_3^6$$

The following questions:

• How many ways are there to select 5 people out of 9 for a survey?

• How many ways are there to select 5 pieces of art out 9 to donate?

and many other questions, all have the same answer:

$$C_5^9$$

Combinations along with permutations is another building blocks of the combinatorics.

This page titled 1.5.2: Permutations and Combinations is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.5.3: Tossing a Coin

Section 1: Tossing a Fair Coin



Next, we will consider the experiment of tossing a fair coin several times and finding the probability of getting a certain number of tails and heads.

Definition: Fair coin

A **fair coin** is a coin with two equally likely outcomes.

Let's denote the probability of having k heads among n tosses as

$$P(kH/n)$$

Let's find the probability of having no heads among 4 tosses. To do that we are going to list the entire sample space that consists of $2^4 = 16$ simple outcomes.

HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT,

THHH, THHT, THTH, THTT, TTHH, TTHT, TTTH, TTTT

The probability of having 0 heads among 4 tosses can be found by dividing the number of outcomes with 0 heads by the size of the sample space.

$$P(0H/4) = \frac{1}{2^4}$$

The probability of having 1 heads among 4 tosses can be found by dividing the number of outcomes with 1 heads by the size of the sample space.

$$P(1H/4) = \frac{4}{2^4}$$

The probability of having 2 heads among 4 tosses can be found by dividing the number of outcomes with 2 heads by the size of the sample space.

$$P(2H/4) = \frac{6}{2^4}$$

The probability of having 3 heads among 4 tosses can be found by dividing the number of outcomes with 3 heads by the size of the sample space.

$$P(3H/4) = \frac{4}{2^4}$$

The probability of having 4 heads among 4 tosses can be found by dividing the number of outcomes with 4 heads by the size of the sample space.

$$P(1H/4) = \frac{4}{2^4}$$

To understand the pattern let's focus on the probability of having 2 heads among 4 tosses that can be found by dividing the number of outcomes with 2 heads by the size of the sample space.

6 is the number of outcomes with 2H/4 and it is the answer to the question:

How many outcomes with 2H out of 4 tosses?

And it is also the answer to the question:

How many ways to create an outcome with 2H out of 4 tosses?

Which is the same as asking:

How many ways to select 2 tosses out of 4 tosses for H?

So, the 6 is the same as the number of ways to select two objects out of four which we denote as:

$$6 = C_2^4$$

It is easy to verify that the numerators in the remaining 4 computations can also be alternatively obtained by using the combinations formula:

$$P(0H/4) = \frac{1}{2^4} = \frac{C_0^4}{2^4}$$

$$P(1H/4) = \frac{4}{2^4} = \frac{C_1^4}{2^4}$$

$$P(3H/4) = \frac{4}{2^4} = \frac{C_3^4}{2^4}$$

$$P(4H/4) = \frac{1}{2^4} = \frac{C_4^4}{2^4}$$

To summarize, the probability of having k heads among 4 tosses can be found by using the formula:

$$P(kH/4) = \frac{C_k^4}{2^4}$$

And in general, the probability of having k heads among n tosses can be found by using the formula:

$$P(kH/n) = \frac{C_k^n}{2^n}$$

Example 1.5.3.1

Find the probability of 5 heads among 9 tosses of a fair coin.

Solution

According to the formula it is

$$P(kH/n) = \frac{C_k^n}{2^n} = \frac{C_5^9}{2^9} = \frac{126}{512} = 24.6\%$$

We discussed the experiment of tossing a fair coin several times and finding the probability of getting a certain number of tails and heads. The significance of this result is that many experiments in real life can be modeled by a toss of a fair coin.

Section 2: Tossing an Unfair Coin



Definition: Unfair coin

An **unfair coin** is a coin with two not equally likely outcomes.

Let's discuss the differences between a fair coin and an unfair coin. The outcomes of a fair coin are equally likely and since there are only two of them, probability of each one is 0.5. The outcomes of an unfair coin are not equally likely and hence the probability of each outcome is unknown unless some information is given.

Example 1.5.3.2.1.1

Let's find the $P(T)$ and $P(H)$ if T is three times more likely than H.

Solution

The given information can be interpreted as the odds in favor of T are 3:1 and therefore the probability of T is 0.75 and the probability of H is 0.25, i.e.:

$P(T)=\frac{3}{3+1}=0.75$ and $P(H)=\frac{1}{3+1}=0.25$

Example 1.5.3.2.1.2

Let's find the $P(T)$ and $P(H)$ if T is four times less likely than H.

Solution

The given information can be interpreted as the odds in favor of T are 1:4 and therefore the probability of T is 0.20 and the probability of H is 0.80, i.e.:

$P(T)=\frac{1}{1+4}=0.20$ and $P(H)=\frac{4}{1+4}=0.80$

Let's consider the sample space of tossing a coin four times:

HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT,

THHH, THHT, THTH, THTT, TTHH, TTHT, TTTH, TTTT

Let's denote $P(H)=p$ and $P(T)=q$. Since the outcomes are not equally likely, let's find the probabilities of each of the simple outcomes. Since the outcomes of different tosses are independent the probability of 4 Heads can be found using the special multiplication rule and is equal to p^4.

$$P(HHHH)=P(H)\cdot P(H)\cdot P(H)\cdot P(H)=p\cdot p\cdot p\cdot p=p^4$$

Similarly, we can find the probabilities of all simple outcomes.

$$P(HHHT) = P(H) \cdot P(H) \cdot P(H) \cdot P(T) = p \cdot p \cdot p \cdot q = p^3 q^1$$
$$P(HHTH) = P(H) \cdot P(H) \cdot P(T) \cdot P(H) = p \cdot p \cdot q \cdot p = p^3 q^1$$
$$P(HHTT) = P(H) \cdot P(H) \cdot P(T) \cdot P(T) = p \cdot p \cdot q \cdot q = p^2 q^2$$
$$\vdots$$
$$P(TTTH) = P(T) \cdot P(T) \cdot P(T) \cdot P(H) = q \cdot q \cdot q \cdot p = p^1 q^3$$
$$P(TTTT) = P(H) \cdot P(H) \cdot P(H) \cdot P(H) = p \cdot p \cdot p \cdot p = p^4$$

Note that the probability of a simple outcome depends only on the number of heads and tails in the simple outcome!

Let's denote the probability of having k heads among n tosses in the following way

$$P(kH/n)$$

Let's find the following probabilities. The probability of having 0 heads among 4 tosses can be found by finding the probability of all tails which is $1q^4$:

$$P(0H/4) = P(TTTT) = 1q^4$$

The probability of having 1 heads among 4 tosses can be found by the special addition rule and is equal to $4pq^3$:

$$P(1H/4) = P(HTTT) + P(THTT) + P(TTHT) + P(TTTH) = 4pq^3$$

Similarly, the probability of having 3 heads among 4 tosses can be found by the special addition rule and is equal to $4p^3q$:

$$P(3H/4) = P(THHH) + P(HTHH) + P(HHTH) + P(HHHT) = 4p^3q$$

Finally, the probability of having 4 heads among 4 tosses can be found by finding the probability of all heads which is p^4.

$$P(4H/4) = P(HHHH) = 1p^4$$

It is not hard to recognize the coefficients of the expressions as we have already seen these numbers before! So we can find the probability of having 2 heads among 4 tosses without writing out all simple outcomes in the following way:

$$P(2H/4) = P(HHTT) + \cdots + P(TTHH) = 6p^2q^2$$

In summary, if $P(H) = p$ and $P(T) = q$ then the probability of having k heads among n tosses is given by the following formula:

$$P(kH/n) = C_k^n p^k q^{n-k}$$

Example 1.5.3.2.2

Find the probability of having 6 heads among 10 tosses if the heads are 7 times likelier than tails.

Solution

First, find $p = P(H)$ and $q = P(T)$ using the fact that the odds in favor of $H$are 7 to 1:

$$P(H) = \frac{7}{7+1} = \frac{7}{8} \text{ and } P(T) = \frac{1}{7+1} = \frac{1}{8}$$

Now, we can find the probability of having 6 heads among 10 tosses by using the formula:

$$P(6H/10) = C_6^{10}(\tfrac{7}{8})^6(\tfrac{1}{8})^{10-6} = 210 \cdot 0.4488 \cdot 0.0002 = 2.3\%$$

In conclusion, if we let the probability of heads and tails to be equal to ½ then the formula for an unfair coin becomes the formula for a fair coin:

$$P(kH/n) = C_k^n(\tfrac{1}{2})^k(\tfrac{1}{2})^{n-k} = \frac{C_k^n}{2^n}$$

We discussed the experiment of tossing an unfair coin several times and finding the probability of getting a certain number of tails and heads. The significance of this result is that many experiments in real life can be modeled by a toss of an unfair coin.

This page titled 1.5.3: Tossing a Coin is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

SECTION OVERVIEW

1.6: Discrete Random Variables

1.6.1: Discrete Random Variables

1.6.2: Applications of Discrete Random Variables

1.6.3: Binomial Random Variables

This page titled 1.6: Discrete Random Variables is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.6.1: Discrete Random Variables

Section 1: Introduction to Random Variables



Definition: Random Variable

A **random variable** represents an unknown quantity whose value depends on chance or the outcome of an experiment.

Algebraic Variables	Random Variables
• Let x be the number of siblings that Bobert has. • We do not know x, but we can definitely find out by asking Bobert. • x is an example of an algebraic variable.	• Let X be the number of siblings of a randomly selected person in the class. • We do not know X until the person is randomly selected. • X is an example of a random variable.

To distinguish between the random and algebraic variables we are going to label a random variable with a capital letter. Now, it is important to distinguish between the events and random variables.

Events	Random Variables
• A, B, C, D, E, F, etc. • $P(A)$ makes sense and means the probability of event A.	• Q, R, S, T, U, V, W, etc. • $P(X)$ makes no sense, but $\{X = 1\}$ is an event, and thus $P(X = 1)$ makes sense.

Example 1.6.1.1.1

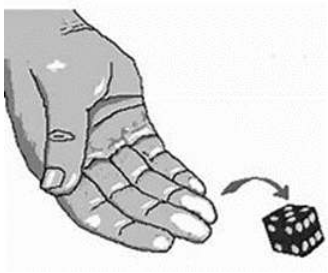
If X is the number of siblings of a randomly selected student then $\{X = 2\}$ is the event in which the number of siblings of a randomly selected person is 2, and $P(X = 2)$ denotes the probability of such event.

We are still going to perform the same experiments as before but from now on we will keep an eye on the outcomes and events whose verbal descriptions can be quantified as in the following examples:

- The product of the dice when two fair dice are rolled.
- The number of sixes when four fair dice are rolled.
- The number of cars in a randomly selected family.
- The household income of a randomly selected household.

- The lifetime of a randomly selected bulb etc.

Note that we can use a random variable to denote the outcome of a roll of a die but the same cannot be done with a toss of a coin because the outcome of the experiment is not a number, but if we assign 0 to heads and 1 to tails (or any other combination of numbers) then we can introduce a random variable.





Since we are dealing with quantities there are two types of random variables: discrete and continuous.

Definition: Discrete Random Variable

A **discrete random variable** is a random variable whose all possible values can be listed.

Definition: Continuous Random Variable

A **continuous random variable** is a random variable whose all possible values form an interval.

Discrete Random Variable	Continuous Random Variable
• All possible values are listable.	• All possible values form a continuous interval.

Now, we can classify the variables as discrete or continuous:

- The number of credit cards of a randomly selected person is discrete, because the possible values are 0, 1, 2, 3 etc.
- The lifetime of a randomly selected bulb is continuous, because it can be any real number greater than 0 such as 0.765, 4.5632 etc.
- The size of a randomly selected household is discrete, because the possible values are 0, 1, 2, 3 etc.
- The monthly income of a randomly selected household is continuous, because it can be any number greater than 0 such as 45,763.21, 81,034.62, etc.

Example 1.6.1.1.2

Discrete Random Variable	Continuous Random Variable
Examples: • Number of credit cards • Household size	Examples: • Lifetime of a bulb • Monthly income

Now that we introduced the concept of a random variable, we will focus on the properties and applications of discrete and continuous random variables in this exact order.

Section 2: Probability Distribution Table



Let's recall the definition of a **discrete random variable** as a random variable whose possible values can be listed. Consider tossing a coin with 0 assigned to heads and 1 to tails and let X be the outcome of the experiment. By definition, we can **list** the outcomes and their probabilities side-by-side in the form of a table.

Definition: Probability distribution table

The table that summarizes the possible values of a variable X and the corresponding probabilities is called the **probability distribution table for X**.



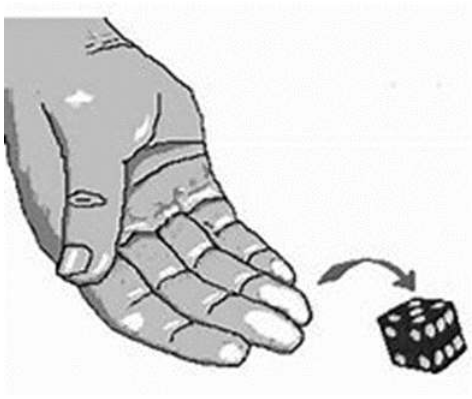
The Probability Distribution Table for X

X	x_i	$P(X = x_i)$
	0	0.5
	1	0.5
	Total:	1

Note that the sum of all probabilities must add up to 1.

Example 1.6.1.2.1

Consider rolling a die and let Y be the outcome of the experiment. We construct the probability distribution table for Y by listing all possible outcomes and their probabilities side-by-side in the form of a table.



The Probability Distribution Table for Y

Y	y_i	$P(Y=y_i)$
	1	1/6
	2	1/6
	3	1/6
	4	1/6
	5	1/6
	6	1/6
	Total:	1

Again, note that the sum of all probabilities must add up to 1.

Sometimes it takes more effort to create the probability distribution table. Consider the following example.

Example 1.6.1.2.2

Let U be the number of Heads out of two tosses of a fair coin. To construct the probability distribution table for U we list all possible outcomes along with their probabilities side-by-side in the form of a table. This time the probability of each outcome is not that easy to observe and has to be computed.

- The probability that $U=0$ is the same as the probability of 0 heads among 2 tosses which we compute using the formula $P(U=0)=\frac{C_0^2}{2^2}=0.25$
- Similarly, the probability that $U=1$ is the same as the probability of 1 heads among 2 tosses which we again compute using the formula $P(U=1)=\frac{C_1^2}{2^2}=0.50$
- Finally, the probability that $U=2$ is the same as the probability of 2 heads among 2 tosses which we again compute using the formula $P(U=2)=\frac{C_2^2}{2^2}=0.25$

Easy to check that the sum of all probabilities is equal to 1.

The Probability Distribution Table for U

U	u_i	$P(U=u_i)$
	0	0.25
	1	0.5
	2	0.25
	Total:	1

Let's increase the number of tosses and consider the following example.

Example 1.6.1.2.3

Let V be the number of Heads out of three tosses of a fair coin. To construct the probability distribution table for V we list all possible outcomes along with their probabilities side-by-side in the form of a table. Again, the probability of each outcome is not that easy to observe and has to be computed.

- The probability that $V=0$ is the same as the probability of 0 heads among 3 tosses which we compute using the formula $P(V=0)=\frac{C_0^3}{2^3}=0.125$

- Similarly, the probability that $V = 1$ is the same as the probability of 1 heads among 3 tosses which we again compute using the formula $P(V = 1) = \frac{C_1^3}{2^3} = 0.375$
- Similarly, the probability that $V = 2$ is the same as the probability of 2 heads among 3 tosses which we again compute using the formula $P(V = 2) = \frac{C_2^3}{2^3} = 0.375$
- Finally, the probability that $V = 3$ is the same as the probability of 3 heads among 3 tosses which we again compute using the formula $P(V = 3) = \frac{C_3^3}{2^3} = 0.125$

Easy to check that the sum of all probabilities is equal to 1.

The Probability Distribution Table for V

V	v_i	$P(V = v_i)$
	0	0.125
	1	0.375
	2	0.375
	3	0.125
	Total:	1

Consider tossing an unfair coin.

Example 1.6.1.2.4

Example: Let W be the number of Heads out of two tosses of an unfair coin for which the tails are twice as likely as heads. To construct the probability distribution table for W we list all possible outcomes along with their probabilities side-by-side in the form of a table. As in the previous two examples, the probability of each outcome is not that easy to observe and has to be computed.

- The probability that $W = 0$ is the same as the probability of 0 heads among 2 tosses which we compute using the formula $P(W = 0) = C_0^2(\frac{1}{3})^0(\frac{2}{3})^2 = 0.44$
- Similarly, the probability that $W = 1$ is the same as the probability of 1 heads among 2 tosses which we again compute using the formula $P(W = 1) = C_1^2(\frac{1}{3})^1(\frac{2}{3})^1 = 0.44$
- Finally, the probability that $W = 2$ is the same as the probability of 2 heads among 2 tosses which we again compute using the formula $P(W = 2) = C_2^2(\frac{1}{3})^2(\frac{2}{3})^0 = 0.11$

Easy to check that the sum of all probabilities is equal to 1.

The Probability Distribution Table for W

W	w_i	$P(W = w_i)$
	0	0.44
	1	0.44
	2	0.11
	Total:	1

Sometimes the table can be filled out using the theoretical probabilities and sometimes the frequency distribution table can be used as an approximation for probabilities. In the next example, we discuss how to use data to define a random variable.

Example 1.6.1.2.5

Assume that the following data was collected and organized into the frequency distribution table.

Number of cars	Frequency	Relative Frequency
0	6	0.24
1	16	0.64
2	2	0.08
3	0	0.00
4	1	0.04
Total:	25	1

Although the data only represents a sample, assuming that the sample is representative, we can use it to approximate the distribution of the population. If we let Y be the number of cars owned by a randomly selected student, the relative frequency distribution table above can be easily converted into the probability distribution for Y:

Y	y_i	$P(Y = y_i)$
	0	0.24
	1	0.64
	2	0.08
	3	0.00
	4	0.04
	Total:	1

Section 3: Probability Histogram



Let X be the outcome of tossing a fair coin with 0 for heads and 1 for tail.



As previously discussed, there is the probability distribution table for X.

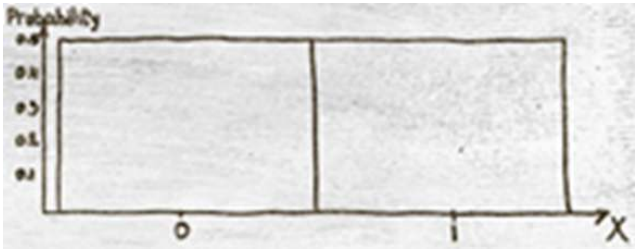
Definition: Probability Histogram

One way to visualize this information about X is to draw the horizontal axis for all values of X, and the vertical axis for the probabilities. Above each value of X we draw the rectangle of height equal to the corresponding probability. Such summary is called the **probability histogram** for X.

The Probability Distribution Table for X

X	x_i	$P(X = x_i)$
	0	0.5
	1	0.5
	Total:	1

The Probability Histogram for X



Similarly, we can construct the probability histogram for variables U, V, W, and Y from the previous examples.

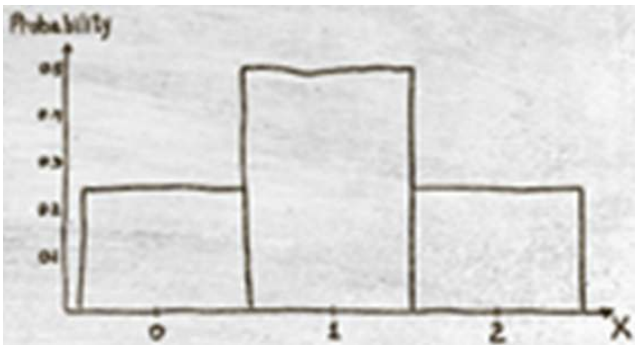
Example 1.6.1.3.1.1

Let U be the number of Heads out of two tosses of a fair coin.

The Probability Distribution Table for U

U	u_i	$P(U = u_i)$
	0	0.25
	1	0.50
	2	0.25
	Total:	1

The Probability Histogram for U



Example 1.6.1.3.1.2

Example: Let V be the number of Heads out of three tosses of a fair coin.

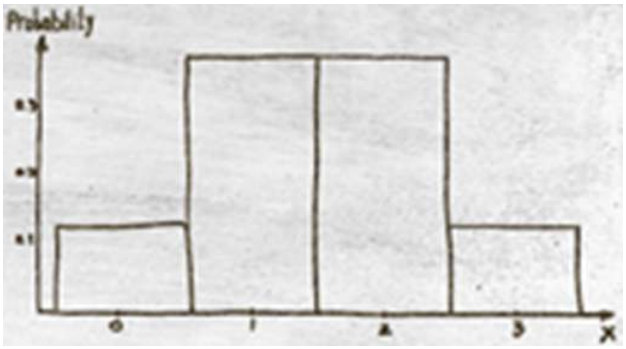
The Probability Distribution Table for V

The Probability Histogram for V

The Probability Distribution Table for V

V	v_i	$P(V = v_i)$
	0	0.125
	1	0.375
	2	0.375
	3	0.125
	Total:	1

The Probability Histogram for V



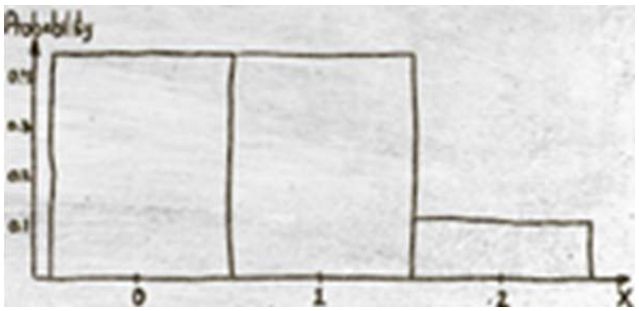
Example 1.6.1.3.1.3

Example: Let W be the number of Heads out of two tosses of an unfair coin with $P(H) = \frac{1}{3}$ and $P(T) = \frac{2}{3}$.

The Probability Distribution Table for W

W	w_i	$P(W = w_i)$
	0	0.44
	1	0.44
	2	0.11
	Total:	1

The Probability Histogram for W



Example 1.6.1.3.1.4

Let Y be the number of cars owned by a randomly selected student.

The Probability Distribution Table for Y

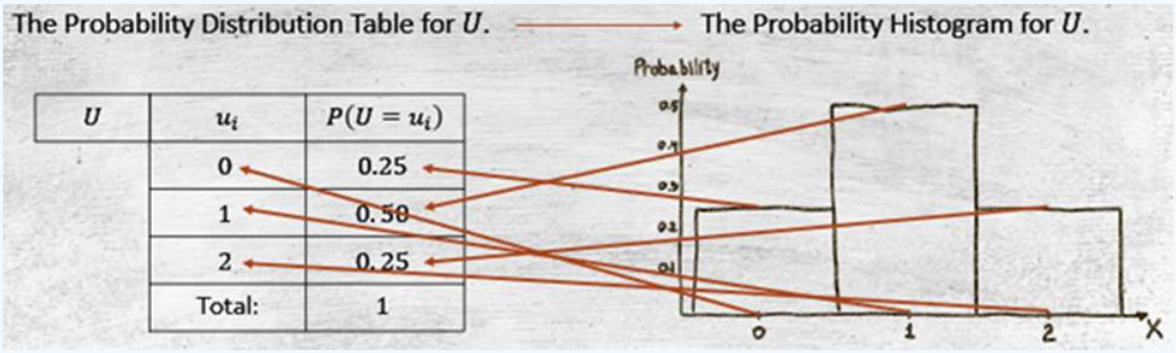
Y	y_i	$P(Y = y_i)$
	0	0.24
	1	0.64
	2	0.08
	3	0.00
	4	0.04
	Total:	1

The Probability Histogram for Y



Just like we can turn the probability distribution table into the probability histogram, we can reverse the process, and construct the probability distribution table from the probability histogram by reading the chart and interpreting the numbers correctly.

Example 1.6.1.3.2



This means that the probability distribution table and the probability histogram contain exactly the same information about the random variable.

Section 4: Probability Rules



Since random variables represent quantities, many events can be expressed in the form of an inequality, and therefore as a result we are interested in being able to compute the probabilities of such events.

$$P(X = c)$$
$$P(X < a)$$
$$P(X \leq a)$$
$$P(X > b)$$
$$P(X \geq b)$$
$$P(a < X < b)$$
$$P(a \leq X < b)$$
$$P(a < X \leq b)$$
$$P(a \leq X \leq b)$$

Previously we learned the probability rules for working with events in general. Next, we will adopt the rules of probability in the context of discrete random variables.

For a discrete random variable X, in addition to the sum of all probabilities being equal to 1:

1. The inequality matters:

$$P(X \leq a) = P(X < a) + P(X = a)$$

2. The complementary rule:

$$P(X \leq a) + P(X > a) = 1$$
$$P(X > a) = 1 - P(X \leq a)$$

3. The subdivision rule:

$$P(X \leq a) + P(a < X \leq b) = P(X \leq b)$$
$$P(a < X \leq b) = P(X \leq b) - P(X \leq a)$$

This page titled 1.6.1: Discrete Random Variables is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.6.2: Applications of Discrete Random Variables

Section 1: Expected Value and Variance



Definition: Mean of a discrete random variable

The **mean of a discrete random variable** X, denoted by $E[X]$ or μ_X, is

$$\sum x_i P(X = x_i)$$

Another name for the mean is the **expected value**.

Definition: Variance and standard deviation of a discrete random variable

The **variance of a discrete random variable** X, denoted by $VAR[X]$ or σ_X^2, is

$$\sum (x_i - \mu_X)^2 \cdot P(X = x_i)$$

and the **standard deviation of a discrete random variable** X, denoted by $SD[X]$ or σ_X, is

$$\sqrt{VAR[X]}$$

Example 1.6.2.1.1

Let Y be the number of cars owned by a randomly selected student with the following probability distribution table for Y:

Y	y_i	$P(Y = y_i)$
	0	0.24
	1	0.64
	2	0.08
	3	0.00
	4	0.04
	Total:	1

Solution

To compute the expected value of Y we can use the probability distribution table in the following way:

- We add another column on the right in which for each outcome we write the product of the outcome by its probability.
- Finally, we find the total 0.96.

Y	y_i	$P(Y=y_i)$	$y_i \cdot P(Y=y_i)$
	0	0.24	$0 \cdot 0.24 = 0$
	1	0.64	$1 \cdot 0.64 = 0.64$
	2	0.08	$2 \cdot 0.08 = 0.16$
	3	0.00	$3 \cdot 0.00 = 0$
	4	0.04	$4 \cdot 0.04 = 0.16$
	Total:	1	0.96

The total is labeled μ_Y, the mean of Y, or $E[Y]$, the expected value of Y.

Example 1.6.2.1.2

To compute the variance of Y we modify the probability distribution table in the following way.

- First, we compute the expected value as previously discussed.
- Next, we add another column on the right in which for each outcome we write the squared difference between each outcome and the expected value.
- Then, we add another column on the right in which for each outcome we write the product of the squared difference by the probability of the outcome.
- Finally, we find the total 0.6784 and the square root of the total $\sqrt{0.6784} = 0.82$.

Y	y_i	$P(Y=y_i)$	$(y_i - E[Y])^2$	$(y_i - E[Y])^2 \cdot P(Y=y_i)$
	0	0.24	$(0-0.96)^2 = 0.9216$	$0.9216 \cdot 0.24 = 0.2212$
	1	0.64	$(1-0.96)^2 = 0.0016$	$0.0016 \cdot 0.64 = 0.0010$
	2	0.08	$(2-0.96)^2 = 1.0816$	$1.0816 \cdot 0.08 = 0.0865$
	3	0.00	$(3-0.96)^2 = 4.1616$	$4.1616 \cdot 0.00 = 0.0000$
	4	0.04	$(4-0.96)^2 = 9.2416$	$9.2416 \cdot 0.04 = 0.3697$
	$E[Y] = 0.96$	1		$VAR[Y] = 0.6784$
				$SD[Y] = \sqrt{0.6784} = 0.82$

The total is labeled σ_Y^2, the variance of Y, and its square root is labeled σ_Y, the standard deviation of Y.

In summary, for a discrete random variable X we can compute the mean and the standard deviation. The process can be summarized with the following formulas:

Expected Value, $E[X] = \mu_X$	Standard Deviation, $SD[X] = \sigma_X$
$\mu_X = \sum x_i P(X = x_i)$	$\sigma_X = \sqrt{\sum (x_i - \mu_X)^2 \cdot P(X = x_i)}$
We interpret the mean of X as the average value of the variable in a long run. In other words, if we perform the experiment many times, the average value of many outcomes can be predicted by the mean of X. So the mean of X sets up the "expectations" from the experiment before the experiment is actually performed.	We interpret the standard deviation of X as the expected deviation of the outcomes from the expected value in a long run.

However, the standard deviation is better interpreted when combined with the mean of X using the three standard deviations rule in the following way:

Range	Interpretation
within $2\sigma_x$ from μ_X	"expected"
more than $2\sigma_x$ from μ_X	"unusual"
more than $3\sigma_x$ from μ_X	"abnormal"

Example 1.6.2.1.3.1

For a discrete random variable X with $\mu_X = 10$ and $\sigma_X = 2$ then by the three standard deviations rule, we may conclude that:

Range	Numerical Range	Interpretation
within $2\sigma_x$ from μ_X	between 6 and 14	"expected"
more than $2\sigma_x$ from μ_X	less than 6 or more than 14	"unusual"
more than $3\sigma_x$ from μ_X	less than 4 or more than 16	"abnormal"

Example 1.6.2.1.3.2

For Y, the number of cars owned by a randomly selected student, with $\mu_Y = 0.96$ and $\sigma_Y = 0.82$, we may conclude that:

Range	Numerical Range	Interpretation
within $2\sigma_x$ from μ_X	between -0.68 and 2.60	"expected"
more than $2\sigma_x$ from μ_X	less than -0.68 or more than 2.60	"unusual"
more than $3\sigma_x$ from μ_X	less than -1.50 or more than 3.42	"abnormal"

Section 2: Applications of Discrete Random Variables



Gambling

Let's play an American roulette! In American roulette there are 18 red slots, 18 black slots, and 2 green slots.



So, the probability of the ball landing on red is 18/38 and the probability of the ball landing on green or black is 20/38. The rules of the game are such that if you place a $1 bet on red, you will win a $1 if the result of the spin is red and lose $1 if the result of the spin is not red. Placing a bet can be thought of as an experiment whose outcome depends on a chance, therefore we can introduce a random variable X representing the payoff of such a bet. Then the following probability distribution table can be constructed for X by listing the possible values of X, for winning and losing scenarios, along with their probabilities.

X	outcome	x_i	$P(X = x_i)$
	Red	+1	18/38
	Black or Green	-1	20/38
		Total:	1

We can now compute the expected value by adding another column on the right and computing the product of each possible value by its probability. The sum of the column is $-\frac{2}{38} = -0.05$ and this is the value that we call the expected value.

X	outcome	x_i	$P(X = x_i)$	$x_i \cdot P(X = x_i)$
	Red	+1	18/38	18/38
	Black or Green	-1	20/38	-20/38
		Total:	1	-2/38

So what does this number mean in the context of this problem? We interpret the mean as the average winnings of placing a $1 bet on red! In other words, on average you would lose 5 cents per game. Of course, every game you either win or lose a dollar but on average it is always a loss! Every game is set up in such a way that the expected payoff of a bet is negative! Do not gamble!

Insurance



Another application of a probability is in the insurance industry. Let's say Bobert just purchased a new car and wants to purchase an additional insurance policy against totaling the car. So Bobert goes to insurance agent's office and says, "I want to purchase a policy against totaling my $20000 car". What happens next? The agent would collect the data about Bobert such as age, driving experience etc. and send it to an actuary. Actuary is a person whose entire job is to evaluate the probabilities of different events. In this case, the actuary will look at the data on how many people just like Bobert had totaled their cars annually in the past and using the data they can estimate the probability of Bobert totaling his car in the next year. So, let us say the probability of him totaling his car during in the next year is estimated to be 3.5%. Using the complimentary rule, we can compute the probability of him not totaling the car which is 96.5%. Let's say that Bobert ended up purchasing the policy for a lump sum of $1000 called premium, and in exchange he received an assurance that in case of him totaling the car he will get paid $20000.

The purchase of an insurance can be thought of as an experiment whose outcome depends on a chance, therefore we can introduce a random variable X representing the profit of the insurance company from selling the insurance to Bobert. Then the following probability distribution table can be constructed for X by listing the possible values of X, for not totaling and totaling scenarios, along with their probabilities.

X	outcome	x_i	$P(X = x_i)$
	Not Totaling	+1000	0.965
	Totaling	-20000+1000	0.035
		Total:	1

We can now compute the expected value by adding another column on the right and computing the product of each possible value by its probability. The sum of the column is 300 and this is the value that we call the expected value.

X	outcome	x_i	$P(X = x_i)$	$x_i \cdot P(X = x_i)$
	Not Totaling	+1000	0.965	965
	Totaling	-20000+1000	0.035	-665
		Total:	1	300

So what does this number mean in the context of this problem? We interpret the mean as the average profit of the insurance company from insuring drivers like Bobert. From insuring 1000 people like Bobert they expect to make 300,000 dollars.

In summary, buying an insurance policy is equivalent to making a bet with the insurance company that a certain event will happen and insurance company would not sell you a policy unless they had a positive expected profit. Insurance business is a very profitable business.

This page titled 1.6.2: Applications of Discrete Random Variables is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.6.3: Binomial Random Variables

Section 1: Introduction to Binomial Random Variables



Definition: Binomial experiment

A **binomial experiment** is an experiment that has exactly two outcomes.

Example 1.6.3.1.1.1

Tossing a coin is a binomial experiment with two outcomes: Heads and Tails.

However, rolling a die and observing an outcome is not a binomial experiment because there are six possible outcomes.

Example 1.6.3.1.1.2

Going to work and observing the tardiness is a binomial experiment with two outcomes - Late and On-Time.

However, asking a randomly selected person how many credit cards they have is not a binomial experiment with two outcomes because a person may have any number of cards: 0, 1, 2, 3, etc.

Why binomial experiments are so important? Because any experiment with any number of outcomes can be thought of as a binomial experiment by considering an event E, then for that experiment we have two outcomes - E and $(\text{not } E)$.

Example 1.6.3.1.2.1

Rolling a die can be thought of as a binomial experiment by considering the two complementary events: Even and Odd, (Six) and (Other than Six) etc.

Example 1.6.3.1.2.2

Asking how many credit cards a randomly selected person has can be thought of as a binomial experiment by considering two complementary events: (None) and (At least one)

Example 1.6.3.1.2.3

More examples of binomial experiments:

- A randomly selected patient dies or survives;
- A randomly selected store visitor makes a purchase or not;
- A randomly selected passenger arrives on-time or late;
- A randomly selected applicant has a master's degree or not;
- A randomly selected student graduates in four years or more, etc.

For a binomial experiment we usually label one outcome a success, and the other one is a failure which is the complement of success. The probability of success is labeled p and the probability of failure is labeled q which by the complementary rule can always be computed by subtracting p from 1. Note that success may not be associated with a literal success but in any case, the failure is the opposite or the complementary event of success.

Example 1.6.3.1.3

For a randomly selected patient, we can label S the event in which the patient dies, then F is the event in which the patient survives.

For a randomly selected store visitor, we can label S the event in which a purchase is made, then F is the event in which the purchase was not made.

For a randomly selected passenger, we can label S the event in which the passenger arrives on-time, then F is the event in which the passenger is late.

Definition: Bernoulli trials

If the same binomial experiment is performed as a sequence of n independent experiments, then such sequence is called n **Bernoulli trials**. The key word here is independent, which means that there should not be any dependency between the different instances of the experiment.

Example 1.6.3.1.4

If we consider a group of 10 patients and track whether they survive or not, we have 10 Bernoulli trials.

If we consider a group of 100 store visitors and track whether they make a purchase or not, we have 100 Bernoulli trials.

If we consider 200 passengers and track whether they show up on-time or late, we have 200 Bernoulli trials.

If we consider 1000 students and track whether they graduate in four years or more, we have 1000 Bernoulli trials.

Next, we going to establish the following two facts.

Fact#1

A binomial experiment can be simulated by a toss of an unfair coin for which S is the event in which the toss results in Heads and F is the event in which the toss results in Tails, and

$P(S) = P(H) = p$ and $P(F) = P(T) = 1 - p = q$

Fact#2

A sequence of n Bernoulli trials can be simulated by n tosses of an unfair coin for which we have established the following formula:

$$P(kH/n) = C_k^n p^k q^{n-k}$$

for finding the probability of observing k heads among n tosses.

Next, let's consider n Bernoulli trials of a binomial experiment with $P(S)=p$ and $P(F)=1-p=q$. Before performing the binomial experiment n times, it is impossible to know the number of successes. Therefore, the number of successes depends on a chance. If we let X be the number of successes among n Bernoulli trials, then X satisfies the definition of a discrete random variable. Why discrete? Because the possible number of successes can be $0, 1, 2, \ldots$ and any number up to n. Just like for any discrete random variable we can construct the probability distribution table for X by listing all of its possible values side-by-side with their probabilities

X	k	$P(X=k)$
	0	$P(X=0)$
	1	$P(X=1)$
	2	$P(X=2)$
	$\ldots$	$\ldots$
	n	$P(X=n)$

If a binomial experiment can be simulated by a toss of a fair coin with $P(S)=P(H)=p$ and $P(F)=P(T)=1-p$ then $P(X=k)=P(kH/n)=C_k^n p^k q^{n-k}$. The significance of the formula is that now we can compute the probabilities of each outcome by simply evaluating the formula for different values of k. So, the entire distribution table can be computed from just two values n and p.

X	k	$P(X=k)=C_k^n p^k q^{n-k}$
	0	$P(X=0)=C_0^n p^0 q^{n-0}$
	1	$P(X=1)=C_1^n p^1 q^{n-1}$
	2	$P(X=2)=C_2^n p^2 q^{n-2}$
	$\ldots$	$\ldots$
	n	$P(X=n)=C_n^n p^n q^{n-n}$

Definition: Binomial Random Variable

This variable X representing the number of successes among n Bernoulli trials with the probability of success p is called a **Binomial Random Variable** with parameters n and p and is denoted in the following way:

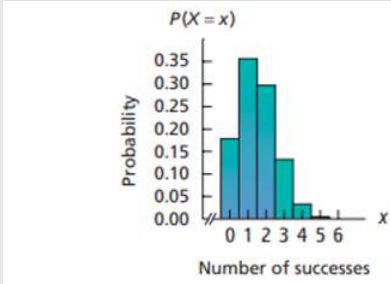
$$X \sim B(n,p)$$

Example 1.6.3.1.7

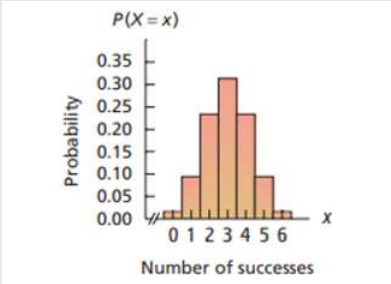
For a binomial random variable with parameters $n=3$ and $p=0.6$, we construct the probability distribution table by replacing n with 3 and p with 0.6 and k with every value from zero to 3.

X	k	$P(X=k)=C_k^3 p^k q^{3-k}$
	0	$P(X=0)=C_0^3 0.6^0 0.4^{3-0}=0.064$
	1	$P(X=0)=C_1^3 0.6^1 0.4^{3-1}=0.288$
	2	$P(X=2)=C_2^3 0.6^2 0.4^{3-2}=0.432$
	3	$P(X=3)=C_3^3 0.6^3 0.4^{3-3}=0.216$

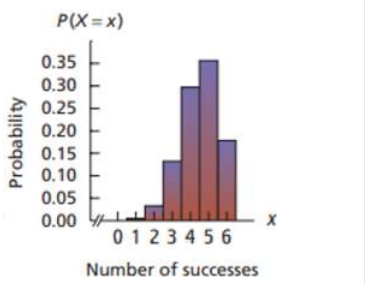
The following relation can be observed between the shape of the histogram and the value of p, one of the parameters of the variable.



Right Skewed, p=0.25



Symmetric, p=0.5



Left Skewed, p=0.75

Now we can use the probability distribution table and the histogram to find probabilities of a variety of events.

Example 1.6.3.1.8

Let's find the probability of X being greater than or equal to 2.

Solution

To find the probability we first identify the event in the probability distribution table and/or the histogram and then add the corresponding probabilities to get the answer, in this case 0.648.

X	k	$P(X=k)=C_k^3 p^k q^{3-k}$
	0	$P(X=0)=C_0^3 0.6^0 0.4^{3-0}=0.064$
	1	$P(X=0)=C_1^3 0.6^1 0.4^{3-1}=0.288$
	2	$\boldsymbol{P(X=2)=C_2^3 0.6^2 0.4^{3-2}=0.432}$
	3	$\boldsymbol{P(X=3)=C_3^3 0.6^3 0.4^{3-3}=0.216}$

The good news is that rarely one must compute these probabilities by hand.

Another task that we can do for a binomial random variable is to compute the mean, variance and standard deviation. We can compute these values using the same approach as previously developed, however, luckily for us, there is an easier way that allows us to simply use the following formulas to compute the mean, variance, and standard deviation.

Theorem: Mean, variance, and standard deviation of a Binomial RV

For $X \sim B(n,p)$:

$$\mu_X = E[X] = np$$
$$\sigma_X^2 = VAR[X] = npq$$
$$\sigma_X = SD[X] = \sqrt{npq}$$

Example 1.6.3.1.9

In the above example,

$$E[X] = np = 3 \cdot 0.6 = 1.8$$
$$VAR[X] = npq = 3 \cdot 0.6 \cdot 0.4 = 7.2$$
$$SD[X] = \sqrt{npq} = \sqrt{7.2} = 2.68$$

Interpretation: we expect 1.8 successes among 3 Bernoulli trials with $p = 0.6$.

Section 2: Applications of Binomial Random Variables



What makes binomial random variables quite useful is the following fact: if a certain ratio of the population possesses a certain property and a sample of size less than 5% of the population is obtained, then the number of people that possess the property in the sample is a binomial random variable. Note that when the sample size is more than 5% of the population the same holds only if the sample is obtained by sampling with replacement.

Example 1.6.3.2.1

Given that 86% of the population have a HS diploma, find the probability that among a group of 10 applicants all have a HS diploma.

Solution

To start the solution, we introduce a variable. Let X be the number of people out of 10 that have a HS diploma. Since the sample size is 10 which is much less than 5% of the US population therefore X is a binomial random variable with parameters $n = 10$ and $p = 0.86$.

$$X \sim B(n = 10, p = 0.86)$$

The probability that among a group of 10 applicants all have a HS diploma can be found by using the formula for $k = 10$.

$$P(X = 10) = C_{10}^{10} 0.86^{10} 0.14^{0} = 0.2213 = 22.13\%$$

So the probability that all 10 applicants have a HS diploma is roughly 22%.

Example 1.6.3.2.1

(... continued)

With enough effort we can create the frequency distribution table for X:

X	k	$P(X = k) = C_k^n p^k q^{n-k}$
	0	$P(X = 0) = C_0^{10} 0.86^0 0.14^{10} = 2.89255 \cdot 10^{-9}$
	1	$P(X = 1) = C_1^{10} 0.86^1 0.14^9 = 1.77685 \cdot 10^{-7}$
	2	$P(X = 2) = C_2^{10} 0.86^2 0.14^8 = 4.91172 \cdot 10^{-6}$
	3	$P(X = 3) = C_3^{10} 0.86^3 0.14^7 = 8.04587 \cdot 10^{-5}$
	4	$P(X = 4) = C_4^{10} 0.86^4 0.14^6 = 0.000864931$

X	k	$P(X=k)=C_k^n p^k q^{n-k}$
	5	$P(X=5)=C_5^{10}0.86^5 0.14^5=0.006375775$
	6	$P(X=6)=C_6^{10}0.86^6 0.14^4=0.032637895$
	7	$P(X=7)=C_7^{10}0.86^7 0.14^3=0.114565673$
	8	$P(X=8)=C_8^{10}0.86^8 0.14^2=0.263910212$
	9	$P(X=9)=C_9^{10}0.86^9 0.14^1=0.360258384$
	10	$P(X=10)=C_{10}^{10}0.86^{10} 0.14^0=0.221301579$

And use the table to find the following probabilities:

- the probability that more than half have a HS diploma, $P(X>5)=0.9927$.
- the probability that at most 8 have a HS diploma, $P(X\leq 8)=0.4184$.
- the probability that between 4 and 7 have a HS diploma, $P(4\leq X\leq 7)=0.1544$.

Example 1.6.3.2.1

(... continued)

In addition to finding the probabilities we can also find and interpret the expected value and the standard deviation.

$$\mu_X = np = 10\cdot 0.86 = 8.6$$

$$\sigma_X = \sqrt{npq} = \sqrt{10\cdot 0.86\cdot 0.14} = 1.1$$

In this example, we expect 8 or 9 applicants out of 10 to have a HS diploma.

Section 3: Overbooking



Example 1.6.3.3.1.1

An airline sold all tickets on a 150-seat plane. Given that only 98% of the people show up for a flight, find the probability that 150 people will show up.

Solution

To start the solution, we introduce a variable. Let X be the number of people out of 150 that will show up, then

$$X \sim B(n = 150, p = 0.98)$$

The probability that all 150 people that purchased a ticket will show up for the flight can be expressed as

$$P(X = 150) = C_{150}^{150} 0.98^{150} 0.02^{0} = 0.0483 \approx 4.8\%$$

In addition to finding this probability we can also find and interpret the expected value which is 147:

$$E[X] = np = 150 \cdot 0.98 = 147$$

In this context, we expect only 147 passengers to show up for the flight. That means that the airline can sell more than 150 tickets for a flight and since not everyone is going to show up, the airline can make more money by selling more tickets. This practice is called **overbooking**.

Example 1.6.3.3.1.2

An airline sold an extra 3 tickets on a 150-seat plane. Given that only 98% of the people show up for a flight, find the probability that at most 150 people will show up. This is the probability of the event in which the airline will have a seat for everyone who shows up. This is the best possible outcome for the airline.

Solution

To start the solution, we again introduce a variable. Let Y be the number of people out of 153 that purchased a ticket and showed up for the flight. Then Y is a binomial random variable with parameters $n = 153$ and $p = 0.98$.

$$Y \sim B(n = 153, p = 0.98)$$

The probability that less than or equal to 150 passengers will show up for the flight can be found by using technology.

$$P(Y \leq 150) = <\text{via technology}> = 0.5925 = 59.25\%$$

Thus, the probability that the airline will have a seat for everyone who shows up for the flight is roughly 60%.

So, what happens when too many passengers show up for a flight? They will be offered a compensation for their trouble. Every airline decides how many tickets to sell for every flight in such a way to maximize the profit and minimize the probability of getting in trouble!

This page titled 1.6.3: Binomial Random Variables is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

SECTION OVERVIEW

1.7: Continuous Random Variables

1.7.1: Continuous Random Variables

1.7.2: Standard Types of Continuous Random Variables

1.7.3: Other Continuous Random Variables

This page titled 1.7: Continuous Random Variables is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.7.1: Continuous Random Variables

Section 1: Introduction to Continuous Random Variables



Previously, we introduced a new framework, which is built around the concept of a random variable. We already discussed the discrete random variables. Let's summarize everything we know about discrete random variables and see what can be used for studying continuous random variables:

1. Every discrete random variable X has the probability distribution table and the probability histogram.
2. Using the probability distribution table or histogram we can find the probabilities of events that are in the form of an inequality that involves X.
3. Also, we are able to find and interpret the mean and standard deviation of a discrete random variable.

Definition: Continuous random variable

A **continuous random variable** is a random variable whose all possible values form an interval.

Example 1.7.1.1.1

For example, the following variables are continuous:

- The income of a randomly selected household;
- The lifetime of a randomly selected machinery, etc.
- The wait time in a drive-through of a randomly selected customer;
- The incubation period of a virus for a randomly selected patient, and there are many others.

We want to be able to answer the same questions about ANY continuous random variable. But right away we run into a problem. The fact that it is impossible to list all values of a continuous random variable makes it impossible to construct a probability distribution table, so instead, we are going to focus on its visual representation. An analog of a histogram for a continuous random variable is a **probability density curve**.

Definition: Probability density curve

A **probability density curve** is a curve that has the following properties:

1. always above or on the horizontal axis;
2. the total area under the curve is equal to 1.

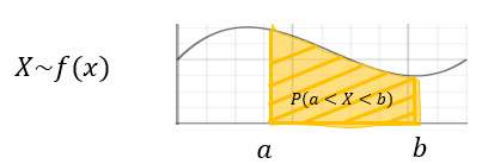
Example 1.7.1.1.2

Determine if a curve is a probability density curve or not.

- $y = 2 - \frac{1}{2}x^2, 0 \leq x \leq 3$ is not a probability density curve;
- $y = \frac{1}{4}, -1 \leq x \leq 3$ is a probability density curve;
- $x^2 + y^2 = 1, y \geq 0$ is not a probability density curve;
- $y = \frac{1}{8}x, 0 \leq x \leq 4$ is a probability density curve.

Definition:

We say that **a continuous random variable X has a given probability density curve $f(x)$** if the probability that X is between a and b, i.e., $P(a \leq X \leq b)$, is equal to the area under $f(x)$ between a and b for any a, b.



Notation: $X \sim f(x)$ means "X has the probability density curve $f(x)$".

In other words, if X has the probability density curve $f(x)$ the probability that X is between a and b is equal to the area under the curve between a and b for any a, b.

In general, it is not always possible to find the mean and standard deviation of a continuous random variable, but when it is possible, we interpret the numbers the same way as for discrete random variables.

Section 2: Probability Rules



For a continuous random variable X, in addition to the total probability under the curve being equal to 1:

1. The inequality doesn't matter:

$$P(X \leq c) = P(X < c)$$
$$P(X \geq c) = P(X > c)$$

2. The complementary rule:

$$P(X < c) + P(X > c) = 1$$

3. The subdivision rule:

$$P(a < X < b) = P(X < b) - P(X < a)$$

We derived the probability rules in the context of continuous random variables. It is easy to notice that with a few nuances these rules are the same as the ones derived for discrete random variables.

Section 3: Probability as Areas

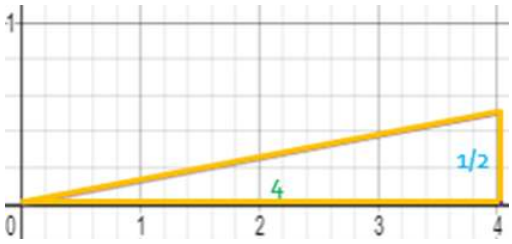


We will consider an example of a continuous random variable and practice some skills.

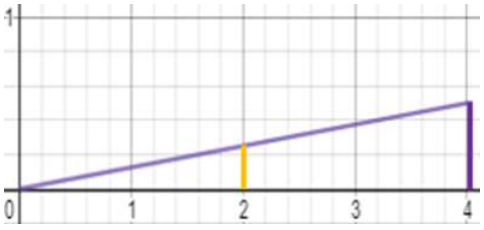
Let X be a continuous random variable with the probability density curve described by the graph of a linear function on the interval from 0 to 4:

$$X \sim f(x) = \frac{1}{8}x, 0 \le x \le 4$$

Easy to see that the graph satisfies the criteria for a probability density curve as it is above the x-axis and its area is the area of a right triangle with the base 4 and the height $\frac{1}{2}$ which can be found using the formula for the area of a right triangle $A = \frac{1}{2}bh = \frac{1}{2} \cdot 4 \cdot \frac{1}{2} = 1$.

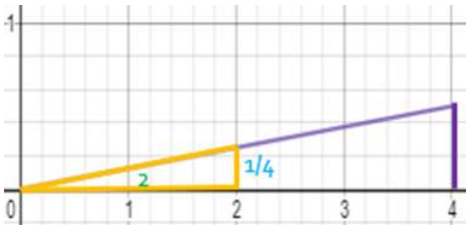


Let's find the probability of $X = 2$. As we previously discussed, the probability of a continuous random variable being equal to a single value is equal to 0, $P(X = 2) = 0$.



To find the probability that X is less than 2 we will look for the area under the curve between 0 and 2. This region is a triangle with the base 2 and the height $\frac{1}{4}$. The area can be found using the formula for the area of a right triangle and is equal to

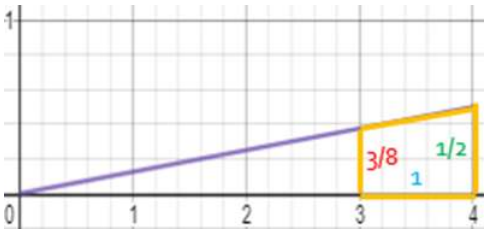
$P(X<2)=0.25$ or 25%.



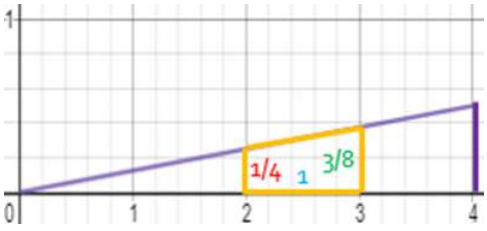
To find the probability that X is less than 3 we will look for the area under the curve between 0 and 3. This region is a triangle with the base 3 and the height $\frac{3}{8}$. The area can be found using the formula for the area of a right triangle and is equal to $P(X<3)=0.5625$ or 56.25%



To find the probability that X is greater than 3 we will look for the area under the curve between 3 and 4. This region is a trapezoid with the bases $\frac{3}{8}$ and $\frac{1}{2}$ and the height 1. The area can be found using the formula for the area of a right trapezoid and is equal to $P(3<X<4)=0.4375$ or 43.75% Alternatively, we could have used the complementary rule to compute the same probability as the complement of the probability of X being less than 3 which we already computed previously, so the answer is $P(3<X<4)=1-P(X<3)=1-0.5625=0.4375$ which is again 43.75%



To find the probability that X is between 2 and 3 we will look for the area under the curve between 2 and 3. This region is a trapezoid with the bases $\frac{1}{4}$ and $\frac{3}{8}$ and the height 1. The area can be found using the formula for the area of a right trapezoid and is equal to $P(2<X<3)=0.3125$ or 31.25% Alternatively, we could have used the subdivision rule to compute the same probability as the difference between the probability of X being less than 3 and the probability of X being less than 2 both of which we already computed previously, so the answer is $P(2<X<3)=P(X<3)-P(X<2)=0.3125$ which is again 31.25%



It is worth noting that the probabilities of X being greater than 5 or less than -2 are both 0 because the probability density curve has no area in that region. Also, the probability of X being greater than -2 or less than 5 are both 1 because the probability density curve has its entire area in that region which is equal to 1 by definition.

This page titled 1.7.1: Continuous Random Variables is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.7.2: Standard Types of Continuous Random Variables

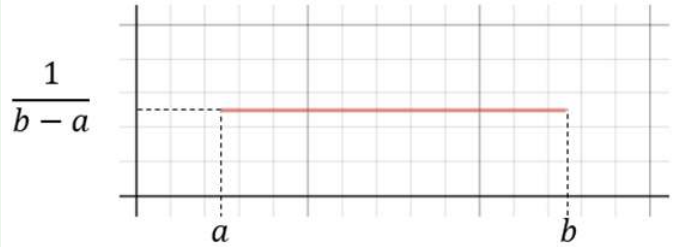
Section 1: Uniform Random Variable



Next, we will consider one of the standard continuous random variables called uniform and how to work with it.

Definition: Uniform probability density curve

A **uniform probability density curve** with parameters a and b is a horizontal line $y = \frac{1}{b-a}$ from $x = a$ to $x = b$.



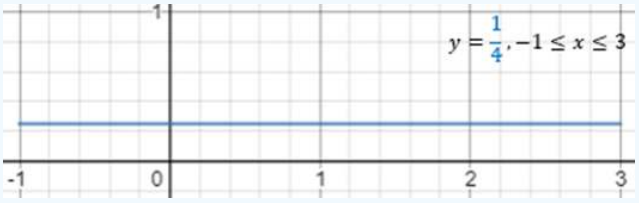
Definition: Uniform random variable

A random variable with a uniform probability density curve is called a **uniform random variable** with parameters a and b. We adopt the following notation to denote a uniform random variable with parameters a and b:

$$X \sim U(a, b)$$

Example 1.7.2.1.1.1

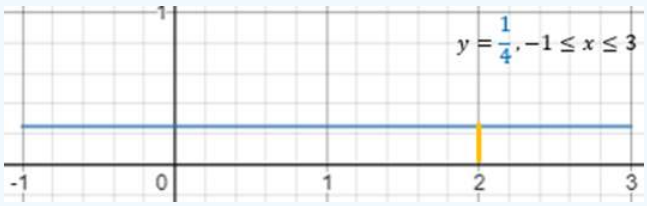
If $X \sim U(-1, 3)$ is a uniform random variable with parameters -1 and 3 then its probability density curve will look like a horizontal line spanning from -1 to 3.



Easy to see that the graph satisfies the criteria for a probability density curve as it is above the x-axis and its area is the area of a rectangle with the base 4 and the height $\frac{1}{4}$ which can be found using the formula for the area of a rectangle and is equal to 1.

Example 1.7.2.1.1.2

Let's find the probability of $X = 2$. As we previously discussed, the probability of a continuous random variable being equal to a single value is equal to $P(X = 2) = 0$.



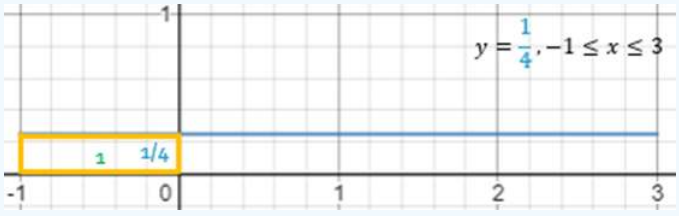
Example 1.7.2.1.1.3

To find the probability that X is less than 2 we will look for the area under the curve between -1 and 2. This region is a rectangle with the base 3 and the height $\frac{1}{4}$. The area can be found using the formula for the area of a rectangle and is equal to $P(X < 2) = 0.75$ or 75%.



Example 1.7.2.1.1.4

To find the probability that X is less than 0 we will look for the area under the curve between -1 and 0. This region is a rectangle with the base 1 and the height $\frac{1}{4}$. The area can be found using the formula for the area of a rectangle and is equal to $P(-1 < X < 0) = 0.25$ or 25%.



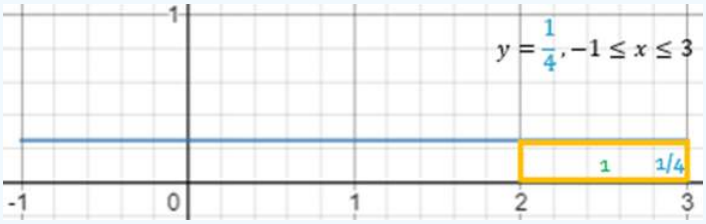
Example 1.7.2.1.1.5

To find the probability that X is between 0 and 2 we will look for the area under the curve between 0 and 2. This region is a rectangle with the base 2 and the height $\frac{1}{4}$. The area can be found using the formula for the area of a rectangle and is equal to $P(0 < X < 2) = 0.5$ or 50%. Alternatively, we could have used the subdivision rule to compute the same probability as the difference between the probability of X being less than 2 and the probability of X being less than 0 both of which we already computed previously, so the answer is again 50%.



Example 1.7.2.1.1.6

To find the probability that X is greater than 2 we will look for the area under the curve between 2 and 3. This region is a rectangle with the base 1 and the height $\frac{1}{4}$. The area can be found using the formula for the area of a rectangle and is equal to $P(X > 2) = 0.25$ or 25%. Alternatively, we could have used the complementary rule to compute the same probability as the complement of the probability of X being less than 2 which we already computed previously, so the answer is again 25%.



Theorem: Formulas for Continuous Uniform Random Variables

In general, for $X \sim U(a, b)$ and any c, d such that $a < c < d < b$:

$$P(c < X < d) = \frac{d-c}{b-a}$$

$$P(X < d) = \frac{d-a}{b-a}$$

$$P(X > c) = \frac{b-c}{b-a}$$

Also,

$$\mu_X = \frac{a+b}{2}$$

$$\sigma_X = \frac{b-a}{\sqrt{12}}$$

Example 1.7.2.1.2

For example, for $X \sim U(-1, 3)$:

$$P(0 < X < 2) = \frac{2-0}{3-(-1)} = \frac{2}{4} = 0.50$$

$$P(X < 0) = \frac{0-(-1)}{3-(-1)} = \frac{1}{4} = 0.25$$

$$P(X > 2) = \frac{3-2}{3-(-1)} = \frac{1}{4} = 0.25$$

Also,

$$\mu_X = \frac{-1+3}{2} = 2$$

$$\sigma_X = \frac{3-(-1)}{\sqrt{12}} = 1.15$$

Example 1.7.2.1.3

A city metro train runs every 15 minutes.

1. Find the probability that the waiting time is:
   i. less than 8 minutes;
   ii. more than 3 minutes;
   iii. between 3 and 7 minutes.
2. Compute the following:
   i. the expected waiting time;
   ii. the standard deviation of waiting time.

Solution

Waiting time is frequently estimated by a uniform random variable. Let X be the waiting time of a randomly selected passenger then it can be assumed uniform with parameters 0 and 15, i.e.:

$$X \sim U(0, 15)$$

Using the formulas we can compute all the quantities.

1. Find the probability that the waiting time is:
   i. less than 8 minutes: $P(X < 8) = \frac{8-0)}{15-0} = \frac{8}{15} = 0.53$
   ii. more than 3 minutes: $P(X > 3) = \frac{15-3}{15-0} = \frac{12}{15} = 0.80$
   iii. between 3 and 7 minutes: $P(3 < X < 7) = \frac{7-3}{15-0} = \frac{4}{15} = 0.27$
2. Compute the following:
   i. the expected value of X: $\mu_X = \frac{0+15}{2} = 7.5$
   ii. the standard deviation of X: $\sigma_X = \frac{15-0}{\sqrt{12}} = 4.33$

Section 2: Standard Normal Variable



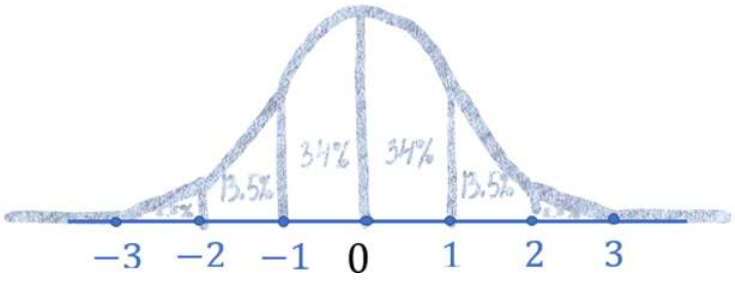
Definition: Standard normal density curve

The **standard normal density curve** is a bell-shaped curve that satisfies the following properties:

1. Has the peak at 0 and symmetric about 0.
2. Extends indefinitely in both directions, approaching but never touching the horizontal axis.

3. The Empirical rule holds that is:
   a. ~68% of the area under the curve is between -1 and +1
   b. ~95% of the area under the curve is between -2 and +2
   c. ~99.7% of the area under the curve is between -3 and +3

The description of the curve should sound familiar as we have seen it before!



Definition: Standard normal probability density curve

We define the standard normal random variable as a random variable with the **standard normal probability density curve**. We denote the standard normal random variable as

$$Z \sim N(0,1)$$

Let's make sure we understand the properties of the standard normal probability density curve and use the empirical rules to find some probabilities.

Example 1.7.2.2.1

- To find the probability that X is less than 0 we will look for the area under the curve to the left of 0. This region's area can be found by adding the areas of included regions together to get $P(X < 0) = 0.50$ or 50%.
- To find the probability that X is between -1 and 1 we will look for the area under the curve between -1 and 1. This region's area can be found by adding the areas of included regions together to get $P(-1 < X < 1) = 0.68$ or 68%.
- To find the probability that X is between 0 and 2 we will look for the area under the curve between 0 and 2. This region's area can be found by adding the areas of included regions together to get $P(0 < X < 2) = 0.475$ or 47.5%.
- To find the probability that X is between -1 and 2 we will look for the area under the curve between -1 and 2. This region's area can be found by adding the areas of included regions together to get $P(-1 < X < 2) = 0.815$ or 81.5%.
- To find the probability that X is greater than 1 we will look for the area under the curve to the right of 1. This region's area can be found by adding the areas of included regions together to get $P(X > 1) = 0.16$ or 16%.
- To find the probability that X is less than -2 we will look for the area under the curve to the left of -2. This region's area can be found by adding the areas of included regions together to get $P(X < -2) = 0.025$ or 2.5%.
- To find the probability that X is greater than 4 we will look for the area under the curve to the right of 4. This region's area is essentially equal to 0 so the probability is nearly 0%, that is $P(X > 4) = 0$.

To find the probability that X is less than 1.23 we will look for the area under the curve to the left of 1.23. While it is clear what this region's area looks like it cannot be found by using the empirical rule.

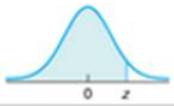
So how do we find the probabilities that involve values that are not covered by the Empirical rule such as $P(Z < 1.23)$, $P(Z > 1.23)$, $P(0 < Z < 1.23)$. The good news is that the latter two probabilities can be related to the first one by using the complementary rule and the subdivision rule:

$$P(Z > 1.23) = 1 - P(Z < 1.23)$$

$$P(0 < Z < 1.23) = P(Z < 1.23) - 0.5$$

But how do we find the probability $P(Z < 1.23)$? Luckily all such probabilities have been computed and put together in the form of a table (part 1, part 2) that looks like this!

Areas under the standard normal curve*



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177

To find the probability that Z is less than 1.23 we split the number 1.23 into 1.2 and 0.03 then find the row corresponding to 1.2 and the column corresponding to 0.03. In the intersection, we find the desired probability, so the probability is $P(Z<1.23)=0.8907$ or 89.07%.

Now that we found the probability $P(Z<1.23)$, we can find the other probabilities using the complementary rule and the subdivision rule:

- $P(Z<1.23)=0.8907$
- $P(Z>1.23)=1-P(Z<1.23)=1-0.8907=0.1093$
- $P(0<Z<1.23)=P(Z<1.23)-0.5=0.8907-0.5=0.3907$

Let's practice using the table!

Example 1.7.2.2.2.1

To find the probability that Z is less than 0.71, we split the number 0.71 into 0.7 and 0.01, find the row corresponding to 0.7, find the column corresponding to 0.01. In the intersection, we find the desired probability, so the probability is $P(Z<0.71)=0.7611$ or 76.11%. Also,

- $P(Z<0.71)=0.7611$
- $P(Z>0.71)=1-P(Z<0.71)=1-0.7611=0.2389$
- $P(0<Z<0.71)=P(Z<0.71)-0.5=0.7611-0.5=0.2611$
- $P(0.71<Z<1.23)=P(Z<1.23)-P(Z<0.71)=0.8907-0.7611=0.1296$

Example 1.7.2.2.2.2

Example: To find the probability that Z is less than -1.02, we split the number -1.02 into -1.0 and -0.02, find the row corresponding to -1.0, find the column corresponding to -0.02. In the intersection, we find the desired probability, so the probability is $P(Z<-1.02)=0.1539$ or 15.39%. Also,

- $P(Z<-1.02)=0.1539$
- $P(Z>-1.02)=1-P(Z<-1.02)=1-0.1539=0.8461$

Example 1.7.2.2.2.3

To find the probability that Z is less than -0.47, we split the number -0.47 into -0.4 and -0.07, find the row corresponding to -0.4, find the column corresponding to -0.07. In the intersection, we find the desired probability, so the probability is $P(Z<-0.47)=0.3192$ or 31.92%. Also,

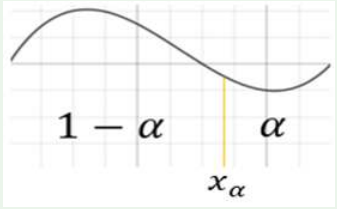
- $P(Z<-0.47)=0.3192$
- $P(Z>-0.47)=1-P(Z<-0.47)=1-0.3192=0.6808$
- $P(-1.02<Z<-0.47)=P(Z<-0.47)-P(Z<-1.02)=0.3192-0.1539=0.1653$

Section 3: α-notation



Definition: α-notation

Consider a random variable X and its probability density curve. A value the area to the right of which is equal to alpha is called x_α.



This fact can be expressed as the following probability statement:

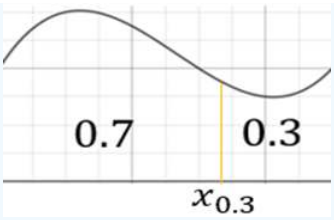
$$P(X>x_\alpha)=\alpha$$

As a result, the area to the left of x_α is by the complementary rule must be $1-\alpha$. This fact can be expressed as the following probability statement:

$$P(X<x_\alpha)=1-\alpha$$

Example 1.7.2.3.1

For a continuous random variable X, the value the area to the right of which is equal to 0.3 is called $x_{0.3}$.



This fact can be expressed as the following probability statement:

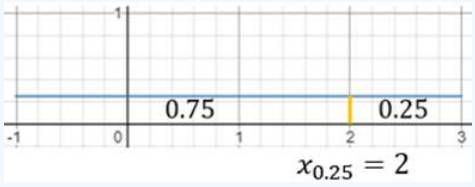
$$P(X > x_{0.3}) = 0.3$$

As a result, the area to the left of X is by the complementary rule must be $1 - 0.3 = 0.7$. This fact can be expressed as the following probability statement:

$$P(X < x_{0.3}) = 0.7$$

Example 1.7.2.3.2.1

Consider a uniform random variable with parameters -1 and 3, i.e., $X \sim U(-1, 3)$. The value the area to the right of which is equal to 0.25 is called $x_{0.25}$. In this case, we happened to know that $x_{0.25} = 2$.

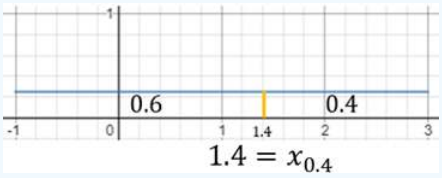


This fact can be expressed in two ways:

$$P(X < x_{0.25}) = 0.75 \text{ and } P(X > x_{0.25}) = 0.25$$

Example 1.7.2.3.2.2

Consider a uniform random variable with parameters -1 and 3, i.e., $X \sim U(-1, 3)$. The value the area to the left of which is equal to 0.6 is called $x_{0.4}$. In this case, we happened to know that $x_{0.40} = 1.4$.



This fact can be expressed in two ways:

$$P(X < x_{0.4}) = 0.6 \text{ and } P(X > x_{0.4}) = 0.4$$

Theorem: Finding x_α for a Uniform Random Variable

In general, for a uniform random variable with parameters a and b, to find x_α means to find the value the area to the right of which is α and the area to the left is $1 - \alpha$. The following formula can be used to find the x_α:

$$x_\alpha = a \cdot \alpha + b \cdot (1 - \alpha)$$

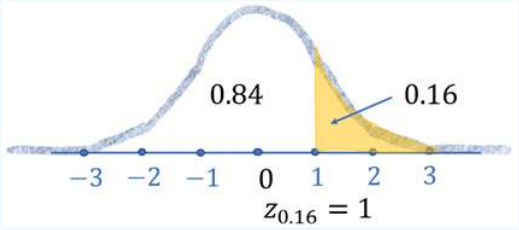
Example 1.7.2.3.3

Again, consider a uniform random variable with parameters -1 and 3, i.e. $X \sim U(-1, 3)$.

1. $x_{0.25} = -1 \cdot 0.25 + 3 \cdot 0.75 = 2$
2. $x_{0.4} = -1 \cdot 0.4 + 3 \cdot 0.6 = 1.4$

Now that we know what the alpha-notation is and how to find u_α for uniform random variables let's consider the standard normal random variable, Z, and learn how to do the same.

Example 1.7.2.3.4.1



In this case, we happened to know that according to the empirical rule the area to the right of 1 is 16%. A value the area to the right of which under the Z-curve is equal to 0.16 is called $z_{0.16}$. Therefore

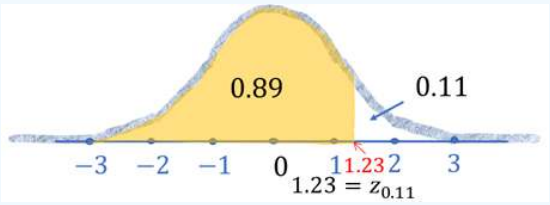
$$z_{0.16} = 1$$

and this fact can be expressed as the following probability statements:

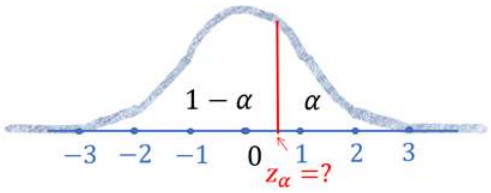
$$P(Z < 1) = 0.84 \text{ and } P(Z > 1) = 0.16$$

Example 1.7.2.3.4.2

Let's consider the standard normal random variable, Z, and interpret the previously discovered probability statement $P(Z < 1.23) = 0.89$ as $z_{0.11} = 1.23$.



In general, we want to be able to find z_α for any α.



To do that we can use the Z-table (part 1, part 2) :

Areas under the standard normal curve*



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177

Example 1.7.2.3.4.3

Find $z_{0.38}$ using the Z-table.

Solution

1. We identify alpha as $\alpha = 0.38$ and $1 - \alpha = 0.62$. In the z-table, the closest value to 0.62 is 0.6217.
2. We read the corresponding numbers 0.3 on the left and 0.01 on top and put them together to obtain the answer.
3. $z_{0.38} = 0.31$

Example 1.7.2.3.4.4

Find $z_{0.10}$ using the Z-table.

Solution

1. We identify alpha as $\alpha = 0.10$ and $1 - \alpha = 0.90$. In the z-table, the closest value to 0.90 is 0.8997.
2. We read the corresponding numbers 1.2 on the left and 0.08 on top and put them together to obtain the answer.
3. $z_{0.10} = 1.28$

Theorem: Property of z-scores for symmetric distributions

For Z and other distributions symmetric about zero:

$$x_{1-\alpha} = -x_{\alpha}$$

Example 1.7.2.3.4.5

$$z_{0.9} = -z_{0.10} = -1.28$$

$$z_{0.95} = -z_{0.05} = -1.645$$

$$z_{0.99} = -z_{0.01} = -2.33$$

Section 4: Percentiles



Definition: p-th percentile

The p-th percentile of a random variable X is the value $x_{p\%}$ that is greater than $p\%$ of the observations if the experiment is repeated many times.

In other words,

$$P(X < x_{p\%}) = p\%$$

The relation between the percentiles and α-notation:

$$x_{p\%} = x_{\alpha=1-\frac{p}{100}}$$

Example 1.7.2.4.1

Find the 80-th percentile for Z.

Solution

$z_{80\%} = z_{0.2} = 0.84$

Example 1.7.2.4.2

Find the 55-th percentile for $X \sim U(1, 5)$.

Solution

$x_{55\%} = x_{0.45} = 1 \cdot 0.45 + 5 \cdot 0.55 = 3.2$

Example 1.7.2.4.3

$z_{10\%} = z_{0.9} = -1.28$ is the 10-th percentile for Z.

$z_{62\%} = z_{0.38} = 0.31$ is the 62-nd percentile for Z.

$x_{75\%} = x_{0.25} = 2$ is the 75-th percentile for $X \sim U(-1, 3)$.

$x_{60\%} = x_{0.40} = 1.4$ is the 60-th percentile for $X \sim U(-1, 3)$.

This page titled 1.7.2: Standard Types of Continuous Random Variables is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.7.3: Other Continuous Random Variables

Section 1: Introduction to Other Continuous Random Variables



What does it mean to know a certain distribution? First of all, this means to know what the probability density curve of the distribution looks like and its parameters and properties. Secondly, to know a distribution also means that one knows how to answer the following two questions:

1. How to find the probability that X is less than a given value c or in other words how to find the area under the probability density curve to the left of some value c.
2. How to find the x_α for any given alpha or in other words how to find the value the area to the right of which is a given α.

Why knowing the answers to just these two questions is sufficient? Because if we know the probability that X is less than a value we can find any other probability by using the complementary and subdivision rules. Also, if we know how to find x_α's would allow us to compute the percentiles as well by using the relationship between the two.

For example, we already know the uniform random variables with parameters a and b because:

- we know its probability density curve and its properties;
- we have the formula for computing the probability that X is less than c:

$$P(X < c) = \frac{c-a}{b-a}$$

- we have the formula for computing the x_α:

$$x_\alpha = a \cdot \alpha + b \cdot (1 - \alpha)$$

For example, we already know the standard normal random variables because:

- we know its probability density curve and its properties;
- we have the method for computing the probability that X is less than c using the Z-table (part 1, part 2):



- we have the method for computing the z_α using the Z-table (part 1, part 2):

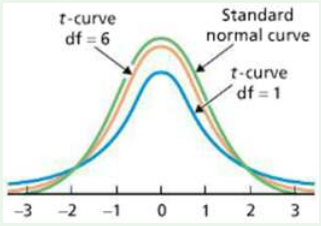


For every type of a new distribution, we will follow the following template. First, we will discuss the parameters and the properties of its probability density curve. Then we will discuss how to compute the probabilities and find x_α's.

Section 2: Student T-Distribution



Definition: Probability density curve of a T-distribution



The **probability density curve of a T-distribution** looks a lot like the standard normal probability density curve, and it has the following properties:

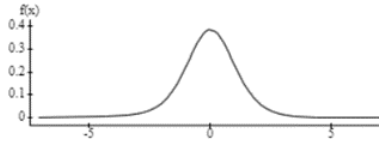
Property 1: The total area equals 1.

Property 2: Extends indefinitely in both directions, approaching, but never touching, the horizontal axis as it does so.

Property 3: Symmetric about 0.

Property 4: As the number of degrees of freedom becomes larger, t-curves look increasingly like the standard normal curve.

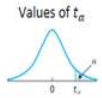
For example, let T be a random variable that has T-distribution with 7 degrees of freedom, then the probability density curve of T takes a very specific shape determined by its parameter.



Next, we want to be able to find $P(T < c)$ for all c and find the t_α's for all α. Let's make sure we understand the questions well:

- To find the probability $P(T < c)$ means to find the area under the T-curve to the left of c.
- To find t_α means to find the value the area under the T-curve to the right of which is α.

Luckily, there are many ways to use technology to answer these two questions. In reality, while it is not too hard to use technology to find the probabilities, very rarely one would have to do this task in the applications. However, while it is also not too hard to use technology to find t_α's, this task will be very common in the applications, so it is worth learning how to perform this task by hand. The most frequently used t_α's are summarized in the following table (t-table):



df	$t_{0.25}$	$t_{0.2}$	$t_{0.15}$	$t_{0.1}$	$t_{0.05}$	$t_{0.025}$	$t_{0.02}$	$t_{0.01}$	$t_{0.005}$	$t_{0.0025}$	$t_{0.001}$	$t_{0.0005}$
1	1.000	1.376	1.963	3.078	6.314	12.706	15.895	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221

We only need to identify α and df, the degrees of freedom, then look for the value in the intersection.

Example 1.7.3.2.1

For $T \sim T_{df=10}$, $t_{0.05} = 1.812$.

For $T \sim T_{df=11}$, $t_{0.01} = 2.718$.

How about $\alpha > 0.5$? Turns out that while it is possible to create a table to summarize such t_α's there is no need for it. Since the T-distribution is symmetric about zero we can use the property of alpha-notation and conclude that

$$t_\alpha = -t_{1-\alpha}$$

Example 1.7.3.2.2

For $T \sim T_{df=10}$, $t_{0.95} = -t_{0.05} = -1.895$.

For $T \sim T_{df=11}$, $t_{0.99} = -t_{0.01} = -2.718$.

Example 1.7.3.2.3 with technology

The z-scores of the test results of a sample of students follow the T-distribution with 10 degrees of freedom. Let T be the z-score of a randomly selected student.

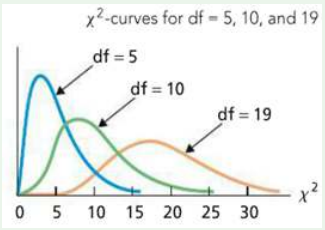
$$T \sim T_{df=10}$$

1. Find the probability that the z-score is:
   i. less than -0.75: $P(T < -0.75) = 0.235 = 23.5\%$
   ii. more than 2.5: $P(T > 2.5) = 0.016 = 1.6\%$
   iii. between -1 and 1.33: $P(-1 < T < 1.33) = 0.723 = 72.3\%$
2. Compute the 95-th percentile: $t_{95\%-th} = t_{0.05} = 1.812$

Section 3: X^2-Distribution



Definition: Probability density curve of a X^2-distribution



The **probability density curve of a X^2-distribution with df degrees of freedom** has the following properties:

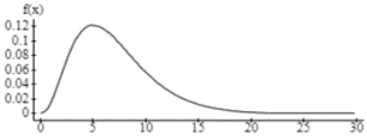
Property 1: The total area equals 1.

Property 2: Starts at the origin and extends indefinitely to the right, approaching, but never touching, the horizontal axis.

Property 3: Right-skewed.

Property 4: As the number of degrees of freedom becomes larger, X^2-curves look increasingly like normal curves.

For example, let X^2 be a random variable that has X^2-distribution with 7 degrees of freedom, then the probability density curve of X^2 takes a very specific shape determined by its parameter.

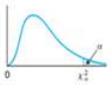


Next, we want to be able to find $P(X^2 < c)$ for all c and find the χ^2_α's for all α. Let's make sure we understand the questions well:

- To find the probability $P(X^2 < c)$ means to find the area under the X^2-curve to the left of c.
- To find χ^2_α means to find the value the area under the X^2-curve to the right of which is α.

Luckily, there are many ways to use technology to answer these two questions. In reality, while it is not too hard to use technology to find the probabilities, very rarely one would have to do this task in the applications. However, while it is also not too hard to use technology to find χ^2_α, this task will be very common in the applications, so it is worth learning how to perform this task by hand. The most frequently used χ^2_α's are summarized in the following table (chi2-table):

Values of χ^2_α



df	$\chi^2_{0.995}$	$\chi^2_{0.99}$	$\chi^2_{0.975}$	$\chi^2_{0.95}$	$\chi^2_{0.9}$	$\chi^2_{0.1}$	$\chi^2_{0.05}$	$\chi^2_{0.025}$	$\chi^2_{0.01}$	$\chi^2_{0.005}$	df
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879	1
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597	2
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838	3
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860	4
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750	5
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548	6
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278	7
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955	8
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589	9
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188	10
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757	11
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300	12
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819	13

We only need to identify α and df, the degrees of freedom, then look for the value in the intersection.

Example 1.7.3.3.1.1

For $X^2 \sim X^2_{df=7}$, $\chi^2_{0.05} = 14.067$.

For $X^2 \sim X^2_{df=12}$, $\chi^2_{0.01} = 26.217$.

How about $\alpha > 0.5$? Since X^2 is not symmetric about zero it is absolutely necessary to have a table to summarize such χ^2_α's.

Example 1.7.3.3.1.2

For $X^2 \sim X^2_{df=10}$, $\chi^2_{0.99} = 2.558$.

Example 1.7.3.3.2 with technology

A randomly selected car salesman's annual salary (in thousands) can be estimated by a random variable that follows X^2-distribution with $df = 40$. Let X be the salary of a randomly selected car salesman.

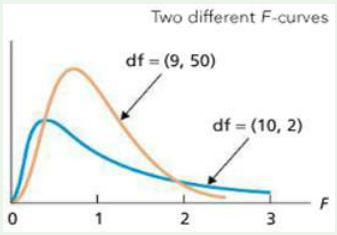
$$X \sim X^2_{df=40}$$

1. Find the probability that the commission is:
   i. less than \$35K: $P(X < 35) = 0.305 = 30.5\%$
   ii. more than \$50K: $P(X > 50) = 0.134 = 13.3\%$
   iii. between \$40K and \$60K: $P(40 < X < 60) = 0.448 = 44.8\%$
2. Compute the 90-th percentile:$x_{90\%-th} = x_{0.1} = 51.8$

Section 4: F-Distribution



Definition: Probability density curve of an F-distribution with parameters (dfn, dfd)



The **probability density curve of an F-distribution with parameters (dfn, dfd)** has the following properties:

Property 1: The total area equals 1.

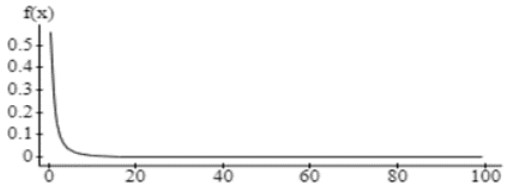
Property 2: Starts at the origin and extends indefinitely to the right, approaching, but never touching, the horizontal axis.

Property 3: Right-skewed.

Property 4: The Reciprocal Property of F-Curves:

$$f_\alpha = \frac{1}{f^*_{1-\alpha}} \text{ where } f^* \text{ has } df = (dfd, dfn)$$

For example, let F be a random variable that has F-distribution with $(6, 2)$ degrees of freedom, then the probability density curve of F takes a very specific shape determined by its parameter.

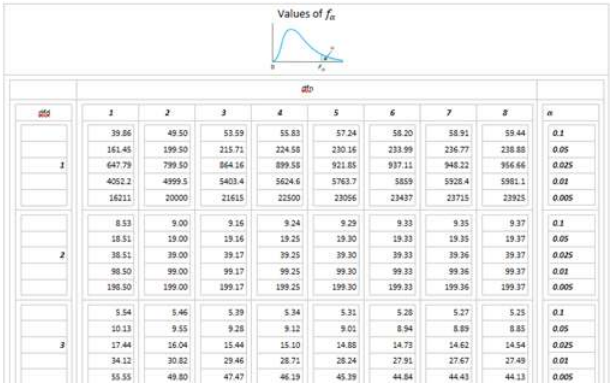


Next, we want to be able to find $P(F < c)$ for all c and find the f_α's for all α. Let's make sure we understand the questions well:

- To find the probability $P(F < c)$ means to find the area under the F-curve to the left of c.
- To find f_α means to find the value the area under the F-curve to the right of which is α.

Luckily, there are many ways to use technology to answer these two questions. In reality, while it is not too hard to use technology to find the probabilities, very rarely one would have to do this task in the applications. However, while it is also not too hard to use

technology to find f_α, this task will be very common in the applications, so it is worth learning how to perform this task by hand. The most frequently used f_α's are summarized in the following table (f-table):



We only need to identify dfn, dfd, and α, then look for the value in the intersection.

Example 1.7.3.4.1

For $F \sim F_{(6,2)}$, $f_{0.05} = 19.330$.

For $F \sim F_{(4,3)}$, $f_{0.01} = 28.71$

How about $\alpha > 0.5$? Since F is not symmetric about zero we can't use the property of symmetric distributions, however we can use the reciprocal property of F-curves.

Example 1.7.3.4.2

For $F \sim F_{(2,3)}$, $f_{0.99} = \frac{1}{f^*_{0.01}} = \frac{1}{99.17} = 0.010$

Example 1.7.3.4.3 with technology

A randomly selected realtor's monthly commission (in thousands) can be estimated by a random variable that follows F-distribution with parameters $(15, 3)$. Let F be the commission of a randomly selected realtor.

$$F \sim F_{(15,3)}$$

1. Find the probability that the commission is:
   i. less than \$1500: $P(F < 1.5) = 0.585 = 58.5\%$
   ii. more than \$8000: $P(F > 8) = 0.056 = 5.6\%$
   iii. between \$3000 and \$7000: $P(3 < F < 7) = 0.131 = 13.1\%$
2. Compute the median: $f_{50\%-th} = f_{0.5} = 1.21$

This page titled 1.7.3: Other Continuous Random Variables is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

SECTION OVERVIEW

1.8: Normal Distributions

1.8.1: Normal Random Variables

1.8.2: Applications of Normal Random Variables

1.8.3: Normality Plots

1.8.4: Approximating Binomial Random Variables

This page titled 1.8: Normal Distributions is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.8.1: Normal Random Variables

Section 1: Introduction

Page ID

202061

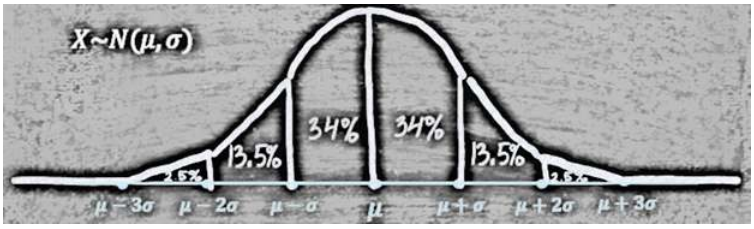


Definition: Normal probability density curve

A **normal probability density curve** with parameters μ and σ is a bell-shaped curve that satisfies the following properties:

1. Has the peak at μ and symmetric about μ.
2. Extends indefinitely in both directions, approaching but never touching the horizontal axis.
3. The Empirical rule holds that is:
   a. ~68% of the area under the curve is between $\mu - 1\sigma$ and $\mu + 1\sigma$;
   b. ~95% of the area under the curve is between $\mu - 2\sigma$ and $\mu + 2\sigma$;
   c. ~99.7% of the area under the curve is between $\mu - 3\sigma$ and $\mu + 3\sigma$.

The description of the curve should sound familiar as we have seen it before!



Definition: Normal random variable with parameters μ and σ

We define a **normal random variable with parameters μ and σ** as a continuous random variable with a normal probability density curve with parameters μ and σ. We denote such a random variable in the following way:

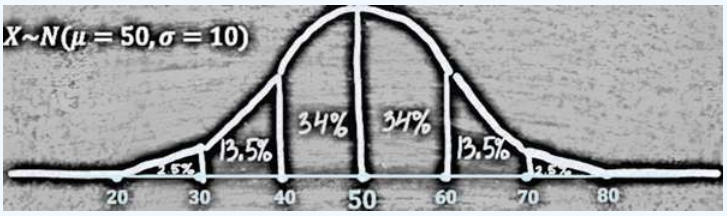
$$X \sim N(\mu, \sigma)$$

Definition: Parameters of a normal distribution

The values, μ and σ, are called the **parameters of a normal distribution** and each pair of such values determines a unique shape of the probability density curve.

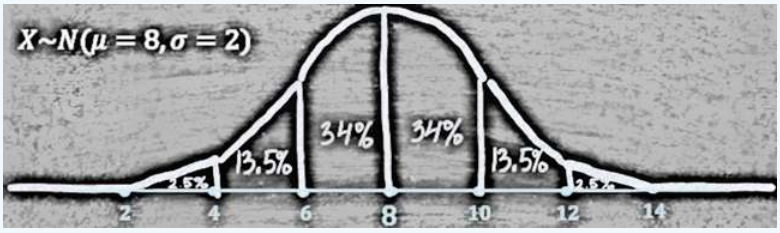
Example 1.8.1.1.1.1

When $\mu = 50$ and $\sigma = 10$, we get the following curve:



Example 1.8.1.1.1.2

When $\mu = 8$ and $\sigma = 2$, we get the following curve:



Let's make sure we understand the properties of normal probability density curves and use the empirical rules to find some probabilities for any $X \sim N(\mu, \sigma)$.

- To find the probability that X is less than μ we will look for the area under the curve to the left of μ. This region's area can be found by adding the areas of included regions together to get $P(X < \mu) = 0.50$ or 50%.
- To find the probability that X is between $\mu - \sigma$ and $\mu + \sigma$ we will look for the area under the curve between $\mu - \sigma$ and $\mu + \sigma$. This region's area can be found by adding the areas of included regions together to get $P(\mu - \sigma < X < \mu + \sigma) = 0.68$ or 68%.
- To find the probability that X is between μ and $\mu + 2\sigma$ we will look for the area under the curve between μ and $\mu + 2\sigma$. This region's area can be found by adding the areas of included regions together to get $P(\mu < X < \mu + 2\sigma) = 0.475$ or 47.5%.
- To find the probability that X is between $\mu - \sigma$ and $\mu + 2\sigma$ we will look for the area under the curve between $\mu - \sigma$ and $\mu + 2\sigma$. This region's area can be found by adding the areas of included regions together to get $P(\mu - \sigma < X < \mu + 2\sigma) = 0.815$ or 81.5%.
- To find the probability that X is greater than $\mu + \sigma$ we will look for the area under the curve to the right of $\mu + \sigma$. This region's area can be found by adding the areas of included regions together to get $P(X > \mu + \sigma) = 0.16$ or 16%.
- To find the probability that X is less than $\mu - 2\sigma$ we will look for the area under the curve to the left of $\mu - 2\sigma$. This region's area can be found by adding the areas of included regions together to get $P(X < \mu - 2\sigma) = 0.025$ or 2.5%.
- To find the probability that X is greater than $\mu + 4\sigma$ we will look for the area under the curve to the right of $\mu + 4\sigma$. This region's area is essentially equal to 0 so the probability is nearly 0%, that is $P(X > \mu + 4\sigma) = 0$.

Example 1.8.1.1.2.1

For $X \sim N(\mu = 50, \sigma = 10)$, let's find the following probabilities:

- To find the probability that X is less than 50 we will look for the area under the curve to the left of 50. This region's area can be found by adding the areas of included regions together to get $P(X < 50) = 0.50$ or 50%.
- To find the probability that X is between 40 and 70 we will look for the area under the curve between 40 and 70. This region's area can be found by adding the areas of included regions together to get $P(40 < X < 70) = 0.815$ or 81.5%.

Example 1.8.1.1.2.2

For $X \sim N(\mu = 8, \sigma = 2)$, let's find the following probabilities:

- To find the probability that X is between 8 and 12 we will look for the area under the curve between 8 and 12. This region's area can be found by adding the areas of included regions together to get $P(8 < X < 12) = 0.475$ or 47.5%
- To find the probability that X is between 6 and 12 we will look for the area under the curve between 6 and 12. This region's area can be found by adding the areas of included regions together to get $P(6 < X < 12) = 0.815$ or 81.5%

Section 2: Working with Normal Random Variables



For a normal random variable with parameters μ and σ, we want to be able to find the probabilities that involve the values a, b, and c that are not covered by the Empirical rule:

- $P(X < c)$
- $P(X > c)$
- $P(a < X < b)$

But first let's make sure we understand the questions!

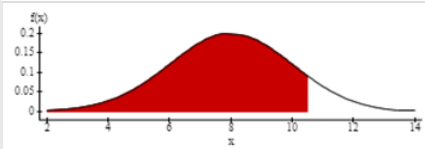
- To find the probability that X is less than c, we will look for the area under the curve to the left of c.
- To find the probability that X is greater than c, we will look for the area under the curve to the right of c.
- To find the probability that X is between a and b, we will look for the area under the curve to the right of a and to the left of b.

While it is clear what the regions' areas look like they cannot be found by using the empirical rule. So how do we find the probabilities that involve values that are not covered by the Empirical rule such as $P(X < c)$, $P(X > c)$, $P(a < X < b)$?

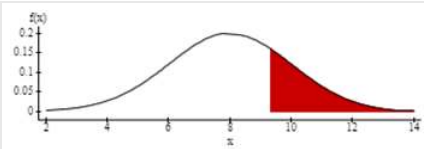
For $X \sim N(\mu = 8, \sigma = 2)$, let's find the following probabilities:

- $P(X < 10.5)$
- $P(X > 9.3)$
- $P(7.1 < X < 8.7)$

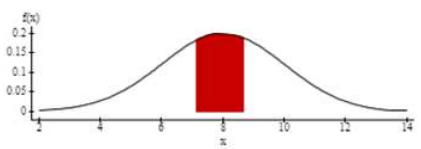
Luckily, there are ways to use technology to answer these questions:



Normal Distribution
Mean:8 Std. Dev.:2
P(X≤10.5) = 0.89435023



Normal Distribution
Mean:8 Std. Dev.:2
P(X≥9.3) = 0.25784611



Normal Distribution
Mean:8 Std. Dev.:2
P(7.1≤X≤8.7) = 0.31047543

However, it is worth learning how to perform this task by hand. To be able to answer the same questions manually we have to review the concept of a z-score.

Definition: z-score

The ***z*-score** of an observation x_i is computed by using the formula:

$$z_{\text{score}} = \frac{x_i - \mu}{\sigma}$$

Example 1.8.1.2.1

Let's find some z-scores for a normal random variable X with $\mu = 8$ and $\sigma = 2$:

- the z-score of $x_1 = 7.1$ is $\frac{7.1-8}{2} = -0.45$
- the z-score of $x_2 = 8.7$ is $8.7 - 82 = 0.35$
- the z-score of $x_3 = 9.3$ is $\frac{9.3-8}{2} = 0.65$
- the z-score of $x_4 = 10.5$ is $\frac{10.5-8}{2} = 1.25$

If instead of observing a normal variable with parameters μ and σ we consider the z-scores of the observed values. Then it is not too hard to observe that the z-scores also have a normal distribution with parameters 0 and 1 which we call the standard normal distribution. In other words, the new random variable $\frac{X-\mu}{\sigma}$ has the standard normal distribution, therefore:

$$\frac{X-\mu}{\sigma} = Z \sim N(0,1)$$

By the way, the fact that the standard normal distribution is called Z-distribution explains the origin of the name for a term "a z-score".

Definition: Standardizing a random variable

The process of converting the original normal random variable to Z is called **standardizing the random variable**.
Most importantly, the area between any two values a and b under the normal curve is the same as the area between the z-scores of the a and b under the z-curve:

$$P(a < X < b) = P(z_a < Z < z_b)$$

Example 1.8.1.2.2

For $X \sim N(\nu = 8, \sigma = 2)$, find the following probabilities:

- $P(X < 10.5) = P(Z < \frac{10.5-8}{2}) = P(Z < 1.25) = 0.8944$
- $P(X > 9.3) = P(Z > \frac{9.3-8}{2}) = P(Z > 0.65) = 0.2578$
- $P(7.1 < X < 8.7) = P(\frac{7.1-8}{2} < Z < \frac{8.7-8}{2}) = P(-0.45 < Z < 0.35) = P(Z < 0.35) - P(Z < -0.45) = 0.6368 - 0.3264 = 0.3104$

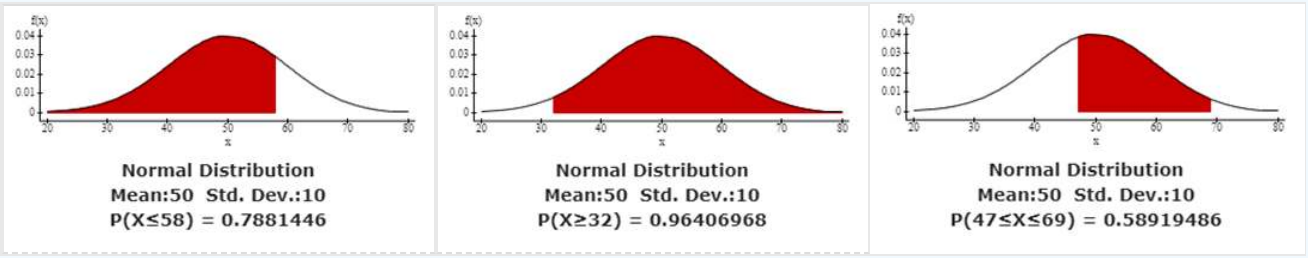
Similarly, we can use the standardization process for any other normal random variable.

Example 1.8.1.2.3

For $X \sim N(\mu = 50, \sigma = 10)$, find the following probabilities:

- $P(X < 58) = P(Z < \frac{58-50}{10}) = P(Z < 0.8) = 0.7881$
- $P(X > 32) = P(Z > \frac{32-50}{10}) = P(Z > -1.8) = 0.9641$
- $P(47 < X < 69) = P(\frac{47-50}{10} < Z < \frac{69-50}{10}) = P(-0.3 < Z < 1.9) = 0.589$

Of course, every result obtained manually can be confirmed using technology:



Next, we will discuss how to find x_α. To find x_α means to find the value the area to the right of which is equal to α. In other words, we are looking for x_α such that

$$P(X > x_\alpha) = \alpha$$

Standardizing the left-hand side of the equation gives us

$$P(X > x_\alpha) = P(Z > \tfrac{x_\alpha - \mu}{\sigma}) = \alpha$$

which implies that

$$\tfrac{x_\alpha - \mu}{\sigma} = z_\alpha$$

therefore

$$x_\alpha = \mu + z_\alpha \cdot \sigma$$

Example 1.8.1.2.4.1

For a normal random variable with parameters $\mu = 8$ and $\sigma = 2$, let's find $x_0.38$ by using the above formula.

$$x_{0.38} = \mu + z_{0.38} \cdot \sigma$$

Now that we related x_{0.38} to z_{0.38} we use the z-table as we have done it before to identify $\alpha = 0.38$ and $1 - \alpha = 0.62$ to obtain the $z_{0.38} = 0.31$. Now we can compute the desired $x_{0.38}$ by finishing the calculations:

$$x_{0.38} = \mu + z_{0.38} \cdot \sigma = 8 + 0.31 \cdot 2 = 8.62$$

Note that the following two statements are true by the definition of the alpha notation:

$$P(X < 8.62) = 0.62 \text{ and } P(X > 8.62) = 0.38$$

Example 1.8.1.2.4.2

For a normal random variable with parameters $\mu = 50$ and $\sigma = 10$, let's find $x_{0.10}$ by using the above formula.

$$x_{0.10} = \mu + z_{0.10} \cdot \sigma$$

Now that we related $x_{0.10}$ to $z_{0.10}$ we use the z-table as we have done it before to identify $\alpha = 0.10$ and $1 - \alpha = 0.90$ to obtain the $z_{0.10} = 1.28$. Now we can compute the desired $x_{0.10}$ by finishing the calculations:

$$x_{0.10} = \mu + z_{0.10} \cdot \sigma = 50 + 1.28 \cdot 10 = 62.8$$

Now that we know how to find the probabilities and x_αs let's discuss the parameters of a normal distribution. It is not a coincidence that the two parameters, μ and σ, for a normal random variable are labeled similar to the mean and the standard

deviation of the data set. The meaning and the interpretation of the mean and standard deviation for a continuous random variable is the same as for discrete random variables. That is, the expectation (or the mean) of a random variable is a measure of the center of the distribution and the variance (and standard deviation) is a measure of the spread of the random variable.

Theorem: Mean, variance, and standard deviation of a Normal RV

For $X \sim N(\mu, \sigma)$:

$$E[N(\mu, \sigma)] = \mu$$
$$VAR[N(\mu, \sigma)] = \sigma^2$$
$$SD[N(\mu, \sigma)] = \sigma$$

Example 1.8.1.2.5

For example, for $X \sim N(10, 2)$ by the three standard deviations rule, we may conclude that:

Range	Numerical Range	Interpretation
within $2\sigma_X$ from μ_X	between 6 and 14	“expected”
more than $2\sigma_X$ from μ_X	less than 6 or more than 14	“unusual”
more than $3\sigma_X$ from μ_X	less than 4 or more than 16	“abnormal”

This page titled 1.8.1: Normal Random Variables is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.8.2: Applications of Normal Random Variables



When working with random variables we adopt the following vocabulary.

Definition:

We say that **a certain population has or follows a normal distribution** with parameters μ and σ when the following is true:

1. If we do the census, find the mean μ and standard deviation σ, and plot the histogram, then the shape of the distribution is a normal probability density curve with parameters μ and σ.
2. If we let X be a randomly selected observation from such population then $X \sim N(\mu, \sigma)$.

Therefore, $P(a < X < b)$, the probability that X is between a and b can be interpreted in the following two ways.

- First, we can literally interpret it as the probability that a randomly selected value is between a and b.
- Alternatively, the same quantity represents the proportion of the population that has a value between a and b.

Similarly, $P(X < c)$, the probability that X is less than c can be interpreted in the following two ways.

- First, we can literally interpret it as the probability that a randomly selected value is less than c.
- Alternatively, the same quantity represents the proportion of the population that has a value less than c.

Finally, $P(X > c)$, the probability that X is greater than c can be interpreted in the following two ways.

- First, we can literally interpret it as the probability that a randomly selected value is greater than c.
- Alternatively, the same quantity represents the proportion of the population that has a value greater than c.

Example 1.8.2.1

Consider the heights of the males in the U.S. We may assume that the population of heights has a normal distribution with $\mu = 175$ cm and $\sigma = 15$ cm.

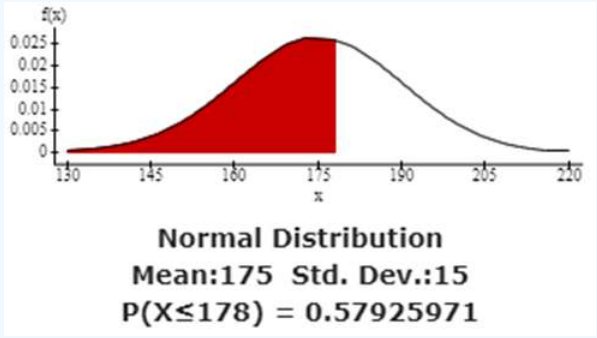
a. Find the proportion of the population that is shorter than 178 cm.
b. Find the probability that a randomly selected US male has the height between 170 cm and 180 cm.
c. Find the 75-th percentile of the population.

Solution

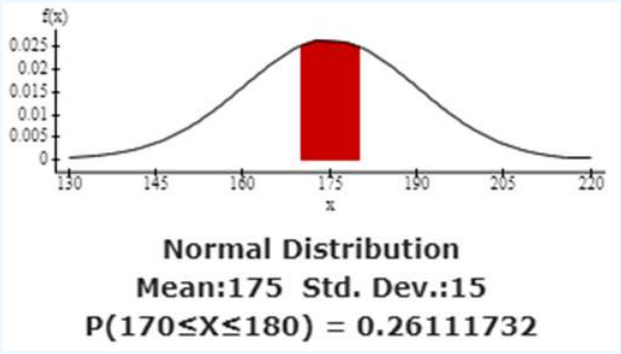
Let's introduce a random variable X to represent the height of a randomly selected male, then

$$X \sim N(175, 15)$$

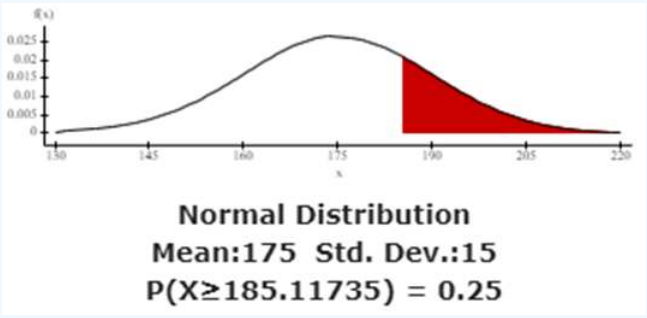
- $P(X<178)=P(\frac{X-\mu}{\sigma}<\frac{178-175}{15})=P(Z<0.20)=0.579$



- $P(170<X<180)=P(\frac{170-175}{15}<\frac{X-\mu}{\sigma}<\frac{180-175}{15})=P(-0.33<Z<0.20)=P(Z<0.20)-P(Z<-0.33)$ $=0.6293-0.3707=0.2586$



- $x_{0.25}=\mu+z_{0.25}\sigma=175+0.67\cdot 15=185.05$



Example 1.8.2.2

As reported in Runner's World magazine, the times of the finishers in the New York City 10-km run are normally distributed with mean 61 minutes and standard deviation 9 minutes.

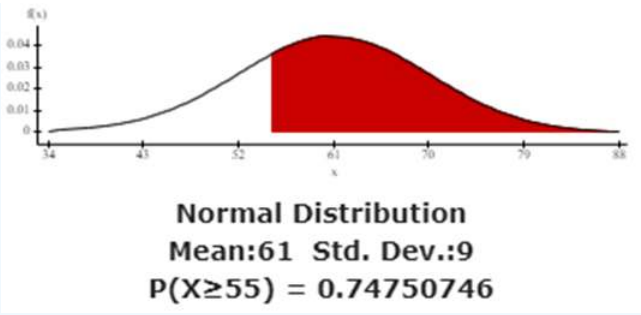
a. Find the probability that a randomly selected runner finished in more than 55 minutes?
b. What proportion of the runners finished in between 60 and 70 minutes?
c. What is the 10-th percentile of the finishing times?

Solution

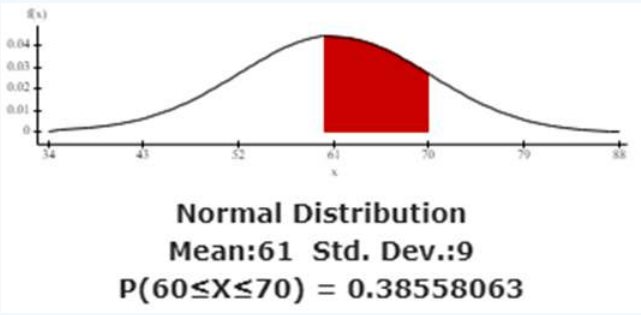
Let's introduce a random variable X to represent the time of a randomly selected marathon finisher, then

$$X\sim N(61,9)$$

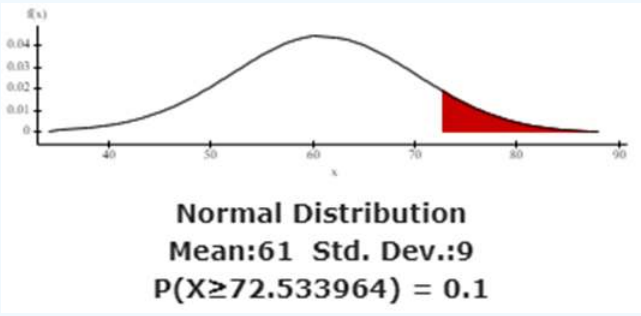
- $P(X>55)=P(\frac{X-\mu}{\sigma}>\frac{55-61}{9})=P(Z>-0.67)=0.7486$



- $P(60 < X < 70) = P(\frac{60-61}{9} < \frac{X-\mu}{\sigma} < \frac{70-61}{9}) = P(-0.11 < Z < 1) = P(Z < 1) - P(Z < -0.11) = 0.8413 - 0.4562 = 0.3851$



- $x_{0.90} = \mu + z_{0.90}\sigma = 61 + 1.28 \cdot 9 = 72.52$



This page titled 1.8.2: Applications of Normal Random Variables is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.8.3: Normality Plots



In a lot of problems in the future, it will be necessary to know whether the sample came from a normal population or not. Next, we will discuss how does one know whether the population is normal or not based on a sample.

We intuitively expect that the shape of the distribution of a sample is approximately the same as the shape of the population distribution. So if one can obtain a large representative sample (or the census) then the shape of its histogram will resemble the shape of the population's distribution. Therefore, if the distribution looks normal then it is an indicator that the sample came from a normal population. But what do we do when only a small sample can be obtained? In such a case, we will use a graph called a normality plot.

Turns out that normal populations are so common and well-studied that we almost know what to expect from a sample of any small size. The expectations are formulated in the following table:

5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
-1.18	-1.28	-1.36	1.43	-1.5	-1.55	-1.59	-1.64	-1.68	-1.71	-1.74	-1.77	-1.8	-1.82	-1.85	-1.87	-1.89	-1.91
-0.5	-0.64	-0.76	-0.85	-0.93	-1	-1.06	-1.11	-1.16	-1.2	-1.24	-1.28	-1.32	-1.35	-1.38	-1.4	-1.43	-1.45
0	-0.2	-0.35	-0.47	-0.57	-0.65	-0.73	-0.79	-0.85	-0.9	-0.94	-0.99	-1.03	-1.06	-1.1	-1.13	-1.16	-1.18
0.5	0.2	0	-0.15	-0.27	-0.37	-0.46	-0.53	-0.6	-0.66	-0.71	-0.76	-0.8	-0.84	-0.88	-0.92	-0.85	-0.98
1.18	0.64	0.35	0.15	0	-0.12	-0.22	-0.31	-0.39	-0.45	-0.51	-0.57	-0.62	-0.66	-0.7	-0.74	-0.78	-0.81
	1.28	0.76	0.47	0.27	0.12	0	-0.1	-0.19	-0.27	-0.33	-0.39	-0.45	-0.5	-0.54	-0.59	-0.63	-0.66
		1.36	0.85	0.57	0.37	0.22	0.1	0	-0.09	-0.16	-0.23	-0.29	-0.35	-0.4	-0.45	-0.49	-0.53
			1.43	0.93	0.65	0.46	0.31	0.19	0.09	0	-0.08	-0.15	-0.21	-0.26	-0.31	-0.36	-0.4
				1.5	1	0.73	0.53	0.39	0.27	0.16	0.08	0	-0.07	-0.13	-0.19	-0.24	-0.28
					1.55	1.03	0.79	0.6	0.45	0.33	0.23	0.15	0.07	0	-0.06	-0.12	-0.17
						1.59	1.11	0.85	0.66	0.51	0.39	0.29	0.21	0.13	0.06	0	-0.06
							1.64	1.16	0.9	0.71	0.57	0.45	0.35	0.26	0.19	0.12	0.06
								1.68	1.2	0.94	0.76	0.62	0.5	0.4	0.31	0.24	0.17
									1.71	1.24	0.99	0.8	0.66	0.54	0.45	0.36	0.28
										1.74	1.28	1.03	0.84	0.7	0.59	0.49	0.4
											1.77	1.32	1.06	0.88	0.74	0.63	0.53
												1.8	1.35	1.1	0.92	0.78	0.66
													1.82	1.38	1.13	0.95	0.81
														1.85	1.4	1.16	0.98
															1.87	1.43	1.18
																1.89	1.45
																	1.91

This table provides the expected z-scores from a sample of each size from 5 to 22 in ascending order. So if we have the sample of size 5, then we expect that one of the observations will be 1.18 standard deviations below the mean, one of the observations will be 0.5 standard deviations below the mean, one of the observations will be equal the mean, one of the observations will be 0.5 standard deviations above the mean, one of the observations will be 1.18 standard deviations above the mean.

Definition: Normality plot

To construct a **normality plot**:

1. Sort the sample of size n in increasing order.
2. In the table of expected z-scores find the column corresponding to the sample size n.
3. List the expected z-scores from the table side-by-side with the observations in the sample.
4. Construct the scatterplot with expected z-scores as the input and the data as the output.

If the plot has a roughly linear pattern, you can assume that the sample came from a normally distributed population. If the plot is not roughly linear, you cannot assume that the sample is from a normal population. These guidelines should be interpreted loosely for smaller samples but usually strictly for larger samples.

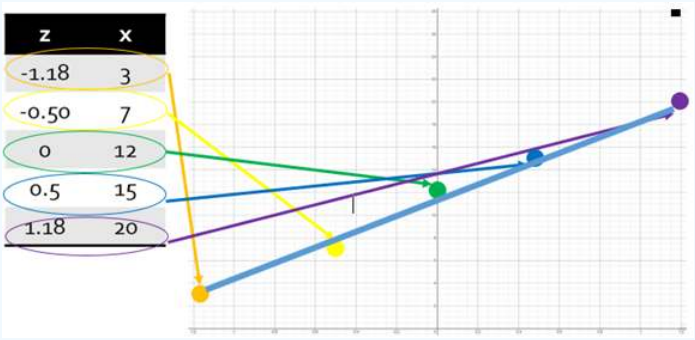
Example 1.8.3.1

A sample of five UBER drivers were asked how many hours they worked last week, and the following answers were obtained:

15, 12, 7, 3, 20

Use the normality plot to determine whether the population of weekly driving time is normal for UBER drivers.

Solution



Based on the shape of the normality plot we conclude that the sample came from a normal population.

Example 1.8.3.2

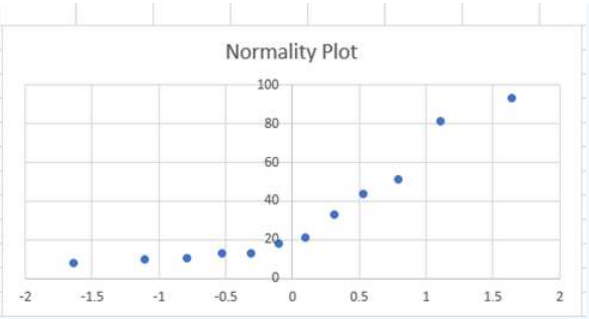
A simple random sample of 12 returns from last year revealed the adjusted gross incomes, in thousands of dollars, shown in the table below.

9.7	93.1	33.0	21.2
81.4	51.1	43.5	10.6
12.8	7.8	18.1	12.7

Construct a normal probability plot for these data and use the plot to assess the normality of adjusted gross incomes.

Solution

z	x
-1.64	7.8
-1.11	9.7
-0.79	10.6
-0.53	12.7
-0.31	12.8
-0.1	18.1
0.1	21.2
0.31	33
0.53	43.5
0.79	51.1
1.11	81.4
1.64	93.1



Based on the shape of the normality plot we conclude that the sample came from a not normal population.

When technology is not available to assess the normality just look at any visual representation of the sample data (dotplot, boxplot, stem-and-leaf, histogram etc.) and check the shape of the distribution for a symmetry, modality, and the outliers.

This page titled 1.8.3: Normality Plots is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.8.4: Approximating Binomial Random Variables



There is roughly an 80% chance that a person of age 20 will be alive at age 65. Suppose that ten people of age 20 are selected at random among those who purchased a life insurance. The insurance company wants to know the probability that the number of policy holders that are alive at age 65 is

a. exactly six.
b. at most four.
c. at least eight.

Let X represent the number of people alive at age 65 among the ten people that were selected, then

$$X \sim B(n = 10, p = 0.8)$$

and the following probability distribution table can be constructed using the formula

$$P(X = k) = C_k^1 0 \cdot 0.8^k \cdot 0.2^{10-k}$$

x_i	$P(X = x_i)$
0	1.024E-07
1	4.096E-06
2	7.3728E-05
3	0.000786432
4	0.005505024
5	0.026424115
6	0.088080384
7	0.201326592
8	0.301989888
9	0.268435456
10	0.107374182

Now, using the probability distribution table we can answer the questions:

a. $P(X=6)=0.0881$
b. $P(X\leq 6)=0.0064$
c. $P(X\geq 8)=0.6778$

For most real-world problems, the number of people under investigation is much larger than ten. Consider the same problem but let's suppose that 100 people of age 20 years are selected at random. The insurance company may want to know the probability that

a. exactly seventy-five of them will be alive at age 65.
b. at least eighty-three of them will be alive at age 65.

Let X represent the number of people alive at age 65 among the hundred people that were selected, then

$$X\sim B(n=100,p=0.8)$$

and the following probability distribution table can be constructed using the formula

$$P(X=k)=C_k^{100}\cdot 0.8^k\cdot 0.2^{100-k}$$

x_i	$P(X=x_i)$
0	$C_0^{100}\cdot 0.8^0\cdot 0.2^{100}$
1	$C_1^{100}\cdot 0.8^1\cdot 0.2^{99}$
2	$C_2^{100}\cdot 0.8^2\cdot 0.2^{98}$
3	$C_3^{100}\cdot 0.8^3\cdot 0.2^{97}$
...	...
75	$C_{75}^{100}\cdot 0.8^{75}\cdot 0.2^{25}$
...	...
83	$C_{83}^{100}\cdot 0.8^{83}\cdot 0.2^{17}$
...	...
99	$C_{99}^{100}\cdot 0.8^{99}\cdot 0.2^{1}$
100	$C_{100}^{100}\cdot 0.8^{100}\cdot 0.2^{0}$

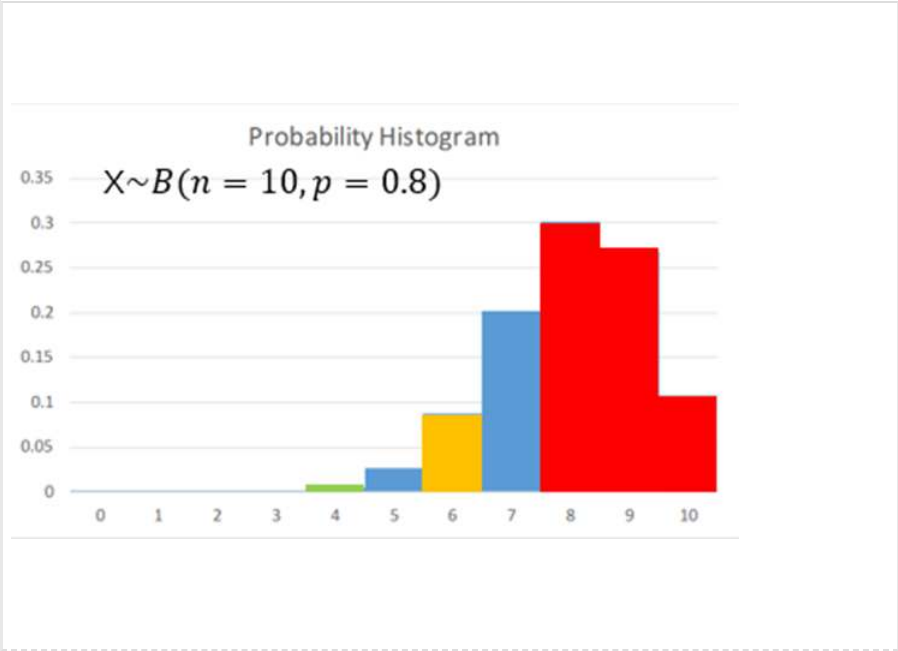
Now, using the probability distribution table we can compute the answers the questions:

a. $P(X=75)=C_{75}^{100}\cdot 0.8^{75}\cdot 0.2^{25}$
b. $P(X\geq 83)=P(X=83)+P(X=84)+\cdots+P(X=100)=C_{75}^{100}\cdot 0.8^{75}\cdot 0.2^{25}$

However, we run into a problem very quick! Finding the probability that $X=60$ requires enough computational power to handle $\frac{100!}{60!40!}$. And computing the probability that $X\geq 80$ requires even more computational power! So we have an interesting situation, when we know how to compute the answer yet cannot compute it without a powerful calculator.

Is there another way to find the desired probability without a powerful calculator? The short answer is yes.

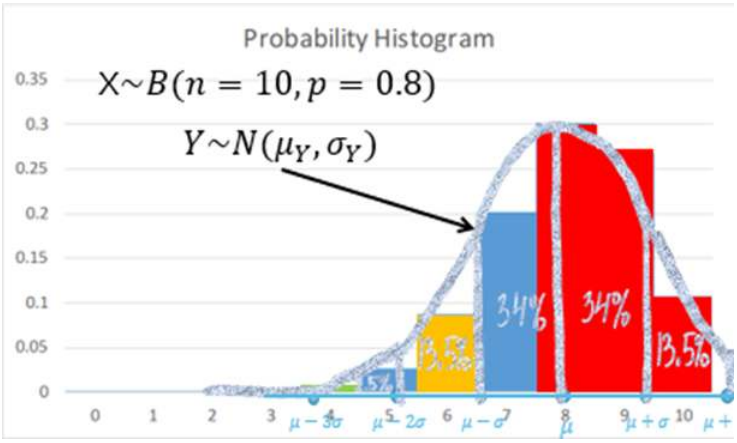
Let's scale back to the problem with the sample of size 10 and let's consider the probability histogram along with the probability distribution table.



x_i	$P(X = x_i)$
0	1.024E-07
1	4.096E-06
2	7.3728E-05
3	0.000786432
4	0.005505024
5	0.026424115
6	0.088080384
7	0.201326592
8	0.301989888
9	0.268435456
10	0.107374182

Recall that the probability histogram serves exactly the same purpose as the probability distribution table, that is we can read the probabilities from the histogram itself. For example, on the histogram, the bar that corresponds to $X = 6$ has the area equal to the $P(X = 6)$. Next, to find the probability $P(X \geq 8)$ is the same as to compute the area of the bars at 8, 9, and 10! Similarly, to find the probability $P(X \leq 4)$ is the same as to compute the area of the bars at 0, 1, 2, 3, 4 which is virtually equal to zero. This principle is true for any discrete random variable of any shape and in particular can be applied for a binomial random variable with any parameters n and p.

Does the overall shape of the histogram remind you of anything? Yes, it does look like a normal probability density curve!



But which values of μ and σ give us the best approximation? Recall that a binomial random variable with parameters n and p has the mean and standard deviation can be computed using the formulas:

$$\mu_X = np$$
$$\sigma_X = \sqrt{np(1-p)}$$

Now it is not surprising that a normal variable Y with the same μ and σ as the binomial random variable X has the probability density curve with the same shape as the histogram of X, in other words, the probability density curve of Y can be used to approximate the area in the histogram for X!

In our case, we have a binomial random variable X with parameters $n = 10$ and $p = 0.8$ whose mean is $\mu_X = np = 10 \cdot 0.8 = 8$ and standard deviation is $\sigma_X = \sqrt{np(1-p)} = \sqrt{10 \cdot 0.8 \cdot 0.2} = \sqrt{1.6} = 1.26$. Therefore, if we let Y be a normal random

variable with parameters $\mu_Y = \mu_X = 8$ and $\sigma_Y = \sigma_X = 1.26$ then Y can be used to approximate X.

So now that we know the parameters of the normal variable that approximates the original binomial variable X, we can focus on using it to answer the original questions about X!

While it is natural to approximate the $P(X \geq 8)$ with the $P(Y > 8)$, by doing so we will be ignoring the area of the left half of the bar at $X = 8$. A better estimate can be obtained by finding the area to the right of 7.5 which is $P(X \geq 8) \approx P(Y > 7.5) = 0.6543$ and is very close to the actual probability that we found earlier.

Definition: Correction for continuity

The adjustment of subtracting ½ from the boundary is called the **correction for continuity** and must be done to obtain the best estimate!

Similarly, while it is natural to approximate the $P(X = 6)$ with the $P(Y = 6)$, by doing so we will be ignoring the fact that the probability of $Y = 6$ is equal to zero. A better estimate can be obtained by finding the area between 5.5 and 6.5 which is $P(X = 6) \approx P(5.5 < Y < 6.5) = 0.0933$ and is very close to the actual probability that we found earlier.

In summary, for $X \sim B(n, p)$ and $Y \sim N(np, \sqrt{np(1-p)})$, the following approximations can be made:

$$P(X = c) \approx P(c - \tfrac{1}{2} < Y < c + \tfrac{1}{2})$$
$$P(X \leq b) \approx P(Y < b + \tfrac{1}{2})$$
$$P(X < b) \approx P(Y < b - \tfrac{1}{2})$$
$$P(X > a) \approx P(Y > a + \tfrac{1}{2})$$
$$P(X \geq a) \approx P(Y > a - \tfrac{1}{2})$$
$$P(a \leq X \leq b) \approx P(a - \tfrac{1}{2} < Y < b + \tfrac{1}{2})$$
$$P(a < X \leq b) \approx P(a + \tfrac{1}{2} < Y < b + \tfrac{1}{2})$$
$$P(a \leq X < b) \approx P(a - \tfrac{1}{2} < Y < b - \tfrac{1}{2})$$
$$P(a < X < b) \approx P(a + \tfrac{1}{2} < Y < b - \tfrac{1}{2})$$

Again, the adjustment of adding or subtracting ½ is called the correction for continuity.

Note that the approximation becomes less accurate for skewed binomial distributions, so only perform it when both $np \geq 5$ and $n(1-p) \geq 5$.

Let's go back to our original problem with 100 people instead of 10.

Example 1.8.4.1

Suppose that 100 people of age 20 years are selected at random. Find the probability that

a. exactly 75 of them will be alive at age 65.
b. at least 83 of them will be alive at age 65.

Solution

Let X represent the number of people alive at age 65 among the hundred people that were selected, then

$$X \sim B(n = 100, p = 0.8)$$

and

$$\mu_X = np = 100 \cdot 0.8 = 80 \text{ and } \sigma_X = \sqrt{np(1-p)} = \sqrt{100 \cdot 0.8 \cdot 0.2} = 4$$

Let's introduce a new random variable

$$Y \sim N(\mu_Y = \mu_X = 80, \sigma_Y = \sigma_X = 4)$$

Now, using the normal random variable Y we can answer the questions:

a. $P(X = 75) \approx P(74.5 < Y < 75.5) = 0.0457$
b. $P(X \geq 83) \approx P(Y > 82.5) = 0.2660$

These approximations are expected to be very close to the true values because both np and $n(1-p)$ are greater than 5.

Conclusion: we discussed how to use normal distributions to approximate the probabilities for binomial distributions. While this result is not as significant now as it was prior to technology becoming accessible, it still highlights the importance of the normal distributions in nature and applications!

This page titled 1.8.4: Approximating Binomial Random Variables is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

SECTION OVERVIEW

1.9: The Central Limit Theorem

1.9.1: The Central Limit Theorem for Sample Means

1.9.2: The Central Limit Theorem for Sample Sums

1.9.3: The Central Limit Theorem for Sample Proportions

This page titled 1.9: The Central Limit Theorem is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.9.1: The Central Limit Theorem for Sample Means

Section 0: Introduction to the Central Limit Theorem



Let's recall the definitions of the parameter and statistic.

Definition: Parameter and statistic

A **parameter** is a numerical summary of a population. A **statistic** is a numerical summary of a sample. The following table summarizes the differences between the two.

Table 1.9.1.0.1 : Table summarizing the differences between a statistic and the parameter.

Statistic	Parameter
sample	population
size	size
varies	constant
can be found	unknown

A parameter is something that is computed from a population, while unknown it remains constant. A statistic is something that is computed from a sample of size n, while it may vary from sample to sample it can be easily found when needed.

The following table summarizes the most common statistics along with their corresponding parameters.

Table 1.9.1.0.2 : The most common statistics along with their corresponding parameters.

Statistic	Parameter
sample mean	μ
sample proportion	p
sample variance	σ^2

One of the most important applications of a sample statistic is to approximate the unknown parameter of the population such as the mean, proportions, and variance.

Example 1.9.1.0.1

To approximate the average age of ALL students on campus, the average height of the class can be used.

To approximate the proportion of males among ALL the students on campus, the proportion of males in the class can be used.

To approximate the variance of the students' GPA in the entire college, the variance of the GPA of the students in the class.

Let's measure a few sample statistics. We may assume that our class is our sample, and it should not be too hard to imagine that the average height of our class is 167.5 cm, the proportion of males is 56%, and the variance of GPA is 0.61. All these numbers approximate the corresponding parameters which are unknown.

Let's take another sample. But before I write the results of the survey, let me ask you a question. Can you predict the outputs of the survey before the survey is performed? Here are the results! The mean is 169.3 cm, the proportion of males is 44% and the variance is 0.52. I am sorry, but I doubt that you guessed any of them correctly!

Table 1.9.1.0.3: The table summarizing the average height, proportion of males, and variance in GPA for two samples.

	Sample 1	Sample 2	Population Parameter
average height	167.5 cm	169.3 cm	μ
proportion of males	56%	44%	p
variance of GPA	0.61	0.52	σ^2

It is not hard to realize that a random observation of any sample statistic is almost never the same as another observation from another sample! Nor is it equal to the population parameter which is in general unknown. This error is unavoidable and is called the **sampling error**. Later, we will learn that the sampling error can be **minimized** by increasing the sample size, but it cannot be eliminated. However, many times the only way to learn anything about the parameter is through observing a statistic.

Let's introduce the notation. A single observation of a sample statistic will depend on the sample and therefore for an i-th sample, let's label the average as $\bar{x}_i$, the proportion as $\hat{p}$, the variance as s_i^2.

Table 1.9.1.0.4: The table summarizing the average height, proportion of males, and variance in GPA for two samples with added notation.

	Sample 1	Sample 2	Sample	Population Parameter
average height	$\bar{x}_1 = 167.5$	$\bar{x}_2 = 169.3$	$\bar{x}_i$	μ
proportion of males	$\hat{p}_1 = 0.56$	$\hat{p}_2 = 0.44$	$\hat{p}_i$	p
variance of GPA	$s_1^2 = 0.61$	$s_2^2 = 0.52$	s_i^2	σ^2

What all these quantities have in common is the fact they are impossible to guess before the actual sample is drawn, i.e., they are unknown quantities whose values depend on chance! Doesn't that sound familiar? Yes, that is the precise definition of a random variable!

Recall that a **random variable** is an unknown quantity whose value depends on chance. A random variable can be classified as discrete or continuous.

Thus, each of the quantities above and some others (such as the sample sum) can be thought of as continuous random variables!

Let's introduce a random variable notation – the average of a sample is $\bar{X}$, the proportion of a sample is $\hat{P}$, and the variance of a sample is S^2.

Why are they all continuous? Because the range of all possible values for a sample statistic is a continuous interval. For example:

- The average height of the students in a random sample can be any number between 140 and 200.
- The proportion of males in a random sample can be any number between 0 and 100%.
- The variance of a random sample can be any positive number.

Table 1.9.1.0.5: Table summarizing the notation associated with the use of random variables for sample statistics.

	Notation	Sample	Population Parameter

	Notation	Sample	Population Parameter
Sample mean	$\bar{X}$	$\bar{x}_i$	μ
Sample proportion	$\hat{P}$ or $\hat{p}$	$\hat{p}_i$	p
Sample variance	S^2 or s^2	s_i^2	σ^2

- $\bar{x}_i$ is a random observation of a random variable called "sample mean", $\bar{X}$.
- $\hat{p}_i$ is a random observation of a random variable called "sample proportion", $\hat{p}$.
- s_i^2 is a random observation of a random variable called "sample variance", s^2.

Next step is to determine the distributions of each of these random variables – are they normal? Student-T? or maybe χ^2? Once we determine the distribution for each of the variables, we would want to know the parameters – mean and standard deviation in case if the distribution is normal, and degrees of freedom if the distribution is T or χ^2 or F.

In summary, what's the Central Limit Theorem all about? It answers the exact questions that we asked earlier – what are the distribution of each sample statistic and its parameters? The Central Limit Theorem has three parts – one for each of the following random variables – the mean, the proportion, and the sum. The distribution of the sample variance can also be determined but it is not part of the Central Limit Theorem.

Section 1: The Central Limit Theorem for Means



Theorem 1.9.1.1.1 Central Limit Theorem for Sample Means

For a relatively large sample size n ($n \geq 30$) from a population X with parameters μ_X and σ_X, the sample mean, $\bar{X}$, is approximately normally distributed with parameters $\mu_{\bar{X}} = \mu_X$ and $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$, regardless of the distribution of X. For normally distributed X, the sample doesn't have to be large.

Symbolically, the statement can be written in the following way:

For $X \sim ?(\mu, \sigma)$ and $n \geq 30$ or for $X \sim N(\mu, \sigma)$ and any n

$$\bar{X} \sim N\left(\mu_{\bar{X}} = \mu_X, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\right)$$

Example 1.9.1.1.1

According to a study, the time between taking a pain reliever and getting relief for a randomly selected patient has unknown distribution with a mean of 46 minutes and a standard deviation of 5.9 minutes. Find the probability that the average time between taking a pain reliever and getting relief for a sample of 36 randomly selected patients is between 45 and 48 minutes.



Solution

Let X be the time between taking a pain reliever and getting relief for a randomly selected patient, then

$$X \sim ?(\mu_X = 46, \sigma_X = 5.9)$$

Let $\bar{X}$ be the average time between taking a pain reliever and getting relief for a random sample of 36 patients, then, by CLT,

$$\bar{X} \sim N\left(\mu_{\bar{X}} = \mu_X = 46, \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{5.9}{\sqrt{36}} = 0.9833\right)$$

Then the probability that the average time between taking a pain reliever and getting relief for a sample of 36 randomly selected patients is between 45 and 48 minutes can be expressed as $P(45 < \bar{X} < 48$:

$$P(45 < \bar{X} < 48) = P\left(\frac{45-46}{0.9833} < \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} < \frac{48-46}{0.9833}\right) = P(-1.02 < Z < 2.03) = 0.8250 = 82.50\%$$

Or using technology

$$P(45 < \bar{X} < 48) = 0.8244 = 82.44\%$$

This page titled 1.9.1: The Central Limit Theorem for Sample Means is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.9.2: The Central Limit Theorem for Sample Sums



Theorem 1.9.2.1 Central Limit Theorem for Sample Sums

For a relatively large sample size n $(n \geq 30)$ from a population X with parameters μ_X and σ_X, the sample sum, S, is approximately normally distributed with parameters $\mu_S = n \cdot \mu_X$ and $\sigma_{\bar{X}} = \sigma_X \cdot \sqrt{n}$, regardless of the distribution of X. For normally distributed X, the sample doesn't have to be large.

Symbolically, the statement can be written in the following way:

For $X \sim ?(\mu, \sigma)$ and $n \geq 30$ or for $X \sim N(\mu, \sigma)$ and any n

$$S \sim N\left(\mu_S = n \cdot \mu_X, \sigma_S = \sigma \cdot \sqrt{n}\right)$$

Example 1.9.2.1

According to a TSA study, the weights of passengers' carry-ons are distributed with a mean of 25 lbs. and a standard deviation of 4.5 lbs. Find the probability that the total weight of 200 randomly selected carry-ons is less than 5200 lbs. - the maximum recommended weight of carry-ons on a plane of such size.



Solution

Let X be the weight of the carry-on of a randomly selected passenger, then

$$X \sim ?(\mu_X = 25, \sigma_X = 4.5)$$

Let S be the total weight of the carry-ons of 200 randomly selected passengers, then, by CLT,

$$S \sim N\left(\mu_S = n \cdot \mu_X = 200 \cdot 25 = 5000, \sigma_S = \sigma_X \cdot \sqrt{n} = 4.5 \cdot \sqrt{200} = 63.64\right)$$

Then the probability that the total weight of 200 randomly selected carry-ons is less than 5200 lbs. can be expressed as

$$P(S < 5200) = P\left(\frac{S - \mu_S}{\sigma_S} < \frac{5200 - 5000}{63.64}\right) = P(Z < 3.14) = 0.9992 = 99.92\%$$

Or using technology,

$$P(S < 5200) = 0.9992 = 99.92\%$$

This page titled 1.9.2: The Central Limit Theorem for Sample Sums is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.9.3: The Central Limit Theorem for Sample Proportions



Theorem 1.9.3.1 Central Limit Theorem for Sample Proportions

If p is the proportion of a population that possesses a certain feature, then $\hat{p}$, the proportion of a sufficiently large sample of size n that possess the same feature, is approximately normally distributed with the mean $\mu_{\hat{p}} = p$ and the standard deviation $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$. In this context, a "large" sample is such that $np \geq 10$ and $n(1-p) \geq 10$.

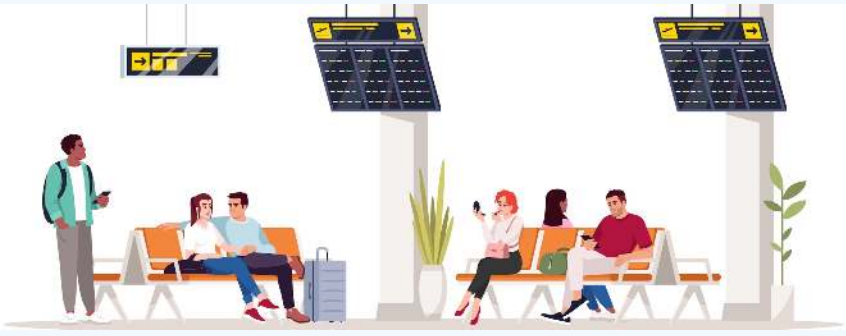
Symbolically, the statement can be written in the following way:

For $X \sim ?(p)$, when $np \geq 10$ and $n(1-p) \geq 10$,

$$\hat{p} \sim N\left(\mu_{\hat{p}} = p, \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\right)$$

Example 1.9.3.1

According to a study, 5.4% of the passengers fly first-class. A random sample of size 215 was obtained by a researcher. Find the probability that at least 10 people in the sample are first-class passengers.



Solution

Let p be the proportion of people that fly first-class, then $p = 0.054$.

Let $\hat{p}$ be the proportion of people that fly first-class in the sample of 215 random passengers. Since the sample is large, i.e.

$$np = 215 \cdot 0.054 = 11.61 \geq 10 \text{ and } n(1-p) = 215 \cdot (1 - 0.054) = 215 \cdot 0.946 = 203.39 \geq 10$$

by CLT,

$$\hat{p} \sim N\left(\mu_{\hat{p}} = p = 0.054, \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.054(1-0.054)}{215}} = 0.0154\right)$$

Then the probability that at least 10 people (i.e., proportion is at least 10/215) in the sample are first-class passengers can be expressed as

$$P(\hat{p} > \tfrac{10}{215}) = P(\hat{p} > 0.0465) = P\left(\frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} > \frac{0.0465 - 0.054}{0.0154}\right) = P(Z > -0.49) = 0.6879 = 68.79\%$$

Or using technology,

$$P(\hat{p} > 0.0465) = 0.6869 = 68.69\%$$

This page titled 1.9.3: The Central Limit Theorem for Sample Proportions is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

SECTION OVERVIEW

1.10: Estimating Parameters

1.10.1: Estimating the Population Mean

1.10.2: The One Mean T Procedure

1.10.3: The One Proportion Z Procedure

1.10.4: The One Variance Chi-Squared Procedure

1.10.5: The Summary of Confidence Intervals

This page titled 1.10: Estimating Parameters is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.10.1: Estimating the Population Mean

Section 0: Introduction to Confidence Intervals



Let's recall the definitions of the parameter and statistic.

Definition: Parameter and statistic

A **parameter** is a numerical summary of a population, and a **statistic** is a numerical summary of a sample. The following table summarizes the differences between the two.

Table 1.10.1.0.1: The side-by-side comparison of statistics and parameters.

Statistic	Parameter
sample	population
size	size
varies	constant
can be found	unknown

A **parameter** is something that is computed from a population, while unknown it remains constant. A **statistic** is something that is computed from a sample of size n, while it may vary from sample to sample it can be easily found when needed.

The most common application of a statistic is to estimate the unknown parameter.

Definition: Point estimate

We define a **point estimate** of a parameter as a single value used to estimate the parameter.

The sample mean, sample proportion, and sample variance can be used as point estimates of the corresponding population parameters.

Table 1.10.1.0.2: Examples of parameters and their corresponding statistics that are used as point estimators of parameters.

Statistic	Relation	Parameter
$\bar{x}$	is a point estimate of	μ
$\hat{p}$	is a point estimate of	p
s^2	is a point estimate of	σ^2

For example, the average age of the students in the class can be used as a point estimate of the average age of all students in the college. But keep in mind that a point estimate does not have to be a statistic - without any survey, one can guesstimate the average age of all college students to be 25 years old. Is there a difference between the two estimates? Intuitively we know that the sample mean is a better estimate, but why?

Definition: Biased and unbiased estimators

We define an **unbiased estimate** as a point estimate whose long-term average equals the parameter, otherwise the point estimate is called a **biased estimate** of the parameter.

All the sample statistics in the table above are unbiased point estimates of their corresponding parameters.

Now that we know that we can estimate an unknown parameter with the appropriate sample statistic we have to address the following questions:

1. How confident can one be when using a sample statistic as a point estimate? In other words, what are the chances that a random observation of a statistic will be equal to the parameter exactly?

   The answer is zero because the sample statistic is a continuous random variable and the probability of any continuous random variable being equal to a single value is always zero. In other words, we have **zero confidence** that a point estimate is equal to the unknown population parameter and vice versa we have zero confidence that the unknown population parameter is equal to one random observation of the sample statistic.

2. If there are zero chances that any point estimate equals exactly the parameter, why do we use the point estimates?

   Consider the following example - let's say we are running late to a meeting, and we receive a call with the question: "How long will it take us to get there?" One of the appropriate and probably the most common answers is "We will be there in, let's say, 10 minutes" This is an example of a point estimate of our arrival time. According to our discussion, we are 0% confident that we will get there in exactly 600 seconds or 10 minutes. Notice that while there are zero chances that it will take us exactly 10 minutes to arrive, the point estimate did answer the question. So often, we use a point estimate because a problem demands some number, and it is necessary to assume some value to continue with the solution. Soon, we will have situations in which we need a population standard deviation or proportion but instead we will be using a sample standard deviation or a sample proportion.

3. Is it possible to increase the confidence level?

   Consider the previous example. We are running late to a meeting, and we receive a call with the question: "How long will it take us to get there?" Is there another way to answer this question? How about instead of saying that we will be there in ten minutes which is obviously wrong we say something like "We will be there between 5 and 15 minutes."? This is an example of an **interval estimate** of our arrival time. Notice that the chances of our estimate being correct are now significantly higher than zero. In other words, with positive confidence, we estimate our arrival time to be between 5 and 15 minutes. In general, we can increase the confidence level by using intervals instead of a single number!

4. Is it possible to increase the confidence level to 100%?

   Continuing with the previous example. We are still running late to a meeting, and we receive a call with the question: "How long will it take us to get here?" Is there a way to answer this question with a 100% confidence level? Turns out that it is not as hard as it seems. How about "We will be there sometime between now and tomorrow."? This is an example of a 100% confidence interval estimate of our arrival time. Why is there 100% confidence? Because the destination is only 5 miles away, thus we can be 100% sure that we will arrive there sometime before the end of the day even if we must walk the rest of the trip. Is there a value in a 100% confidence interval? Not much. In practice, the most practical estimate will be with the confidence level of 95% and we can increase it to 99% or decrease it to 90% depending on the application.

The goal is to create a formal approach for constructing interval estimates with some level of confidence.

Section 1: The Generalized Empirical Rule



Next, we will develop a procedure to create an interval estimate of the unknown parameter with positive confidence.



Let's assume that Bobert's commute time is normally distributed with the mean 25 minutes and standard deviation 2 minutes. In other words, if we let X be a randomly selected commute time of Bobert, then

$$X \sim N(\mu = 25, \sigma = 2)$$

Now, recall that for a normal random variable the Empirical Rule applies and according to the Empirical rule the following probability statements are true!

- The probability that a random observation is within 1 standard deviation from the mean is 68%.
- The probability that a random observation is within 2 standard deviations from the mean is 95%.
- The probability that a random observation is within 3 standard deviations from the mean is 99.7%.

In our cases, where X is a random Bobert's commute time, we can apply and interpret the Empirical rule in the following way:

- The probability that a random Bobert's commute is between 23 and 27 minutes is 68%.
- The probability that a random Bobert's commute is between 21 and 29 minutes is 95%.
- The probability that a random Bobert's commute is between 19 and 31 minutes is 99.7%.

Theorem: The Generalized Empirical Rule

In general, for any continuous random variable X, $(1-\alpha)\cdot 100\%$ of the area under the probability density curve is between $x_{1-\alpha/2}$ and $x_{\alpha/2}$, thus for $X \sim N(\mu, \sigma)$, it can be visualized as:

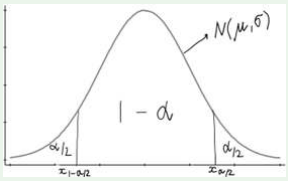


Figure 1.10.1.1.1: Image of a bell curve with the area and the boundaries of the regions labeled using the proper notation.

Alternatively, this fact can be expressed as the following probability statement:

$$P\left(x_{1-\alpha/2} < X < x_{\alpha/2}\right) = 1-\alpha$$

where

$$x_{\alpha/2} = \mu + z_{\alpha/2}\sigma$$

and

$$x_{1-\alpha/2} = \mu + z_{1-\alpha/2}\sigma = \mu - z_{\alpha/2}\sigma$$

thus

$$P\left(\mu - z_{\alpha/2}\sigma < X < \mu + z_{\alpha/2}\sigma\right) = 1-\alpha$$

We will refer to the last statement above as **the generalized empirical rule**.

Let's get used to the generalized empirical rule and practice applying it for some common values of alpha.

Example 1.10.1.1.1.1

When $\alpha = 0.32$:

$$1-\alpha = 0.68$$

$$\alpha/2 = \frac{0.32}{2} = 0.16$$

$$z_{\alpha/2} = z_{0.16} = 0.99 \approx 1$$

Hence,

$$P\left(\mu - 1\sigma < X < \mu + 1\sigma\right) = 0.68$$

which is equivalent to the first part of the Empirical Rule. We may also interpret this probability statement as if we are 68% confident that a random observation is within 1σ from μ.

Example 1.10.1.1.1.2

When $\alpha = 0.10$:

$$1 - \alpha = 0.90$$

$$\alpha/2 = \frac{0.10}{2} = 0.05$$

$$z_{\alpha/2} = z_{0.05} = 1.65$$

Hence,

$$P\left(\mu - 1.65\sigma < X < \mu + 1.65\sigma\right) = 0.90$$

which goes beyond what the Empirical Rule states. We may also interpret this probability statement as if we are 90% confident that a random observation is within 1.65σ from μ.

Note that the generalized empirical rule allows us to make probability statements and interpret them in a certain way for any choice of α.

Table 1.10.1.1.2: Summary of the applications of the Generalized Empirical Rule and interpretations.

Probability statement	Interpretation
$P\left(\mu - 1\sigma < X < \mu + 1\sigma\right) = 0.68$	We are 68% confident that a random observation x_i is within 1σ from μ.
$P\left(\mu - 1.65\sigma < X < \mu + 1.65\sigma\right) = 0.90$	We are 90% confident that a random observation x_i is within 1.65σ from μ.
$P\left(\mu - z_{\alpha/2}\sigma < X < \mu + z_{\alpha/2}\sigma\right) = 1 - \alpha$	We are $(1-\alpha)\cdot 100\%$ confident that a random observation x_i is within $z_{\alpha/2}\sigma$ from μ.

The most common values of α are 0.01, 0.05, and 0.10 to produce 99%, 95%, and 90% confidence levels, respectively. This means that to make a statement with a certain level of confidence one just simply has to pick the right alpha.

Example 1.10.1.1.2

What value of α will result in 80% confidence?

Solution

$\alpha = 0.2$

Each probability statement about X made by applying the generalized empirical rule

$$P\left(\mu - z_{\alpha/2}\sigma < X < \mu + z_{\alpha/2}\sigma\right) = 1 - \alpha$$

can be interpreted in two ways as shown below:

We are $(1-\alpha)\cdot 100\%$ confident that

(1) x_i is within $z_{\alpha/2}\sigma$ from μ

OR

(2) μ is within $z_{\alpha/2}\sigma$ from x_i.

In the first interpretation, we need to know μ to estimate a random observation x_i, and in the second interpretation, we need to know the random observation x_i to estimate the unknown μ. We certainly see the potential in the second interpretation!

Section 2: Constructing a Confidence Interval Using a Single Observation



Let's assume now that the average of Bobert's commute is unknown, that is, X is normally distributed with the unknown mean (μ) and given standard deviation (σ) of 2 minutes. Instead of trying to predict today's commute we can simply measure it. So, let's assume that today's commute (x_i) was 24 minutes. Can we estimate the unknown mean commute time (μ) with 95% confidence?

By the Generalized Empirical Rule for $\alpha = 0.05$:

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 95\%$$

Hence, we are 95% confident that:

$$x_i \text{ is within } 2\sigma \text{ from } \mu$$

which is the same as

$$\mu \text{ is within } 2\sigma \text{ from } x_i$$

which is the same as

$$\mu \text{ is within 4 from 24}$$

which is the same as

$$\mu \text{ is between 20 and 28}$$

As a result, we obtained a 95% confidence interval for the unknown mean commute time based on a single random observation.

Next, we are going to construct an interval estimate with a 90% confidence.

By the Generalized Empirical Rule for $\alpha = 0.1$:

$$P(\mu - 1.65\sigma < X < \mu + 1.65\sigma) = 95\%$$

Hence, we are 90% confident that:

$$x_i \text{ is within } 1.65\sigma \text{ from } \mu$$

which is the same as

$$\mu \text{ is within } 1.65\sigma \text{ from } x_i$$

which is the same as

$$\mu \text{ is within 3.30 from 24}$$

which is the same as

$$\mu \text{ is between 20.7 and 27.3}$$

As a result, we obtained a 90% confidence interval for the unknown mean commute time based on a single random observation.

Next, we are going to introduce some vocabulary.

Definition: Interval estimate

Interval estimate is a range of numbers that we think contains the unknown parameter.

Definition: Confidence level

Confidence level is the confidence that we have that the interval estimate actually contains the parameter.

Definition: Confidence interval

Confidence interval is a way to estimate the parameter using an interval estimate along with some confidence level.

Also, let's recall the properties of an interval.

Definition: Properties of an interval

An interval has the **center**, the **radius**, and the left and right **boundaries**. When constructing a confidence interval, it is common to use a point estimate as the center, the radius is usually called the **margin of error** and the boundaries are referred to as lower and upper bounds.

To construct an interval means to find its boundaries. In our example,

- the point estimate is a single random observation x_i,
- the margin of error can be found using the provided formula

$$ME = z_{\alpha/2}\sigma$$

- the boundaries can be found by adding and subtracting the ME from and to the point estimate

$$x_i \pm ME$$

Definition: Critical values

Note, that in this context, the values $z_{\alpha/2}$ and $-z_{\alpha/2}$ (or $z_{1-\alpha/2}$) are referred to as the left and right **critical values**.

The procedure can be summarized in the form of the table:

Table 1.10.1.2.1: Template of the procedure to construct a confidence interval from a single observation.

Unknown Parameter	μ, population mean
Point Estimate	x_i, a random observation
Confidence Level (of the point estimate)	0%
Confidence Level (for a new interval estimate)	$(1-\alpha)100\%$ α $\alpha/2$
Critical Value(s)	$z_{\alpha/2}$
Margin of Error	$ME = z_{\alpha/2}\sigma$
Lower Bound	$LB = x_i - ME$
Upper Bound	$UB = x_i + ME$
Interpretation	We are $(1-\alpha)100\%$ confident that the unknown population mean μ is between LB and UB.

In our example, we fill out the table in the following way.

Table 1.10.1.2.2: The summary of the procedure to construct a 90% confidence interval for Bobert's commute time from a single observation.

Unknown Parameter	μ, the average commute time
Point Estimate	$x_i = 24$, today's commute time
Confidence Level (of the point estimate)	0%
Confidence Level (for a new interval estimate)	$(1-\alpha)100\% = 90\%$ $\alpha = 0.1$ $\alpha/2 = 0.05$
Critical Value(s)	$z_{\alpha/2} = z_{0.05} = 1.65$
Margin of Error	$ME = 1.65 \cdot 2 = 3.30$

Lower Bound	$LB = 24 - 3.30 = 20.70$
Upper Bound	$UB = 24 + 3.30 = 27.30$
Interpretation	We are 90% confident that the average commute time is between 20.70 and 27.30 minutes.

We are 90% confident that the average commute time is between 20.70 and 27.30 minutes.

Section 3: Constructing a Confidence Interval Using a Sample



Is there a way to improve the accuracy of our existing procedure? Naturally, the smaller the ME the more accurate the interval estimate. The ME formula has only two components: σ and α:

$$ME = z_{\alpha/2} \cdot \sigma$$

σ cannot be chosen as it is the population parameter and therefore must remain constant! If we look at α, we can see that a higher accuracy can be achieved by sacrificing the confidence. In our example above, we obtained the following confidence intervals:

[20,28] is a 95% confidence interval

[20.70,27.30] is a 90% confidence interval

[24,24] is a 0% confidence interval

Easy to see that a 90% confidence interval is narrower than the 95% confidence. This observation is consistent with our understanding of the estimating process. The narrowest confidence interval is the point estimate which has a 0% confidence since there is a 0% chance that the point estimate is correct! So, to answer the question, we **can** improve the accuracy by lowering the confidence. But is there a way to improve the accuracy without sacrificing the confidence? The short answer is YES!

Let's take a look at the same problem once again. Based on the previous procedure, we are 90% confident that is between 20.7 and 27.3. Is there anything in the problem that we could have done better? Does a sample of size one seem large enough? Intuitively we know that a sample of size greater than 1 should have produces a better result but how exactly?

This time let's assume that Bobert's commute is normally distributed with the unknown mean and given standard deviation 2 minutes. Instead of measuring it once we surveyed a sample of 10 commutes, and the results are listed below:

25, 21, 23, 22, 22, 26, 24, 27, 24, 26

Can we improve the estimate of the average commute with 90% confidence level?

Note that the average of the sample is

$$\bar{x}_i = \frac{25+21+23+22+22+26+24+27+24+26}{10} = 24$$

and $\bar{x}_i$ is the best choice for a point estimate of the unknown population mean.

Previously, we have applied the Generalized Empirical Rule on the normal population, in our example, the population of commute times. However, by the Central Limit Theorem the distribution of sample means is also normally distributed with known parameters $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$:

$$\bar{X} \sim N\left(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\right)$$

Thus, the Generalized Empirical Rule can also be applied for a normally distributed sample mean, $\bar{X}$. This results in the following probability statement:

$$P\left(\mu_{\bar{X}} - z_{\alpha/2}\sigma_{\bar{X}} < X < \mu_{\bar{X}} + z_{\alpha/2}\sigma_{\bar{X}}\right) = 1-\alpha$$

Replacing $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$:

$$P\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < X < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Visually, the only difference is in the presence of the square root of the sample size in the denominators. However, this is exactly what will cause the ME to get smaller as the sample size gets larger!

Just like before, this probability statement can be interpreted in two ways as shown below:

We are $(1-\alpha)\cdot 100\%$ confident that

(1) $\bar{x}_i$ is within $z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ from μ

OR

(2) μ is within $z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ from $\bar{x}_i$.

In the first interpretation, we need to know μ to estimate the average of a random sample, $\bar{x}_i$, and in the second interpretation, we need to know the average of a random sample, $\bar{x}_i$, to estimate the unknown population mean, μ. We again will use the second interpretation to develop a formal procedure for estimating the unknown mean using the mean of a random sample.

The table below summarizes the differences between using a single random observation (on the left) and the mean of a random sample of size n (on the right).

Table 1.10.1.3.1: Template of the procedure to construct a confidence interval from a sample of size n.

	Single observation	Sample of size n
Unknown Parameter	μ, population mean	
Point Estimate	x_i, a random observation	$\bar{x}_i$, the mean of a random sample of size n
Confidence Level (of the point estimate)	0%	
Confidence Level (for a new interval estimate)	$(1-\alpha)100\%$ α $\alpha/2$	
Critical Value(s)	$z_{\alpha/2}$	
Margin of Error	$ME = z_{\alpha/2}\sigma$	$ME = z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$
Lower Bound	$LB = x_i - ME$	$LB = \bar{x}_i - ME$
Upper Bound	$UB = x_i + ME$	$UB = \bar{x}_i + ME$
Interpretation	We are $(1-\alpha)100\%$ confident that the unknown population mean μ is between LB and UB.	

Again, it is easy to see that the main difference is in the formula for the margin of error. With the square root of the sample size in the denominator, we found a way to decrease the margin of error without sacrificing the confidence!

In our example, we are going to use this template to construct a 90% confidence interval for the unknown population mean using the average of a random sample of size 10.

Table 1.10.1.3.2: The summary of the procedure to construct a 90% confidence interval for Bobert's commute time from a sample of 10 observations.

Unknown Parameter	μ, population mean
Point Estimate	$\bar{x}_i = 24$, the mean of the sample of size $n = 10$
Confidence Level (of the point estimate)	0%
Confidence Level (for a new interval estimate)	$(1-\alpha)100\% = 90\%$ $\alpha = 0.1$ $\alpha/2 = 0.05$
Critical Value(s)	$z_{\alpha/2} = z_{0.05} = 1.65$
Margin of Error	$ME = 1.65\cdot\dfrac{2}{\sqrt{10}} = 0.73$
Lower Bound	$LB = 24 - 0.73 = 23.27$

Upper Bound	$UB = 24 + 0.73 = 24.73$
Interpretation	We are 90% confident that the average commute time is between 23.27 and 24.73 minutes.

Let's list all the assumptions that were made in order for the designed procedure to work:

- The population standard deviation, σ, is known.
- The sample is simple random.
- The CLT must be applicable, that is at least one of the following must be true:
  - The sample size, n, is greater than 30.
  - The population is normally distributed.

We have developed a procedure for finding a confidence interval for the unknown population mean with known population standard deviation by using the average of a random sample.

Section 4: The Accuracy and Margin of Error

Let's take another look at the following formula.

$$ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Since σ is a constant and cannot be changed, a higher accuracy or smaller ME can be achieved by increasing α and sacrificing the confidence **OR** by increasing the sample size without any sacrifice.

Therefore, if we want to construct a $(-\alpha)\cdot 100\%$ confidence interval with the margin of error no more than E_0, how large should weuld choose the sample size? Since the margin of error can't be greater than E_0 we have the following inequality:

$$ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < E_0$$

By isolating the sample size, we obtain the following solution:

$$n > \left(z_{\alpha/2}\cdot\frac{\sigma}{E_0}\right)^2$$

This means that to find the required sample size we need to know α, σ, and E_0, the desired bound on the margin of error.

For example, let's find the sample size required to construct a 90% confidence interval with the margin of error no more than 0.5, if the population standard deviation is 2?

To apply the formula, we must know E_0, σ, and $z_{\alpha/2}$. From the problem, we know that $E_0 = 0.5$ and $\sigma = 2$. For a confidence level of 90%, $\alpha = 0.1$, therefore (z_{\alpha/2}=z_{0.05}=1.65\).

Finally, we get:

$$n > \left(z_{\alpha/2}\cdot\frac{\sigma}{E_0}\right)^2 = \left(1.65\cdot\frac{2}{0.5}\right)^2 = 43.56 \uparrow 44$$

Remember in such problems to always round the sample size UP!

We just learned how to choose the sample size to achieve a desired accuracy when constructing a confidence interval for the unknown population mean with known population standard deviation using the average of a random sample.

Section 5: One Mean Z Procedure



Consider the following example:

An incubation period is a time between when you contract a virus and when your symptoms start. Assume that the population of incubation periods for a novel coronavirus is normally distributed with the population standard deviation of 2 days. By surveying randomly selected local hospitals, a researcher was able to obtain the following sample of incubation periods of 10 patients:

6, 2, 4, 3, 3, 7, 5, 8, 5, 7

a. Use the sample to construct and interpret a 90% confidence interval for the average incubation period of the novel coronavirus.

First, let's check if all necessary assumptions are satisfied:

- The sample is simple random.
- The population is normally distributed.
- The population standard deviation, σ, is known and $\sigma = 2$.

Also, note that the average of the sample is

$$\bar{x}_i = \frac{6+2+4+3+3+7+5+8+5+7}{10} = 5$$

Once the assumptions are verified, we may apply the procedure by following the template:

Table 1.10.1.5.1: The summary of the procedure to construct a 90% confidence interval for the mean incubation period of a novel coronavirus from a sample of size 10.

	Example	Template
Unknown Parameter	μ, the average incubation period of the novel coronavirus	μ, population mean
Point Estimate	$\bar{x}_i = 5$, the average incubation period in the sample of size $n = 10$	$\bar{x}_i$, the mean of a random sample of size n
Confidence Level (of the point estimate)	0%	0%
Confidence Level (for a new interval estimate)	$(1-\alpha)100\% = 90\%$ $\alpha = 0.1$ $\alpha/2 = 0.05$	$(1-\alpha)100\%$ α $\alpha/2$
Critical Value(s)	$z_{\alpha/2} = z_{0.05} = 1.65$	$z_{\alpha/2}$
Margin of Error	$ME = 1.65 \cdot \frac{2}{\sqrt{10}} = 1.04$	$ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
Lower Bound	$LB = 5 - 1.04 = 3.96$	$LB = \bar{x}_i - ME$
Upper Bound	$UB = 5 + 1.04 = 6.04$	$UB = \bar{x}_i + ME$
Interpretation	We are 90% confident that the average incubation period of the novel coronavirus is between 3.96 and 6.04 days.	We are $(1-\alpha)100\%$ confident that the unknown population mean μ is between LB and UB.

Now we can interpret the results: We are 90% confident that the average incubation period of the novel coronavirus is between 3.96 and 6.04 days.

In addition, we may also answer the following question.

b. How large must the sample be so that the margin of error is less than 0.25?

To answer this question, we will use the formula

$$n > \left(z_{\alpha/2} \cdot \frac{\sigma}{E_0} \right)^2$$

in which we can plug in all the values

$$n > \left(1.65 \cdot \frac{2}{0.25} \right)^2 = 174.24 \uparrow 175$$

In conclusion, 175 is the smallest sample size for which the margin of error will be less than 0.25 for a 90% confidence level.

This page titled 1.10.1: Estimating the Population Mean is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.10.2: The One Mean T Procedure



In One Mean Z Procedure, one major assumption was made that is rarely satisfied - it is very unlikely for anyone to know the population standard deviation without knowing the population mean. So, what do we do in a real-life scenario?

Naturally when a population parameter is not known we can use a sample statistic as an unbiased estimate. It is very natural to use the sample standard deviation, s, instead of the unknown population standard deviation, σ.

As a result, the sample mean of a sample of size n is not normally distributed anymore but rather has another well-studied distribution called Student t-distribution with $n-1$ degrees of freedom.

From the practical point of view, the effects of this change are very simple – we can still use the same procedure but need to replace σ and $z_{\alpha/2}$ with s and $t_{\alpha/2}$ respectively.

For the procedure to work the following assumptions must be made:

- The sample is simple random.
- The CLT must be applicable, that is at least one of the following must be true:
  - The sample size, n, is greater than 30.
  - The population is normally distributed.
- The population standard deviation, σ, is unknown.

Consider the following example.

Example 1.10.2.1



An incubation period is a time between when you contract a virus and when your symptoms start. Assume that the population of incubation periods for a novel coronavirus is normally distributed. By surveying randomly selected local hospitals, a researcher was able to obtain the following sample of incubation periods of 10 patients:

6, 2, 4, 3, 3, 7, 5, 8, 5, 7

Use the sample to construct and interpret a 90% confidence interval for the average incubation period of the novel coronavirus.

Solution

First, let's check if all necessary assumptions are satisfied:

- The sample is simple random.
- The population is normally distributed thus the CLT for sample means is applicable.
- The population standard deviation, σ, is unknown.

Also, note that the average and the standard deviation of the sample are respectively

$$\bar{x}_i = \frac{6+2+4+3+3+7+5+8+5+7}{10} = 5 \text{ and } s = 2$$

We will use the following template to construct and interpret the confidence interval:

Table 1.10.2.1.1: Template of the procedure to construct a confidence interval for the population mean from a sample when the population standard deviation is unknown.

Unknown Parameter	μ, population mean
Point Estimate	$\bar{x}_i$, the mean of a random sample of size n
Confidence Level (of the point estimate)	0%
Confidence Level (for a new interval estimate)	$(1-\alpha)100\%$ α $\alpha/2$
Critical Value(s)	$t_{\alpha/2}$ with $df = n-1$
Margin of Error	$ME = t_{\alpha/2}\frac{s}{\sqrt{n}}$
Lower Bound	$LB = \bar{x}_i - ME$
Upper Bound	$UB = \bar{x}_i + ME$
Interpretation	We are $(1-\alpha)100\%$ confident that the unknown population mean μ is between LB and UB.

In our example, we perform the procedure (fill out the table) in the following way:

Table 1.10.2.1.2: The summary of the procedure to construct a 90% confidence interval for the mean incubation period of a novel coronavirus from a sample of size 10 when the population standard deviation is unknown.

	Example	Template
Unknown Parameter	μ, the mean incubation period	μ, population mean
Point Estimate	$\bar{x}_i = 5$, the mean incubation period in the sample of size $n = 10$	$\bar{x}_i$, the mean of a random sample of size n
Confidence Level (of the point estimate)	0%	0%
Confidence Level (for a new interval estimate)	$CL = 90\%$ $\alpha = 0.10$ $\alpha/2 = 0.05$	$(1-\alpha)100\%$ α $\alpha/2$
Critical Value(s)	When $df = 9$: $t_{\alpha/2} = t_{0.05} = 1.83$	$t_{\alpha/2}$ with $df = n-1$
Margin of Error	$ME = 1.83 \cdot \frac{2}{\sqrt{10}} = 1.16$	$ME = t_{\alpha/2}\frac{s}{\sqrt{n}}$
Lower Bound	$LB = 5 - 1.16 = 3.84$	$LB = \bar{x}_i - ME$
Upper Bound	$UB = 5 + 1.16 = 6.16$	$UB = \bar{x}_i + ME$
Interpretation	We are 90% confident that the average incubation period of the novel coronavirus is between 3.84 and 6.16 days.	We are $(1-\alpha)100\%$ confident that the unknown population mean μ is between LB and UB.

We are 90% confident that the average incubation period of the novel coronavirus is between 3.84 and 6.16 days.

This page titled 1.10.2: The One Mean T Procedure is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.10.3: The One Proportion Z Procedure



In One Mean Z and T Procedures, we estimate the unknown population mean by constructing a confidence interval around the sample mean. But we can estimate other unknown parameters as well. To construct the confidence interval for the population mean we used the Central Limit Theorem for Sample Means. Similarly, we can use the Central Limit Theorem for Sample Proportions to construct a confidence interval for an unknown population proportion.

By the Central Limit Theorem for proportions

$$\hat{p} \sim N\left(\mu_{\hat{p}} = p, \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}\right)$$

thus, we can apply the Generalized Empirical Rule for $\hat{p} \sim N\left(\mu_{\hat{p}}, \sigma_{\hat{p}}\right)$:

$$P\left(\mu_{\hat{p}} - z_{\alpha/2}\sigma_{\hat{p}} < X < \mu_{\hat{p}} + z_{\alpha/2}\sigma_{\hat{p}}\right) = 1 - \alpha$$

After plugging in $\mu_{\hat{p}} = p$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$:

$$P\left(p - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < X < p + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) = 1 - \alpha$$

Therefore, we have the following two interpretations:

We are $(1-\alpha)\cdot 100\%$ confident that

(1) $\hat{p}_i$ is within $z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ from p

OR

(2) p is within $z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ from $\hat{p}_i$.

Hence, we are $(1-\alpha)\cdot 100\%$ confident that p is between $\hat{p}_i - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ and $\hat{p}_i + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$.

Similarly to the other procedures, it appears that it is reasonable to build the confidence interval around the sample proportion $\hat{p}_i$ at the center with the margin of error $z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$. However, the estimate for the unknown value p cannot depend on itself, therefore we cannot use p in $ME = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$ to calculate the margin of error. As in One Mean T Procedure, where we replaced the unknown σ with s, we are going to replace the unknown p with $\hat{p}_i$. Thus, the formula to approximate the margin of error is

$$ME = z_{\alpha/2}\sqrt{\frac{\hat{p}_i(1-\hat{p}_i)}{n}}$$

As a result, the following template will be used to construct and interpret the confidence intervals:

Table 1.10.3.1.1: Template of the procedure to construct a confidence interval for unknown population proportion.

Unknown Parameter	p, population proportion

Point Estimate	$\hat{p}_i = \frac{x}{n}$, the sample proportion where n is the sample size and x is the number of observations that satisfy certain condition.
Confidence Level (of the point estimate)	0%
Confidence Level (for a new interval estimate)	$(1-\alpha)100\%$ α $\alpha/2$
Critical Value(s)	$z_{\alpha/2}$
Margin of Error	$ME = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$
Lower Bound	$LB = \hat{p}_i - ME$
Upper Bound	$UB = \hat{p}_i + ME$
Interpretation	We are $(1-\alpha)100\%$ confident that the unknown population proportion p is between LB and UB.

For the procedure to work the following assumptions must be made:

- The sample is simple random.
- The CLT for sample proportions must be applicable that is $np \geq 10$ and $n(1-p) \geq 10$.

Let's say we want to construct a $(1-\alpha)100\%$ confidence interval with the margin of error no more than E_0, how large we should choose the sample size?

By solving the following inequality:

$$ME = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < E_0$$

We get:

$$n > \left(z_{\alpha/2} \cdot \frac{\sqrt{p(1-p)}}{E_0}\right)^2$$

Note that the right-hand side contains the unknown p which makes the formula useless! The good news is that while p is unknown the following inequality is true for ANY p!

$$p(1-p) \leq \tfrac{1}{4}$$

So that

$$\left(\frac{z_{\alpha/2}}{2E_0}\right)^2 > \left(z_{\alpha/2} \cdot \frac{\sqrt{p(1-p)}}{E_0}\right)^2$$

Thus, we can guarantee the desired margin of error by choosing n to satisfy the following inequality:

$$n > \left(\frac{z_{\alpha/2}}{2E_0}\right)^2$$

Consider the following example.

Example 1.10.3.1



The ABC News/Ipsos poll was conducted by Ipsos Public Affairs' Knowledge Panel® on March 18-19, 2020, in English and Spanish, among a random national sample of 512 adults. In the poll, 282 Americans approve of the president's management of the crisis.

a. Construct and interpret a 90% confidence interval for the proportion of the Americans that approve of the president's management of the crisis.

First, let's check if all necessary assumptions are satisfied:

- The sample is simple random.
- $np = 282 \geq 10$ and $n(1-p) = 230 \geq 10$ thus the CLT for sample proportions is applicable.

Also, note that the sample proportion is

$$\hat{p}_i = \frac{282}{512} = 0.5508 = 55.08\%$$

Once the assumptions are verified, we may apply the procedure by using the template.

Table 10.3.1.2: The summary of the procedure to construct a 90% confidence interval for the population of Americans that approve of the president's management of the crisis.

Table 1.10.3.1.2: The summary of the procedure to construct a 90% confidence interval for the population of Americans that approve of the president's management of the crisis.

	Example	Template
Unknown Parameter	p, the approval rating	p, population proportion
Point Estimate	$\hat{p}_i = 0.5508$, the proportion of the sample that approve the president's handling of the crisis.	$\hat{p}_i = \frac{x}{n}$, the sample proportion where n is the sample size and x is the number of observations that satisfy certain condition.
Confidence Level (of the point estimate)	0%	0%
Confidence Level (for a new interval estimate)	$(1-\alpha)100\% = 90\%$ $\alpha = 0.10$ $\alpha/2 = 0.05$	$(1-\alpha)100\%$ α $\alpha/2$
Critical Value(s)	$z_{\alpha/2} = z_{0.05} = 1.65$	$z_{\alpha/2}$
Margin of Error	$ME = 1.65\sqrt{\frac{0.5508(1-0.5508)}{512}} = 0.0363$	$ME = z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$
Lower Bound	$LB = 0.5508 - 0.0363 = 0.5145$	$LB = \hat{p}_i - ME$
Upper Bound	$UB = 0.5508 + 0.0363 = 0.5871$	$UB = \hat{p}_i + ME$
Interpretation	We are 90% confident that the approval rating is between 51.45% and 58.71%.	We are $(1-\alpha)100\%$ confident that the unknown population proportion p is between LB and UB.

Now we can interpret the results: We are 90% confident that the approval rating is between 51.45% and 58.71%.

In addition, we may also answer the following question:

b. How large must the sample be so that the margin of error is less than 2.5%?

To answer this question, we will use the following formula:

$$n > \left(\frac{z_{\alpha/2}}{2E_0}\right)^2$$

in which we can plug in all the values:

$$n > \left(\frac{z_{\alpha/2}}{2E_0}\right)^2 = \left(\frac{1.65}{2 \cdot 0.025}\right)^2 = 1089.0 \uparrow 1090$$

We conclude that 1090 is the smallest sample size for which the margin of error will be less than 2.5% for a 90% confidence level.

This page titled 1.10.3: The One Proportion Z Procedure is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.10.4: The One Variance Chi-Squared Procedure



To create the One Mean Z and T Procedures and One Proportion Z Procedure, we used the Central Limit Theorem for sample means and proportions. While there is no Central Limit Theorem for sample variances, we can still construct a confidence interval using the fact that $\frac{s^2}{\sigma^2/(n-1)}$ follows a well-studied χ^2-distribution with $n-1$ degrees of freedom, i.e.

$$\frac{s^2}{\sigma^2/(n-1)} \sim \chi^2_{df=n-1}$$

The following is the equivalent of the Generalized Empirical Rule:

$$P\left(\chi^2_{1-\alpha/2} < \frac{s^2}{\sigma^2/(n-1)} < \chi^2_{\alpha/2}\right) = 1-\alpha$$

After manipulating the inequality inside the parenthesis, we can conclude that:

$$P\left(\frac{\sigma^2\chi^2_{1-\alpha/2}}{n-1} < s^2 < \frac{\sigma^2\chi^2_{\alpha/2}}{n-1}\right) = 1-\alpha$$

Therefore, we have the following two interpretations:

We are $(1-\alpha)\cdot 100\%$ confident that

(1) s_i^2 is between $\frac{\sigma^2\chi^2_{1-\alpha/2}}{n-1}$ and $\frac{\sigma^2\chi^2_{\alpha/2}}{n-1}$

OR

(2) σ^2 is between $\frac{s_i^2(n-1)}{\chi^2_{\alpha/2}}$ and $\frac{s_i^2(n-1)}{\chi^2_{1-\alpha/2}}$.

Hence, we are $(1-\alpha)\cdot 100\%$ confident that σ^2 is between $\frac{s_i^2\chi^2_{1-\alpha/2}}{n-1}$ and $\frac{s_i^2\chi^2_{\alpha/2}}{n-1}$.

Unlike the previous procedures, the One Variance Chi-Square Procedure is based on a distribution that is not symmetric. As a result, the confidence interval is not symmetric. While the sample variance, s_i^2, is still an unbiased estimate of the population variance it is not the center of the interval. From the practical point of view, this means that the interval can be found by computing its lower and upper bounds directly with the following formulas:

$$LB = \frac{s_i^2(n-1)}{\chi^2_{\alpha/2}} \text{ and } RB = \frac{s_i^2(n-1)}{\chi^2_{1-\alpha/2}}$$

Note that we can always turn a confidence interval for the population variance into a confidence interval for the population standard deviation by taking the square root of the boundaries!

$$LB = \sqrt{\frac{s_i^2(n-1)}{\chi^2_{\alpha/2}}} \text{ and } RB = \sqrt{\frac{s_i^2(n-1)}{\chi^2_{1-\alpha/2}}}$$

As a result, the following template will be used to construct and interpret the confidence intervals:

Table 1.10.4.1.1: Template of the procedure to construct a confidence interval for an unknown population variance.

Unknown Parameter	σ^2, population variance
Point Estimate	s_i^2, the sample variance
Confidence Level (of the point estimate)	0%
Confidence Level (for a new interval estimate)	$(1-\alpha)100\%$ $\alpha/2$ $1-\alpha/2$
Critical Value(s)	$\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ for $df = n-1$
Margin of Error	n/a
Lower Bound	$LB = \dfrac{s_i^2(n-1)}{\chi^2_{\alpha/2}}$
Upper Bound	$RB = \dfrac{s_i^2(n-1)}{\chi^2_{1-\alpha/2}}$
Interpretation	We are $(1-\alpha)100\%$ confident that the unknown population variance σ^2 is between LB and UB.

For the procedure to work the following assumptions must be made:

- The sample is simple random.
- The population has a normal distribution.

Consider the following example.

Example 1.10.4.1



For a certain drug, based on standards set by the United States Pharmacopeia (USP) - an official public standards-setting authority for all prescription and over-the-counter medicines and other health care products manufactured or sold in the United States, a standard deviation of capsule weights of less than 2 mg is acceptable. A sample of 10 capsules was taken and the weights are provided below:

99.8, 100.6, 101.7, 100.8, 99.5, 98.9, 101.7, 100.8, 99.8, 98.2

Construct and interpret a 90% confidence interval for the standard deviation of the weights of all capsules. You may assume that the weights of capsules are normally distributed.

Solution

First, let's check if all necessary assumptions are satisfied:

- The sample is simple random.
- The population has a normal distribution.

Also, note that the sample mean and variance are respectively:

$$\bar{x}_i = 100.18$$

$$s_i^2 = 1.32$$

Once the assumptions are verified, we may apply the procedure by using the template.

Table 1.10.4.1.2: The summary of the procedure to construct a 90% confidence interval for the population variance of capsule weights.

	Example	Template
Unknown Parameter	σ^2, population variance of capsule weights	σ^2, population variance
Point Estimate	$s_i^2 = 1.32$, the sample variance of a sample of size $n = 10$	s_i^2, the sample variance
Confidence Level (of the point estimate)	0%	0%

	Example	Template
Confidence Level (for a new interval estimate)	$CL = 90\%$ $\alpha/2 = 0.05$ $1-\alpha/2 = 0.95$	$(1-\alpha)100\%$ $\alpha/2$ $1-\alpha/2$
Critical Value(s)	When $df = 9$: $\chi^2_{0.95} = 3.325$ and $\chi^2_{0.05} = 16.919$	$\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ for $df = n-1$
Margin of Error	n/a	n/a
Lower Bound	$LB = \dfrac{1.32 \cdot 9}{16.919} = 0.70$	$LB = \dfrac{s_i^2(n-1)}{\chi^2_{\alpha/2}}$
Upper Bound	$RB = \dfrac{1.32 \cdot 9}{3.325} = 3.57$	$RB = \dfrac{s_i^2(n-1)}{\chi^2_{1-\alpha/2}}$
Interpretation	We are 90% confident that the variance of capsule weights is between 0.70 and 3.57.	We are $(1-\alpha)100\%$ confident that the unknown population variance σ^2 is between LB and UB.

Now we can interpret the results: We are 90% confident that the variance of capsule weights is between 0.70 and 3.57.

Remember that we can always convert a confidence interval for the population variance to a confidence interval for the population standard deviation by taking the square root of the boundaries!

In our example, we are 90% confident that the weight variance of capsules is between 0.70 and 3.57, therefore we are 90% confident that the weight standard deviation of capsules is between 0.83 and 1.89 mg - well within the standard 2 mg.

This page titled 1.10.4: The One Variance Chi-Squared Procedure is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.10.5: The Summary of Confidence Intervals



We learned altogether 4 procedures for constructing confidence intervals to estimate 3 different parameters.

So how do we know when and which procedure to use?

First, we must determine which parameter is being estimated by looking for the key words such as

- Mean/average when estimating the population mean.
- Proportion/percentage when estimating the population proportion.
- Variance/standard deviation when estimating the population variance.

Next, we determine the procedure in the following way:

We always use

- The χ^2 procedure to estimate variances.
- The Z procedure to estimate proportions.

To estimate a sample mean, we need to check whether the population standard deviation is known

- If it is known, then we use One Mean Z Procedure.
- If it is not known, then we use One Mean T Procedure.

Table 1.10.5.1.1: Summary of confidence interval procedures in regard to purposes and assumptions.

Procedure	One mean Z-interval	One mean T-interval	One proportion Z-interval	One variance χ^2-interval
Purpose	To find a confidence interval for a population mean, μ	To find a confidence interval for a population mean, μ	To find a confidence interval for a population proportion, p	To find a confidence interval for a population variance, σ^2
Assumptions	• Simple random sample • Normal population or large sample • σ is known	• Simple random sample • Normal population or large sample • σ is unknown	• Simple random sample • np and $n(1-p)$ are both 10 or greater.	• Simple random sample • Normal population

Remember - before applying the procedure we have to verify all the assumptions!

So before starting the procedure we must have already figured out the unknown parameter along with its point estimate.

Table 1.10.5.1.2: Summary of confidence interval procedures in regard to unknown parameters and point estimates.

Procedure	One mean Z-interval	One mean T-interval	One proportion Z-interval	One variance χ^2-interval
Unknown Parameter	population mean, μ	population mean, μ	population proportion,p	population variance, σ^2
Point Estimate	sample mean, $\bar{x}_i$	sample mean, $\bar{x}_i$	sample proportion, $\hat{p}_i$	sample variance, s_i^2

How confident are we that the point estimate equals exactly the unknown parameter? Zero percent confident. In other words, there are zero chances that the point estimate equals the parameter.

Next, we are going to construct an interval estimate so we need a new confidence level. Usually, it is given in the problem, if not then 95% is the most common confidence level in applications. If you want to be more confident make it 99%, if you want to be less confident then make it 90%. Whatever confidence level you choose, the value α is such that $(1-\alpha)100\%$ is equal to the confidence level.

Table 1.10.5.1.3: Summary of confidence interval procedures in regard to identifying confidence level and alpha.

Procedure	One mean Z-interval	One mean T-interval	One proportion Z-interval	One variance χ^2-interval
Confidence level	95% is the most common confidence level in applications If you want to be more confident then make it 99%, if you want to be less confident then make it 90%.			
α	$CL = (1-\alpha)100\%$			

Next, we determine the critical values. Depending on the procedure you may have to do it in several steps. In Z procedures, we only need $\alpha/2$, in T procedures, we also need the degrees of freedom, and in χ^2 procedures, we need both, $\alpha/2$ and $1-\alpha/2$ and the degrees of freedom.

Table 1.10.5.1.4: Summary of confidence interval procedures in regard to identifying the critical value(s).

Procedure	One mean Z-interval	One mean T-interval	One proportion Z-interval	One variance χ^2-interval
$\alpha/2$ or $1-\alpha/2$	$\alpha/2$	$\alpha/2$	$\alpha/2$	$\alpha/2$, $1-\alpha/2$
Critical Value(s)	$\pm z_{\alpha/2}$	$\pm t_{\alpha/2}$ $df = n-1$	$\pm z_{\alpha/2}$	$\chi^2_{1-\alpha/2}$, $\chi^2_{\alpha/2}$ $df = n-1$

Next, we construct the confidence interval by figuring out the lower and upper bounds. For one mean and one proportion procedures we do that by finding the margin of error first and then adding it to and subtracting it from the point estimate. For the one variance procedure, we compute the lower and upper bounds directly using the formulas due to the fact that the interval is not symmetric around the point estimate and therefore there is no margin of error.

Table 1.10.5.1.5: Summary of confidence interval procedures in regard to identifying the lower and upper bounds and margins of error when applicable.

Procedure	One mean Z-interval	One mean T-interval	One proportion Z-interval	One variance χ^2-interval
Margin of Error	$z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$	$t_{\alpha/2}\dfrac{s}{\sqrt{n}}$	$z_{\alpha/2}\sqrt{\dfrac{p(1-p)}{n}}$	n/a
Lower Bound	$\bar{x}_i - ME$	$\bar{x}_i - ME$	$\hat{p}_i - ME$	$\dfrac{s_i^2(n-1)}{\chi^2_{\alpha/2}}$
Upper Bound	$\bar{x}_i + ME$	$\bar{x}_i + ME$	$\hat{p}_i + ME$	$\dfrac{s_i^2(n-1)}{\chi^2_{1-\alpha/2}}$

Finally, we interpret the interval by writing or saying the following:

“We are (1-alpha)100% confident that the population proportion is between the LB and UB.”

We always include the units in the answer. Also, it is not uncommon to convert the proportion to percentages, and variance to standard deviation by taking the square root of the bounds.

For One Mean and One Proportion Z Procedures it is also possible to derive the formula that relates the size of the sample to the margin of error, so that we can determine the sample size necessary to achieve a certain level of accuracy.

Table 1.10.5.1.6: Summary of confidence interval procedures in regard to identifying the sample size to achieve the desired margin of error when applicable.

Procedure	One mean Z-interval	One mean T-interval	One proportion Z-interval	One variance χ^2-interval
Desired margin of error	E_0	n/a	E_0	n/a
The sample size	$n > \left(z_{\alpha/2} \cdot \frac{\sigma}{E_0}\right)^2$	n/a	$n > \left(\frac{z_{\alpha/2}}{2E_0}\right)^2$	n/a

All four procedures can be put in one table in the following way:

Table 1.10.5.1.7: Summary of all confidence interval procedures.

Procedure	One mean Z-interval	One mean T-interval	One proportion Z-interval	One variance χ^2-interval
Purpose	To find a confidence interval for a population mean, μ	To find a confidence interval for a population mean, μ	To find a confidence interval for a population proportion, p	To find a confidence interval for a population variance, σ^2
Assumptions	• Simple random sample • Normal population or large sample • σ is known	• Simple random sample • Normal population or large sample • σ is unknown	• Simple random sample • np and $n(1-p)$ are both 10 or greater.	• Simple random sample • Normal population
Point estimate	sample mean, $\bar{x}_i$	sample mean, $\bar{x}_i$	sample proportion, $\hat{p}_i$	sample variance, s_i^2
Confidence Level $(1-\alpha)100\%$	$\alpha/2$	$\alpha/2$	$\alpha/2$	$\alpha/2, 1-\alpha/2$
Critical value(s)	$\pm z_{\alpha/2}$	$\pm t_{\alpha/2}$ $df = n-1$	$\pm z_{\alpha/2}$	$\chi^2_{1-\alpha/2}, \chi^2_{\alpha/2}$ $df = n-1$
Margin of Error	$z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$t_{\alpha/2}\frac{s}{\sqrt{n}}$	$z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$	n/a
Lower Bound	$\bar{x}_i - ME$	$\bar{x}_i - ME$	$\hat{p}_i - ME$	$\frac{s_i^2(n-1)}{\chi^2_{\alpha/2}}$
Upper Bound	$\bar{x}_i + ME$	$\bar{x}_i + ME$	$\hat{p}_i + ME$	$\frac{s_i^2(n-1)}{\chi^2_{1-\alpha/2}}$
Interpretation	We are $(1-\alpha)100\%$ confident that the unknown population parameter is between LB [units] and UB [units].			

This page titled 1.10.5: The Summary of Confidence Intervals is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

SECTION OVERVIEW

1.11: Hypothesis Testing

1.11.1: One Mean Z Test

1.11.2: One Mean T Test

1.11.3: One Proportion Z Test

1.11.4: One Variance Chi-Squared Test

1.11.5: The Summary of Hypothesis Testing for One Parameter

This page titled 1.11: Hypothesis Testing is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.11.1: One Mean Z Test

Section 0.1: Introduction to Hypothesis Testing - Fundamentals



Let’s start with a definition of a hypothesis.

Definition: Hypothesis

A **hypothesis** is a testable claim that is either true or false.

For example, someone may claim that Earth is round. To test a hypothesis means to verify its truthfulness in a scientific way that is universally acceptable by both, the opponents and proponents of the hypothesis. How would we test whether Earth is round or not? We may send a sailing expedition that travels only west, and we would agree that if the sailing expedition comes all the way around then Earth is round and if the sailing expedition falls of the edge then Earth is flat.

By the definition, the hypothesis can be either true or false. Testing the hypothesis may also result in two outcomes – the test will either confirm or reject the hypothesis. In case when the expedition sails all the way around - the true hypothesis will be confirmed. In case when the expedition falls off the edge - the false hypothesis will be rejected. If either one of these scenarios was to happen then the testing procedure would serve its true purpose of figuring out the truth. But is it possible that something goes wrong? What if Earth is round but the boat falls off Niagara Falls? Then we would believe that it fell of the edge of Earth, and for a wrong reason we would have to reject the hypothesis that is true. On the other hand, what if Earth is flat, and the expedition only sailed around the continent due to a compass malfunction? Then we would believe that it sailed around Earth, and for a wrong reason we would have to confirm the hypothesis that is false. If either one of these scenarios was to happen then the testing procedure would return the wrong result.

Earth is round	Confirmed	Rejected
True	Sailed around Earth – CORRECT	Fell off Niagara Falls – ERROR!
False	Sailed around a continent – ERROR!	Fell off the edge of Earth - CORRECT

No matter how perfect the testing procedure is – there is always a chance of making an error.

Definition: Type 1 and 2 errors

Rejecting a true hypothesis is called **type 1 error**. Confirming or failing to reject a false hypothesis is called **type 2 error**.

The following table summarizes all possible outcomes of the hypothesis testing:

A hypothesis	Confirmed	Rejected
True	Correct Inference	Error Type 1
False	Error Type 2	Correct Inference

The difference between the two types of errors is clear and depends on the context. Consider the following example. When a suspect is brought to a trial, they are either guilty or not, and there are 2 outcomes of the trial – the suspect is either convicted or cleared. You can clearly see the difference between the two types of errors as it is not the same to clear the guilty and convict the innocent.

A suspect is …	Convicted	Cleared
Guilty	Correct Inference	Error Type 1
Innocent	Error Type 2	Correct Inference

At different times different justice systems may prioritize one type of error over the other one.

In the testing procedure that we will develop, we will set a maximum threshold for type 1 errors and we will call it alpha – **the significance level**.

In this course, we will learn how to make and test statistical claims.

Definition: Statistical claim

A **statistical claim** is a statement about a population parameter in the form of an equation or an inequality.

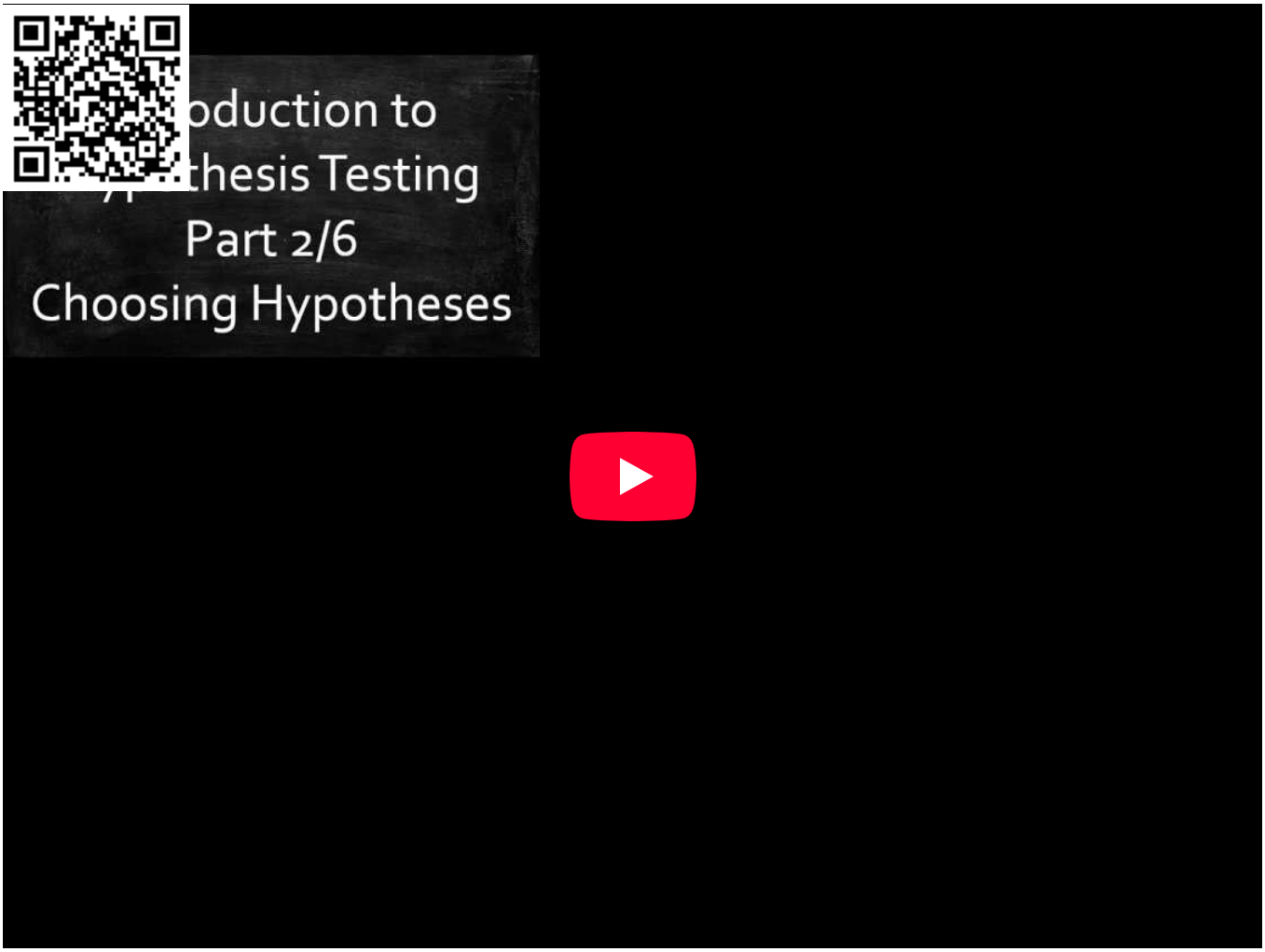
Example 1.11.1.0.1.1

For example,

- "The average height of students is less than 170 cm" is a statistical claim about the population mean, and it can be expressed symbolically as $\mu < 170$.
- "The proportion of females on campus is more than 60%" is a statistical claim about the population proportion, and it can be expressed symbolically as $p > 0.60$.
- "The students' GPA variance is 1.21" is a statistical claim about the population variance, and it can be expressed symbolically as $\sigma^2 = 1.21$.
- "The company's average salary is different from the state's average of $67,565" is a statistical claim about the population mean, and it can be expressed symbolically as $\mu \neq 67565$.

Next, we will develop a formal procedure to test statistical claims.

Section 0.2: Introduction to Hypothesis Testing - Choosing Hypotheses



Due to the nature of the relations that we discovered in statistics we will always be testing a hypothesis in the context of another hypothesis – one in the form of an equation and the other in the form of an inequality.

Definition: Null hypothesis

A **null hypothesis** is a to be tested statistical claim in the form of an equation. It is usually labeled as H_0.

Definition: Alternative hypothesis

An **alternative hypothesis** is a statistical claim in the form of an inequality to be tested as an alternative to the null hypothesis. It is usually labeled as H_a or H_1.

A testing procedure can be classified as **left-tail**, **two-tail**, or **right-tail** depending on the inequality used in the alternative hypothesis:

$<$	$\neq$	$>$
LT	TT	RT

Example 1.11.1.0.2.1

For example,

- "The average height of students is less than 170 cm" can be expressed symbolically as $\mu < 170$. Since the claim is in the form of an inequality, therefore it must be set as an alternative hypothesis, therefore the null hypothesis is $\mu = 170$ and the test is left-tail, that is:

$H_0 : \mu = 170$ $H_a : \mu < 170$	LT

- "The proportion of females on campus is more than 60%" can be expressed symbolically as $p > 0.60$. Since the claim is in the form of an inequality, therefore it must be set as an alternative hypothesis, therefore the null hypothesis is $p = 0.60$ and the test is right-tail, that is:

$H_0 : p = 0.60$ $H_a : p > 0.60$	RT

- "The students' GPA variance is 1.21" can be expressed symbolically as $\sigma^2 = 1.21$. Since the claim is in the form of an equation, therefore it must be set as a null hypothesis, therefore the alternative hypothesis is $\sigma^2 \neq 1.21$ and the test is two-tail, that is:

$H_0 : \sigma^2 = 1.21$ $H_0 : \sigma^2 \neq 1.21$	TT

- "The company's average salary is different from the state's average of \$67,565" can be expressed symbolically as $\mu \neq 67565$. Since the claim is in the form of an inequality, therefore it must be set as an alternative hypothesis, therefore the null hypothesis is $\mu \neq 67565$ and the test is two-tail, that is

$H_0 : \mu = 67565$ $H_a : \mu \neq 67565$	TT

Although right now naming the procedure as left-tail, right-tail, or two-tail may not make sense, soon, we will be using this information as a part of the testing procedure.

The idea is simple – the inequality defines a spectrum which we will use to evaluate the strength of the evidence. For example, in a right-tail test only values on the right side of the spectrum will be considered as a strong evidence. The further to the right the stronger the evidence.

Section 0.3: Introduction to Hypothesis Testing - Test Statistic



Next, we will discuss the concept of a test statistic and what constitutes evidence against the null hypothesis.

In the statistical hypothesis testing:

1. We will treat the null hypothesis as the assumed status quo.
2. The goal of the testing procedure is to evaluate the strength of any evidence against the null hypothesis.
3. We will not reject the null hypothesis unless the evidence strongly suggests the alternative hypothesis.

In other words, we are not going to change the status quo unless the evidence is beyond a reasonable doubt. One may think of it as the presumption of the status quo principle as an equivalent of the presumption of innocence principle in a justice system. Next, we need to clarify what constitutes evidence and what evidence is beyond a reasonable doubt.

What is considered evidence against the status quo? Whether we have evidence or not depends on the inequality in the alternative hypothesis and a particular observation of a sample statistic. Consider the following example in which the null hypothesis assumes that the height is 170 cm. We are going to evaluate whether a particular observation is evidence or not for different alternative hypotheses:

- If the average height of a sample is less than 170 then it is interpreted as evidence against the null hypothesis in LT, TT but not the RT test.
- If the average height of a sample is exactly 170 then it can't be interpreted as evidence against the null hypothesis as it literally confirms it.
- If the average height of a sample is greater than 170 then it is interpreted as evidence against the null hypothesis in TT, RT but not the LT test.

The following table summarizes whether a certain observation may be interpreted as evidence in favor of the alternative hypothesis:

Observation	$H_0 : \mu < 170$	$H_0 : \mu \neq 170$	$H_0 : \mu > 170$
< 170	yes	yes	no
$= 170$	no	no	no
> 170	no	yes	Yes

Definition: Significance level

A **significance level** is the maximum tolerated probability of type 1.

Usually, the significance level is determined from the context of the problem. The most common is 5% but it can be brought up to 10% or down to 1%. The smaller the significance level the stricter the test is. What considered a strong evidence under 5% significance level may not be such at 1% significance level.

Definition: Test statistic

A **test statistic** is a value that measures how far a particular observation is away from what it is expected to be according to the null hypothesis.

The way we compute the test statistic will vary for different procedures. When Z-distribution is involved the test statistic is simply the z-score of the sample statistic. When other distributions are involved, such as T, χ^2, or F it will be analogous to the z-score but it will not be called so. Note that each sample has a test statistic value that measures the strength of it as evidence. So how exactly a test statistic can help us to determine whether we have evidence or not?

The observations that are smaller than expected will have smaller test statistic value than the observations that are larger than expected. So, a small test statistic is evidence that the population parameter is smaller than assumed in the null hypothesis. Similarly, a large test statistic is evidence that the population parameter is larger than assumed in the null hypothesis.

For example, in case when a z-score is used as the test statistic, we know that the average is zero, so negative test statistic means "small" and positive test statistic means "large". The intuition is similar when a student T-distribution is involved. But for other distributions, such as χ^2 and F, the definitions of "small" and "large" are not so intuitive.

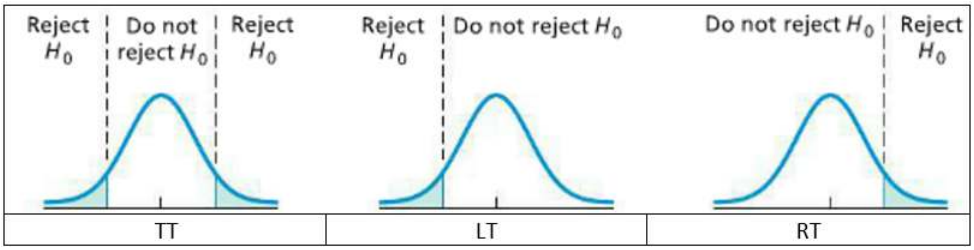
Now that we know what observations constitute evidence against the null hypothesis, we need to decide what a strong evidence is. In other words, at which point the test statistic is far enough from expected for us to reject the null hypothesis. This is when **the significance level** would come to play.

Section 0.4: Introduction to Hypothesis Testing - Critical Value Approach (CVA)



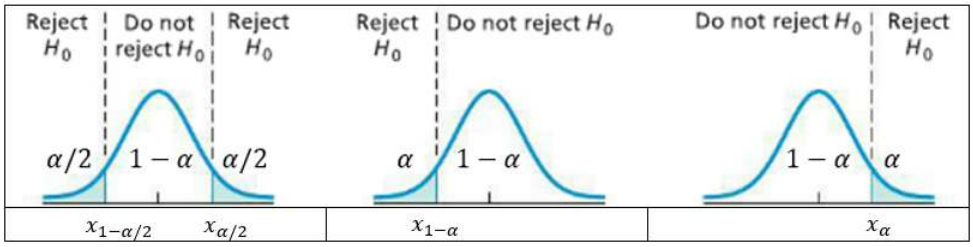
In the **critical value approach** (**CVA**), "too far" is defined by the range of values called the "**rejection region**" (RR). If a statistic falls into the RR, then it is interpreted as a strong evidence against the null hypothesis in favor of the considered alternative.

1. Draw the region according to the type of the test. For demonstration purposes we use a generic probability density curve of some distribution X. But in an actual application it can be Z, T, X^2, or F depending on the procedure.



It should be easy to see the association between the type of the test and the way the rejection region is drawn.

2. The area of the region must be equal to α, the significance level. By having the area of the rejection region equal to α we guarantee that the probability of type 1 error is not greater than a preset significance level.

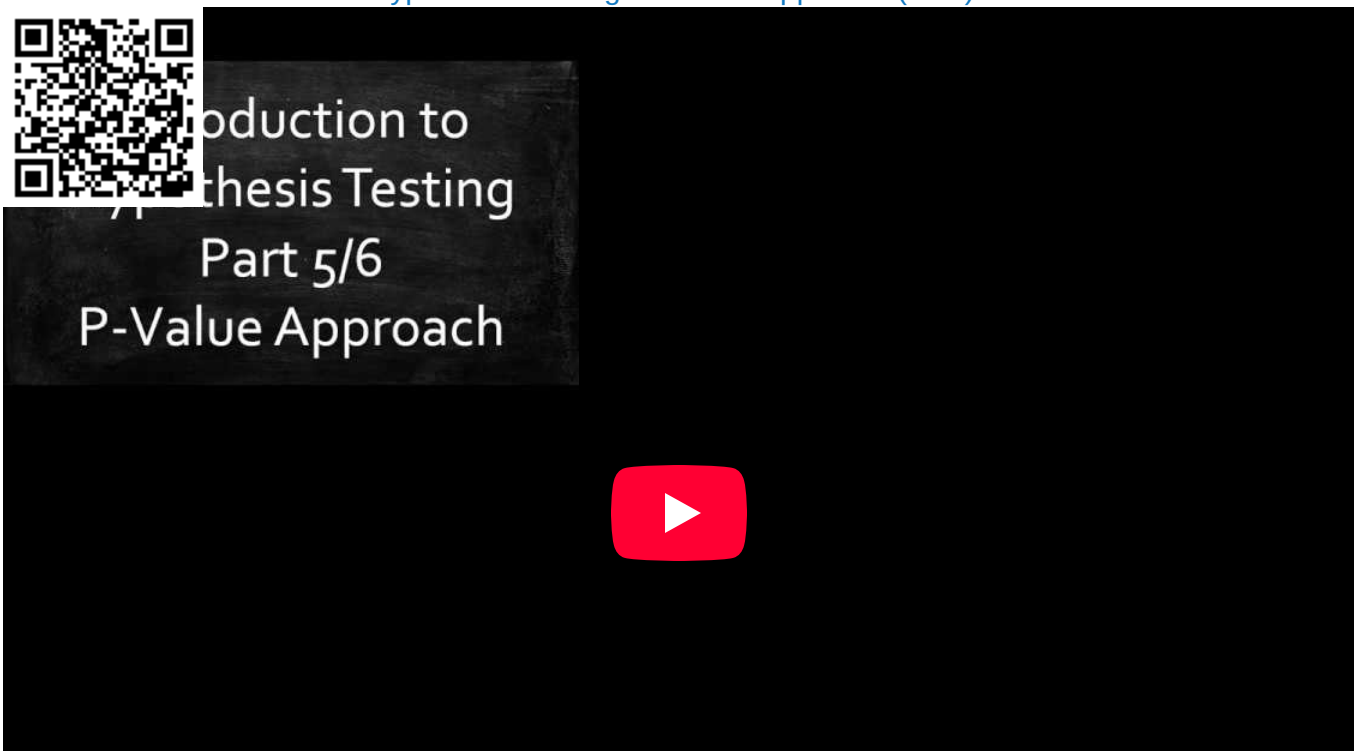


Geometrically the rejection region is defined by its boundaries which we call the **critical values**:

- for a right-tail test, the rejection region is the right tail with the total area α, so it is defined by only the right critical value which can be expressed as x_α
- for a left-tail test, the rejection region is the left tail with the total area α, so it is defined by only the left critical value which can be expressed as $x_{1-\alpha}$
- for a two-tail test, the rejection region with the total area α is split evenly between the left and the right tail, so there are two critical values, the left and the right, $x_{1-\alpha/2}$ and $x_{\alpha/2}$

In the Critical Value Approach, the decision whether to reject the null hypothesis or not is made based on whether the test statistics is in the rejection region or not.

Section 0.5: Introduction to Hypothesis Testing - P-Value Approach (PVA)

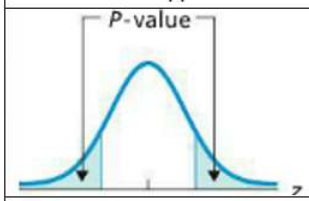
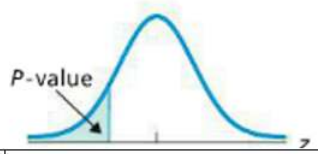
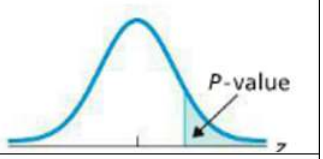




In the **p-value approach** (PVA), "too far" is defined by the probability of a statistic to be at least as far from the expected value as the current observation. We called such probability the **p-value** of the test statistic.

The region the area of which is called p-value is determined by the type of the test and the value of the test statistic x_0 as shown below. Symbolically the p-value can be expressed in the following way:

- In the right-tail test it is the area to the right of the test statistic.
- In the left-tail test it is the area to the left of the test statistic.
- In the two-tail test it is the area to the right or the left of whichever one is smaller multiplied by 2.

TT	LT	RT
		
x_0 or x_0	x_0	x_0
$2P(X < x_0)$ or $2P(X > x_0)$	$P(X < x_0)$	$P(X > x_0)$

Note that X stands for some generic distribution. But in a real application it will be Z, T, χ^2, or F.

In the P-Value Approach, the decision whether to reject the null hypothesis or not is made based on the relation between the p-value and α. If the p-value of the test statistic is less than α, then we consider the test statistic to be a strong evidence against the null hypothesis in favor of the alternative.

In summary, the p-value is a measure of how likely we were to observe such or stronger evidence assuming that the null hypothesis is true. When the p-value is less than a preset threshold we consider our observation a strong evidence against the null hypothesis.

Section 0.6: Introduction to Hypothesis Testing - Conclusion



Let's summarize all the steps:

In step 1, we decide on the hypotheses that need to be tested and the type of the test.

In step 2, we decide on the significance level alpha.

In step 3, we compute the test statistic using the formula that is specific for each procedure.

In step 4, we perform the testing using either the critical value approach or the p-value approach.

In step 5, we will draw the conclusion to the testing procedure and decide whether to reject the null hypothesis or not. Note that the result is the same regardless of the approach.

In step 6, we interpret the results.

Due to the nature of the relations that we discovered in statistics the application of the test is **never** to confirm the null hypothesis but to evaluate the strength of the evidence against it. In case when there is enough evidence against it, the null hypothesis will be rejected in favor of the alternative hypothesis. However, the lack of any evidence against the null hypothesis does not imply that the null hypothesis is true. In other words, if you fail to prove something is false it does not necessarily make it true. For example, when there is not enough evidence to convict a suspect, it does not necessarily imply that they are innocent. As a result, there are only two possible logical conclusions to the hypothesis testing procedure - either **reject** the null hypothesis in favor of the alternative hypothesis or **do not to reject** the null hypothesis due to the lack of evidence. Again, the testing procedure is not intended to prove or verify the null hypothesis.

We summarize the procedure by creating the following template:

Step 1	The null hypothesis The alternative hypothesis
Step 2	The significance level
Step 3	The test statistic
Step 4	Testing
Step 5	Conclusion
Step 6	Interpretation: Under $\alpha \cdot 100\%$ significance level we DO or DO NOT have sufficient evidence to suggest the alternative hypothesis.

Now we have completed the development of the hypothesis testing procedure that can be used to test a variety of statistical claims.

Section 1: Hypothesis Test - One Mean Z Procedure



Next, we will learn how to apply the One Mean Z Procedure to test a statistical claim about a population mean. Consider the following example.

An incubation period is a time between when you contract a virus and when your symptoms start. Assume that the population of incubation periods for a novel coronavirus is normally distributed with the population standard deviation of 2 days. By surveying randomly selected local hospitals, a researcher was able to obtain the following sample of incubation periods of 10 patients:

$$6, 2, 4, 3, 3, 7, 5, 8, 5, 7$$

At 5% significance level, test the claim that the average incubation period of the novel coronavirus is less than 6 days.

Note that the average of the sample is $\bar{x}_i = 5$ and the population standard deviation is given $\sigma = 2$.

Now, let's identify the statistical claim that needs to be tested:

"the average incubation period of the novel coronavirus is less than 6 days"

The key word "average" suggests that the claim is about the parameter μ, so we can symbolically express the claim as

$$\mu < 6$$

Since the claim is about the population mean and the standard deviation is known we will use the one mean Z procedure. Let's check if all necessary assumptions are satisfied:

- The sample is simple random
- The population is normally distributed
- The population standard deviation, σ, is known and $\sigma = 2$

We will use the following template to perform the hypothesis testing:

In step 1, we will set up the hypothesis:

Since the claim is in the form of an inequality, therefore it must be set as an alternative hypothesis, therefore the null hypothesis is $\mu = 6$ and the test is left-tail, that is:

$H_0 : \mu = 6$ $H_a : \mu < 6$	LT

In step 2, we will identify the significance level:

The significance level can always be found in the text of the problem. In our case it is 5%, thus:

$$\alpha = 0.05$$

In step 3, we will find the test statistic using the formula:

$$z_0 = \frac{\bar{x}_i - \mu_0}{\sigma/\sqrt{n}} = \frac{5-6}{2/\sqrt{10}} = -1.58$$

In step 4, we will perform either the critical value approach or p-value approach to test the claim:

- In critical value approach, we construct the rejection region:

$$\text{RR: less than } -z_{0.05} = -1.645$$

- In p-value approach, we compute the p-value:

$$\text{P-Value: } P(Z < -1.58) = 0.057$$

In step 5, we will draw the conclusion:

- In the critical value approach, we must check whether the test statistic is in the rejection region or not. Our test statistic is -1.58 and it is to the right of the critical value -1.645, thus it is not in the rejection region.
- In the p-value approach, we must check whether the p-value is less than the significance level or not. Our p-value is 0.057 and it is greater than α.

Both tests suggest that we DO NOT reject the null hypothesis in favor of the alternative.

In step 6, we will interpret the results:

Under 5% significance level we DO NOT have sufficient evidence to suggest that the mean incubation period is less than 6 days.

We discussed how to apply the One Mean Z Procedure to test a statistical claim about a population mean when the population standard deviation is given.

This page titled 1.11.1: One Mean Z Test is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.11.2: One Mean T Test



Next, we will learn how to apply the One Mean T Procedure to test a statistical claim about a population mean. Consider the following example.

An incubation period is a time between when you contract a virus and when your symptoms start. Assume that the population of incubation periods for a novel coronavirus is normally distributed. By surveying randomly selected local hospitals, a researcher was able to obtain the following sample of incubation periods of 10 patients:

$$6, 2, 4, 3, 3, 7, 5, 8, 5, 7$$

At 5% significance level, test the claim that the average incubation period of the novel coronavirus is less than 6 days.

Note that the average of the sample is $\bar{x}_i = 5$ and since the population standard deviation is not given, we compute the sample standard deviation instead $s = 2$.

Now, let's identify the statistical claim that needs to be tested:

"the average incubation period of the novel coronavirus is less than 6 days"

The key word "average" suggests that the claim is about the parameter μ, so we can symbolically express the claim as

$$\mu < 6$$

Since the claim is about the population mean and the standard deviation is unknown, we will use the one mean T procedure. Let's check if all necessary assumptions are satisfied:

- The sample is simple random
- The population is normally distributed
- The population standard deviation, σ, is unknown but $s = 2$

We will use the following template to perform the hypothesis testing:

In step 1, we will set up the hypothesis:

Since the claim is in the form of an inequality, therefore it must be set as an alternative hypothesis, therefore the null hypothesis is $\mu = 6$ and the test is left-tail, that is:

$H_0 : \mu = 6$ $H_a : \mu < 6$	LT

In step 2, we will identify the significance level:

The significance level can always be found in the text of the problem. In our case it is 5%, thus:

$$\alpha = 0.05$$

In step 3, we will find the test statistic using the formula:

$$t_0 = \frac{\bar{x}_i - \mu_0}{s/\sqrt{n}} = \frac{5-6}{2/\sqrt{10}} = -1.58$$

In step 4, we will perform either the critical value approach or p-value approach to test the claim:

- In critical value approach, we construct the rejection region using the T-curve with $df = n - 1 = 10 - 1 = 9$:

$$\text{RR: less than } -t_{0.05} = -1.833$$

- In p-value approach, we compute the p-value using the T-curve with $df = n - 1 = 10 - 1 = 9$:

$$\text{P-Value: } P(T < -1.58) = 0.074$$

In step 5, we will draw the conclusion:

- In the critical value approach, we must check whether the test statistic is in the rejection region or not. Our test statistic is -1.58 and it is to the right of the critical value -1.833, thus it is not in the rejection region.
- In the p-value approach, we must check whether the p-value is less than the significance level or not. Our p-value is 0.074 and it is greater than α.

Both tests suggest that we DO NOT reject the null hypothesis in favor of the alternative.

In step 6, we will interpret the results:

Under 5% significance level we DO NOT have sufficient evidence to suggest that the mean incubation period is less than 6 days.

We discussed how to apply the One Mean T Procedure to test a statistical claim about a population mean when the population standard deviation is not given.

This page titled 1.11.2: One Mean T Test is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.11.3: One Proportion Z Test



Next, we will learn how to apply the One Proportion Z Procedure to test a statistical claim about a population proportion. Consider the following example.

The ABC News/Ipsos poll was conducted by Ipsos Public Affairs‘ Knowledge Panel® on March 18-19, 2020, in English and Spanish, among a random national sample of 512 adults. In the poll, 282 Americans approve of the president's management of the crisis. At 10% significance level, test the claim that the majority of the Americans approve the president's management of the crisis.

Note that the sample proportion is $\hat{p}_i = \frac{282}{512} = 0.5508$ or 55.08%.

Now, let’s identify the statistical claim that needs to be tested:

“the majority of the Americans approve”

It is not obvious, but the claim is about the population proportion being greater than 50%, so we can symbolically express the claim as

$$p > 0.50$$

Since the claim is about the population proportion, we will use the One Proportion Z Procedure. Let’s check if all necessary assumptions are satisfied:

- The sample is simple random
- The number of positive and negative responses are both greater than 10.

We will use the following template to perform the hypothesis testing:

In step 1, we will set up the hypothesis:

Since the claim is in the form of an inequality, therefore it must be set as an alternative hypothesis, therefore the null hypothesis is $p = 0.5$ and the test is right-tail, that is:

$H_0 : p = 0.5$ $H_a : p > 0.5$	RT

In step 2, we will identify the significance level:

The significance level can always be found in the text of the problem. In our case it is 10, thus:

$$\alpha = 0.10$$

In step 3, we will find the test statistic using the formula:

$$z_0 = \frac{\hat{p}_i - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0.5508 - 0.5}{\sqrt{0.5(1-0.5)/512}} = 2.30$$

In step 4, we will perform either the critical value approach or p-value approach to test the claim:

- In critical value approach, we construct the rejection region:

$$\text{RR: greater than } z_{0.1} = 1.28$$

- In p-value approach, we compute the p-value:

$$\text{P-Value: } P(Z > 2.30) = 0.011$$

In step 5, we will draw the conclusion:

- In the critical value approach, we must check whether the test statistic is in the rejection region or not. Our test statistic is 2.30 and it is to the right of the critical value 1.28, thus it is in the rejection region.
- In the p-value approach, we must check whether the p-value is less than the significance level or not. Our p-value is 0.011 and it is less than α.

Both tests suggest that we DO reject the null hypothesis in favor of the alternative.

In step 6, we will interpret the results:

Under 10% significance level we have sufficient evidence to suggest that the majority of the Americans approve the president's management of the crisis.

We discussed how to apply the One Proportion Z Procedure to test a statistical claim about a population proportion.

This page titled 1.11.3: One Proportion Z Test is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.11.4: One Variance Chi-Squared Test



Next, we will learn how to apply the One Variance X^2 Procedure to test a statistical claim about a population variance. Consider the following example.

For a certain drug, based on standards set by the United States Pharmacopeia (USP) - an official public standards-setting authority for all prescription and over-the-counter medicines and other health care products manufactured or sold in the United States, a standard deviation of capsule weights of less than 2 mg is acceptable. A sample of 10 capsules was taken and the weights are provided below:

99.8, 100.6, 101.7, 100.8, 99.5, 98.9, 101.7, 100.8, 99.8, 98.2

At 1% significance level, test the claim that the standard deviation is 0.7.

Note that the sample variance is $s_i^2 = 1.32$.

Now, let's identify the statistical claim that needs to be tested:

"the standard deviation is 0.7"

The key words "standard deviation" suggest that the claim is about the parameter σ, so we can symbolically express the claim as

$$\sigma = 0.7$$

which is equivalent to the claim that the variance is 0.49 that is

$$\sigma^2 = 0.7^2 = 0.49$$

Since the claim is about the population variance, we will use the One Variance X^2 Procedure. Let's check if all necessary assumptions are satisfied:

- The sample is simple random
- The population is normally distributed

We will use the following template to perform the hypothesis testing:

In step 1, we will set up the hypothesis:

Since the claim is in the form of an equation, therefore it must be set as a null hypothesis. The alternative hypothesis, which is always in the form of an inequality, must be carefully chosen depending on the context. Since there is no any indication of one-sided inequality, we set the alternative hypothesis as $\sigma^2 \neq 0.49$ and the test is two-tail, that is:

$H_0 : \sigma^2 = 0.49$ $H_a : \sigma^2 \neq 0.49$	TT

In step 2, we will identify the significance level:

The significance level can always be found in the text of the problem. In our case it is 1%, thus:

$$\alpha = 0.01$$

In step 3, we will find the test statistic using the formula:

$$\chi_0^2 = \frac{s_1^2(n-1)}{\sigma_0^2} = \frac{1.32 \cdot 9}{0.49} = 24.24$$

In step 4, we will perform either the critical value approach or p-value approach to test the claim:

- In critical value approach, we construct the rejection region using the X^2-curve with $df = n - 1 = 10 - 1 = 9$:

$$\text{RR: less than } \chi^2_{0.995} = 1.73$$
$$\text{or greater than } \chi^2_{0.005} = 23.59$$

- In p-value approach, we compute the p-value using the X^2-curve with $df = n - 1 = 10 - 1 = 9$:

$$\text{P-Value: } 2P(X^2 > 24.24) = 0.0078$$

In step 5, we will draw the conclusion:

- In the critical value approach, we must check whether the test statistic is in the rejection region or not. Our test statistic is 24.24 and it is to the right of the critical value 23.59, thus it is in the rejection region.
- In the p-value approach, we must check whether the p-value is less than the significance level or not. Our p-value is $0.0078$and it is less than α.

Both tests suggest that we DO reject the null hypothesis in favor of the alternative.

In step 6, we will interpret the results:

Under 1% significance level we DO have sufficient evidence to suggest that the variance is not 0.49, and therefore the standard deviation is not 0.7.

We discussed how to apply the One Variance X^2 Procedure to test a statistical claim about a population variance (and standard deviation).

This page titled 1.11.4: One Variance Chi-Squared Test is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.11.5: The Summary of Hypothesis Testing for One Parameter



We learned altogether 4 procedures for testing statistical claims about 3 different parameters. So how do we know when and which procedure to use? Next, we will summarize all the procedures for testing hypotheses for one parameter.

First, we must determine the statistical claim that is being tested in two steps:

1. Look for the key words to identify the parameter in the claim:
   - Mean/average when testing a claim about a population mean.
   - Proportion/percentage when testing a claim about a population proportion.
   - Variance/standard deviation when testing a claim about a population variance.
2. Look for the key words that identify the inequality in the claim:
   - Less than/at most for LT
   - Not equal/different for TT
   - Greater than/at least for RT

   The statistical claim in the form of an inequality is called the alternative hypothesis, and the statistical claim in the form of an equation is called the null hypothesis.

Next, we determine the procedure in the following way:

- Use the X^2 procedure for testing claims about variances.
- Use the Z procedure for testing claims about proportions.
- For testing claims about means, we need to check whether the population standard deviation is known:
  - If it is known, then we use the Z procedure.
  - If it is not known, then we use the T procedure.

Remember - before applying the procedure verify all the assumptions!

So before starting the procedure we must have already figured out the significance level α which is the maximum tolerated type 1 error. The most common values for alpha are 1, 5, and 10%.

Next, we compute the test statistic using a formula that is specific to each procedure. Note that we label the test statistic using the lower-case letter of the name of the involved distribution with the index zero. That is in a Z-test the test statistic is called z_0, in the T-test the test statistic is called t_0, and in the X^2-test the test statistic is called χ_0^2.

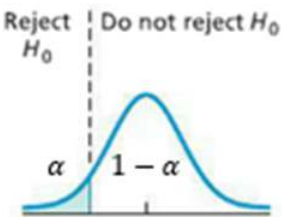
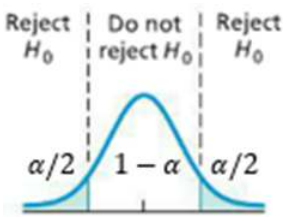
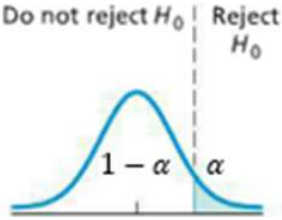
Hypothesis Test	One Mean Z Test	One Mean T Test	One Proportion Z Test	One Variance X^2 Test

Hypothesis Test	One Mean Z Test	One Mean T Test	One Proportion Z Test	One Variance X^2 Test
Test Statistic:	$z_0 = \frac{\bar{x}_i - \mu_0}{\sigma/\sqrt{n}}$	$t_0 = \frac{\bar{x}_i - \mu_0}{s_i/\sqrt{n}}$ with $df = n-1$	$z_0 = \frac{\hat{p}_i - p_0}{\sqrt{p_0(1-p_0)/n}}$	$\chi_0^2 = \frac{s_i^2(n-1)}{\sigma_0^2}$ with $df = n-1$

To apply the formula, one must be able to understand the notation. In this step, it may be necessary to reread the question several times to extract the information.

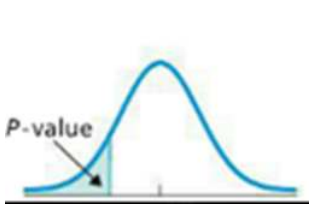
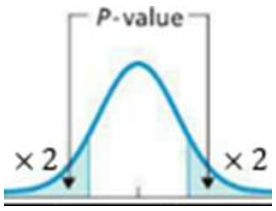
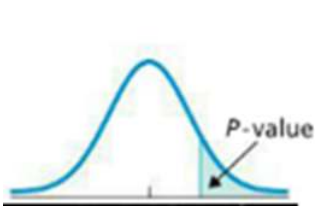
Hypothesis Test	One Mean Z or T Test	One Proportion Z Test	One Variance X^2 Test
Notation:	μ_0 - mean from H_0 $\bar{x}_i$ - sample mean n – sample size σ^2 – population variance s_i^2 – sample variance	p_0 - proportion from H_0 $\hat{p}_i$ - sample proportion n – sample size	σ_0^2 - variance from H_0 s_i^2 - sample variance n – sample size

To test the claim, we can use the critical value approach. In this approach, we construct the rejection region under the probability density curve of the involved distribution based on the type of the test and the significance level alpha. The boundaries of the region are called the critical values and can be found in a similar way for all procedures. We reject the null hypothesis if the test statistic is in the rejection region.

Test Type	LT	TT	RT
RR			
Critical Values	LCV: $x_{1-\alpha}$ RCV: DNE	LCV: $x_{1-\alpha/2}$ RCV: $x_{\alpha/2}$	LCV: DNE RCV: x_α
Reject if ...	$x_0 < x_{1-\alpha}$	$x_0 < x_{1-\alpha/2}$ or $x_0 > x_{\alpha/2}$	$x_0 > x_\alpha$

Note that x_α and x_0 can be z_α and z_0, t_α and t_0, or χ_α^2 and χ_0^2.

Alternatively, we can use the P-value approach. In this approach, we find the p-value depending on the test statistic and the type of the test. We reject the null hypothesis if the p-value is less than the significance level α.

Test Type	LT	TT	RT
P-Value			
	$P(X < x_0)$	$2P(X < x_0)$ or $2P(X > x_0)$	$P(X > x_0)$
Reject if ...	P-value $< \alpha$		

Note that x_α and x_0 can be z_α and z_0, t_α and t_0, or χ^2_α and χ^2_0.

Due to the logic of a hypothesis testing procedure, there are only two outcomes of the procedure:

reject or not reject the null hypothesis

Note that the double negative in this case is never interpreted as a positive that is, we never-never accept the null hypothesis! Therefore, the interpretation is that we either do or do not have sufficient evidence to suggest the alternative hypothesis.

All four procedures can be summarized in the following table!

Procedure	One Mean Z Test	One Mean T Test	One Proportion Z Test	One Variance X^2 Test
Purpose:	To test a statistical claim about a population mean, μ	To test a statistical claim about a population mean, μ	To test a statistical claim about a population proportion, p	To test a statistical claim about a population variance, σ^2
Assumptions:	• Simple random sample • Normal population or large sample • σ is known	• Simple random sample • Normal population or large sample • σ is unknown	• Simple random sample • np and $n(1-p)$ are both 10 or greater.	• Simple random sample • Normal population
Hypotheses:	$H_0 : \mu = \mu_0$ $H_a : \mu ? \mu_0$		$H_0 : p = p_0$ $H_a : p ? p_0$	$H_0 : \sigma^2 = \sigma_0^2$ $H_a : \sigma^2 ? \sigma_0^2$
Test Statistic:	$z_0 = \frac{\bar{x}_i - \mu_0}{\sigma/\sqrt{n}}$	$t_0 = \frac{\bar{x}_i - \mu_0}{s/\sqrt{n}}$ with $df = n-1$	$z_0 = \frac{\hat{p}_i - p_0}{\sqrt{p_0(1-p_0)/n}}$	$\chi^2_0 = \frac{s_i^2(n-1)}{\sigma_0^2}$ with $df = n-1$
Test Type:	$<$	$\neq$	$>$	
	LT	TT	RT	
CVA: Reject if …	$x_0 < x_{1-\alpha}$	$x_0 < x_{1-\alpha/2}$ or $x_0 > x_{\alpha/2}$	$x_0 > x_\alpha$	
PVA: Reject if …	$P(X < x_0) < \alpha$	$2P(X < x_0) < \alpha$ or $2P(X > x_0) < \alpha$	$P(X > x_0) < \alpha$	
Interpretation:	We DO (NOT) have sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis.			

It is easy to see that the procedures not only have a lot in common among themselves but also with the procedures for constructing confidence intervals when estimating population parameters.

This page titled 1.11.5: The Summary of Hypothesis Testing for One Parameter is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

SECTION OVERVIEW

1.12: Inferences about Two Parameters

1.12.1: Two Variances F Test

1.12.2: Two Proportions Z Test

1.12.3: Two Means Z Test

1.12.4: Two Means T Pooled Test

1.12.5: Two Means T Non-Pooled Test

1.12.6: Two Paired Means T Test

1.12.7: The Summary of Hypothesis Testing for Two Parameters

This page titled 1.12: Inferences about Two Parameters is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.12.1: Two Variances F Test

Section 0: Introduction to Hypothesis Testing for Two Parameters



The hypothesis testing procedures for testing claims about two population parameters is performed in the same way as the hypothesis testing procedures for one population parameter. As previously, we will check the assumptions, setup the hypotheses and compute the test statistic using a specific formula, and then use the critical value approach or the p-value approach to decide whether to reject or not reject the null hypothesis. Next, we will discuss the nuances of setting up and testing the hypothesis to test the claims about two population parameters. We first introduce a few new terms.

Definition: Independent samples

Independent samples – a pair of samples in which the observations in one sample do not influence the observations in the other.

Definition: Dependent samples

Dependent samples – a pair of samples in which the observations in one sample somehow influence the observations in the other.

Definition: Paired samples

Paired samples – a pair of samples in which the observations are paired in a distinct way.

Consider the following three different ways to pick a sample of ten males and a sample of ten females. It can be done on two different occasions or it can be done simultaneously by picking male-female couples or rotating between males and females. Which of the following samples are independent, dependent, and paired?

1. A sample of 10 males and a sample of 10 females were obtained by randomly picking a male first, then a female second, then another male third, and then another female until the numbers were reached.
2. A sample of 10 males and a sample of 10 females were obtained by randomly picking 10 male-female couples.
3. A sample of 10 males and a sample of 10 females were obtained by picking ten random males on one occasion and then picking ten random females on the other.

Ideally, we want to separate the two samples, so the bottom one is the example of independent samples, the top one is the example of dependent samples. Note that dependency does not have to be explicit, and the middle one is the example of paired samples.

Other common examples of paired samples are the samples obtained for the same subjects before and after some event.

We will learn how to test statistical claims that compare a pair of independent parameters of two populations such as:

- Two independent variances
- Two independent proportions
- Two independent means
  - Population standard deviations are known
  - Population standard deviations are not known
    - Can be assumed equal
    - Cannot be assumed equal

We will also learn how to test statistical claims about

- Two paired means

Consider the following examples of statistical claims that involve the parameters about two populations.

- "The average GPA of female students is greater than the average GPA of male students" can be symbolically written as

$$\mu_1 > \mu_2 \text{ or } \mu_1 - \mu_2 > 0$$

- "The percentage of baby-boomers that support the democratic party is less than the percentage of millennials" can be symbolically written as

$$p_1 < p_2 \text{ or } p_1 - p_2 < 0$$

- "The income variation is not the same among those who graduated from college and those who didn't" can be symbolically written as

$$\sigma_1^2 \neq \sigma_2^2 \text{ or } \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

We can create more interesting statements by replacing 0 and 1 with a number!

- by changing 0 to 0.5 we obtain a slightly different claim

$$\mu_1 - \mu_2 > 0.5$$

  which can be interpreted as "the average GPA of female students is at least 0.5 greater than the average GPA of male students".

- by changing 0 to -0.1 we obtain a slightly different claim

$$p_1 - p_2 < -0.1$$

  which can be interpreted as "the percentage of baby-boomers that support the democratic party is less than the percentage of millennials by 10% or more".

- by changing 1 to 2 we obtain a slightly different claim

$$\frac{\sigma_1^2}{\sigma_2^2} \neq 2$$

  which can be interpreted as "the income variation is not twice as large among those who graduated from college and those who didn't".

The following table summarizes the null and alternative hypothesis for which we have developed the testing procedures.

Testing Procedure	Two Means	Two Proportions	Two Variances
Null Hypothesis, H_0:	$\mu_1 - \mu_2 = \mu_0$	$p_1 - p_2 = p_0$	$\frac{\sigma_1^2}{\sigma_2^2} = k$
Alternative Hypothesis, H_a:	$\mu_1 - \mu_2 < \mu_0$	$p_1 - p_2 \neq p_0$	$\frac{\sigma_1^2}{\sigma_2^2} > k$

A few more terms:

Definition: Pooled statistic

A **pooled statistic** – a statistic computed from two sample as if it was one sample.

For example, for a **pooled proportion** we have the following formula:

$$\hat{p}_p = \frac{x_1+x_2}{n_1+n_2}$$

If the proportion of males in one class is 9 out of 26 and in the other class, it is 12 out of 24, then the pooled proportion is

$$\hat{p}_p = \frac{9+12}{26+24} = \frac{21}{50} = 0.42$$

For example, for a **pooled standard deviation** we have the following formula:

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

If sample 1 of size 8 has the sample variance 3.84 and sample 2 of size 16 has the sample variance 6.12 the pooled standard deviation is

$$s_p = \sqrt{\frac{(8-1)\cdot 3.84+(16-1)\cdot 6.12}{8+16-2}} = 2.32$$

Section 1: Two Variances F Test



Next, we will learn how to apply the Two Variances F Procedure to test a statistical claim about 2 population variances. Consider the following example.

A pharmaceutical company is about to launch a new manufacturing process in addition to the existing one. The quality control manager believes that the new method results in a different variation in the weights of the capsules. To verify the claim, the samples from each production line were obtained and the results are below:

Sample 1: 99.48 98.79 99.14 97.16 99.88 98.74 100.10

Sample 2: 99.32 99.93 99.86 99.89 99.46 97.67 101.09 100.75 102.03 100.73 101.27

At 5% significance level, test the claim that the two production methods have different variations. You may assume that the distributions of weights are normal in both production lines.

Note that the sample variances of the samples are $s_1^2 = 0.95$ and $s_2^2 = 1.39$, and the sizes of the samples are $n_1 = 7$ and $n_2 = 10$.

Now, let's identify the statistical claim that needs to be tested:

"variations are different"

The key word “variations” suggests that the claim is about the two variances, so we can symbolically express the claim as

$$\sigma_1^2 \neq \sigma_2^2 \text{ or } \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

Since the claim is about the 2 population variances we will use the 2 Variances F Procedure. Let’s check if all necessary assumptions are satisfied:

- The samples are simple random and independent
- The populations are normally distributed

We will use the following template to perform the hypothesis testing:

In step 1, we will set up the hypothesis:

Since the claim is in the form of an inequality, therefore it must be set as an alternative hypothesis, therefore the null hypothesis is $\frac{\sigma_1^2}{\sigma_2^2} = 1$ and the test is two-tail, that is:

$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ $H_a : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$	TT

In step 2, we will identify the significance level:

The significance level can always be found in the text of the problem. In our case it is 5%, thus:

$$\alpha = 0.05$$

In step 3, we will find the test statistic using the formula:

$$f_0 = \frac{s_1^2}{s_2^2} = \frac{0.95}{1.39} = 0.683$$

In step 4, we will perform either the critical value approach or p-value approach to test the claim:

- In critical value approach, we construct the rejection region using the F-curve with $(dfn = n_1 - 1 = 7 - 1 = 6, dfd = n_2 - 1 = 10 - 1 = 9$:

$$\text{RR: less than } f_{0.975} = 0.181 \text{ or greater than } f_{0.025} = 4.320$$

- In p-value approach, we compute the p-value:

$$\text{P-Value: } 2P(F < 0.683) = 0.662$$

In step 5, we will draw the conclusion:

- In the critical value approach, we must check whether the test statistic is in the rejection region or not. Our test statistic is 0.683 and it is between the critical values 0.181 and 4.320, thus it is not in the rejection region.
- In the p-value approach, we must check whether the p-value is less than the significance level or not. Our p-value is 0.662 and it is greater than $\alpha = 0.05$.

Both tests suggest that we DO NOT reject the null hypothesis in favor of the alternative.

In step 6, we will interpret the results:

Under 5% significance level we DO NOT have sufficient evidence to suggest that the two production methods have different variations.

We discussed how to apply the Two Variances F Procedure to test a statistical claim about two population variances or standard deviations.

This page titled 1.12.1: Two Variances F Test is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.12.2: Two Proportions Z Test



Next, we will learn how to apply the Two Proportions Z Procedure to test a statistical claim about 2 population proportions. Consider the following example.

A group of researchers wants to know whether men are more likely than women to contract a novel coronavirus. They surveyed two random samples of 350 men and 340 women who were tested and found that 70 men and 36 women in the samples tested positive.

At 1% significance level, test the claim that men are more likely to have the novel coronavirus than women.

Note that the sample proportions are $\hat{p}_1 = \frac{70}{350} = 0.2000$ and $\hat{p}_2 = \frac{36}{340} = 0.1059$ for men and women respectively, also the pooled proportion is $\hat{p}_p = \frac{70+36}{350+340} = 0.1536$.

Now, let's identify the statistical claim that needs to be tested:

"men are more likely to have the novel coronavirus than women"

It is not obvious, but the claim is about the population proportion of positive tested men being greater than the proportion of positive tested women, so we can symbolically express the claim as

$$p_1 > p_2 \text{ or } p_1 - p_2 > 0$$

Since the claim is about the 2 population proportions we will use the 2 Proportions Z Procedure. Let's check if all necessary assumptions are satisfied:

- The samples are simple random and independent
- The number of positive and negative responses are both greater than 10 for both samples.

We will use the following template to perform the hypothesis testing:

In step 1, we will set up the hypothesis:

Since the claim is in the form of an inequality, therefore it must be set as an alternative hypothesis, therefore the null hypothesis is $p_1 - p_2 = 0$ and the test is right-tail, that is:

$H_0 : p_1 - p_2 = 0$ $H_a : p_1 - p_2 > 0$	RT

In step 2, we will identify the significance level:

The significance level can always be found in the text of the problem. In our case it is 1%, thus:

$$\alpha = 0.01$$

In step 3, we will find the test statistic using the formula:

$$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_p \cdot (1-\hat{p}_p)} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{0.2000 - 0.1059}{\sqrt{0.1536 \cdot 0.8464} \sqrt{\frac{1}{350} + \frac{1}{340}}} = 3.427$$

In step 4, we will perform either the critical value approach or p-value approach to test the claim:

- In critical value approach, we construct the rejection region using the Z-curve:

$$\text{RR: greater than } z_{0.01} = 2.326$$

- In p-value approach, we compute the p-value:

$$\text{P-Value: } P(Z > 3.427) = 0.0003$$

In step 5, we will draw the conclusion:

- In the critical value approach, we must check whether the test statistic is in the rejection region or not. Our test statistic is 3.427 and it is to the right of the critical value 2.326, thus it is in the rejection region.
- In the p-value approach, we must check whether the p-value is less than the significance level or not. Our p-value is $0.0003$and it is less than $\alpha = 0.01$.

Both tests suggest that we DO reject the null hypothesis in favor of the alternative.

In step 6, we will interpret the results:

Under 1% significance level we DO have sufficient evidence to suggest that men are more likely to have the novel coronavirus than women.

We discussed how to apply the Two Proportions Z Procedure to test a statistical claim about two population proportions.

This page titled 1.12.2: Two Proportions Z Test is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.12.3: Two Means Z Test



Next, we will learn how to apply the Two Means Z Procedure to test a statistical claim about 2 population means when population standard deviations are known. Consider the following example.

A group of researchers wants to know whether men's life expectancy is shorter than women's life expectancy in the US in 2018. Historically, the standard deviation of men's life expectancy is known to be 15.8 years and of women's life expectancy is 13.4 years. The average life expectancy among 68 men in the sample was found to be 76.3 years and the average life expectancy among 54 women in the sample was found to be 81.4 years.

At 10% significance level, test the claim that men had a shorter life expectancy than women in the US in 2018.

Note that the sample means are $\bar{x}_1 = 76.3$ years and $\bar{x}_2 = 81.4$ years for men and women respectively, the sample sizes are $n_1 = 68$ for men and $n_2 = 54$ for women, and the population standard deviations are given $\sigma_1 = 15.8$ years and $\sigma_2 = 13.4$ years for men and women respectively.

Now, let's identify the statistical claim that needs to be tested:

"men's life expectancy is shorter than women's"

The claim is that the men's life expectancy is shorter than women's, so we can symbolically express the claim as

$$\mu_1 < \mu_2 \text{ or } \mu_1 - \mu_2 < 0$$

Since the claim is about the 2 population means and the population standard deviations are known we will use the 2 Means Z Procedure. Let's check if all necessary assumptions are satisfied:

- The samples are simple random and independent
- The CLT must be applicable i.e. the populations are normal or the samples are greater than 30
- The population standard deviations are known

We will use the following template to perform the hypothesis testing:

In step 1, we will set up the hypothesis:

Since the claim is in the form of an inequality, therefore it must be set as an alternative hypothesis, therefore the null hypothesis is $\mu_1 - \mu_2 = 0$ and the test is left-tail, that is:

$H_0 : \mu_1 - \mu_2 = 0$ $H_a : \mu_1 - \mu_2 < 0$	LT

In step 2, we will identify the significance level:

The significance level can always be found in the text of the problem. In our case it is 10%, thus:

$$\alpha = 0.10$$

In step 3, we will find the test statistic using the formula:

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{76.3 - 81.4}{\sqrt{\frac{15.8^2}{68} + \frac{13.4^2}{54}}} = -1.928$$

In step 4, we will perform either the critical value approach or p-value approach to test the claim:

- In critical value approach, we construct the rejection region using the Z-curve:

$$\text{RR: less than } -z_{0.10} = -1.282$$

- In p-value approach, we compute the p-value:

$$\text{P-Value: } P(Z < -1.928) = 0.027$$

In step 5, we will draw the conclusion:

- In the critical value approach, we must check whether the test statistic is in the rejection region or not. Our test statistic is -1.928 and it is to the left of the critical value -1.282, thus it is in the rejection region.
- In the p-value approach, we must check whether the p-value is less than the significance level or not. Our p-value is 0.027 and it is less than $\alpha = 0.10$.

Both tests suggest that we DO reject the null hypothesis in favor of the alternative.

In step 6, we will interpret the results:

Under 10% significance level we DO have sufficient evidence to suggest that men's life expectancy was less than women's life expectancy in the US in 2018.

We discussed how to apply the Two Means Z Procedure to test a statistical claim about two population means with known population standard deviations.

This page titled 1.12.3: Two Means Z Test is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.12.4: Two Means T Pooled Test



Next, we will learn how to apply the Two Means T Pooled Procedure to test a statistical claim about 2 population means when population standard deviations are unknown but can be assumed equal. Consider the following example.

A group of activists wants to the verify the claim that in a certain Fortune 500 company the average salary of female employees is different from the average salary of male employees. Two samples of male and female workers were obtained independently and analyzed. The sample of 19 women had the average salary \$64,491 and the standard deviation \$13,211. The sample of 24 men had the average salary \$67,541 and the standard deviation \$19,274. Test the claim at 10% significance level.

Note that the sample means are $\bar{x}_1 = 64491$ dollars and $\bar{x}_2 = 67541$ dollars, the sample standard deviations are $s_1 = 13211$ dollars and $s_2 = 19274$ dollars, and the sample sizes are $n_1 = 19$ and $n_2 = 24$ for women and men respectively.

Now, let's identify the statistical claim that needs to be tested:

"women's average salary is different from men's"

The claim is that women's average salary is different from men's, so we can symbolically express the claim as

$$\mu_1 \neq \mu_2 \text{ or } \mu_1 - \mu_2 \neq 0$$

Since the claim is about the 2 population means and the population standard deviations are unknown, we will use the 2 Means T Procedure but now we need to choose between pooled or non-pooled? We use the T **Pooled Procedure** when the standard deviations can be assumed equal, otherwise we use the T **Non-Pooled Procedure**.

To test the claim whether the standard deviations are the same we setup the 2 Variances F Test with $\alpha = 10\%$ (for convenience we chose the same alpha as in the problem but in general it doesn't have to match and depends on the context).

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_a : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

$$f_0 = \frac{s_1^2}{s_2^2} = \frac{13211^2}{19274^2} = 0.470$$

$$\text{p-value} = 0.106 > \alpha = 10\%$$

The conclusion is that we don't have sufficient evidence to reject the null hypothesis, therefore we will continue assuming that the variances are the same and choose the T-Pooled Procedure.

Let's check if all necessary assumptions are satisfied:

- The samples are simple random and independent

- The CLT must be applicable i.e. the populations are normal, or the samples are greater than 30
- The population standard deviations are unknown but are assumed equal

To use the T Pooled Procedure, we must compute the pooled standard deviation using the formula:

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

$$s_p = \sqrt{\frac{(19-1)\cdot 13211^2+(24-1)\cdot 19274^2}{19+24-2}} = 16485.20$$

Also, the degrees of freedom of the involved T distribution is

$$df = n_1 + n_2 - 2 = 19 + 24 - 2 = 41$$

We will use the following template to perform the hypothesis testing:

In step 1, we will set up the hypothesis:

Since the claim is in the form of an inequality, therefore it must be set as an alternative hypothesis, therefore the null hypothesis is $\mu_1 - \mu_2 = 0$ and the test is two-tail, that is:

$H_0 : \mu_1 - \mu_2 = 0$ $H_a : \mu_1 - \mu_2 \neq 0$	TT

In step 2, we will identify the significance level:

The significance level can always be found in the text of the problem. In our case it is 10%, thus:

$$\alpha = 0.10$$

In step 3, we will find the test statistic using the formula:

$$t_0 = \frac{\bar{x}_1-\bar{x}_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{64491-67541}{16485.20\sqrt{\frac{1}{19}+\frac{1}{24}}} = -0.602$$

In step 4, we will perform either the critical value approach or p-value approach to test the claim:

- In critical value approach, we construct the rejection region using the T-curve with $df = 41$:

  RR: less than $-t_{0.05} = -1.683$ or greater than $t_{0.05} = 1.683$

- In p-value approach, we compute the p-value:

P-Value: $2P(T < -0.602) = 0.550$

In step 5, we will draw the conclusion:

- In the critical value approach, we must check whether the test statistic is in the rejection region or not. Our test statistic is -0.602 and it is between the critical values -1.683 and 1.683, thus it is NOT in the rejection region.
- In the p-value approach, we must check whether the p-value is less than the significance level or not. Our p-value is 0.550 and it is greater than $\alpha = 0.10$.

Both tests suggest that we DO NOT reject the null hypothesis in favor of the alternative.

In step 6, we will interpret the results:

Under 10% significance level we DO NOT have sufficient evidence to suggest that women's salaries are different from men's salaries.

We discussed how to apply the Two Means T Pooled Procedure to test a statistical claim about two population means with population standard deviations that are unknown but assumed equal.

This page titled 1.12.4: Two Means T Pooled Test is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.12.5: Two Means T Non-Pooled Test



Next, we will learn how to apply the Two Means T Non-Pooled Procedure to test a statistical claim about 2 population means when population standard deviations are unknown and cannot be assumed equal. Consider the following example.

A group of economists wants to the verify the claim that US workers have the average annual leave less than EU workers. Two samples of US and EU workers were obtained independently and analyzed. The sample of 32 US workers had the average annual leave of 20.4 days and the standard deviation 8.5 days. The sample of 38 EU workers had the average annual leave of 24.9 days and the standard deviation 4.3 days. Test the claim at 1% significance level.

Note that the sample means are $\bar{x}_1 = 20.4$ days and $\bar{x}_2 = 24.9$ days, the sample standard deviations are $s_1 = 8.5$ days and $s_2 = 4.3$ days, and the sample sizes are $n_1 = 32$ and $n_2 = 38$ for US and EU respectively.

Now, let's identify the statistical claim that needs to be tested:

"US workers' annual leave is less than EU workers'"

The claim is that the US workers' annual leave is less than EU workers', so we can symbolically express the claim as

$$\mu_1 < \mu_2 \text{ or } \mu_1 - \mu_2 < 0$$

Since the claim is about the 2 population means and the population standard deviations are unknown, we will use the 2 Means T Procedure but now we need to choose between pooled or non-pooled? We use the T **Pooled Procedure** when the standard deviations can be assumed equal, otherwise we use the T **Non-Pooled Procedure**.

To test the claim whether the standard deviations are the same we setup the 2 Variances F Test with $\alpha = 1\%$ (for convenience we chose the same alpha as in the problem but in general it doesn't have to match and depends on the context).

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

$$H_a : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

$$f_0 = \frac{s_1^2}{s_2^2} = \frac{8.5^2}{4.3^2} = 3.908$$

$$\text{p-value} = 0.001 < \alpha = 1\%$$

The conclusion is that we do have sufficient evidence to reject the null hypothesis, therefore we will believe that the variances are not the same and choose the T Non-Pooled Procedure.

Let's check if all necessary assumptions are satisfied:

- The samples are simple random and independent

- The CLT must be applicable i.e. the populations are normal, or the samples are greater than 30
- The population standard deviations are unknown and not assumed equal

To use the T Non-Pooled Procedure, we must compute the degrees of freedom of the involved T distribution:

$$df = \frac{[\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}]^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1-1}+\frac{(\frac{s_2^2}{n_2})^2}{n_2-1}} = \frac{[\frac{8.5^2}{32}+\frac{4.3^2}{38}]^2}{\frac{(\frac{8.5^2}{32})^2}{32-1}+\frac{(\frac{4.3^2}{38})^2}{38-1}} = 44.086 \downarrow 44$$

We will use the following template to perform the hypothesis testing:

In step 1, we will set up the hypothesis:

Since the claim is in the form of an inequality, therefore it must be set as an alternative hypothesis, therefore the null hypothesis is $\mu_1 - \mu_2 = 0$ and the test is left-tail, that is:

$H_0 : \mu_1 - \mu_2 = 0$ $H_a : \mu_1 - \mu_2 < 0$	LT

In step 2, we will identify the significance level:

The significance level can always be found in the text of the problem. In our case it is 1%, thus:

$$\alpha = 0.01$$

In step 3, we will find the test statistic using the formula:

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_1^2}{n_2}}} = \frac{20.4-24.9}{\sqrt{\frac{8.5^2}{32}+\frac{4.3^2}{38}}} = -2.716$$

In step 4, we will perform either the critical value approach or p-value approach to test the claim:

- In critical value approach, we construct the rejection region using the T-curve with $df = 44$:

$$\text{RR: less than } -t_{0.01} = -2.414$$

- In p-value approach, we compute the p-value:

$$\text{P-Value: } P(T < -2.716) = 0.0047$$

In step 5, we will draw the conclusion:

- In the critical value approach, we must check whether the test statistic is in the rejection region or not. Our test statistic is -2.716 and it is to the left of -2.414, thus it is in the rejection region.
- In the p-value approach, we must check whether the p-value is less than the significance level or not. Our p-value is 0.0047 and it is less than $\alpha = 0.01$.

Both tests suggest that we DO reject the null hypothesis in favor of the alternative.

In step 6, we will interpret the results:

Under 1% significance level we DO have sufficient evidence to suggest that US workers have less annual leave than EU workers.

We discussed how to apply the Two Means T Non-Pooled Procedure to test a statistical claim about two population means with population standard deviations that are unknown and not assumed equal.

This page titled 1.12.5: Two Means T Non-Pooled Test is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.12.6: Two Paired Means T Test



Next, we will learn how to apply the Two Paired Means T Procedure to test a statistical claim about 2 paired population means. Consider the following example.

A local group of economists wants to study the effects of the pandemic on the small businesses in their city. A survey was sent out to a random sample of 100 small businesses to anonymously report their weekly earnings throughout March and April. Only 10 businesses provided enough data to compute their average weekly earnings for two months:

	1	2	3	4	5	6	7	8	9	10
March	8.1	7.4	7.2	6.9	7.3	6.6	6.1	5.9	7.5	7.7
April	4.8	2.4	3.2	4.1	1.9	2.1	4.6	2.6	1.6	4.2

Use 5% significance level to test the claim that the weekly earnings of small businesses have decreased during the pandemic.

Note that the sample means and the sample standard deviations for each sample can be easily computed. However, it appears that there is a clear pairing between the samples since it is the same 10 businesses for which the data was obtain in March and April, so we are not interested in the differences between the two samples but in the sample of differences.

	1	2	3	4	5	6	7	8	9	10	Avg	St. dev.
March-April	8.1-4.8	7.4-2.4	7.2-3.2	6.9-4.1	7.3-1.9	6.6-2.1	6.1-4.6	5.9-2.6	7.5-1.6	7.7-4.2		
Difference	3.3	5	4	2.8	5.4	4.5	1.5	3.3	5.9	3.5	3.92	1.32

As a result, we turned two samples into one with the sample mean $\bar{x}_d = 3.92$ thousand of dollars and sample standard deviation $s_d = 1.32$ thousand of dollars.

Now, let's identify the statistical claim that needs to be tested:

"the weekly earnings of small businesses have decreased during the pandemic"

The claim is that the weekly earnings of small businesses in March is greater than the weekly earnings of small businesses in April, so we can symbolically express the claim as

$$\mu_1 > \mu_2 \text{ or } \mu_1 - \mu_2 > 0 \text{ or } \mu_d > 0$$

Since the claim is about the paired means and the standard deviation of the population of differences is unknown, we will use the 1 Mean T Procedure.

Let's check if all necessary assumptions are satisfied:

- The sample is simple random
- The population is normally distributed, or the samples size is greater than 30
- The population standard deviation, σ_d, is unknown but $s_d = 1.32$

We will use the following template to perform the hypothesis testing:

In step 1, we will set up the hypothesis:

Since the claim is in the form of an inequality, therefore it must be set as an alternative hypothesis, therefore the null hypothesis is $\mu_d = 0$ and the test is right-tail, that is:

$H_0 : \mu_d = 0$ $H_a : \mu_d > 0$	RT

In step 2, we will identify the significance level:

The significance level can always be found in the text of the problem. In our case it is 5%, thus:

$$\alpha = 0.05$$

In step 3, we will find the test statistic using the formula:

$$t_0 = \frac{\bar{d} - \mu_0}{s_d/\sqrt{n}} = \frac{3.92 - 0}{1.32/\sqrt{10}} = 9.391$$

In step 4, we will perform either the critical value approach or p-value approach to test the claim:

- In critical value approach, we construct the rejection region using the T-curve with $df = n - 1 = 10 - 1 = 9$:

$$\text{RR: greater than } t_{0.05} = 1.833$$

- In p-value approach, we compute the p-value:

P-Value: $P(T > 9.391) = 0.000$

In step 5, we will draw the conclusion:

- In the critical value approach, we must check whether the test statistic is in the rejection region or not. Our test statistic is 9.391 and it is to the right of the critical value 1.833, thus it is in the rejection region.
- In the p-value approach, we must check whether the p-value is less than the significance level or not. Our p-value is 0.000 and it is less than $\alpha = 0.05$.

Both tests suggest that we DO reject the null hypothesis in favor of the alternative.

In step 6, we will interpret the results:

Under 5% significance level we have sufficient evidence to suggest that the weekly earnings of small businesses have decreased during the pandemic.

We discussed how to apply the Two Paired Means T Procedure to test a statistical claim about two paired population means with population standard deviations unknown.

This page titled 1.12.6: Two Paired Means T Test is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.12.7: The Summary of Hypothesis Testing for Two Parameters



We learned altogether 6 procedures for testing statistical claims about 3 different pairs of parameters. So how do we know when and which procedure to use? Next, we will summarize all the procedures for testing hypotheses for two parameters.

First, we must determine the statistical claim that is being tested in two steps:

1. Look for the key words to identify the parameters in the claim:
   - A pair of means/averages when testing a claim about two population means.
   - A pair of proportions/percentages when testing a claim about two population proportions.
   - A pair of variances/standard deviations when testing a claim about two population variances.
2. Look for the key words that identify the inequality in the claim:
   - Less than/at most for LT
   - Not equal/different for TT
   - Greater than/at least for RT

   The statistical claim in the form of an inequality is called the alternative hypothesis, and the statistical claim in the form of an equation is called the null hypothesis.

Next, we determine the procedure in the following way:

- Use the F procedure for testing claims about 2 variances.
- Use the Z procedure for testing claims about 2 proportions.
- For testing claims about two means, we first need to check whether the samples are paired or independent:
  - If they are paired, then we use One Mean T Procedure.
  - If they are independent, then we check whether the population standard deviations are known:
    - If the population standard deviations are known, then we use the Z procedure.
    - If the population standard deviations are unknown, then we check whether they can be assumed equal:
      - If unknown population standard deviations can be assumed equal, then we use the T pooled procedure.
      - If unknown population standard deviations cannot be assumed equal, then we use T non-pooled procedure.

Remember - before applying the procedure verify all the assumptions!

So before starting the procedure we must have already figured out the significance level α which is the maximum tolerated type 1 error. The most common values for alpha are 1, 5, and 10%.

Next, we compute the test statistic using a formula that is specific to each procedure. Note that we label the test statistic using the lower-case letter of the name of the involved distribution with the index zero. That is in a Z-test the test statistic is called z_0, in the

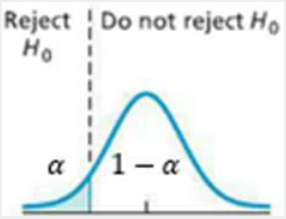
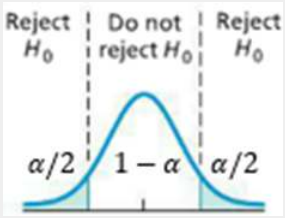
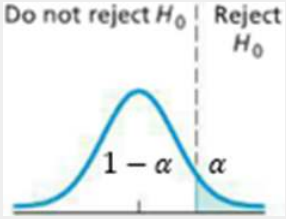
T-test the test statistic is called t_0, and in the F-test the test statistic is called f_0.

Procedure	Two Variances F Test	Two Proportions Z Test	Two Means Z Test	Two Means T Non-Pooled Test	Two Means T Pooled Test	Two Paired Means T Test
Test Statistic:	$f_0 = \frac{s_1^2}{s_2^2}$ $df = (n_1 - 1, n_2 - 1)$	$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_p \cdot (1-\hat{p}_p)} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $\hat{p}_p = \frac{x_1 + x_2}{n_1 + n_2}$	$z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}}$	$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $df = n_1 + n_2 - 2$	$t_0 = \frac{\bar{d} - \mu_0}{s_d / \sqrt{n}}$ $df = n - 1$

To apply the formula, one must be able to understand the notation. In this step, it may be necessary to reread the question several times to extract the information.

Procedure	Two Variances F Test	Two Proportions Z Test	Two Means Z Test	Two Means T Non-Pooled Test	Two Means T Pooled Test	Two Paired Means T Test
Notation:	s_i^2 sample variances	$\hat{p}_i$ sample proportions $\hat{p}_p$ pooled proportion x_i number of positive responses n_i sample sizes	$\bar{x}_i$ sample means σ_i^2 population variances n_i sample sizes	$\bar{x}_i$ sample means s_i^2 sample variances n_i sample sizes	$\bar{x}_i$ sample means s_i^2 sample variances s_p^2 pooled variance n_i sample sizes	$\bar{d}$ average difference σ_d^2 variance of differences μ_0 mean from H_0 n sample size

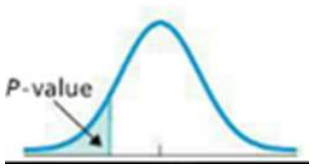
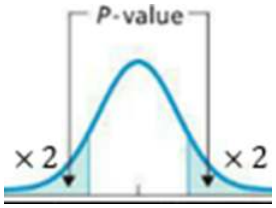
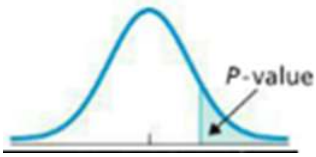
To test the claim, we can use the critical value approach. In this approach, we construct the rejection region under the probability density curve of the involved distribution based on the type of the test and the significance level alpha. The boundaries of the region are called the critical values and can be found in a similar way for all procedures. We reject the null hypothesis if the test statistic is in the rejection region.

Test Type	LT	TT	RT
RR			
Critical Values	LCV: $x_{1-\alpha}$ RCV: DNE	LCV: $x_{1-\alpha/2}$ RCV: $x_{\alpha/2}$	LCV: DNE RCV: x_α
Reject if ...	$x_0 < x_{1-\alpha}$	$x_0 < x_{1-\alpha/2}$ or $x_0 > x_{\alpha/2}$	$x_0 > x_\alpha$

Note that x_α and x_0 can be z_α and z_0, t_α and t_0, or χ^2_α and χ^2_0.

Alternatively, we can use the P-value approach. In this approach, we find the p-value depending on the test statistic and the type of the test. We reject the null hypothesis if the p-value is less than the significance level α.

Test Type	LT	TT	RT

Test Type	LT	TT	RT
P-Value			
	$P(X < x_0)$	$2P(X < x_0)$ or $2P(X > x_0)$	$P(X > x_0)$
Reject if ...	P-value $< \alpha$		

Note that x_α and x_0 can be z_α and z_0, t_α and t_0, or χ^2_α and χ^2_0.

Due to the logic of a hypothesis testing procedure, there are only two outcomes of the procedure:

reject or not reject the null hypothesis

Note that the double negative in this case is never interpreted as a positive that is, we never-never accept the null hypothesis! Therefore, the interpretation is that we either do or do not have sufficient evidence to suggest the alternative hypothesis.

All six procedures can be summarized in the following table!

Procedure	Two Variances F Test	Two Proportions Z Test	Two Means Z Test	Two Means T Non-Pooled Test	Two Means T Pooled Test	Two Paired Means T Test
Purpose:	To test a statistical claim about two population variances	To test a statistical claim about two population proportions	To test a statistical claim about two population means			
Assumptions:	• Simple random sample • Independent samples • Normal populations	• Simple random samples • Independent samples • $n_i\hat{p}_i \geq 10$ and $n_i(1-\hat{p}_i$ geq 10\) for $i = 1, 2$	• Simple random samples • Independent samples • Normal population or large samples • Population standard deviations are known	• Simple random samples • Independent samples • Normal population or large samples • Population standard deviations are unknown	• Simple random samples • Independent samples • Normal population or large samples • Population standard deviations are unknown but assumed to be equal	• Simple random samples • Paired samples • Normal population or large sample
Hypotheses:	$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = k$ $H_a : \frac{\sigma_1^2}{\sigma_2^2} ? k$	$H_0 : p_1 - p_2 = \Delta_0$ $H_a : p_1 - p_2 ? \Delta_0$	$H_0 : \mu_1 - \mu_2 = \Delta_0$ $H_a : \mu_1 - \mu_2 ? \Delta_0$			$H_0 : \mu_d = \Delta_0$ $H_a : \mu_d ? \Delta_0$
Test Statistic:	$f_0 = \frac{s_1^2}{s_2^2}$ $df = (n_1 - 1, n_2 - 1)$	$z_0 = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_p \cdot (1-\hat{p}_p)} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ $\hat{p}_p = \frac{x_1 + x_2}{n_1 + n_2}$		$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_1^2}{n_2}}}$	$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $df = n_1 + n_2 - 2$	$t_0 = \frac{\bar{d} - \Delta_0}{s_d / \sqrt{n}}$ $df = n - 1$
Test Type:	$<$	$\neq$			$>$	

Procedure	Two Variances F Test	Two Proportions Z Test	Two Means Z Test	Two Means T Non-Pooled Test	Two Means T Pooled Test	Two Paired Means T Test
	LT		TT		RT	
CVA: Reject if …	$x_0 < x_{1-\alpha}$		$x_0 < x_{1-\alpha/2}$ or $x_0 > x_{\alpha/2}$		$x_0 > x_\alpha$	
PVA: Reject if …	$P(X < x_0)$		$2P(X < x_0)$ or $2P(X > x_0)$		$P(X > x_0)$	
Interpretation:	We DO (NOT) have sufficient evidence to reject the null hypothesis in favor of the alternative hypothesis.					

It is easy to see that the procedures not only have a lot in common among themselves but also with the procedures for testing the hypotheses about one parameter.

This page titled 1.12.7: The Summary of Hypothesis Testing for Two Parameters is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

SECTION OVERVIEW

1.13: Chi-Squared Procedures

1.13.1: The Goodness-of-Fit Test

1.13.2: Homogeneity Test

1.13.3: Test for Independence

This page titled 1.13: Chi-Squared Procedures is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.13.1: The Goodness-of-Fit Test



Next, we will learn how to apply the goodness-of-fit test to test a statistical claim about the distribution of a categorical variable.

The table below shows the distribution of ER visits across the days of the week in a random sample of 434 visits at a random local hospital in a large city:

Days	Sun	Mon	Tue	Wed	Thu	Fri	Sat
Visits	40	50	63	77	77	78	49

Use 10% significance level to test the claim that the ER visits are evenly distributed across the days of the week.

Now, let's identify the statistical claim that needs to be tested:

"the ER visits are evenly distributed across the days of the week"

What exactly does it mean to be evenly or uniformly distributed? If that was true, we would expect all days to have the same number of visits! The total is 434, and to have a total of 434 while uniformly distributed would require a 62 visits per day. As a result, we expect the following outcomes:

Days	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Total
Visits	62	62	62	62	62	62	62	434

This table is called the expected (E) distribution, and what we obtained from the sample is called the observed (O) distribution.

Let's check if all necessary assumptions are satisfied:

- Simple random sample
- All expected frequencies are 1 or greater
- At most 20% of the expected frequencies are less than 5

We will use the following template to perform the hypothesis testing:

In step 1, we will set up the hypothesis:

The null and alternative hypotheses are always the same in a goodness-of-fit test. The null hypothesis is that the variable follows a specific distribution, and the alternative hypothesis is that the variable does not follow the specific distribution. In our case, the variable is the days on which an ER was visited, and the type of the distribution is uniform. So, the null hypothesis is that the

number of visits to ER is uniformly distributed, and the alternative hypothesis is that the number of visits is not uniformly distributed! The goodness of fit test is always(!) right-tail.

H_0: the number of ER visits has a uniform distribution H_a: the number of ER visits doesn't have a uniform distribution	RT

In step 2, we will identify the significance level:

The significance level can always be found in the text of the problem. In our case it is 10%, thus:

$$\alpha = 0.10$$

In step 3, we will find the test statistic using the formula:

$$\chi_0^2 = \sum \frac{(O-E)^2}{E}$$

O	E	$O-E$	$(O-E)^2$	$\frac{(O-E)^2}{E}$
40	62	$40-62=-22$	$(-22)^2=484$	$\frac{484}{62}=7.806$
50	62	$50-62=-12$	$12^2=144$	$\frac{144}{62}=2.323$
63	62	$63-62=1$	$1^2=1$	$\frac{1}{62}=0.016$
77	62	$77-62=15$	$15^2=225$	$\frac{225}{62}=3.629$
77	62	$77-62=15$	$15^2=225$	$\frac{225}{62}=3.629$
78	62	$79-62=16$	$16^2=256$	$\frac{256}{62}=4.129$
49	62	$49-62=-13$	$13^2=169$	$\frac{169}{62}=2.726$
434	434		$\chi_0^2=$	24.26

In step 4, we will perform either the critical value approach or p-value approach to test the claim:

- In critical value approach, we construct the rejection region using the X^2-curve with $df = c-1 = 7-1 = 6$, where $c=7$ is the number of categories:

$$\text{RR: greater than } \chi^2_{0.10} = 10.645$$

- In p-value approach, we compute the p-value:

$$\text{P-Value: } P(X^2 > 24.26) = 0.0005$$

In step 5, we will draw the conclusion:

- In the critical value approach, we must check whether the test statistic is in the rejection region or not. Our test statistic is 24.26 and it is to the right of the critical value 10.645, thus it is in the rejection region.
- In the p-value approach, we must check whether the p-value is less than the significance level or not. Our p-value is $0.0005$and it is less than $\alpha = 0.10$.

Both tests suggest that we DO reject the null hypothesis in favor of the alternative.

In step 6, we will interpret the results:

Under 10% significance level we have sufficient evidence to suggest that the ER visits are not uniformly distributed throughout a week.

We discussed how to apply the X^2 Goodness-of-Fit Procedure to test a statistical claim whether a variable follows a specific distribution or not.

This page titled 1.13.1: The Goodness-of-Fit Test is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.13.2: Homogeneity Test



This page titled 1.13.2: Homogeneity Test is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.13.3: Test for Independence



This page titled 1.13.3: Test for Independence is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

SECTION OVERVIEW

1.14: ANOVA Procedures

1.14.1: One-Way ANOVA

1.14.2: Two-Way ANOVA

This page titled 1.14: ANOVA Procedures is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.14.1: One-Way ANOVA



Next, we will learn how to apply the one-way analysis of variance to test a statistical claim about more than two population means.

A dean wants to check the consistency of teaching quality among the faculty in one of the departments. A sample of students was obtained from three classes taught by three different faculty, and their common final exam scores were recorded.

Class	Final Exam Scores							
1	62	66	73	73	91	68	72	
2	78	74	76	71	70	77	68	71
3	82	80	69	77	68	86		

Use 10% significance level to test the claim that the teaching quality is consistent among the three courses.

Let's compute the summaries for each sample:

Class	Final Exam Scores								$\bar{x}_i$	s_i	n_i
1	62	66	73	73	91	68	72		72.14	9.26	7
2	78	74	76	71	70	77	68	71	73.13	3.64	8
3	82	80	69	77	68	86			77.00	7.21	6

Let's compute the summary of all sample combined:

Sample	$\bar{x}_i$	s_i	n_i
Combined	73.90	6.90	21

Now, let's identify the statistical claim that needs to be tested:

"the teaching quality is consistent among the three courses"

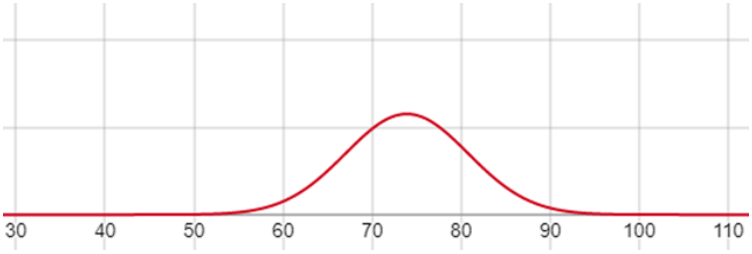
A one way to measure the quality is by the average grade on the common exam thus we are testing the claim that

"all means are the same"

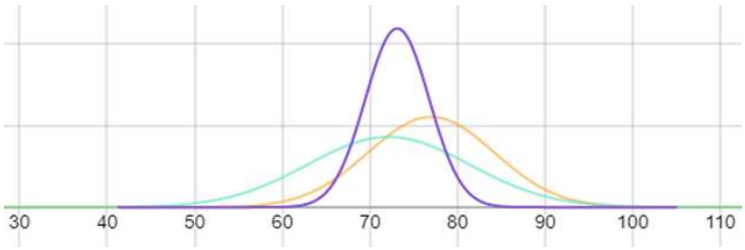
We can symbolically express the claim as

$$\mu_1 = \mu_2 = \mu_3$$

The way we are going to measure the likeliness of three samples is by computing the test statistic. But let's gather some intuition about what we are trying to achieve. This is the distribution of all test results as if they were in one sample – we can call it the average distribution.

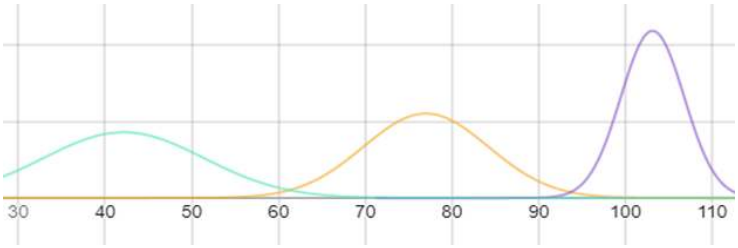


These are the graphs of three sample distributions that we obtained.



Do they look the same or in other words to they look like the average distribution? The answer is not obvious.

It would be an obvious "no" if they looked like this:



In such a case, we would expect the test statistic to be very high and if all distributions were overlapping the average distribution, then the answer will be an obvious "yes" and the test statistic will be equal to 0.

Let's check if all necessary assumptions are satisfied:

- Simple random and independent samples
- Normal populations
- Equal population standard deviations

We will use the following template to perform the hypothesis testing:

In step 1, we will set up the hypothesis:

The null and alternative hypothesis are always the same in an ANOVA test. The null hypothesis is that the population means are the same and the alternative hypothesis is that at least one of the means is not the same as the others. The one-way ANOVA test is always (!) right-tail.

H_0: all the means are equal H_a: not all the means are equal	RT

In step 2, we will identify the significance level:

The significance level can always be found in the text of the problem. In our case it is 10%, thus:

$$\alpha = 0.10$$

In step 3, we will find the test statistic by finding the ratio of the variation between the samples and the variation within the samples:

$$f_0 = \frac{\text{VAR}_{\text{between}}}{\text{VAR}_{\text{within}}} = \frac{42.04}{48.21} = 0.872$$

The table summarizes all the computations necessary to compute the test statistic f_0 which follows F distribution with $dfn = k-1 = 2$ degrees of freedom in the numerator and $dfd = n-k$ degrees of freedom in the denominator, where $k = 3$ is the number of samples and $n = 7+8+6 = 21$ is the size of all samples combined.

Sample Stats	Sample 1	Sample 2	Sample 3		
Size (n_i)	7	8	6	Mean ($\bar{x}$)	
Mean ($\bar{x}_i$)	72.14	73.13	77.00	**73.90**	
Variance (s_i^2)	85.81	13.27	52.00		
	n_i	$\bar{x}_i$	$n_i \cdot (\bar{x}_i - \bar{x})^2$		
Sample 1	7	72.14	$7 \cdot (72.14 - 73.90)^2 =$	**21.68**	
Sample 2	8	73.13	$8 \cdot (73.13 - 73.90)^2 =$	4.74	
Sample 3	6	77.00	$6 \cdot (77.00 - 73.90)^2 =$	**57.66**	
$k-1 =$	$3-1=2$		$\text{Total}_1 =$	**84.08**	$\text{VAR}_{\text{between}} = \frac{\text{Total}_1}{k-1} = \frac{84.08}{2} = 42.04$
	n_i	s_i^2	$(n_i - 1) \cdot s_i^2$		
Sample 1	7	85.81	$(7-1) \cdot 85.81 =$	514.86	
Sample 2	8	13.27	$(8-1) \cdot 13.27 =$	**92.89**	
Sample 3	6	52.00	$(6-1) \cdot 52.00 =$	260	
$n-k =$	$21-3=18$		$\text{Total}_1 =$	867.75	$\text{VAR}_{\text{within}} = \frac{\text{Total}_2}{n-k} = \frac{867.75}{18} = 48.21$

In step 4, we will perform either the critical value approach or p-value approach to test the claim:

- In critical value approach, we construct the rejection region using the F-curve with $dfn = k-1 = 3-1 = 2$ and $dfd = n-k = 21-3 = 18$, where $n = 21$ is the total number of observations in $k = 3$ samples:

$$\text{RR: greater than } f_{0.10} = 2.624$$

- In p-value approach, we compute the p-value:

$$\text{P-Value: } P(F > 0.872) = 0.435$$

In step 5, we will draw the conclusion:

- In the critical value approach, we must check whether the test statistic is in the rejection region or not. Our test statistic is 0.872 and it is to the left of the critical value 2.624, thus it is NOT in the rejection region.
- In the p-value approach, we must check whether the p-value is less than the significance level or not. Our p-value is 0.435 and it is greater than $\alpha = 0.10$.

Both tests suggest that we DO NOT reject the null hypothesis in favor of the alternative.

In step 6, we will interpret the results:

Under 10% significance level we DO NOT have sufficient evidence to suggest that the quality of teaching is inconsistent among the faculty in the department.

We discussed how to apply the ANOVA Procedure to test a statistical claim whether the population means are the same or not. This procedure is similar in spirit to the 2 means T POOLED procedure except that it can be done for three and more means.

This page titled 1.14.1: One-Way ANOVA is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.14.2: Two-Way ANOVA



This page titled 1.14.2: Two-Way ANOVA is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

SECTION OVERVIEW

1.15: Regression Analysis

1.15.1: Introduction to Regression Analysis

1.15.2: Linear Regression Analysis

This page titled 1.15: Regression Analysis is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.15.1: Introduction to Regression Analysis

Section 1: Fundamentals



Next, we will lay the foundation for introducing the concept of linear regression and correlation.

Let's start with a few definitions:

Definition: Relation

A **relation** is any form of connection between the variables usually can be expressed as an equation.

For example, $a+b=2$ is an equation that describes a relation between the two variables a and b.

Definition: Input variable (predictor)

We call a variable an **input variable** or **predictor** if its value does not depend on the value of the other variable but can be chosen independently.

Definition: Output variable (response variable)

We call a variable an **output variable** or **response variable** if its value depends on the value of the other variable and can be computed via formula.

For example, consider the following relation:

$$V+2000t=20000$$

where V is the value (\$) and t is the age of a vehicle (years).

Depending on what variable is dependent and independent, we can use one of the following equations:

$$V=-2000t+20000 \text{ or } t=-\tfrac{1}{2000}V+10$$

Now, is it the age of the vehicle that depends on the value or the value depends on the age? The answer is obvious so in the given relation the better formula for interpretation is

$$V=-2000t+20000$$

We continue with more definitions.

Definition: Linear relation

A relation is called **linear** if it can be written in the form

$$y = mx + b$$

where x is the input variable and y is the output variable.

Definition: Slope and y-intercept

In a linear relation, $y = mx + b$, m is called the **slope**, and b is called the **y-intercept**.

Example 1.15.1.1.1

The relation $V = -2000t + 20000$ is linear, with the slope $m = -2000$ and the y-intercept $b = 20000$.

Definition: Linear graph

A **graph** of a linear equation is a straight line that passes through the point $(0, b)$ and rises $|m|$ units (or drops $|m|$ units if $m < 0$) vertically for every horizontal change of 1 unit.

Example 1.15.1.1.2

The graph of $V = -2000t + 20000$ is shown below.



Definition: Intercept

An **intercept** is a point where the graph intersects the axis.

Example 1.15.1.1.3

The graph of $V = -2000t + 20000$ has the y-intercept 20000 and the x-intercept 10.

To find the y-intercept, set the input to 0 and evaluate the right-hand side:

$$V = -2000 \cdot 0 + 20000$$

$$V = 0 + 20000$$

$$V = 20000$$

To find the x-intercept, set the output to 0 and solve the equation:

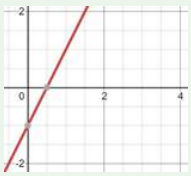
$$0 = -2000 \cdot t + 20000$$

$$2000 \cdot t = 20000$$

$$t = 10$$

Definition: Positive relation

A **positive relation** is a relation in which the output increases as the input increases.



Definition: Negative relation

A **negative relation** is a relation in which the output decreases as the input increases.



Example 1.15.1.1.4

The relation $V = -2000t + 20000$ is a negative relation because as the age of the vehicle increases the value decreases.

In applications, the input and output variables have units, therefore everything in the equation and the graph will have units and its own interpretation:

- The y-intercept, b, has the same units as the output variable and is interpreted as the output when input is zero.
- The x-intercept has the same units as the input variable and is interpreted as the input for which the output is zero.
- The slope, m, has the units that are the units of the output divided by the units of the input and is interpreted as the rate of change of the output with respect to the input.

Example 1.15.1.1.5

In $V = -2000t + 20000$

- The units of the input variable t are years and the units of the output variable V are dollars.
- The y-intercept is $b = 20000$ dollars which is interpreted as the value of the vehicle when it was new ($t = 0$).
- The x-intercept is $t = 10$ years which is interpreted as the age when the vehicle is worthless ($V = 0$).
- The slope is $m = -2000$ \$/year which is interpreted as the rate of depreciation of the vehicle.

We can use the equation in two different ways:

- to find the output for a given input and
- to find the input with a given output.

Task:	Find the output for a given input	Find the input with a given output
Given:	$t = 5.5$	$V = 4000$
Solution:	$V = -2000 \cdot 5.5 + 20000$ $V = -11000 + 20000$ $V = 9000$	$4000 = -2000 \cdot t + 20000$ $2000 \cdot t = 16000$ $t = 8$
Interpretation:	A 5.5-year-old vehicle worth \$9000.	For \$4000, you can get an 8-year old vehicle.

Imagine what you can do if such function was given for every vehicle! We CAN

- Predict the future resale value of a vehicle;
- Write-off the annual depreciation of a vehicle from taxes;
- Compare vehicles and decide which vehicle better retains its value.

However, the important question here is where does this function comes from, or in other words how does one obtain such a function or a relation?

Next, we will develop a formal procedure that will allow us to find the relation between two quantitative variables such as the value and the age of a vehicle!

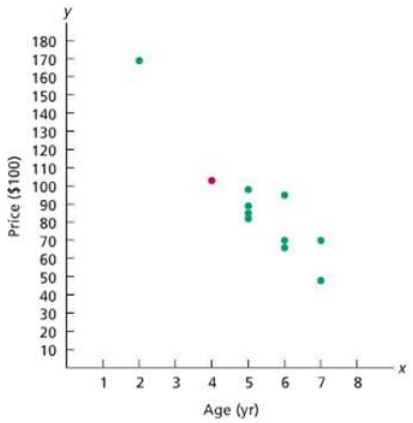
Section 2: Development



So how does one reveal a relation between the two variables such as the price and the age of a car? The answer of course is in the data. To study the relation between the two variables we must collect and analyze bivariate data, in our case we must collect a sample of cars and record the age and the price of each car. We can organize such bivariate data in the form of the table:

Car	Age (yr)	Price ($100)
1	5	85
2	4	103
3	6	70
4	5	82
5	5	89
6	5	98
7	6	66
8	6	95
9	2	169
10	7	70
11	7	48

The best way to see if there is any relation is to visualize the data! In other words, we must construct a scatterplot. When graphing the scatter plot, we decide which variable to present as independent and which as dependent. In this case, the context provides clarification – the price depends on the age of the vehicle.



Next let's answer some questions about the data.

Q1: Does the picture suggest that there is a relation between the two variables?

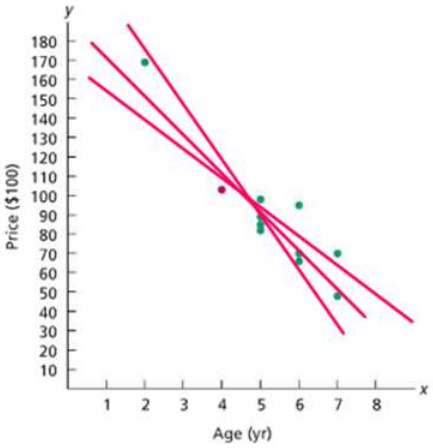
A1: The image clearly shows a negative relation.

Q2: Does it look like the relation is linear?

A2: Yes. It certainly looks like a straight line can describe the relation.

Q3: What's the graph of the linear relation?

A3: We can approximately the relation with many lines



Q4: Is there one line that is better than others?

A4: Need a criterion. The criterion that proved itself practical is **the least-squares criterion**.

Definition: Least-squares criterion

The **least-squares criterion** is that the line that best fits a set of data points is one having the smallest possible sum of squared errors.

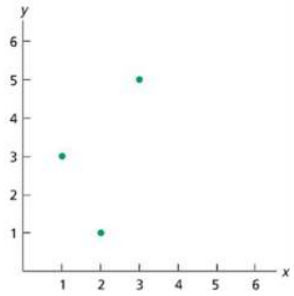
Definition: Regression line and regression equation

The line that describes a linear relation between any two variables is called the **regression line**, and its equation is called the **regression equation**.

So, our goal is to learn how to construct the regression line and find its equation from the data set such as in the example above.

To learn the procedure, let's consider the following small dataset along with its scatterplot:

Sample	x	y
1	1	3
2	2	1
3	3	5

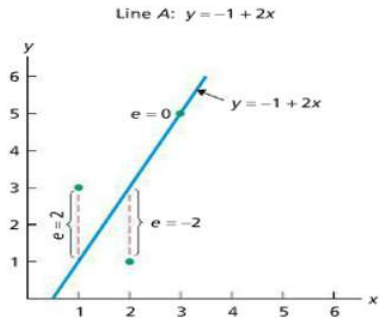


Let's try to fit a few lines $y = 2x - 1$ and $y = x + 1$. How do we decide which of the above lines is a better fit for the data according to the least–sum of squared errors criterion? Let's compare the "fits" of two different lines by finding the sum of squared errors for each line.

For each point (x_i, y_i), we find the predicted output $\hat{y}_i$ by evaluating the regression line for each input x_i; then we find the error e_i by subtracting the predicted output from the actual output observed; and then we find the squared error by squaring the error; and finally for each line we compute the sum of squared errors ($\sum$).

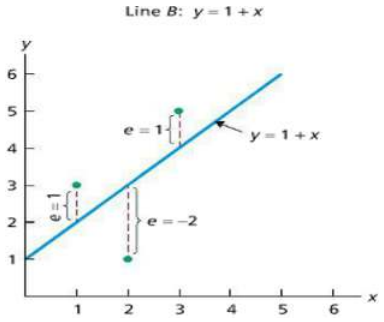
Line A

x_i	y_i	$\hat{y}_i = 2x_i - 1$	$e_i = y_i - \hat{y}_i$	e_i^2
1	3	1	$3 - 1 = 2$	$2^2 = 4$
2	1	3	$1 - 3 = -2$	$(-2)^2 = 4$
3	5	5	$5 - 5 = 0$	$0^2 = 0$
				$\sum = 8$



Line B

x_i	y_i	$\hat{y}_i = 2x_i - 1$	$e_i = y_i - \hat{y}_i$	e_i^2
1	3	2	$3 - 2 = 1$	$1^2 = 1$
2	1	3	$1 - 3 = -2$	$(-2)^2 = 4$
3	5	4	$5 - 4 = 1$	$1^2 = 1$
				$\sum = 6$



Based on this criterion we know that Line B is a better fit than Line A, but is it the best one?

Assume $y = mx + b$ is the line of best fit then the slope and the intercept can be determined by using the following formulas:

$$m = \frac{S_{xy}}{S_{xx}} \text{ and } b = \bar{y} - m\bar{x}$$

where

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xx} = \sum (x_i - \bar{x})^2$$

$$\bar{x} = \frac{\sum x_i}{n}$$

$$\bar{y} = \frac{\sum y_i}{n}$$

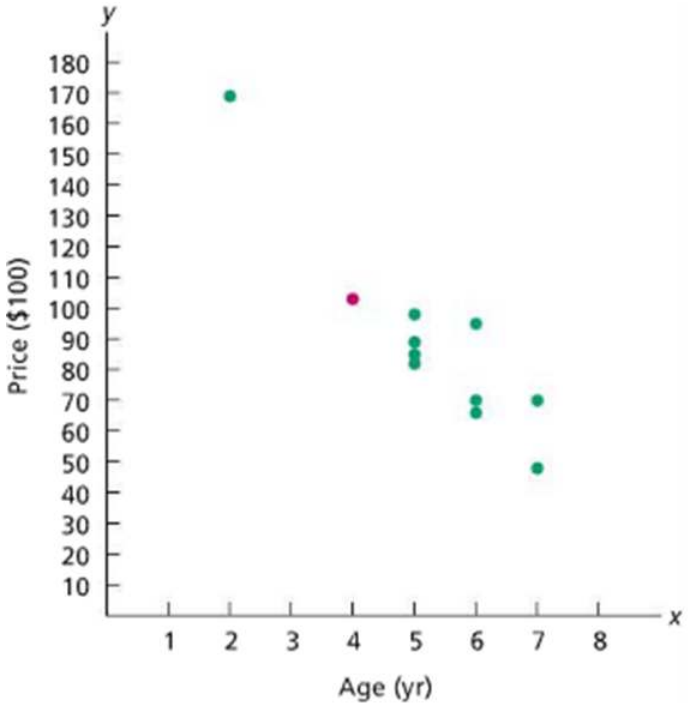
The process is outlined below in the form of the table:

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
1	3	-1	1	0	0
2	1	0	0	-2	0
3	5	1	1	2	2
$\bar{x}=2$	$\bar{y}=3$		$S_{xx}=\sum=2$		$S_{xy}=\sum=2$

$$m=\frac{S_{xy}}{S_{xx}}=\frac{2}{2}=1$$
$$b=\bar{y}-m\bar{x}=3-1\cdot 2=1$$
$$y=mx+b\rightarrow y=1\cdot x+1$$

Now let's use the same process to find the line of best fit for our original data set.

Car	Age (yr)	Price ($100)
1	5	85
2	4	103
3	6	70
4	5	82
5	5	89
6	5	98
7	6	66
8	6	95
9	2	169
10	7	70
11	7	48



The process will be summarized in the form of the table:

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
5	85	-0.27	0.07	-3.64	0.99
4	103	-1.27	1.62	14.36	-18.28
6	70	0.73	0.53	-18.64	-13.55
5	82	-0.27	0.07	-6.64	1.81
5	89	-0.27	0.07	0.36	-0.10
5	98	-0.27	0.07	9.36	-2.55
6	66	0.73	0.53	-22.64	-16.46
6	95	0.73	0.53	6.36	4.63

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
2	169	-3.27	10.71	80.36	-263.01
7	70	1.73	2.98	-18.64	-32.19
7	48	1.73	2.98	-40.64	-70.19
$\bar{x}=5.27$	$\bar{y}=8.64$	$S_{xx}=20.18$		$S_{xy}=-408.91$	

$$m=\frac{S_{xy}}{S_{xx}}=\frac{-408.91}{20.18}=-20.26$$

$$b=\bar{y}-m\bar{x}=88.636-(-20.26)\cdot 5.273=195.47$$

$$y=mx+b\rightarrow y=-20.26\cdot x+195.47$$

$$V=-20.26t+195.47$$

	Units	Interpretation
$m=-20.26$	\$100/year	depreciation rate is \$2026/year
$b=195.47$	\$100	the price of a new car is \$19547

We can use the equation in two different ways – to find the output for a given input and to find the input with a given output.

Task:	Find the output for a given input	Find the input with a given output
Given:	$t=5.5$	$V=40.00$
Solution:	$V=-20.26\cdot 5.5+195.47$ $V=-111.43+195.47$ $V=84.04$	$40.00=-20.26\cdot t+195.47$ $20.26\cdot t=195.47$ $t=7.7$
Interpretation:	A 5.5-year-old vehicle worth \$8404.	For \$4000, you can get a 7.7-year old vehicle.

A few more definitions.

Definition: Extrapolation

An **extrapolation** is trying to use the regression line outside of the domain which may lead to inaccurate results.

So technically, in our case the x and y intercepts are examples of extrapolation. Extrapolation may or may not make sense and it depends on the context.

Definition: Influential observation

An **influential observation** is an observation whose removal causes the regression equation to change considerably.

In our example the point $(2,169)$ is such an observation because if it is removed the equation of the regression line changes significantly.

Definition: Outlier

An **outlier** is an observation that lies too far from the regression line, relative to other data points.

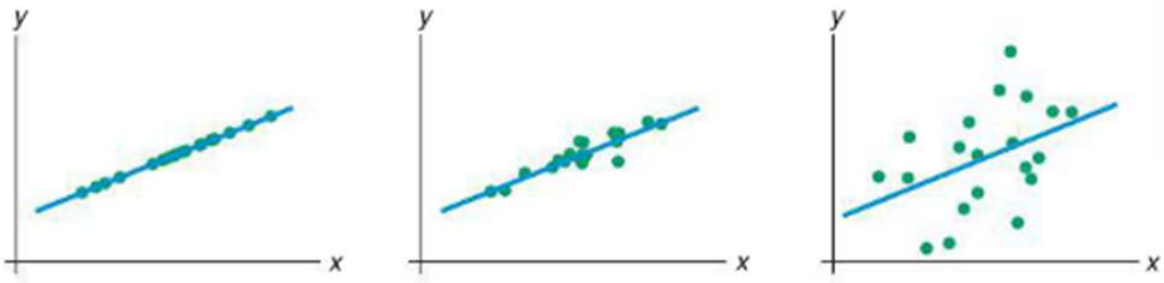
It is important to distinguish between influential observations and outliers!

Section 3: Correlation

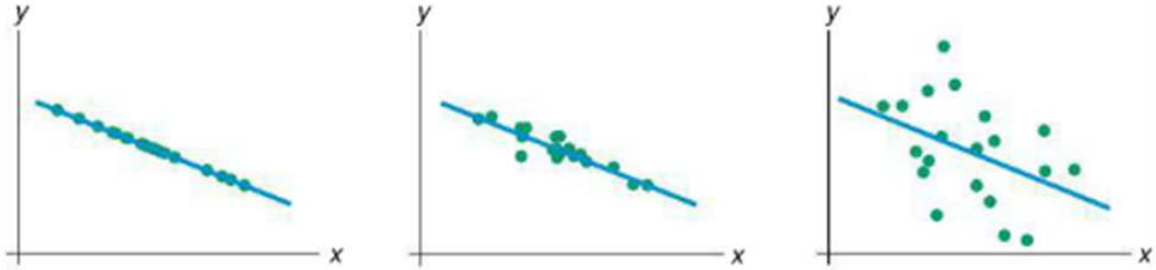


Now that we know how to find a regression line, we want to answer some interesting questions.

It appears that the formula can be applied to any data set, and it is true - here are the examples of the regression lines superimposed on various data sets.



As you can see all three data sets have the same linear regression, however there are some clear distinctions between the data sets. Here is another example of such behavior.



This means that some data sets have "stronger"/"more obvious" linear relations than others.

Definition: Coefficient of determination

To determine how strong the linear relation between the variables we introduce the coefficient of determination, r^2. To compute the coefficient of determination, we have the formula:

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

Example 1.15.1.3.1

Find the coefficient of determination of the line of best fit for the following dataset:

x	y
1	3
2	1
3	5

Solution

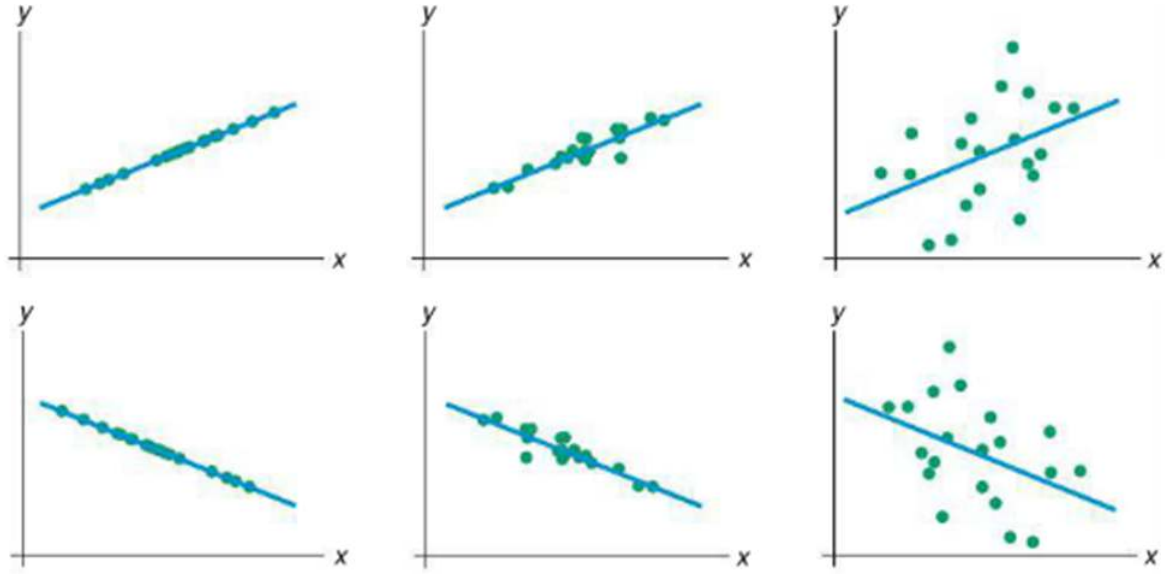
To find the coefficient of determination of the line of best fit we set up the following table:

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	3	-1	1	0	0	0
2	1	0	0	-2	4	0
3	5	1	1	2	4	2
$\bar{x}=2$	$\bar{y}=3$		$S_{xx}=\sum=2$		$S_{yy}=\sum=8$	$S_{xy}=\sum=2$

$$r^2=\frac{S_{xy}^2}{S_{xx}S_{yy}}=\frac{2^2}{2\cdot 8}=0.25$$

The coefficient of determination is a measure of how strong is the linear relation between the variables: $r=1$ corresponds to a perfect linear relation, and $r=0$ corresponds to the absence of any linear relation. In our example above, $r^2=0.25$ which is an indicator that linear relation between the variables for which the data were presented is not strong.

Can you guess which scatterplots below have the highest r^2? How about the lowest?



The charts on the left demonstrate a perfect linear relation so the coefficient of determination is equal to 1. The charts in the center and on the right demonstrate a less than perfect linear relation so the coefficient of determination is much less than 1.

While the scatterplots on the left have the same $r^2=1$ there is a clear difference between them. To quantify this difference, we use the linear correlation coefficient, r.

To compute the linear correlation coefficient, we have the formula:

$$r = \frac{\frac{1}{n-1} S_{xy}}{S_x S_y}$$

Example 1.15.1.3.2

Find the coefficient of determination of the line of best fit for the following dataset:

x	y
1	3
2	1
3	5

Solution

To find the linear correlation coefficient of the line of best fit we set up the following table:

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	3	-1	1	0	0	0
2	1	0	0	-2	4	0
3	5	1	1	2	4	2
$\bar{x}=2$	$\bar{y}=3$	$s_x = \sqrt{S_{xx}/(n-1)} = 1$		$s_y = \sqrt{S_{yy}/(n-1)} = 2$		$S_{xy} = \sum = 2$

$$r = \frac{\frac{1}{n-1} S_{xy}}{S_x S_y} = \frac{\frac{1}{3-1} \cdot 2}{1 \cdot 2} = 0.5$$

The linear correlation coefficient is another measure of how strong is the linear relation between the variables: $r = +1$ corresponds to a perfect positive linear relation, and $r = -1$ corresponds to a perfect negative linear relation. And the reason why it is denoted by r is because of the relation between the linear correlation coefficient and coefficient of determination:

$$(r)^2 = r^2$$

In our example above, $r = 0.5$ which is an indicator that we have a positive linear relation between the variables for which the data were presented.

We finished the development of the procedure that allows us to find and interpret the linear regression line along with the coefficient of determination and the linear correlation coefficient.

This page titled 1.15.1: Introduction to Regression Analysis is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

1.15.2: Linear Regression Analysis



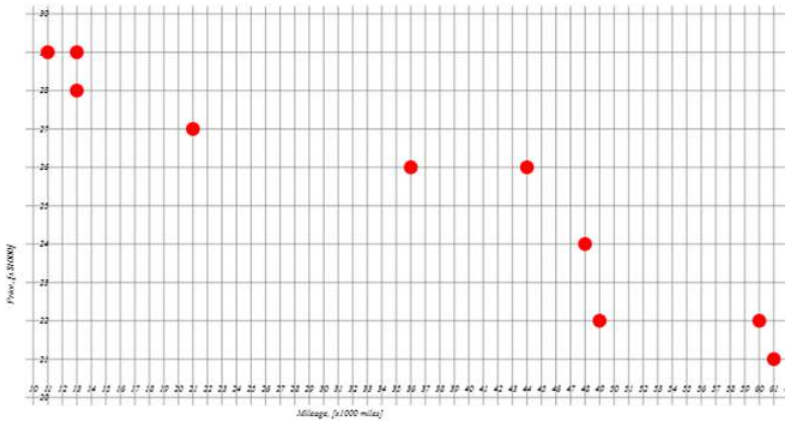
Consider the following data set that contains information about a sample of ten 2016 Subaru Outback offered for sale:

Vehicle	Mileage (x1000miles)	Price (x$1000)
1	11	29
2	13	29
3	13	28
4	21	27
5	36	26
6	44	26
7	48	24
8	49	22
9	60	22
10	61	21

Let's see what kind of questions we can answer using the linear regression analysis.

1. Is there an association between the two variables?

To see if there is an association and whether it is positive or negative, we need to construct a scatterplot – this can be easily done by hand or with technology. From the context it is clear that the price of a vehicle depends on the miles and not the other way around, so the mileage is the input variable on the horizontal axis, and the price is the output variable on the vertical axis. The scatter plot below summarizes the data regarding the mileage as the input variable and the price as the output variable:



From the scatterplot it can be observed that there is a negative association since the increase in mileage is associated with the decrease in value.

2. Find the slope, the y-intercept, and the coefficient of determination of the linear regression line.

For that we are going to fill out the template table:

x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
11	29	-24.6	605.16	3.6	12.96	-88.56
13	29	-22.6	510.76	3.6	12.96	-81.36
13	28	-22.6	510.76	2.6	6.76	-58.76
21	27	-14.6	213.16	1.6	2.56	-23.36
36	26	0.4	0.16	0.6	0.36	0.24
44	26	8.4	70.56	0.6	0.36	5.04
48	24	12.4	153.76	-1.4	1.96	-17.36
49	22	13.4	179.56	-3.4	11.56	-45.56
60	22	24.4	595.36	-3.4	11.56	-82.96
61	21	25.4	645.16	-4.4	19.36	-111.76
$\bar{x}=35.6$	$\bar{y}=25.4$	$S_{xx}=$	3484.4	$S_{yy}=$	80.4	$S_{xy}=-504.4$

Now that we have the entire table complete, we can find the equation of the linear regression line.

$$m=\frac{S_{xy}}{S_{xx}}=-\frac{504.4}{3484.4}=-0.145$$

$$b=\bar{y}-m\bar{x}=25.4-(-0.145)\cdot 35.6=30.553$$

$$y=mx+b\rightarrow y=-0.145\cdot x+30.553$$

We can also find the coefficient of determination using the formula

$$r^2=\frac{S_{xy}^2}{S_{xx}S_{yy}}=\frac{(-504.4)^2}{3484.4\cdot 80.4}=0.908$$

Next, we can interpret the results:

	Units	Interpretation
$m=-0.145$	\$/mile	depreciation rate is 14.5c/mile
$b=30.553$	\$1000	the price of a new car is \$30553

	Units	Interpretation
$r^2 = 0.902$	n/a	linear relation is strong

Now that we confirmed the strong relation and found the equation that estimates the relation between the two variables, we can answer the next question.

3. Estimate

   a. the price of a car with 27500 miles.

$$y = -0.145 \cdot 27.5 + 30.553 = 26.565$$

   The units of the output variable in the equation are 1000$, so

$$26.565 \text{is equivalent to } \$26566$$

   Thus, the price of a 2016 Subaru Outback with 27500 miles is $26566.

b. the mileage of a car whose price is $25000.

   Solve:

$$25 = -0.145x + 30.553$$

$$x = 38.297$$

   The units of the input variable in the equation are 1000 miles, so

$$38.297 \text{is equivalent to } 38297 \text{miles}$$

   Thus, the expected mileage of a 2016 Subaru Outback that cost $25000 is 38297 miles.

We just learned how to find and interpret the linear regression line and the coefficient of determination and use it in applications.

This page titled 1.15.2: Linear Regression Analysis is shared under a not declared license and was authored, remixed, and/or curated by Anton Butenko.

Index

A

accessible
InfoPage

C

collaborative
InfoPage

E

education
InfoPage

I

innovation
InfoPage

L

LibreTexts
InfoPage

O

OER
InfoPage

onlinelearning
InfoPage

T

textbooks
InfoPage

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: STAT C1000

Webpages: 88

All licenses found:

- Undeclared: 98.9% (87 pages)
- CC BY-SA 4.0: 1.1% (1 page)

By Page

- STAT C1000 - *CC BY-SA 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Interactive Lecture Notes for Introductory Statistics - *Undeclared*
    - 1.1: Introduction to Statistics - *Undeclared*
      - 1.1.1: Introduction to Statistics - *Undeclared*
      - 1.1.2: Sampling Designs - *Undeclared*
      - 1.1.3: Experimental Designs - *Undeclared*
      - 1.1.4: Types of Data - *Undeclared*
    - 1.2: Visual Summaries of Data - *Undeclared*
      - 1.2.1: Visual Summaries of Qualitative Data - *Undeclared*
      - 1.2.2: Visual Summaries of Quantitative Data - *Undeclared*
      - 1.2.3: Other Charts - *Undeclared*
      - 1.2.4: Distribution Shapes - *Undeclared*
      - 1.2.5: Organizing Bivariate Data - *Undeclared*
      - 1.2.6: Misleading Graphs - *Undeclared*
    - 1.3: Numerical Summaries of Data - *Undeclared*
      - 1.3.1: Measures of Center - *Undeclared*
      - 1.3.2: Parameter and Statistic - *Undeclared*
      - 1.3.3: Measures of Variation - *Undeclared*
      - 1.3.4: Interpreting the Mean and Standard Deviation - *Undeclared*
      - 1.3.5: Measures of Relative Standing - *Undeclared*
      - 1.3.6: Five-Number Summaries and Boxplots - *Undeclared*
    - 1.4: Probability Concepts - *Undeclared*
      - 1.4.1: Introduction to Probability - *Undeclared*
      - 1.4.2: Finding the Probability - *Undeclared*
      - 1.4.3: Probability Rules - *Undeclared*
      - 1.4.4: Probabilities from Frequency Tables - *Undeclared*
      - 1.4.5: Conditional Probability and Multiplication Rules - *Undeclared*
    - 1.5: Combinatorics Concepts - *Undeclared*
      - 1.5.1: Introduction to Combinatorics - *Undeclared*
      - 1.5.2: Permutations and Combinations - *Undeclared*
      - 1.5.3: Tossing a Coin - *Undeclared*
    - 1.6: Discrete Random Variables - *Undeclared*
      - 1.6.1: Discrete Random Variables - *Undeclared*
      - 1.6.2: Applications of Discrete Random Variables - *Undeclared*
      - 1.6.3: Binomial Random Variables - *Undeclared*
    - 1.7: Continuous Random Variables - *Undeclared*
      - 1.7.1: Continuous Random Variables - *Undeclared*
      - 1.7.2: Standard Types of Continuous Random Variables - *Undeclared*
      - 1.7.3: Other Continuous Random Variables - *Undeclared*
    - 1.8: Normal Distributions - *Undeclared*
      - 1.8.1: Normal Random Variables - *Undeclared*
      - 1.8.2: Applications of Normal Random Variables - *Undeclared*
      - 1.8.3: Normality Plots - *Undeclared*
      - 1.8.4: Approximating Binomial Random Variables - *Undeclared*
    - 1.9: The Central Limit Theorem - *Undeclared*
      - 1.9.1: The Central Limit Theorem for Sample Means - *Undeclared*
      - 1.9.2: The Central Limit Theorem for Sample Sums - *Undeclared*
      - 1.9.3: The Central Limit Theorem for Sample Proportions - *Undeclared*
    - 1.10: Estimating Parameters - *Undeclared*
      - 1.10.1: Estimating the Population Mean - *Undeclared*
      - 1.10.2: The One Mean T Procedure - *Undeclared*

- 1.10.3: The One Proportion Z Procedure - *Undeclared*
- 1.10.4: The One Variance Chi-Squared Procedure - *Undeclared*
- 1.10.5: The Summary of Confidence Intervals - *Undeclared*

- 1.11: Hypothesis Testing - *Undeclared*
  - 1.11.1: One Mean Z Test - *Undeclared*
  - 1.11.2: One Mean T Test - *Undeclared*
  - 1.11.3: One Proportion Z Test - *Undeclared*
  - 1.11.4: One Variance Chi-Squared Test - *Undeclared*
  - 1.11.5: The Summary of Hypothesis Testing for One Parameter - *Undeclared*
- 1.12: Inferences about Two Parameters - *Undeclared*
  - 1.12.1: Two Variances F Test - *Undeclared*
  - 1.12.2: Two Proportions Z Test - *Undeclared*
  - 1.12.3: Two Means Z Test - *Undeclared*
  - 1.12.4: Two Means T Pooled Test - *Undeclared*
  - 1.12.5: Two Means T Non-Pooled Test - *Undeclared*
  - 1.12.6: Two Paired Means T Test - *Undeclared*
  - 1.12.7: The Summary of Hypothesis Testing for Two Parameters - *Undeclared*
- 1.13: Chi-Squared Procedures - *Undeclared*
  - 1.13.1: The Goodness-of-Fit Test - *Undeclared*
  - 1.13.2: Homogeneity Test - *Undeclared*
  - 1.13.3: Test for Independence - *Undeclared*
- 1.14: ANOVA Procedures - *Undeclared*
  - 1.14.1: One-Way ANOVA - *Undeclared*
  - 1.14.2: Two-Way ANOVA - *Undeclared*
- 1.15: Regression Analysis - *Undeclared*
  - 1.15.1: Introduction to Regression Analysis - *Undeclared*
  - 1.15.2: Linear Regression Analysis - *Undeclared*

- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*
  - Detailed Licensing - *Undeclared*

Detailed Licensing

Overview

Title: STAT C1000

Webpages: 88

All licenses found:

- Undeclared: 98.9% (87 pages)
- CC BY-SA 4.0: 1.1% (1 page)

By Page

- STAT C1000 - *CC BY-SA 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Interactive Lecture Notes for Introductory Statistics - *Undeclared*
    - 1.1: Introduction to Statistics - *Undeclared*
      - 1.1.1: Introduction to Statistics - *Undeclared*
      - 1.1.2: Sampling Designs - *Undeclared*
      - 1.1.3: Experimental Designs - *Undeclared*
      - 1.1.4: Types of Data - *Undeclared*
    - 1.2: Visual Summaries of Data - *Undeclared*
      - 1.2.1: Visual Summaries of Qualitative Data - *Undeclared*
      - 1.2.2: Visual Summaries of Quantitative Data - *Undeclared*
      - 1.2.3: Other Charts - *Undeclared*
      - 1.2.4: Distribution Shapes - *Undeclared*
      - 1.2.5: Organizing Bivariate Data - *Undeclared*
      - 1.2.6: Misleading Graphs - *Undeclared*
    - 1.3: Numerical Summaries of Data - *Undeclared*
      - 1.3.1: Measures of Center - *Undeclared*
      - 1.3.2: Parameter and Statistic - *Undeclared*
      - 1.3.3: Measures of Variation - *Undeclared*
      - 1.3.4: Interpreting the Mean and Standard Deviation - *Undeclared*
      - 1.3.5: Measures of Relative Standing - *Undeclared*
      - 1.3.6: Five-Number Summaries and Boxplots - *Undeclared*
    - 1.4: Probability Concepts - *Undeclared*
      - 1.4.1: Introduction to Probability - *Undeclared*
      - 1.4.2: Finding the Probability - *Undeclared*
      - 1.4.3: Probability Rules - *Undeclared*
      - 1.4.4: Probabilities from Frequency Tables - *Undeclared*
      - 1.4.5: Conditional Probability and Multiplication Rules - *Undeclared*
    - 1.5: Combinatorics Concepts - *Undeclared*
      - 1.5.1: Introduction to Combinatorics - *Undeclared*
      - 1.5.2: Permutations and Combinations - *Undeclared*
      - 1.5.3: Tossing a Coin - *Undeclared*
    - 1.6: Discrete Random Variables - *Undeclared*
      - 1.6.1: Discrete Random Variables - *Undeclared*
      - 1.6.2: Applications of Discrete Random Variables - *Undeclared*
      - 1.6.3: Binomial Random Variables - *Undeclared*
    - 1.7: Continuous Random Variables - *Undeclared*
      - 1.7.1: Continuous Random Variables - *Undeclared*
      - 1.7.2: Standard Types of Continuous Random Variables - *Undeclared*
      - 1.7.3: Other Continuous Random Variables - *Undeclared*
    - 1.8: Normal Distributions - *Undeclared*
      - 1.8.1: Normal Random Variables - *Undeclared*
      - 1.8.2: Applications of Normal Random Variables - *Undeclared*
      - 1.8.3: Normality Plots - *Undeclared*
      - 1.8.4: Approximating Binomial Random Variables - *Undeclared*
    - 1.9: The Central Limit Theorem - *Undeclared*
      - 1.9.1: The Central Limit Theorem for Sample Means - *Undeclared*
      - 1.9.2: The Central Limit Theorem for Sample Sums - *Undeclared*
      - 1.9.3: The Central Limit Theorem for Sample Proportions - *Undeclared*
    - 1.10: Estimating Parameters - *Undeclared*
      - 1.10.1: Estimating the Population Mean - *Undeclared*
      - 1.10.2: The One Mean T Procedure - *Undeclared*

- 1.10.3: The One Proportion Z Procedure - *Undeclared*
- 1.10.4: The One Variance Chi-Squared Procedure - *Undeclared*
- 1.10.5: The Summary of Confidence Intervals - *Undeclared*

- 1.11: Hypothesis Testing - *Undeclared*
  - 1.11.1: One Mean Z Test - *Undeclared*
  - 1.11.2: One Mean T Test - *Undeclared*
  - 1.11.3: One Proportion Z Test - *Undeclared*
  - 1.11.4: One Variance Chi-Squared Test - *Undeclared*
  - 1.11.5: The Summary of Hypothesis Testing for One Parameter - *Undeclared*
- 1.12: Inferences about Two Parameters - *Undeclared*
  - 1.12.1: Two Variances F Test - *Undeclared*
  - 1.12.2: Two Proportions Z Test - *Undeclared*
  - 1.12.3: Two Means Z Test - *Undeclared*
  - 1.12.4: Two Means T Pooled Test - *Undeclared*
  - 1.12.5: Two Means T Non-Pooled Test - *Undeclared*
  - 1.12.6: Two Paired Means T Test - *Undeclared*
  - 1.12.7: The Summary of Hypothesis Testing for Two Parameters - *Undeclared*
- 1.13: Chi-Squared Procedures - *Undeclared*
  - 1.13.1: The Goodness-of-Fit Test - *Undeclared*
  - 1.13.2: Homogeneity Test - *Undeclared*
  - 1.13.3: Test for Independence - *Undeclared*
- 1.14: ANOVA Procedures - *Undeclared*
  - 1.14.1: One-Way ANOVA - *Undeclared*
  - 1.14.2: Two-Way ANOVA - *Undeclared*
- 1.15: Regression Analysis - *Undeclared*
  - 1.15.1: Introduction to Regression Analysis - *Undeclared*
  - 1.15.2: Linear Regression Analysis - *Undeclared*

- Back Matter - *Undeclared*
  - Index - *Undeclared*
  - Glossary - *Undeclared*
  - Detailed Licensing - *Undeclared*
  - Detailed Licensing - *Undeclared*