THE STORIES BEHIND SOCIAL STATISTICS: DATA ANALYSIS, INTERPRETATION, AND COMMUNICATION



Ezra Temko
Southern Illinois University Edwardsville

LibreTexts™

Southern Illinois University Edwardsville

The Stories Behind Social Statistics: Data Analysis, Interpretation, and Communication

Ezra Temko

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

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden 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 04/16/2026

TABLE OF CONTENTS

Licensing

1: Data, Datasets, SPSS, and Frequency Distribution Tables

2: Levels of Measurement; Five Point Summaries

3: More Measures of Central Tendency and Spread- Mean, Standard Deviation, Z-scores

4: Histograms, Normal Distributions, and the Central Limit Theorem

5: Confidence Intervals

6: Comparing Confidence Intervals

7: Hypothesis Tests, T-test, & ANOVA

8: Crosstabs and Chi-square

- 8.1: Crosstabs for Research Methods

9: Elaborated Crosstabs

10: Chapter 10. Scatterplots and Correlation

11: Regression Equations

12: Linear Regression, continued

13: Linear Regression with Reference Group Set Independent Variables

14: Multivariate Linear Regression

15: Interactions

16: Appendix- Resources

Index

Glossary

Detailed Licensing

Licensing

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

1: Data, Datasets, SPSS, and Frequency Distribution Tables

A. Welcome to statistics!

1.1. Statistics in action: movies and monarchy



I have an 8 year old, and so watch my fair share of kids movies. I kept feeling annoyed that these movies featured monarchies and not democracies. Even the happy endings did not seem to problematize rule by monarch. Why was this being valorized in the United States, and in 2023? However, perhaps as a sociologist I was just more sensitized to this, and/or it irked me so it stood out. How common is it really for these movies to feature royaldoms? If I wanted an answer, I knew… I'd have to turn to **statistics**!

Definition: Statistics

Statistics: data, the practice of using data for research

Statistics is useful for unveiling group patterns and variability. Statistics is an important gateway into understanding and describing the social world. Statistics can advance our knowledge, improve decision-making, and enable us to make statements about populations. Statistics are used in all kinds of literature, and by all kinds of people trying to influence you, so being able to evaluate and understand statistical claims being made is also a critical skill.

B. Unit of Analysis

Definition: Unit of analysis

Unit of analysis: Who or what you are analyzing

What is the entity your data is about (what does each case represent)? What are you trying to make claims about? These answers need to match, and their answer is your unit of analysis.

1.2. Statistics in Action: Movies and Monarchy

What would I need to analyze to find out how common it is for kids movies to feature royaldoms?

Would I be looking at the characteristics of countries? No!

Individual people? No!

I would need to analyze movies. The **unit of analysis** for my study is individual movies.

When you look at a dataset, the unit of analysis is what each row (or case) represents—be it individual people, states, books, countries, songs, or something else. Often survey data is going to be of individual people, but even then the data may have been aggregated (grouped) such that you have data at a different level of analysis. For example, if you have graduation rates for different schools, the school is the unit of analysis.

Example: Unit of Analysis Runaway argument for \eatSpaces?

The Institute on Taxation and Economic Policy (ITEP) analyzed the federal income taxes paid by S&P 500 and Fortune 500 corporations. They found that 39 profitable corporations collectively generated $122 billion in profit from 2018 to 2020 yet paid no federal income tax. According to ITEP, "T-Mobile reported the largest profits. It reported $11.5 billion in profits

over this time but had a federal income tax liability of negative $80 million, meaning the company received $80 million in tax refunds."

39 Corporations Avoided Federal Income Tax from 2018 through 2020

figures in millions of dollars

Company Name	U.S. Pre-Tax Income	Current Federal Income Tax	Effective Federal Tax Rate	U.S. Pre-Tax Income			Current Federal Income Tax			Effective Federal Tax Rate		
	2018-2020	2018-2020	2018-2020	2020	2019	2018	2020	2019	2018	2020	2019	2018
Advanced Micro Devices	$1,655	$-1	-0.1%	$1,208	$333	$114	$0	$-13	$12	$0%	$-4%	11%
Agilent Technologies	402	-165	-41.0%	49	185	168	5	-191	21	10%	-103%	13%
Alliant Energy	1,746	-7	-0.4%	559	612	575	1	-7	-1	0%	-1%	0%
Ally Financial	4,823	-14	-0.3%	1,334	1,902	1,587	0	-2	-12	0%	0%	-1%
Ameren	3,064	-12	-0.4%	1,032	997	1,035	2	-4	-10	0%	0%	-1%
American Electric Power	5,936	-177	-3.0%	2,163	1,830	1,943	-138	-7	-32	-6%	0%	-2%
Archer Daniels Midland	2,130	-2	-0.1%	438	745	947	-164	37	125	-37%	5%	13%
Atmos Energy	1,989	-10	-0.5%	747	642	600	0	0	-10	0%	0%	-2%
Ball	598	-11	-1.8%	193	217	188	-33	-1	23	-17%	0%	12%
Booz Allen Hamilton Holding	1,673	-196	-11.7%	623	561	489	-227	-3	34	-36%	0%	7%
Cabot Oil & Gas	1,828	-157	-8.6%	240	896	691	-32	-30	-95	-13%	-3%	-14%
Celanese	2,262	-135	-6.0%	1,530	252	480	13	-6	-142	1%	-2%	-30%
CMS Energy	2,488	-133	-5.3%	885	829	774	-35	-31	-67	-4%	-4%	-9%
DISH Network	6,577	-13	-0.2%	2,532	1,900	2,145	-231	173	44	-9%	9%	2%
Dominion Resources	6,137	-327	-5.3%	1,492	1,624	3,021	-314	32	-45	-21%	2%	-1%
DTE Energy	4,063	-448	-11.0%	1,531	1,317	1,215	-247	-184	-17	-16%	-14%	-1%
Duke Energy	7,908	-1,227	-15.5%	826	4,053	3,029	-281	-299	-647	-34%	-7%	-21%
Edison International	3,309	-44	-1.3%	588	1,121	1,600	13	0	-57	2%	0%	-4%
Evergy	2,077	-134	-6.4%	722	758	597	-27	-40	-67	-4%	-5%	-11%
FedEx	8,247	-138	-1.7%	4,716	1,218	2,312	199	-230	-107	4%	-19%	-5%
FirstEnergy	3,696	-46	-1.2%	1,108	1,093	1,495	-14	-16	-16	-1%	-1%	-1%
Juniper Networks	617	-47	-7.5%	184	282	152	73	6	-126	40%	2%	-83%
Kinder Morgan	4,910	-44	-0.9%	654	2,472	1,784	-20	-2	-22	-3%	0%	-1%
Mohawk Industries	611	-20	-3.3%	87	151	373	-34	19	-6	-39%	13%	-1%
NRG Energy	1,958	0	0.0%	727	769	462	0	0	0	0%	0%	0%
Oneok	3,967	0	0.0%	801	1,650	1,516	1	-1	0	0%	0%	0%
Penske Automotive Group	1,322	-68	-5.1%	505	424	393	-78	24	-13	-15%	6%	-3%
PPL	2,933	-38	-1.3%	878	945	1,110	-9	-10	-19	-1%	-1%	-2%
Principal Financial	4,306	-7	-0.2%	1,318	1,347	1,641	16	32	-55	1%	2%	-3%
Salesforce.com	4,083	-4	-0.1%	2,630	653	800	-12	8	0	0%	1%	0%
Sanmina-SCI	262	0	-0.1%	95	151	16	-1	1	0	-1%	1%	-1%
SpartanNash	123	-1	-0.5%	80	3	40	2	-1	-2	2%	-33%	-4%
Telephone & Data Systems	589	-150	-25.5%	241	170	178	-175	15	10	-73%	9%	6%
Textron	1,402	-46	-3.3%	202	652	548	-1	-48	3	0%	-7%	1%
T-Mobile US	11,519	-80	-0.7%	3,409	4,487	3,623	-17	-24	-39	0%	-1%	-1%
UGI	1,130	-36	-3.2%	420	264	446	-85	52	-3	-20%	20%	-1%
Westlake Chemical	1,737	-30	-1.7%	227	451	1,059	-208	20	158	-92%	4%	15%
Williams	3,187	-153	-4.8%	278	1,075	1,834	-29	-41	-83	-10%	-4%	-5%
Xcel Energy	4,395	-63	-1.4%	1,465	1,496	1,434	-13	-16	-34	-1%	-1%	-2%
TOTAL	**$121,659**	**$ -4,182**	**-3.4%**									

Source: Institute on Taxation and Economic Policy analysis of Securities and Exchange Commission filings of publicly traded corporations

Source: https://itep.org/corporate-tax-avoid...-and-jobs-act/

What is the unit of analysis for this research?

What is ITEP analyzing? Here it is individual companies. Each row of the table is a company. They collected data on federal income taxes for each company.

While a study might give you ideas for future research at a different unit of analysis, you can only make claims at the unit of analysis you are using. For example, if I am studying university graduation rates (unit of analysis = individual universities), I may know that Harvard University has one of the highest graduation rates, around 98%. However, if I meet someone who goes to Harvard, I should not assume they will graduate, or if someone tells me they studied at Harvard, I should not assume they graduated (e.g., Bill Gates, Matt Damon, Mark Zuckerberg, William Randolph Hearst, and Pete Seeger all studied at Harvard but did not graduate). Likewise, I should not assume that the U.S. higher education graduation rate is comparable. (In reality, 64% of first-time, full-time undergraduate students who began seeking a bachelor's degree at a 4-year degree-granting institution in Fall 2014 had their degree within six years.)

C. Variables

Definition: Variable

Variable: Anything that can vary (differ), as opposed to something that must be constant (the same).

Other words you will come across that have a form of the word variable in them (variate) include:
univariate: having to do with one (uni) variable
bivariate: having to do with two (bi) variables
multivariate: having to do with multiple (multi) variables

Variables can be all kinds of things, such as height, temperature, birth weight, favorite color, income, partisan affiliation, number of pets in household, state voter turnout, etc. Generally large social datasets about individuals will all have variables like age, gender, and race. Age is a variable because not everyone has to have the same age.

If you are trying to ask a research question to find out if a relationship exists, you need to have variables that vary. If your variable is a constant, you won't be able to find out relationships. For example, if you want to know if there is a relationship between age and political ideology, but all of the people in your dataset are the same age, you won't be able to tell if political ideology varies as age changes.

1.3. Statistics in Action: Movies and Monarchy

For my study on movies and monarchy, the **variables** are the items I am going to collect attributes/characteristics regarding for each of the movies. This includes things like movie name, type of government represented in the movie, company that made the movie, and year the movie was released.

D. Data & Datasets

- It is estimated that less than 10% of U.S. drivers don't wear seatbelts, but that drivers without seatbelts constitute 50% of all vehicle passenger deaths.
- The average daily intake of plain water in the United States is about 4 cups per person.
- If current trends continue, within the next half-century Christians will constitute a minority of U.S.-Americans.

You have probably come across statistics that sound like these throughout your life. But where did they come from? These statistics were all the result of analyzing **data**, factual information that could be analyzed. The driver statistics came from the National Highway Traffic Safety Administration's (NHTSA) National Occupant Protection Use Survey and from Fatality Analysis Reporting System data NHTSA collects from every state government with information on all fatal crashes. The water statistic came from the National Health and Nutrition Examination Survey. The religion statistic was based on Pew Research Center analysis of their American Trends Panel, other Pew surveys, and NORC's General Social Survey. To do statistical analysis, first you need collected data. You can collect your own data through original research, or use data that has already been collected and conduct secondary data analysis.

Definition: Data

Data: Information. In statistics, usually the values for variable(s).
Data is a plural word. The singular is datum.

Example: Raw data Runaway argument for \eatSpaces?

If you asked 10 friends what religion they are, their responses constitute your raw data. For example, your raw data might look like this: {Unitarian, secular, refused to say, Protestant, Reconstructionist Jewish, Christian, recovering Catholic, Hindu, Muslim, half-Catholic half-Protestant}.

1.4. Statistics in Action: Movies and Monarchy

Back to my movies and monarchy study. I took a look at the top 20 kids films between 2015 (when my kiddo was born) and August 2023 (when I'm writing this). I determined what I'm considering the top 20 films for the purposes of this study by using Nash Information Services' data on the 25 top-grossing children's films ever in the United States, adjusted for ticket-price inflation (11 were from 2015-2023), and then added 9 more movies from their data on the top-grossing children's films worldwide (I added the first 9 films not already listed that were from 2015-2023, which were in their top 47 list). I collected information for each of these movies. I put my raw data into a spreadsheet using Microsoft Excel to keep track of them. Here's what my data looks like:

Movie Name	Monarchy Govt.?	Year	Company
The Boss Baby	no	2017	Dreamworks, Fox
Coco	no	2017	Pixar, Disney
Despicable Me 3	no	2017	Universal, Illumination
Finding Dory	no	2016	Pixar, Disney
Frozen II	yes	2019	Disney
Hotel Transylvania 3: Summer Vacation	no	2018	Sony
How to Train Your Dragon: The Hidden World	yes	2019	Dreamworks, Universal Pictures
Incredibles 2	no	2018	Pixar, Disney
Inside Out	no	2015	Pixar, Disney
The Lion King	yes	2019	Disney, Fairview
The Little Mermaid	yes	2023	Disney, Lucamar Productions, Marc Platt Productions
Minions	yes	2015	Universal, Illumination
Minions: The Rise of Gru	no	2022	Universal, Illumination
Moana	comparable	2016	Disney
Ralph Breaks The Internet	appearance	2018	Disney
The Secret Life of Pets	no	2016	Universal, Illumination
Sing	no	2016	Universal, Illumination
The Super Mario Bros. Movie	yes	2023	Universal, Illumination, Nintendo
Toy Story 4	no	2019	Pixar, Disney
Zootopia	no	2016	Disney

When I put a collection of data together in an organized fashion that can be used for analysis, it is called a **dataset**. In statistics, your dataset will be one file in spreadsheet form. A qualitative dataset could consist of something different like a set of interview transcripts.

Definition: Dataset

Dataset: A collection of data organized for analysis

Because we're dealing with a small population of movies, I could do my own mathematical calculations, for example to figure out what percent of the movies featured monarchies. However, you are limited in what analyses you can do by hand, and what happens

when you have a dataset with hundreds or thousands of cases? Either way, I would rather make the computer do the math for me. It's easier and more likely to be accurate. There are a number of statistical software programs. In sociology, the most popular ones include: SPSS, Stata, SAS, and R. One advantage of R is that it is free to use. This textbook uses SPSS. Specifically, the instructions for SPSS and the outputs shared here use SPSS Version 29. Other versions will be similar but sometimes have small differences. I like SPSS because I think it has the easiest learning curve. Its interface looks very similar to other applications you might use, and you can use drop-down menus to run commands and/or use its programming language. While you need a license to use SPSS, as an alternative, you can do almost all the analyses in this textbook using an open-source statistical software called PSPP, which is very similar to SPSS. It's available for download here.

When you use statistical software, you want a dataset where all of your data is represented by numbers, not labels like "refused to say" or "Disney." The numbers that represent certain characteristics, values, or categories are called **codes**.

Definition: Codes

Codes: Numbers that represent variable attributes
Codebook: A document that tells you the codes for each variable in your dataset and what attributes they represent

When we enter data into a dataset to use in SPSS, we engage in coding, meaning we assign codes for each category of our variable. We then enter numbers into our dataset. For any given variable, a code matches particular numbers to particular responses. A **codebook** tells you the coding scheme for each variable—what the codes are for each variable and what they represent.

1.5 Statistics in Action

For my movies and monarchy study, I needed to get my dataset ready to analyze using SPSS. I replaced all of the text responses with numbers. I did not have to replace anything for year, because that was already a number. Here is my codebook:

Codebook

Title
Name of Movie
1=The Boss Baby
2=Coco
3=Despicable Me 3
4=Finding Dory
5=Frozen II
6=Hotel Transylvania 3: Summer Vacation
7=How to Train Your Dragon: The Hidden World
8=Incredibles 2
9=Inside Out
10=The Lion King
11=The Little Mermaid
12=Minions
13=Minions: The Rise of Gru
14=Moana
15=Ralph Breaks The Internet
16=The Secret Life of Pets
17=Sing
18=The Super Mario Bros. Movie
19=Toy Story 4
20=Zootopia

MonarchyPresent
Is monarchy a central form of government for the film's setting? Is there monarchy in the film?

0=no
1=appearance
2=comparable
3=yes

Year
Year of movie's release

Company
Company/companies that produced and/or released the film
1=Dreamworks and Fox
2=Pixar and Disney
3=Universal and Illumination
4=Disney
5=Sony
6=Dreamworks and Universal
7=Disney and Fairview
8=Disney, Lucamar, and Marc Platt
9=Universal, Illumination, and Nintendo

So for example, for the Company variable, for Hotel Transylvania 3, I replaced "Sony" with "5".

Here's what my spreadsheet looked like after that:

Title	MonarchyPresent	Year	Company
1	0	2017	1
2	0	2017	2
3	0	2017	3
4	0	2016	2
5	3	2019	4
6	0	2018	5
7	3	2019	6
8	0	2018	2
9	0	2015	2
10	3	2019	7
11	3	2023	8
12	3	2015	3
13	0	2022	3
14	2	2016	4
15	1	2018	4
16	0	2016	3
17	0	2016	3
18	3	2023	9
19	0	2019	2
20	0	2016	4

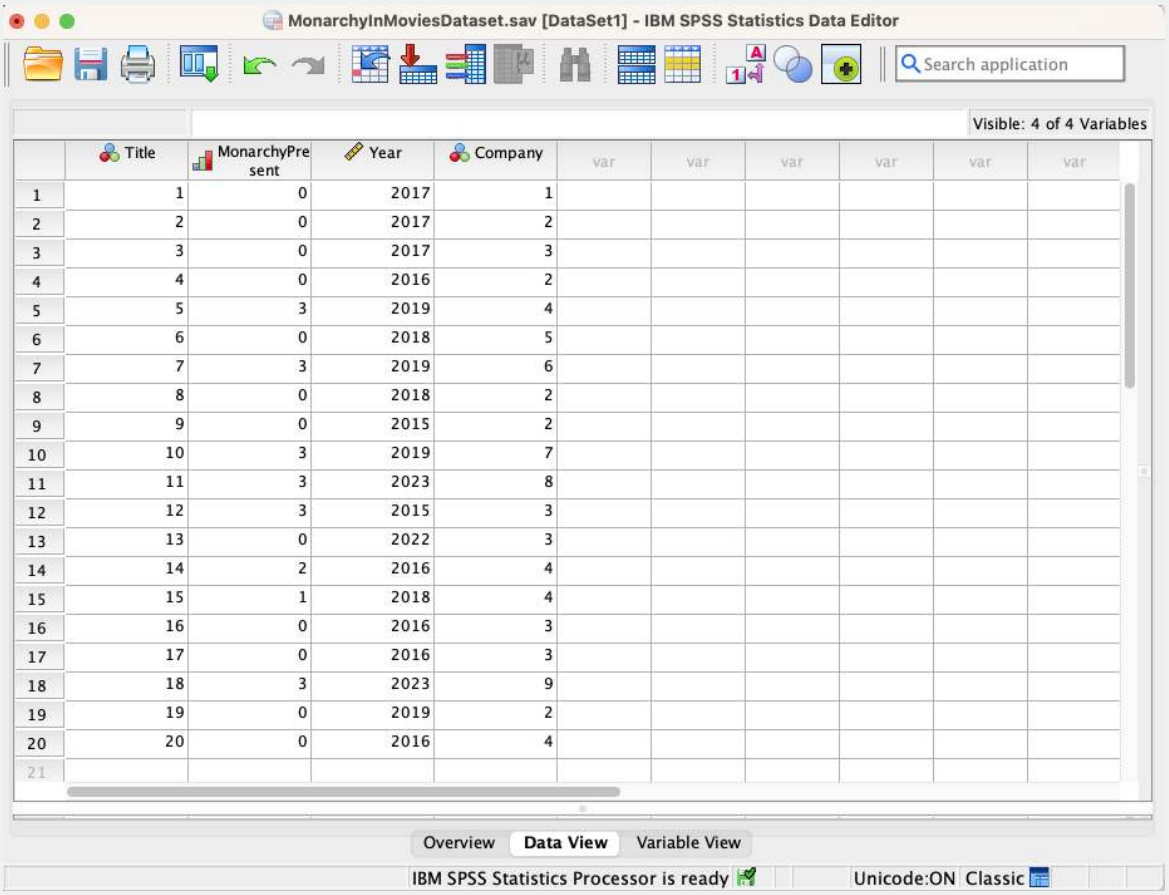
SPSS datasets are files that end in .sav (System for Analysis of Variables format). I can open up my Excel spreadsheet in SPSS as well, but all the data files specific to SPSS will end in .sav. To open up a dataset in SPSS, when you are in SPSS, go to File → Open → Data, and select your file.

SPSS datasets have a tab on the bottom that allows you to toggle between different views of the dataset. One view is called the Data View, which looks similar to the spreadsheet above. Each column is a unique variable and each row is a unique case. The other view is called Variable View, which gives you information about the variables: their name, "label," which is a description of the variable that will appear in many outputs, "values," which is the coding for the variable, and "missing," which will specify if any codes are supposed to be excluded from analyses.

1.6 Statistics in Action: Monarchy in Movies

Here is what my Monarchy in Movies dataset looks like in SPSS, in Data View and Variable View.

Data View:



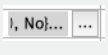
	Title	MonarchyPresent	Year	Company
1	1	0	2017	1
2	2	0	2017	2
3	3	0	2017	3
4	4	0	2016	2
5	5	3	2019	4
6	6	0	2018	5
7	7	3	2019	6
8	8	0	2018	2
9	9	0	2015	2
10	10	3	2019	7
11	11	3	2023	8
12	12	3	2015	3
13	13	0	2022	3
14	14	2	2016	4
15	15	1	2018	4
16	16	0	2016	3
17	17	0	2016	3
18	18	3	2023	9
19	19	0	2019	2
20	20	0	2016	4

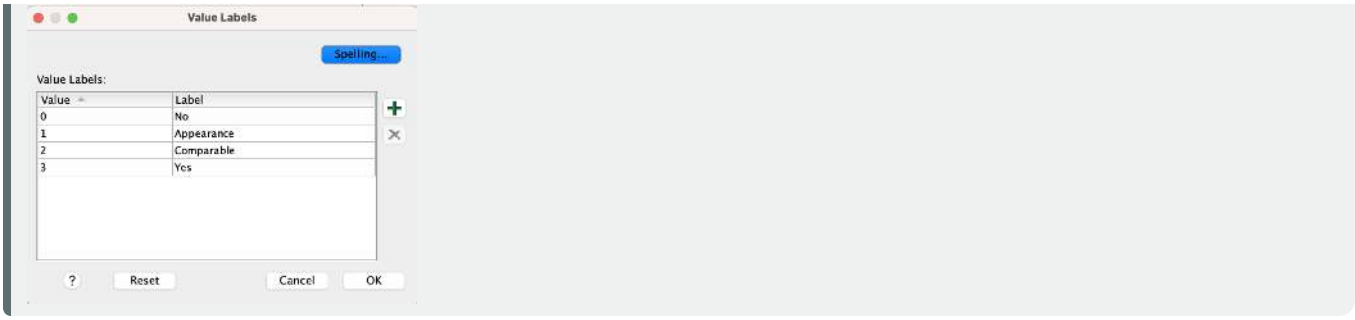
Using my codebook, I know that, for example, when I see a 2 in the Title column, that's Coco.

Variable View:



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Title	Numeric	3	0	Name of Movie	{1, The Bo...	None	11	Right	Nominal	Input
2	MonarchyP...	Numeric	2	0	Is monarchy a central form of government for the film's setting? Is there monarchy in the film?	{0, No}...	None	11	Right	Ordinal	Input
3	Year	Numeric	5	0	Year of movie's release	None	None	11	Right	Scale	Input
4	Company	Numeric	2	0	Company/companies that produced and/or released the film	{1, Dream...	None	11	Right	Nominal	Input

In Variable view, if I go to one of the Values column cells, I can click on a grey box with an ellipsis (three dots) in it. Here is the Values cell for the MonarchyPresent variable: {0, No}... . Once I clicked on the cell, it looked like this:  I clicked on the ellipsis, and a window popped up that showed me the coding for the variable:



When there is no data for a particular case for a particular variable, in SPSS Data View you will see a . instead of a number in that particular cell. Answers can also be coded as "missing" if they are supposed to be excluded from your analysis.

Most of the time when you are told stories about data, you are not going to encounter the actual datasets. You are going to hear a story about aggregate, analyzed data. However, if you want to do your own data analysis, you need a dataset. You are learning social statistics, so you are going to spend a lot of time with datasets. Datasets can come from data you collect yourself (e.g., your own original survey or content analysis), or you can engage in secondary data analysis—when you use existing data someone else already collected, but do your own analysis. There are plenty of public datasets available. Some with great social data include: the General Social Survey, U.S. Census Bureau datasets, the World Values Survey, Pew Research Center datasets, Uniform Crime Reporting Program data, the International Social Survey Programme, and American National Election Studies data. One popular source for finding data sets is the Interuniversity Consortium for Political and Social Research (ICPSR); ICPSR has over 3500 data sets researchers can search through to potentially find one that will address their research question.

E. Frequency Distributions

In social statistics, frequency is synonymous with count. In some disciplines, the term frequency refers only to rates.

A **frequency distribution** table shows you each category of a variable and the number of cases for each category.

Title: Frequency Distribution of Variable

Categories	Frequency	Percentage	Cumulative %
Category A	w	h	h
Category B	x	i	h+i
Category C	y	j	h+i+j
Category D	z	k	100%
Total	sum	100%	

Source: abc

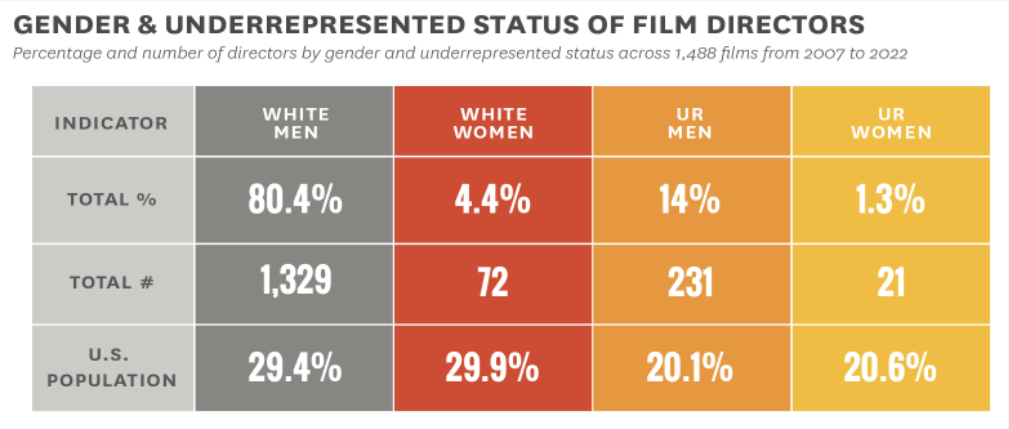
In addition to the count or frequency, frequency distribution tables have percentages and cumulative percentages. **Percent** means "out of 100." **Cumulative** is the total amount when you add everything together. So above, the cumulative percentage for Category B would be the total percentage for both Category A and Category B.

Example: Frequency tables Runaway argument for \eatSpaces?

When you see frequency tables in publications, they may be formatted differently.

Here is an example from a USC Annenberg Inclusion Initiative report by Stacy L. Smith, Katherine Pieper, and Sam Wheeler with an intersectional analysis of gender and race for directors of the 100 top-grossing domestic fictional films in North

America from 2007 to 2022 (fewer films were included in 2020 and 2021 due to pandemic-related performance).



In this table, you can see the total number of directors for each demographic (frequencies), what percentage of directors were from that demographic, and also the percentage of the U.S. population for that demographic, for comparison purposes. This table quickly captures social patterns of the underrepresentation of women and people of color among director positions.

The table is presented in the full report as well, in a less stylized format:

Table 6
Gender & Underrepresented Status of Directors from 2007-2022

Directors	White Males	White Females	UR Males	UR Females
Total %	80.4%	4.4%	14%	1.3%
Total #	1,329	72	231	21
U.S. Census[7]	29.4%	29.9%	20.1%	20.6%

7. U.S. Census (2021). Quick facts. Retrieved December 7, 2022 from https://www.census.gov/quickfacts/fa...e/US/PST045221.

Source: https://assets.uscannenberg.org/docs/aii-inclusion-directors-2023.pdf

When you create a frequency table, you want to select a variable that does not have too many categories. Imagine if the Annenberg Report had included a frequency table with statistics for each of the 1,488 films! That would be a really long table. If you were making a frequency table broken down by age, and have a range of adult ages, you would likely want to collapse the age variable into age groups (e.g., Google Analytics does age reporting based on the groups 18-24, 25-34, 35-44, 45-54, 55-64, and 65+). If you were looking at voting preferences, you might combine any candidates with less than 1% support into an "other" category. If there are too many categories, you should likely either collapse them into fewer grouped categories. There are times you may want a larger frequency table, such as if you are doing a state-by-state comparison and analysis. However, even then such a table is probably more useful to have and provide for reference and less useful for highlighting the main story of your findings.

SPSS: Frequency table Runaway argument for \eatSpaces?

To make a frequency distribution table in SPSS, you can either use the menus in the SPSS application (i.e., the drag-and-drop / point-and-click method), or you can use **syntax**, SPSS's programming language. Once you tell SPSS to generate the frequency table, it will open in a new *Output* window where all your outputs will go.

Menu method: In SPSS, go to Analyze → Descriptive Statistics → Frequencies. Make sure there is a check mark next to "Display frequency tables." Select your variable from the left window/box and either drag it to the right window ("Variable(s):") or click the blue arrow to bring it over. Then click OK.

Syntax method: You can open the provided syntax file (see online resources, and go File → Open → Syntax, then select the .sps file) and edit the variable name. Alternatively, you can create a new syntax file (go File → New → Syntax) and then type

in the language below, substituting the actual name of the variable for "VariableName". One nice feature of SPSS syntax language is that variable names are not case-sensitive.

FREQUENCIES VARIABLES=VariableName
/ORDER=ANALYSIS.

Click the green play button to run the command.

Syntax

Notes to Self: *Want to make your own notes right on your syntax file?*
*When you precede your writing with an asterisk and end with a period, that portion of your writing will not be interpreted as programming language to run a command and simply as a note for you to have for yourself.

```
*This will not be part of the command.
FREQUENCIES VARIABLES=VariableName
/ORDER=ANALYSIS.
```

Syntax is a great way to keep track of your work and allow transparency into it as well as being able to check that you did things accurately. Also, if you do something by mistake that messes up your dataset, you can always start again with a clean dataset and just run through your syntax.

Generating syntax logs with your output
While you have a syntax file provided, what happens if you don't know the syntax, don't have the file or want to open it, or don't want to look it up? Or what happens if you use the menus to run a command, but want to keep track of your work?

You can generate syntax with your output, both when using syntax or when using the drop-down menus. The default for SPSS Versions 28 and 29 is for this feature to be off. To turn it on, either:

- Go to Edit → Options → Viewer, and check the box labeled "Display commands in the log," or
- Run the following syntax: SET PRINTBACK=ON.

If you don't know the syntax for something, you can always click Help → Command Syntax Reference, but the syntax manual it will bring you gives you over 2500 pages to search through! An easier way is to use a syntax file you already have and just copy-and-paste then edit the syntax, or to use the drop-down menus and then copy-and-paste the syntax that generates in the output window.

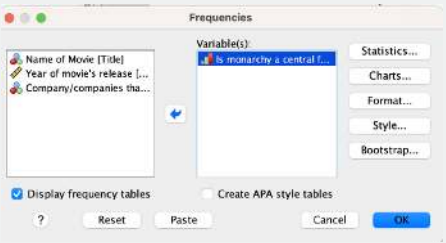
The exception to this is that if you edit the actual dataset (the Data View or Variable View windows, e.g., adding variable descriptions, coding schemes, deleting a variable, etc.), this will not generate a syntax log.

1.7 Statistics in Action: Monarchy in Movies

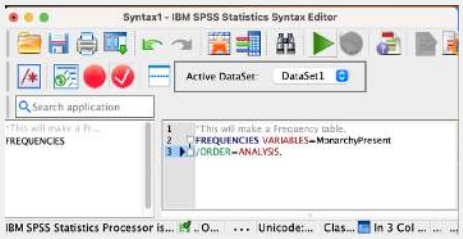
I am ready to analyze my data and answer my research question!

First, I need to tell SPSS to run a frequency table for the government/monarchy variable.

Menu method:



Syntax method:



Once I run either of these, a frequency table pops up in my output window:

Statistics

Is monarchy a central form of government for the film's setting? Is there monarchy in the film?

N	Valid	20
	Missing	0

Is monarchy a central form of government for the film's setting? Is there monarchy in the film?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	12	60.0	60.0	60.0
	Appearance	1	5.0	5.0	65.0
	Comparable	1	5.0	5.0	70.0
	Yes	6	30.0	30.0	100.0
	Total	20	100.0	100.0	

At the top, the Valid N tells me the number of cases in my dataset that had data being analyzed for this variable. The 20 reflects that there were 20 movies where I had an answer about form of government. Below that it indicates there are 0 missing cases, meaning there were not any that were excluded from the analysis.

The frequency table looks similar to the template above, except there is an extra column, "Valid Percent." This is usually the column you want to look at for percent, as it excludes the missing cases. "Percent" is the percent out of all cases, including missing cases. In our case these columns are identical since there were no missing cases. (If you want to delete the "percent" column, you can edit the table by double-clicking on it. Delete the column, close the pop-up window, and your table will update.)

So, 30% of the films had monarchy as a central form of government, and 35% had monarchy or something comparable. That's over 1/3 of the films! Monarchy movies may be money-makers.

Interpreting frequency tables

If I told you "30% of cases were a yes; 35% were yes or comparable" that would not mean much of anything. Maybe it would to you since you've been following along with me as I conducted this study, but otherwise it would be pretty meaningless. **When you interpret statistics, it's important to provide their story in real-life context, so that it makes sense to someone not familiar with your data or dataset, coding, or analysis.** In this case, I might say something like, "I looked at 20 of the most popular kids movies from the past 8 years. About 1/3 of them featured monarchies."

F. Pie Charts

Pie charts are useful for showing relationships of parts to the whole (slices of the pie) when you do not have too many categories.

SPSS: Pie chart Runaway argument for \eatSpaces?

Menu: Graphs → Pie → Summaries for groups of cases; Define. Under "Slices Represent" check the circle for "% of cases." Put your variable into "Define Slices by:". Click OK.

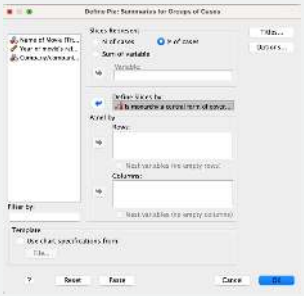
Syntax: Replace "VariableName" with your variable's actual name.

GRAPH
/PIE=PCT BY VariableName.

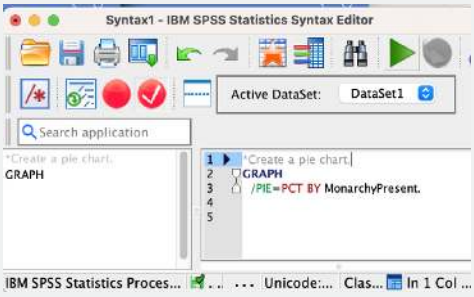
1.8 Statistics in Action: Monarchy in Movies

Let's put the same data we put into a frequency table into a pie chart.

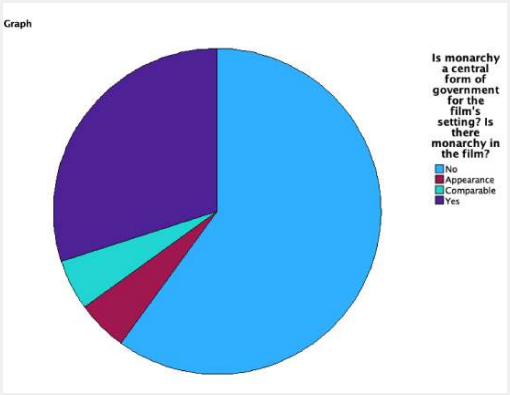
Menu:



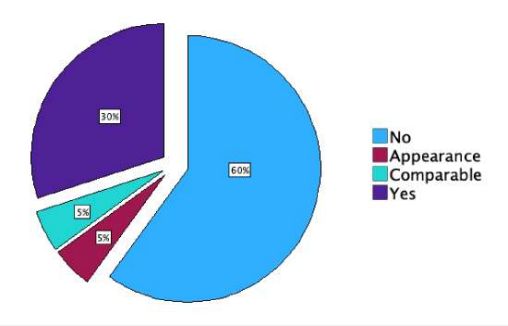
Syntax:



Output:



If you double-click on a graph (or table) in the output, it will pop out into a new window and you can make edits to it. Once you're done editing, just exit the pop-up window and the graph in the Output window will update. I decided to add data labels to the pie chart, explode the pieces, and make the legend text bigger.



G. Bar Graphs

Use bar graphs when you have frequencies of categories. If you use a bar graph on a variable with a Strongly Disagree to Strongly Agree scale, each bar will get a label stating what category it is. In Chapter 3 we will look at histograms. Those are a form of bar graph for when you have frequencies for numerical rather than categorical data (e.g., salary, number of pets, etc.).

SPSS: Bar graph Runaway argument for \eatSpaces?

Menu: Graphs → Bar → Simple; Summaries for groups of cases; Define. Under "Bars Represent" check the circle for "% of cases." Put your variable into "Category Axis:". Click OK.

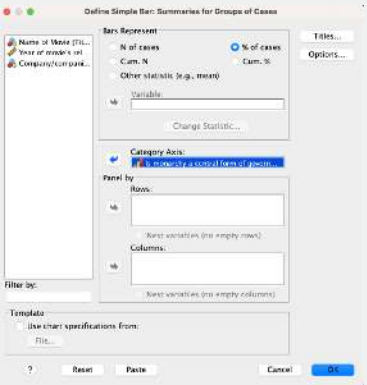
Syntax: Replace "VariableName" with your variable's actual name.

```
GRAPH
/BAR(SIMPLE)=PCT BY VariableName.
```

1.9 Statistics in Action: Monarchy in Movies

Let's put the same data we put into a frequency table into a bar graph.

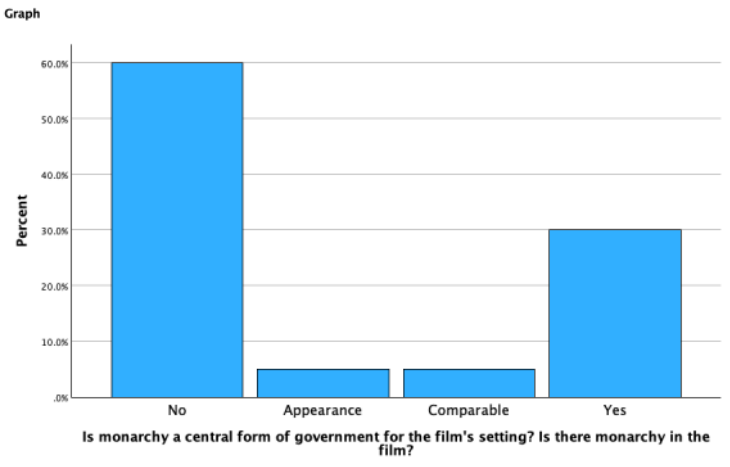
Menu:



Syntax:

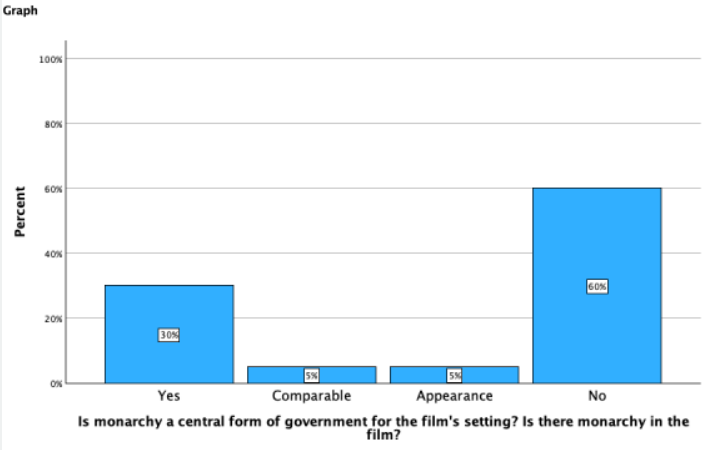
```
*Make a bar graph. Sub COUNT for PCT if you want frequency instead of percentage for the y-axis.
GRAPH
  /BAR(SIMPLE)=PCT BY VariableName.
```

Output:



I double-clicked on the graph and made a few edits.

Revised output:



H. Critical Consumption

I wouldn't write a statistics book if I didn't think statistics have awesome utility. But with a powerful tool comes responsibility. You should be accurate and transparent when you share your own data analysis. You also have to be a critical consumer of statistics.

Example: Data in action Runaway argument for \eatSpaces?

I was in a university meeting and we were looking at differences over time for the proportion of the university student population that is non-Hispanic white. It looked like the school had almost 1 in 10 more graduate students of color in 2010 than 2009.

However, data analysis is only as good as one's data. In 2010, there were new race/ethnicity reporting requirements, which included the addition of "unknown" to the university's race/ethnicity data. There were not actually any substantial changes in the racial-ethnic make-up of the university between 2009 and 2010.

1.9 Statistics in Action: Monarchy in Movies

My conclusion that about 1/3 of the top 20 kids movies from the past 8 years feature monarchies seems trustworthy. But if you just hear a statistics like that, you do not know if it is accurate unless you know it came from valid data and that the analysis

was done correctly. Even then, you need to be careful about generalizations and how the constructs in the claim were operationalized, and how that impacts its meaning.

My analysis gave preference to U.S. rankings. It measured popularity based on revenue. Originally I said I was interested in whether the movies my kid and I were watching really did have as many monarchies as it seemed. To answer a question at that individual level, I could probably go through my viewing history on streaming services and list the movies I've gone to in the theatre. My child and I have watched many more films than the ones that made it to this top 20 list (and there are a few on the list I haven't seen). We've watched other movies featuring monarchies (e.g., Aladdin, Cinderella) and others that do not. The list excludes films that are originals by streaming services. If I expanded or changed my criteria, the distribution of monarchy to non-monarchy movies might look different. Furthermore, how did I classify whether or not a movie featured a monarchy? If a movie is set in Canada or the United Kingdom, should that count? To what extent does the monarchy need to be central to the film for me to classify it as being a movie with a monarchy government. What happens if the film takes place in multiple locations and some are and some are not monarchies? Without clear criteria for how I went about classifying the films, you may not want to trust the dataset I made. Finally, it is very easy to make simple mistakes when recording and entering data, entering codes, etc. A few typos and the results could be very different. There are ways to safeguard against these issues and improve validity, but this is also one reason replication is important.

You should also pay attention to the phrasing and framing of the story I am telling about my data, and what claims I am actually making. It is just as technically accurate for me to say, "Most of the top 20 kids movies do not feature monarchies." That tells a very different story. Even specifying "About 2/3 of the top 20 kids movies from the past 8 years did not include monarchies" focuses your attention on most movies not having monarchies, whereas "about 1/3... did include monarchies" focuses your attention on the fact that a substantial number of the films did.

This page titled 1: Data, Datasets, SPSS, and Frequency Distribution Tables is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Ezra Temko (Consortium of Academic and Research Libraries in Illinois (CARLI)) .

2: Levels of Measurement; Five Point Summaries

Chapters 2 and 3 focus on measures of central tendency and spread, with Chapter 2 focused on median and quartiles and Chapter 3 focused on mean and standard deviations.

2A.1 Statistics in Action: Defining Democracy

In Chapter 1, I looked at whether movies featured monarchies. It seemed strange to me that the United States would both valorize democracy and keep having popular movies that present monarchies without problematizing the concept. These next two chapters will continue with investigating forms of government. When I say the U.S. seems to valorize democracy, what do I actually mean by democracy? "Democracy" comes from the Greek words "demos" and "kratia" for "the people" and "power, rule, strength." Because we elect people to represent us, the United States has a representative democracy, or republic, not a direct democracy. Or do we? Is the United States an oligarchy, where there is a small elite that holds the power and governs? An autocracy, where one person holds all the power? A theocracy, where religion governs civil life? Recently we have heard some right-wing U.S.-Americans insisting the United States is not a democracy---and was never intended to be one. The International Institute for Democracy and Electoral Assistance classifies the U.S.A. as a "backsliding democracy" and the Economic Intelligence Unit labels the U.S.A. a "flawed democracy." The World Values Survey asks residents how democratically their country is being governed. However, as noted at the end of Chapter 1, how we conceptualize and operationalize constructs matters. What is democracy? Do people even agree on what democracy is?

The World Values Survey asks a series of 9 questions about what democracy is. The interviewer asks, "Please tell me for each of the following things how essential you think it is as a characteristic of democracy. Use this scale where 1 means 'not at all an essential characteristic of democracy' and 10 means it definitely is 'an essential characteristic of democracy'."

Not an essential characteristic of democracy									An essential characteristic of democracy
1	2	3	4	5	6	7	8	9	10

While not asked, if the respondent instead volunteers that the characteristic is against democracy, that is coded as a 0.

The different characteristics asked about include:

- Governments tax the rich and subsidize the poor.
- Religious authorities ultimately interpret the laws.
- People choose their leaders in free elections.
- People receive state aid for unemployment.
- The army takes over when government is incompetent.
- Civil rights protect people from state oppression.
- The state makes people’s incomes equal.
- People obey their rulers.
- Women have the same rights as men.

These questions were asked to a representative sample of U.S. adults in 2017, and to residents of 90 other countries between 2017 and 2022.

To me it seems fairly obvious that democracy requires free elections. Do others agree? Was I socialized into this view as a U.S.-American and share it with most U.S.-Americans, or is this a more universal truth of what democracy means? We'll take a look in this chapter.

A. Levels of Measurement

Before we can answer our question, we need to learn about **levels of measurement.** Level of measurement will be an important concept throughout social statistics, because different techniques and measurements can only be used with variables at certain

levels of measurement. There are different things you can do with variables depending on its level of measurement.

Definition: Level of measurement

Level of measurement: a variable characteristic classifying the relationship among a variable's values

We will deal with four main levels of measurement: **nominal, ordinal, ratio,** and **indicator.**

Definition: Nominal

Nominal: Nominal variables have values that are qualitative categories. They do not have any meaningful rank or order.

Nominal comes from the Latin word "nomen," meaning "name." The values for nominal variables have names for their labels.

Some common nominal variables include: race and ethnicity, gender, religion, occupation, and geographic area.

For example, U.S. Census questions on race include White, Black or African American, American Indian or Alaska Native, Asian, and Native Hawaiian or Other Pacific Islander. These are categories that have name labels. Any "ordering" or "ranking" would be arbitrary. For example, if we coded White=1, Black/African American=2, American Indian/Alaska Native=3, Asian=4, Native Hawaiian or Other Pacific Islander=5, and then found a middle value of 2.3, what would that mean? It is not the middle of anything because there are not ordered categories.

Even nominal variables have codes, numbers assigned for each category. For example, for the above, the coding scheme could be White=1, Black or African American=2, American Indian or Alaska Native=3, Asian=4, Native Hawaiian or Other Pacific Islander=5, multi-racial=6. However, just because the codes are still numbers does not mean the labels aren't categories. Each category could be assigned any number and the variable would still have the same meaning. And just because numbers are ordered does not mean the categories are ordered (e.g., Asian is not lower or higher than multiracial).

Definition: Ordinal

Ordinal: Ordinal variables have values that are qualitative categories. They do have a meaningful rank or order.

Ordinal comes from the Latin word "ordo," meaning "order." Ordinal variables have named categories, like nominal variables, but the categories can be rank-ordered.

Some common ordinal variables include social class, ideology, and *Likert scale* variables that ask for the level of agreement/disagreement with a particular statement.

For example, the General Social Survey (GSS) has a variable *spanking* with the question, "Do you strongly agree, agree, disagree, or strongly disagree that it is sometimes necessary to discipline a child with a good, hard spanking?" The categorical responses and their coding are 1 strongly agree, 2 agree, 3 disagree, and 4 strongly disagree. These answers are still categories, not numbers, but here they are ordered from most agreement to least agreement, which makes the variable ordinal. (If you're curious, 434 respondents strongly agreed, 887 agreed, 634 disagreed, and 389 strongly disagreed. While a minority, 44%, disagreed or strongly disagreed, when this question was first on the GSS in 1986 only 16% of respondents disagreed or strongly disagreed.)

Ordinal variables are often treated like numeric ratio-level variables. While this has limitations to its validity, it usually gives us some insights and a better picture than if we did not do this. The more categories an ordinal-level variable has, the more ratio-like it is. (Generally you want at least 5 categories, and if you have at least 10 you can usually confidently proceed with using the variable as ratio-like.) Furthermore, to make an ordinal variable ratio-like, it can be useful to re-code it to approximate equal distance. For the *spanking* variable, it could be re-coded as -2 strongly disagree, -1 disagree, 1 agree, and 2 strongly agree. This more closely approximates equal distancing because each unit is one movement towards agreement, with 0 being neutral / neither agree nor disagree, even though there are no cases with that answer. (Actually, in 2022 there were 23 respondents who volunteered that they could not choose or did not know, so those respondents could be re-coded as neutral). Centering on zero also makes interpretation easier---anything negative leans towards disagreeing, and anything positive leans towards agreeing.

Definition: Ratio

Ratio: Ratio-level variables have values that are numeric. They have a true zero point, meaningful ratios, consistent intervals between them, and the numbers function arithmetically.

Income (in dollars) and age (in years) are common ratio-level variables.

Let's elaborate on the criteria, using *age* as an example.

- Numeric: When expressed in years, the values are numbers. They do not need separate coding. 10 means 10 years old. 60 means 60 years old.
- True zero point: Zero years old is the lack or absence of any years of age (birth). (For weight this would be weightless, for salary it would mean $0).
- Meaningful ratios and arithmetically functional: Someone who is 20 is twice as old as someone who is 10 and half as old as someone who is 40. Someone who is 5 years older than 20 will be 25 years old. Someone who is 5 years younger than 20 will be 15 years old.
- Consistent intervals: Each unit is one year apart. 15 is one year from 16 is one year from 17, etc. (Note: Technically the age in years variable in the General Social Survey is ratio-like and not ratio-level, because the top age is 89+, which violates the consistent intervals criterion.)

Definition: Interval

Interval: Interval-level variables have equal intervals between values, but do not meet all the criteria to be classified as ratio-level variables. Some definitions of interval-level variables require interval-level variables to have numeric values.

We will not be discussing interval-level variables further in this book because they are somewhat rare in the social sciences, and for the purposes of the statistical techniques in this book, all interval variables are ratio-like, meaning they share enough in common with ratio-level variables to be used for the same statistical procedures.

A common example of an interval-level variable is temperature in degrees Fahrenheit. Degrees is numeric (e.g., 60 means 60°F; it is not a code for something else). There is an equal distance between each degree (one degree each), but there is no absolute 0 (0°F does not indicate lack of temperature, lack of warmth, or the like) and temperatures do not have meaningful ratios (50°F is not twice as warm as 25°F or half as warm as 100°F). Another example are standardized tests like the ACT and SAT. The ACT goes from 1-36 and the SAT goes from 200-800. There are no meaningful zeros.

Definition: Indicator

Indicator: Indicator variables, also called dummy variables, are dichotomous/binary, meaning they only have two categories, and these two values are coded as 0 and 1.

Dichotomous variables are those with only two options. The word comes from the Greek words "dikho" for "in two" and "temnein" for "to cut.

Often we would think of dichotomous variables as nominal (e.g., do you prefer Skittles or M&Ms; are you married or not married, agree/disagree or yes/no questions), but sometimes they could be considered ordinal (e.g., student vs. alum; secondary school vs. postsecondary school). Regardless, we separate them out here because when dichotomous variables are coded as 0 and 1, they have special properties that will be revisited throughout this book.

Exercise 1

What is the level of measurement for job satisfaction, when the responses include: very satisfied, somewhat satisfied, indifferent, somewhat unsatisfied, very unsatisfied?

Is it a nominal, ordinal, ratio, or indicator variable?

Answer

These are categories, and they are rank-ordered from most to least satisfied. This makes it ordinal.

? Exercise 2

What is the level of measurement for trees, when the responses include types of trees like: redwood, sequoia, fir, pine, birch, spruce, oak, and maple?

Is it a nominal, ordinal, ratio, or indicator variable?

Answer

These are un-ranked names. This makes it nominal.

? Exercise 3

What is the level of measurement for number of children in home, where the respondent writes how many children are in their home?

Answer

Number of children is numeric. It has an absolute zero (0 means the absence of children in the home). 4 children is double 2 children and half of 8 children, and each unit increase is one more child. This makes it ratio.

? Exercise 4

What is the level of measurement for a survey question asking how old someone is, when the options are age groups such as 0-7, 8-15, 16-19, 20-23, 24-27, 28-31, 32-35, 36-39, etc.?

Answer

While the answers have numbers in them, the attribute labels are actually categories, not numeric. They are age groups, not actual ages in years. They are ordered from youngest to oldest. This makes the variable ordinal. Note that there is not consistent intervals/distance between each category. The first two have an 8-year span, whereas the rest have a 4-year span. Before using this variable as ratio-like, it should be re-coded so that they are all in consistent 8-year spans (e.g., combine together 16-19 and 20-23 into 16-23, combine together 24-27 and 28-31 into 24-31, etc.).

? Exercise 5

What is the level of measurement for a survey question asking for a respondent's political ideology, if the options and coding are: 1: very conservative, 2: somewhat conservative, 3: moderate, 4: somewhat liberal, 5: very liberal, 6: I don't know.

Answer

The answers are named categories, not numeric. At first it appears that the variable is ordinal, because it is ordered (the higher the code, the less conservative/more liberal the respondent). However, the addition of "don't know" makes this variable nominal.

? Exercise 6

What is the level of measurement for a survey question asking individuals how they felt about the year 2022 on a scale of 0 to 10, with 0 being awful, 5 being neutral/mixed, and 10 being awesome?

Answer

While the answer the respondents give is a number, don't get fooled. These numbers are actually categories. One respondent's 8 could mean something different than another respondent's 8. There is not equal distance between the numbers. One respondent could say a 3, another a 1, and yet they could both actually have the same feelings about 2022. Also, it is not ratio-level: a 10 means awesome, not double neutral/mixed (double 5). Because these are ordered categories, the variable is ordinal.

The same construct can have different levels of measurement depending on its coding.

The General Social Survey (GSS) has a series of questions asking about different problems facing the country, and asks respondents whether they think "we're spending too much money on it, too little money, or about the right amount."

This variable's original coding below is nominal --- these are categories and they are not ordered. The ordinal coding below shows how I re-coded the variable to be ordinal, as now as the coded numbers increase, the values go from the respondent thinking we spend too little to too much. The indicator coding below is an indicator variable, a dichotomous variable coded with 0 and 1, where those who responded "too little" are classified as supporting more spending on the problem, while those who said we either spend "about right" or "too much" are classified as not supporting spending more on the problem.

Nominal/Ordinal Labels	Nominal Coding	Ordinal Coding	Indicator Coding	Indicator Labels
Too much	1	1	0	Not supporter
Too little	2	-1	1	Supporter
About right	3	0	0	Not supporter

Definition: Re-coding

Recoding is the process of changing a variable's number-attribute assignments so that it has new/different coding.

Another GSS variable, partyID, is based on the question, "Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or what?" The original coding is nominal, because the "other party," "don't know," and "refused" responses do not follow the rest of the values that are ordered from more Democrat to more Republican. For the first ordinal-level variable ("Ordinal 1 Coding" below), these three values were re-coded as missing, removed from the variable, making it rank-ordered. In the second ordinal-level variable ("Ordinal 2 Coding"), the Ordinal 1 coding is collapsed to make three categories, classifying respondents either as Democrat, Independent, or Republican. In Chapter 13, you'll learn how this second ordinal variable can be re-coded into a a reference group set made up of indicator variables.

Nominal / Ordinal 1 Labels	Nominal Coding	Ordinal 1 Coding	Ordinal 2 Coding	Ordinal 2 Labels
Strong Democrat	1	-3	-1	Democrat
Weak Democrat	2	-2		
Independent, close to Democrat	3	-1	0	Independent
Independent	4	0		
Independent, close to Republican	5	1		
Weak Republican	6	2	1	Republican
Strong Republican	7	3		
Other party	8	missing	missing	
DK/Not sure	98	missing	missing	
NA/Refused	99	missing	missing	

While ratio-level variables are often the most desired, sometimes your interests may lead you to re-code a variable into a different level of measurement. Let's say I have a ratio-level variable of birth weight, with the coding being weight in pounds. I might be interested in re-coding birth weight into an ordinal-level variable to be able to differentiate between low, normal, and high birth weights and then research potential determinants into these categories.

Ordinal Labels	Nominal Coding	Original Coding
Low birthweight	-1	<5.5 pounds
Normal birthweight	0	5.5 - 8.5 pounds
High birthweight	1	>8.5 pounds

Indexes

Another type of variable you might come across is an index.

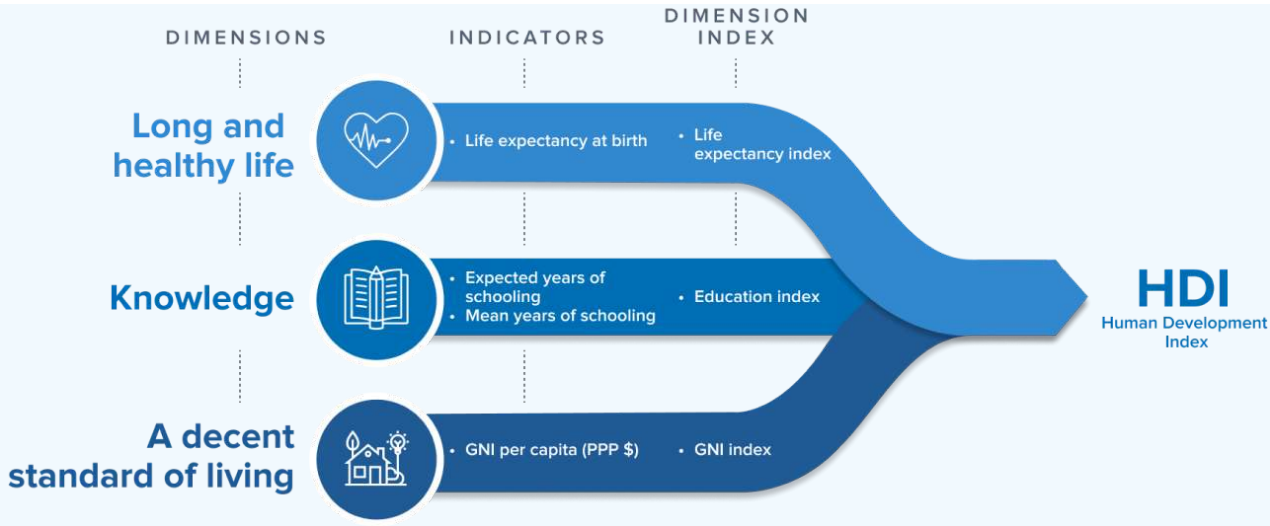
Definition: Index

An index is when multiple variables are combined together into one construct.

An index is usually treated as a ratio-level variable.

There are many indexes that are used for real-life evaluation purposes, such as the Human Development Index, Consumer Confidence Index, and Global Peace Index.

- Human Development Index (HDI)
  The United Nations (UN) introduced the HDI to measure a country's development based on people's capabilities, rather than past measures focused solely on economic growth. It takes into account five different variables that together form one summary measure of human development.

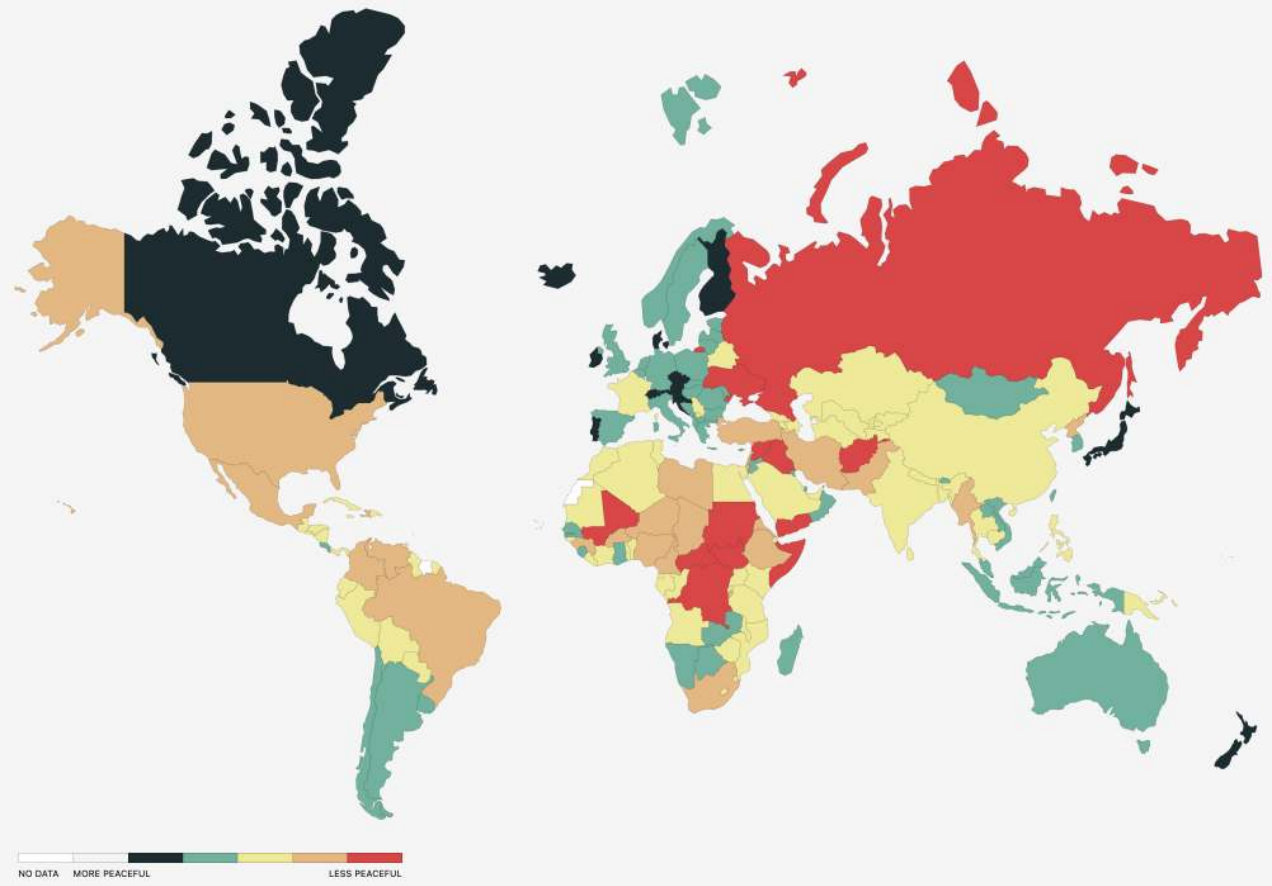


Currently the United States (0.921) is ranked 21st on the list of 191 countries, with Switzerland (0.962), Norway, Iceland, Hong Kong, and Australia taking the top 5 spots, and South Sudan (0.385), Chad, Niger, and the Central African Republic the bottom 5 spots.

The UN also has an inequality-adjusted HDI that takes each of the variables and adjusts them based on levels of inequality for those particular measures. The United States (0.819) is tied with Cyprus for 25th, while Iceland (0.915), Norway, Denmark, Switzerland, and Finland take the top 5 spots, and Central African Republic (0.404), South Sudan, Chad, Mali, and Niger take the bottom 5 spots.

- Global Peace Index (GPI)
  The Institute for Economic and Peace's Vision of Humanity's GPI is based on 23 indicators, each weighted on a scale of 1 to 5, with 1 being the most peaceful. The 23 indicators include: perceived criminality in society, security officers and police, homicides, jailed population, access to weapons, organized internal conflict, violent demonstrations, violent crime, political instability, political terror, weapons imports, terrorist activity, deaths from internal conflicts, military expenditure, armed services personnel, UN peacekeeping funding, nuclear and heavy weapons, neighboring country relations, deaths from external conflicts, external conflicts fought, internal conflicts fought, domestic and international conflict, safety and security, and militarization.

  For 2023, the United States ranked 131st (2.448), with the most peaceful countries being Iceland (1.124), Denmark, Ireland, New Zealand, and Austria, and the least peaceful countries being Afghanistan (3.448), Yemen, Syria, South Sudan, and the Democratic Republic of the Congo.



- Consumer Confidence Index (CCI)
  The Organization for Economic Cooperation and Development (OECD)'s CCI is a composite of questions about one's expected financial situation, sentiment about the general economic situation, unemployment, and capability of savings. An index score above 100 signals positive consumer confidence toward the economic future, whereas under 100 signals a more pessimistic outlook. The CCI is intended to predict spending compared to saving.

  The United States ranked 28th (based on its most recent July 2023 score of 97.42). The highest ranked countries were Mexico (103.78), Indonesia, Costa Rica, Lithuania, and Brazil. The lowest were Estonia (94.18), China, Columbia, New Zealand, and South Africa.

Cronbach's alpha, often notated with the Greek lowercase alpha letter α, is a common measure of index reliability.

Definition: Cronbach's alpha

A common measure of index reliability (from 0 unreliable to 1 perfect reliability).

Do the variables that make up the index all measure the same overall construct? You will likely come across Chronbach's alpha when you read quantitative academic studies, though it is much more frequently used in psychology statistics compared to sociology statistics. Overall, just know that 0 means not reliable, 1 means perfect reliability, and anything at 0.7 or higher is good enough.

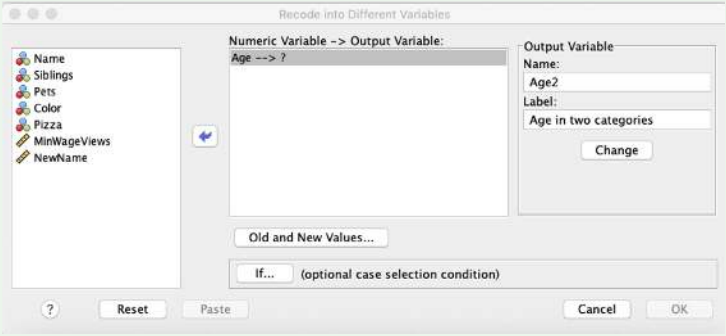
Recoding with SPSS

There are two ways to re-code using SPSS.

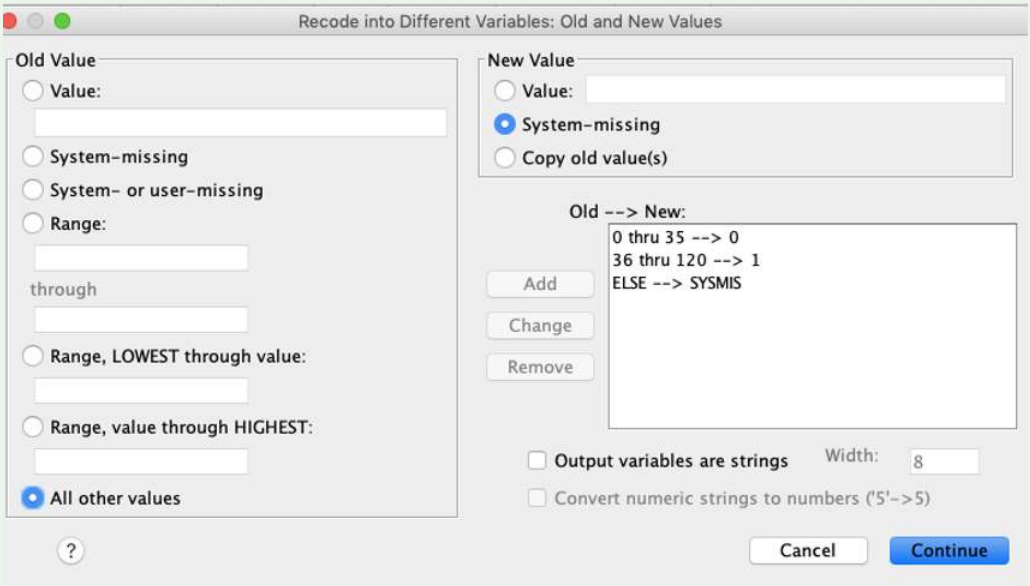
Method #1: Recoding by listing old and new values

Menu method: In SPSS, go to Transform → Recode into Different Variables...
Choose this one instead of "Recode into Same Variables..." so that you are not overwriting the old variable, which can get confusing.

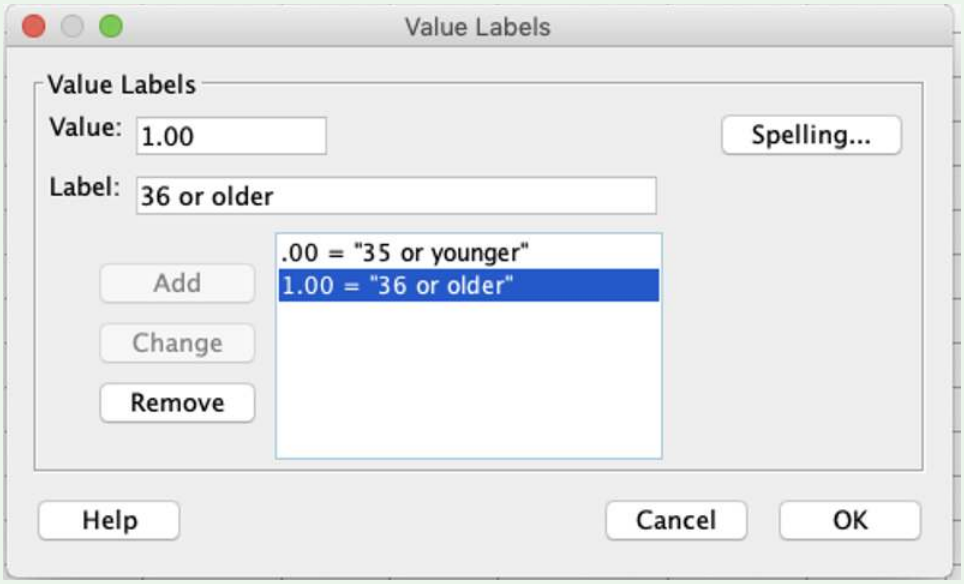
1. Put your original variable into the middle box. On the right side, where it says "Output Variable," give the variable you are creating (the re-coded variable) a name and a label/description. Make sure to click "Change"! This is especially important if you are re-coding multiple variables, because otherwise your new coding scheme will apply to whatever you had in there before, which is not the correct variable.



2. Click the "Old and New Values" button.
In the window that pops up, define what you want the new codes to be. On the left side, you select the old codes, and on the new side, you name what the new code will be. You can do this for one value at a time (e.g., you can state that a 1 should become a 5), or you can use the ranges to change multiple old codes at the same time into a new code (e.g., if you are collapsing a variable). In the example below, age was collapsed into an indicator variable, with 0 coded as those 35 and younger, and 1 coded as those 36 and older. (The last line, else=sysmis, said that if there were any values outside of 0 through 120, they would be coded as missing.) When you are done specifying how the old codes should be re-coded into new codes, click Continue, then in the original window, click OK.



3. Go to the Variable view of the dataset, and add labels to define your coding scheme (as applicable).



4. Advanced/more complex recoding: In the original recoding window, there was also an "IF" button. This can be useful if you have more complicated re-codes. For example, you can recode values contingent on a different variable having a certain value (e.g., if Hispanic=2). If you are using this to combine multiple variables together, sometimes it is best to start by recoding into a different variable, and then using the "recode into same variables" drop-down so that you can keep making changes until you get to the coding you want for the new variable.

Syntax:

Edit the syntax below as applicable. You can see options for changing a range of variables into a new code or just changing them one-for-one. Add and delete as needed.

*First line RECODE old codes into new codes.
RECODE OriginalVariableName (1=0) (2=1) (3=3) (4 thru 6 = 2) (7 thru 9 = 3) (ELSE=SYSMIS) INTO NewVariableName.
VARIABLE LABELS NewVariableName 'Descriptive title for variable'.
VALUE LABELS NewVariableName 0 'Code description' 1 'Code description' 2 'Code description' 3 'Code description'.

Advanced/more complex recoding:

If you want to do contingent coding (like how you could select "IF" in the recoding window), use the syntax below to re-code into the same variable. Again, I would suggest re-coding into a different variable first, then editing and overwriting that newly created variable:

DO IF (VariableAName=#).
RECODE VariableBName (OldCode#=NewCode#).
END IF.

Example

In practice, re-coding can be a multi-step practice if you are doing more complicated re-codes. I've included below an example of syntax for creating a four-category race variable for the 2022 General Social Survey dataset. While GSS provides a "race" variable (1=White, 2=Black, 3=Other), it is separate from the "hispanic" variable (0=No, 1=Yes). I wanted to combine these together into one variable.

```
*Recode Hispanic into Race4Cat.
*Not Hispanic becomes 2, Hispanic becomes 3, missing stays missing.
RECODE RaceACS16 (0=2) (1=3) (ELSE=SYSMIS) INTO Race4Cat.
VARIABLE LABELS Race4Cat 'Race in 4 categories'.
VALUE LABELS Race4Cat 0 'NH White only' 1 'NH Black only' 2 'NH Other/2+' 3 'Hispanic'.
EXECUTE.

*If NH white, change NH coding to 0.
DO IF (RaceACS1=1).
RECODE Race4Cat (2=0).
END IF.

*If NH black, change NH coding to 1.
DO IF (RaceACS2=1).
RECODE Race4Cat (2=1).
END IF.

*If black and white and NH, change coding to 2+.
DO IF (RaceACS1=1 & RaceACS2=1).
RECODE Race4Cat (0=2) (1=2).
END IF.

*If American Indian or Alaskan Native, and also white or black, and already coded as white only or black only, change to 2+.
DO IF (RaceACS3=1).
RECODE Race4Cat (0=2) (1=2).
END IF.

*If Asian Indian, and also white or black, and already coded as white only or black only, change to 2+.
DO IF (RaceACS4=1).
RECODE Race4Cat (0=2) (1=2).
END IF.

*Same as above for other racial categories that are not Hispanic or white or black.
DO IF (RaceACS5=1).
RECODE Race4Cat (0=2) (1=2).
END IF.

DO IF (RaceACS6=1).
RECODE Race4Cat (0=2) (1=2).
END IF.

DO IF (RaceACS7=1).
RECODE Race4Cat (0=2) (1=2).
END IF.

DO IF (RaceACS8=1).
RECODE Race4Cat (0=2) (1=2).
END IF.
```

```
DO IF (RaceACS9=1).
RECODE Race4Cat (0=2) (1=2).
END IF.

DO IF (RaceACS10=1).
RECODE Race4Cat (0=2) (1=2).
END IF.

DO IF (RaceACS11=1).
RECODE Race4Cat (0=2) (1=2).
END IF.

DO IF (RaceACS12=1).
RECODE Race4Cat (0=2) (1=2).
END IF.

DO IF (RaceACS13=1).
RECODE Race4Cat (0=2) (1=2).
END IF.

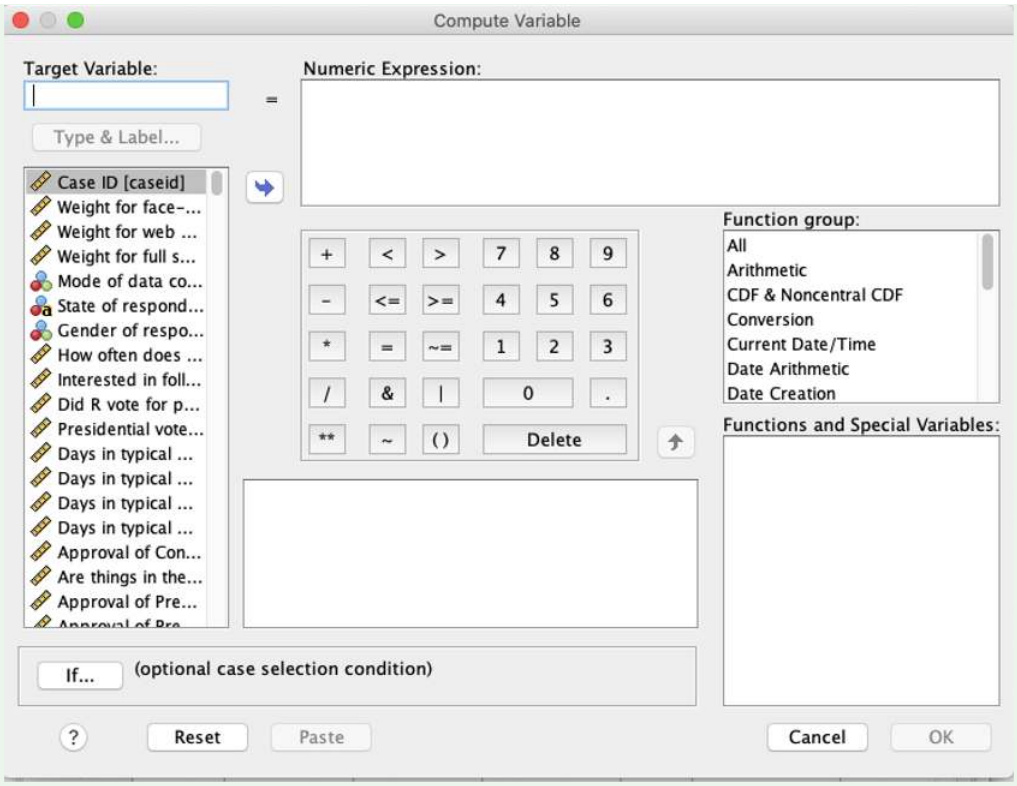
DO IF (RaceACS14=1).
RECODE Race4Cat (0=2) (1=2).
END IF.

DO IF (RaceACS15=1).
RECODE Race4Cat (0=2) (1=2).
END IF.
```

Method #2: Recoding using formulas (compute)

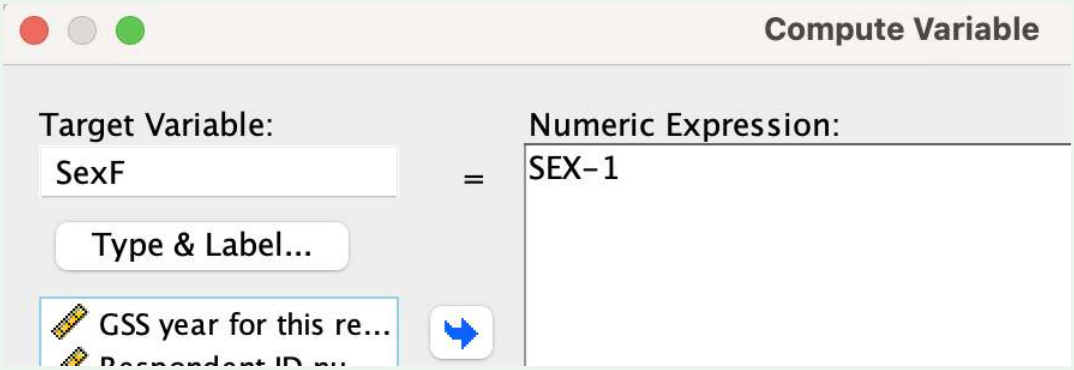
Depending on the re-code, I may also be able to use Compute, which allows me to use arithmetical formulas to change variables. This is especially useful for indexes (where we generally add variables together) and interactions (where we multiply variables together). It can also be used for some simple re-codes. For example, in GSS, the variable "sex" is coded as 1=male, 2=female. While I could use Method #1 and re-code the variable (1=0) (2=1), the other option is just to change the coding by subtracting one.

Menu:
In SPSS, go to Transform → Compute Variable.



In "Target Variable," write the name you want for your new re-coded variable.

In the "Numeric Expression" box, type the equation that should be used to re-code the variable, including the name(s) of the old variables you are using for the re-code (type them in or find them in the box on the left and click the blue arrow to put them into the Numeric Expression box). Note that there are specific symbols you need to use (e.g., * for multiplying, / for dividing), so if you're not sure on which symbol to use, you can just use the buttons.



Here you can see I am creating a new variable "SexF" (F to designate that in the indicator variable, female is coded as 1). The "numeric expression" is to take the old codes for "SEX" and subtract 1, so 2 would be 2-1=1 and 1 would be 1-1=0. 2 becomes 1 and 1 becomes 0, making it an indicator variable.

After using compute, go into the Variable View on the dataset and add your variable description and coding labels as appropriate.

Syntax:
Below is just one example. The Numeric Expression below is where it says "VariableOne+VariableTwo" which would be if you were adding two variables together. Change this based on whatever you want to do, e.g., above I would have typed "sex-1". Make sure to leave the period at the end of the line.

```
COMPUTE NewVariableName=VariableOne+VariableTwo.
VARIABLE LABELS NewVariableName 'Descriptive title for variable'.
VALUE LABELS NewVariableName 0 'Code description' 1 'Code description' 2 'Code description' 3 'Code description'.
```

B. Splitting Data into Quarters (5 Point Summary)



In this section we will go through a few methods to split data into four ordered **quartiles**, where an ordered distribution is split into four equal groups.

Level of Measurement

As previously noted, different things can be done with variables with different levels of measurement.

Nominal: Nominal-level variables cannot be split into ordered quartiles, because they are not ordered.

Ordinal & Ratio: Ordinal-level and ratio-level variables can be split into ordered quartiles.

Indicator: While indicator variables can be split into ordered quartiles, doing so is not very useful. If you already know what percentage of cases are at each of the two values, splitting the data into quarters does not give you additional useful information.

2B. Statistics in Action: Pets in the White House

Did you know that President Teddy Roosevelt had dozens of pets in the White House, including a bear and a hyena, and President Calvin Coolidge had a lion and bobcat? Only three presidents have had no pets with them while they were in the White House.

1. **Order and group or cross-out method**

Let's take a look at the number of pets the 8 most recent presidents have had in the White House, and divide the number of presidents into quartiles.

Term	Name	Number of Pets
2021-present	Joe Biden	4
2017-2021	Donald Trump	0
2009-2017	Barack Obama	2
2001-2009	George W. Bush	5
1993-2001	Bill Clinton	2
1989-1993	George H. W. Bush	2
1981-1989	Ronald Reagan	6
1977-1981	Jimmy Carter	3

The first step is to make this an ordered list.

Screenshot 2023-08-27 at 4.03.23 PM.png

Next this list can be separated into equal quarters. With 8 total presidents, each quarter will have 2 presidents in it.

Term	Name	Number of Pets
2017-2021	Donald Trump	0
2009-2017	Barack Obama	2
1993-2001	Bill Clinton	2
1989-1993	George H. W. Bush	2
1977-1981	Jimmy Carter	3
2021-present	Joe Biden	4
2001-2009	George W. Bush	5
1981-1989	Ronald Reagan	6

A 5-point summary consists of the lowest value (minimum), value that is higher than 1/4 and lower than 3/4 of values (lower quartile), middle value that is higher than 2/4 or 1/2 and lower than 2/4 or 1/2 of values (median), value that is higher than 3/4 and lower than 1/4 of values (upper quartile), and the highest value (maximum).

For the data above, separated into quarters, the lowest value is 0 and the highest is 6. The middle value is more complicated, as both 2 and 3 straddle the middle. In this case, find their average or midpoint, adding the two numbers together and dividing by two (2+3=5, 5/2=2.5). The lower quartile is 2. The upper quartile is between 4 and 5, so 4.5.

Now my data is separated into quarters along these five points: 0, 2, 2.5, 4.5, 5.

While this data separated evenly into four quarters, what happens if your dataset does not have an even multiple of 4?
Let's try adding the 9th most recent president to the dataset and find out. President Gerald Ford (1974-1977) had 4 pets in the White House.

Term	Name	Number of Pets
2021-present	Joe Biden	4
2017-2021	Donald Trump	0
2009-2017	Barack Obama	2
2001-2009	George W. Bush	5
1993-2001	Bill Clinton	2
1989-1993	George H. W. Bush	2
1981-1989	Ronald Reagan	6
1977-1981	Jimmy Carter	3
1974-1977	Gerald Ford	4

Again, the first step is to put these in order.

Term	Name	Number of Pets
2017-2021	Donald Trump	0
2009-2017	Barack Obama	2
1993-2001	Bill Clinton	2
1989-1993	George H. W. Bush	2
1977-1981	Jimmy Carter	3
2021-present	Joe Biden	4
1974-1977	Gerald Ford	4
2001-2009	George W. Bush	5
1981-1989	Ronald Reagan	6

I can see that the lowest number is 0 and the highest number is 6.

Next, I cross out an equal number of cases from both sides.

Term	Name	Number of Pets
~~2017-2021~~	~~Donald Trump~~	~~0~~
~~2009-2017~~	~~Barack Obama~~	~~2~~
~~1993-2001~~	~~Bill Clinton~~	~~2~~
~~1989-1993~~	~~George H. W. Bush~~	~~2~~
1977-1981	Jimmy Carter	3
~~2021-present~~	~~Joe Biden~~	~~4~~
~~1974-1977~~	~~Gerald Ford~~	~~4~~
~~2001-2009~~	~~George W. Bush~~	~~5~~
~~1981-1989~~	~~Ronald Reagan~~	~~6~~

Here I can see my middle number (median) is 3. To separate the data into quarters, I will find the middle number for the bottom half, and the middle number for the bottom half.

Unlike when the data split evenly into two halves, with an odd number of cases we have to either exclude or include the median from the two halves. Excluding tends to be more common, but both are acceptable methods (which will sometimes produce different results).

Inclusive method:

Here, the middle case (President Carter, with 3 pets) is included in both the lower half and upper half.

Term	Name	Number of Pets
2017-2021	Donald Trump	0
2009-2017	Barack Obama	2
1993-2001	Bill Clinton	2
1989-1993	George H. W. Bush	2
1977-1981	Jimmy Carter	3

1977-1981	Jimmy Carter	3
2021-present	Joe Biden	4
1974-1977	Gerald Ford	4
2001-2009	George W. Bush	5
1981-1989	Ronald Reagan	6

Now I can find the middle case for each of these halves:

Term	Name	Number of Pets
~~2017-2021~~	~~Donald Trump~~	~~0~~
~~2009-2017~~	~~Barack Obama~~	~~2~~
1993-2001	Bill Clinton	2
~~1989-1993~~	~~George H. W. Bush~~	~~2~~
~~1977-1981~~	~~Jimmy Carter~~	~~3~~

~~1977-1981~~	~~Jimmy Carter~~	~~3~~
~~2021-present~~	~~Joe Biden~~	~~4~~
1974-1977	Gerald Ford	4
~~2001-2009~~	~~George W. Bush~~	~~5~~
~~1981-1989~~	~~Ronald Reagan~~	~~6~~

Now my data is separated into quarters along these five points: 0, 2, 3, 4, and 6.

Exclusive method:

Here, the middle case (President Carter, with 3 pets) is excluded from both the lower half and upper half.

Term	Name	Number of Pets
2017-2021	Donald Trump	0
2009-2017	Barack Obama	2
1993-2001	Bill Clinton	2
1989-1993	George H. W. Bush	2
~~1977-1981~~	~~Jimmy Carter~~	~~3~~
2021-present	Joe Biden	4
1974-1977	Gerald Ford	4
2001-2009	George W. Bush	5
1981-1989	Ronald Reagan	6

Now I can find the middle case for each of these halves:

Term	Name	Number of Pets
~~2017-2021~~	~~Donald Trump~~	~~0~~
2009-2017	Barack Obama	2
1993-2001	Bill Clinton	2
~~1989-1993~~	~~George H. W. Bush~~	~~2~~
~~1977-1981~~	~~Jimmy Carter~~	~~3~~
~~2021-present~~	~~Joe Biden~~	4
1974-1977	Gerald Ford	4
2001-2009	George W. Bush	5
~~1981-1989~~	~~Ronald Reagan~~	~~6~~

With each half, I crossed out an equal number of cases from both sides. Because there were two middle numbers, their average or midpoint gives me the lower and upper quartiles. For the lower half, the middle number is 2. For the upper half, it's the midpoint/average of 4 and 5, which is 4.5.

Now my data is separated into quarters along these five points: 0, 2, 3, 4.5, and 6.

2. Percentiles

Definition: Percentile

Percentile: The proportion of data that falls below a given value. The nth percentile is the value below which n% of observations fall.

Just like we have been splitting data into four parts, percentiles split data into 100 parts (since "percent" means "out of 100"). In addition to the minimum and maximum values, to split data into quarters we find the 25th, 50th, and 75th percentiles.

Let's return to our ordered list of White House pets from the most recent eight presidents.

Term	Name	Number of Pets
2017-2021	Donald Trump	0
2009-2017	Barack Obama	2
1993-2001	Bill Clinton	2
1989-1993	George H. W. Bush	2
1977-1981	Jimmy Carter	3
2021-present	Joe Biden	4
2001-2009	George W. Bush	5
1981-1989	Ronald Reagan	6

- One way to calculate percentiles is to use the formula p%*(n+1) to figure out which case number the percentile is. This will tell you which case is greater than p% of other cases.

  For the 25th percentile this would be 25%*(n+1), for the 50th percentile 50%*(n+1), and for the 75th percentile 75%*(n+1). Since there are 8 cases (presidents), n+1=9.

25%*9=2.25 The 2nd case is 2 and the 3rd case is 2, so the 2.25 case must also be 2. The 25th percentile is 2.

50%*9=4.5 The 4th case is 2 and the 5th case is 3, so the 4.5 case is in between, estimated at 2.5. The 50th percentile is 2.5.

75%*9=6.75 The 6th case is 4 and the 7th case is 5, so the 6.75 case is estimated at 4.75. The 75th percentile is 4.75.

Now my data is separated into quarters along these five points: 0, 2, 2.5, 4.75, and 6.

Here is the same technique again, but with the 9 most recent presidents (n+1=9+1=10):

Term	Name	Number of Pets
2017-2021	Donald Trump	0
2009-2017	Barack Obama	2
1993-2001	Bill Clinton	2
1989-1993	George H. W. Bush	2
1977-1981	Jimmy Carter	3
2021-present	Joe Biden	4
1974-1977	Gerald Ford	4
2001-2009	George W. Bush	5
1981-1989	Ronald Reagan	6

25%*10=2.5 The 2nd case is 2 and the 3rd case is 2, so the 2.5 case must also be 2. The 25th percentile is 2.

50%*10=5 The 5th case is 3. The 50th percentile is 3.

75%*10=7.5 The 7th case is 4 and the 8th case is 5, so the 7.5 case is estimated at 4.5. The 75th percentile is 4.5.

Now my data is separated into quarters along these five points: 0, 2, 3, 4.5, and 6.

- Another way to calculate percentiles is from a frequency table. In addition to identifying the minimum and maximum values that have cases, here we can look at the cumulative percentage column to find the 25th percentile, 50th percentile, and 75th percentile.

Number of pets the most recent 8 presidents had in the White House

# of Pets	Frequency	Percent	Cumulative percent
0	1	12.50%	12.50%
2	3	37.50%	50.00%
3	1	12.50%	62.50%
4	1	12.50%	75.00%
5	1	12.50%	87.50%
6	1	12.50%	100.00%
Total	**8**	100.00%	

Here the first 12.5% of cases are at 0 pets. The next 37.5% of cases are all at 2, which crosses the point where 25% of cases are below it and 75% above it. We can see that 2 pets gets to a cumulative percentage of 50%, so 50% of the presidents were at 2 or fewer and 50% were at 3 or more. Since the 50th percentile means it needs to be more than half the cases and less than half the cases, the 50th percentile is between 2 and 3. Similarly, we can see that from 0 to 4 encompasses 75% of cases, meaning 5 to 6 encompasses 25% of cases. Therefore the 75th percentile has to be between 4 and 5.

The exact number of pets for the 50th and 75th percentiles is theoretical. The former is somewhere between 2 and 3 and the latter somewhere between 4 and 5. There are different ways to calculate percentiles; SPSS actually has 5 different ways it calculates/estimates percentiles.

Returning to our example with the 9 most recent presidents provides more concrete quartiles.

Number of pets the most recent 9 presidents had in the White House

# of Pets	Frequency	Percent	Cumulative percent
0	1	11.11%	11.11%
2	3	33.33%	44.44%
3	1	11.11%	55.56%
4	2	22.22%	77.78%
5	1	11.11%	88.89%
6	1	11.11%	100.00%
Total	**9**	100.00%	

Here the 25%, 50%, and 75% thresholds are all crossed at a particular value. For example, at 3 pets the cumulative percentage goes from less than 50% to more than 50%, indicating that 3 pets is both greater than half of the cases and less than half of the cases. Here the 5-number summary is: 0, 2, 3, 4, 6.

2A.2 Statistics in Action: Defining Democracy

Now let's take a look at an example with a much larger number of cases. We'll return to the question of how people define democracy, and here focus on U.S.-American adults.

1. Percentiles

Here is a weighted frequency distribution. Weighting is covered in Chapter 4, but for now just note that the analysis is slightly adjusted to more accurately reflect the U.S. adult population. You can weight using syntax (the weight variable's name was W_WEIGHT and my syntax was "WEIGHT BY W_WEIGHT") or by using the drop-down menus (Data → Weight Cases. Click "Weight cases by," put the variable in the box, and click OK).

Next, I either use the drop-down menu to run a frequency table or I use the following syntax (the variable name is Q243).

```
FREQUENCIES VARIABLES=Q243
/ORDER=ANALYSIS.
```

Frequencies

Statistics

Democracy: People choose their leaders in free elections

N	Valid	2534
	Missing	62

Democracy: People choose their leaders in free elections

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Not an essential characteristic of democracy	38	1.5	1.5	1.5
	2	9	.3	.3	1.8
	3	28	1.1	1.1	2.9
	4	54	2.1	2.1	5.1
	5	378	14.5	14.9	20.0
	6	112	4.3	4.4	24.4
	7	121	4.6	4.8	29.1
	8	167	6.4	6.6	35.7
	9	196	7.6	7.7	43.5
	An essential characteristic of democracy	1432	55.2	56.5	100.0
	Total	2534	97.6	100.0	
Missing	No answer	58	2.3		
	Don't know	4	.1		
	Total	62	2.4		
Total		2596	100.0		

The minimum here is clearly 1 and the maximum 10. I could still use the above formula to figure out which cases are at the 25th, 50th, and 75th percentile. (For example, 25%*(n+1) = 25%*2535 = the 633.75th case).

However, just looking at the frequencies I can tell that the 25% cumulative percent threshold gets crossed at 7, the 50% at 10, and the 75% at 10.

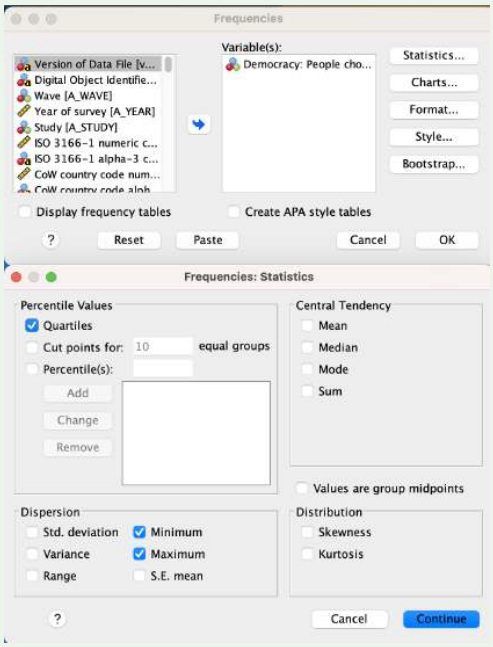
Now my data is separated into quarters along these five points: 1, 7, 10, 10, and 10.

Among the U.S. respondents, the lowest quarter were between evaluating free elections as not an essential characteristic of democracy to leaning towards it being essential, the next quarter were between leaning towards it being essential and evaluating it as essential, and then over half of respondents evaluated free elections as essential for democracy.

2. SPSS Summary Statistics

I can also have SPSS figure out the five-point summary for me.

Menu:
I start just like I am going to create a frequency table, going to Analyze → Descriptive Statistics → Frequencies. However, this time I do not need the frequency table checked (unless I want to make one). Instead I click the "Statistics..." box in the upper-right corner. In the pop-up labeled "Frequencies: Statistics," I make sure to check minimum, maximum, and quartiles.



Syntax

For the syntax, I substitute VariableName with the variable name, in this case Q243.

```
FREQUENCIES VARIABLES=VariableName
/FORMAT=NOTABLE
/NTILES=4
/STATISTICS=MINIMUM MAXIMUM
/ORDER=ANALYSIS.
```

Output:
Here is what the result looks like:

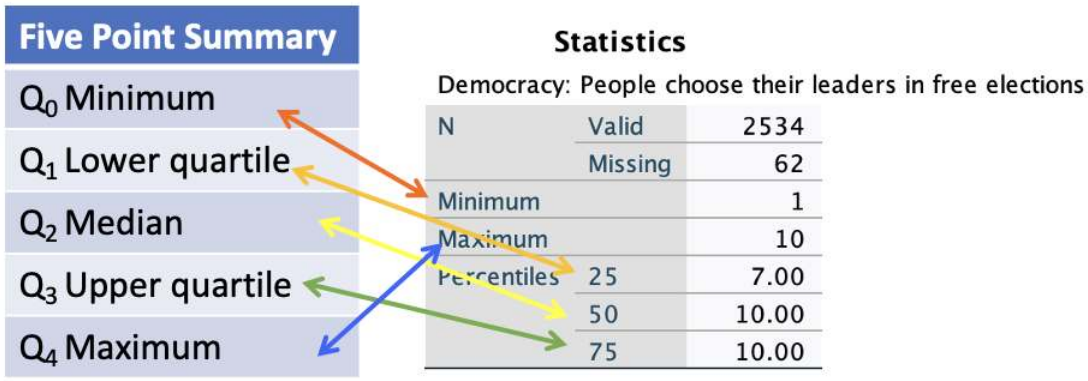
Frequencies

Statistics

Democracy: People choose their leaders in free elections

N	Valid	2534
	Missing	62
Minimum		1
Maximum		10
Percentiles	25	7.00
	50	10.00
	75	10.00

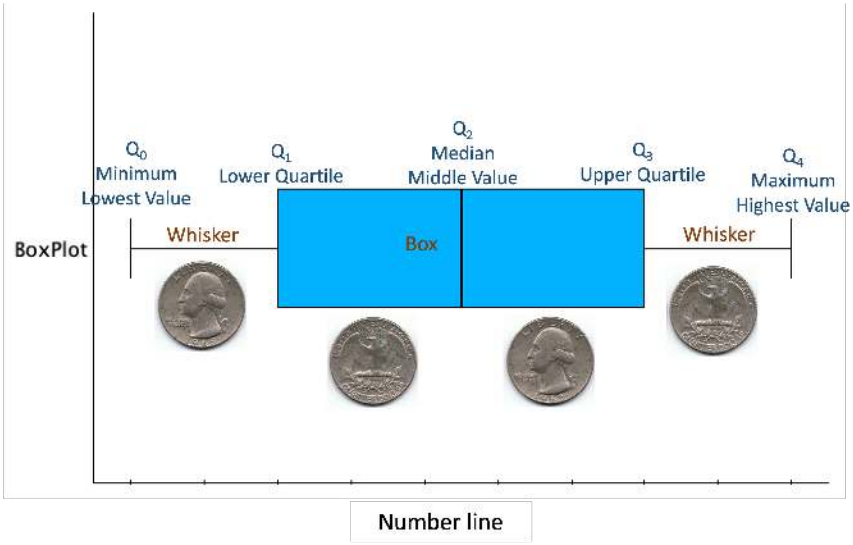
Here is how to translate that into a five point summary:



Now my data is separated into quarters along these five points: 1, 7, 10, 10, and 10.

C. Box and Whisker Plots

Box and whisker plots, or box plots, are a graph that visualize the distribution of ordered data split into four equal quarters on a number line. They show both central tendency and spread; what is typical and what variation there is.



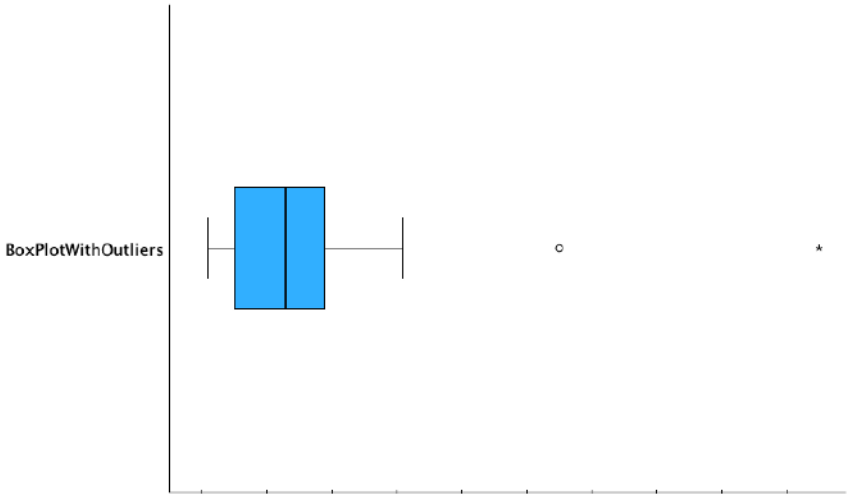
The five points from the five point summary are placed on a number line. Each whisker contains 1/4 of the data, and each portion of the box (separated by a vertical line for the median) contains 1/4 of the data.

In SPSS, the default is for box plots to also show **outliers**, and then for the box and whiskers to contain all the data except for the outliers.

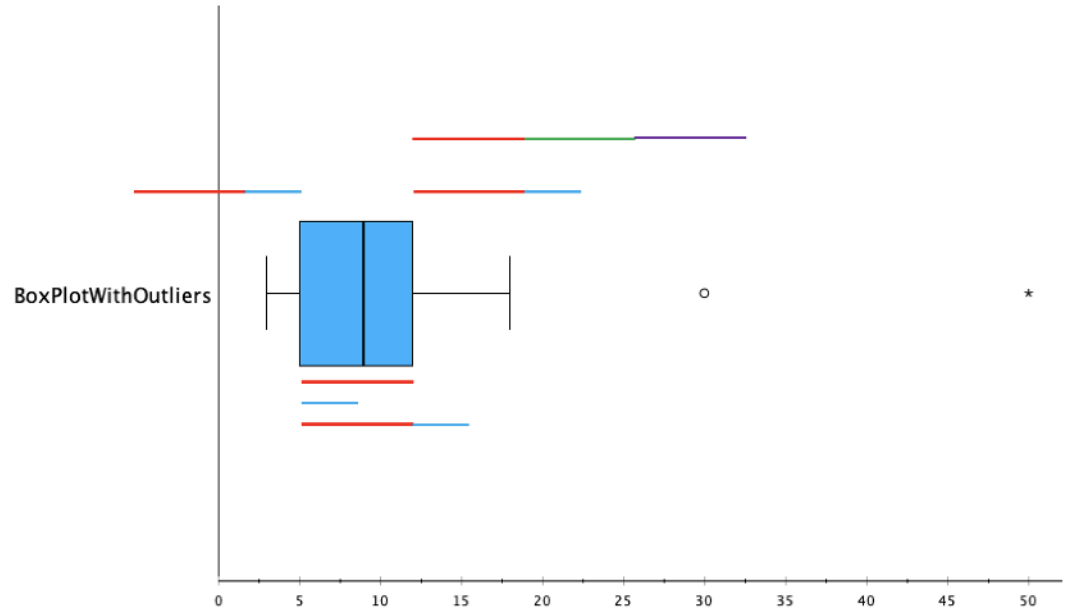
Definition: Outlier

Outlier: An extreme value; a case with a value that substantively differs from most other values.

The box plot below shows two outliers. Circles are closer to the rest of the data, while stars/asterisks are "extreme outliers" that are further away.

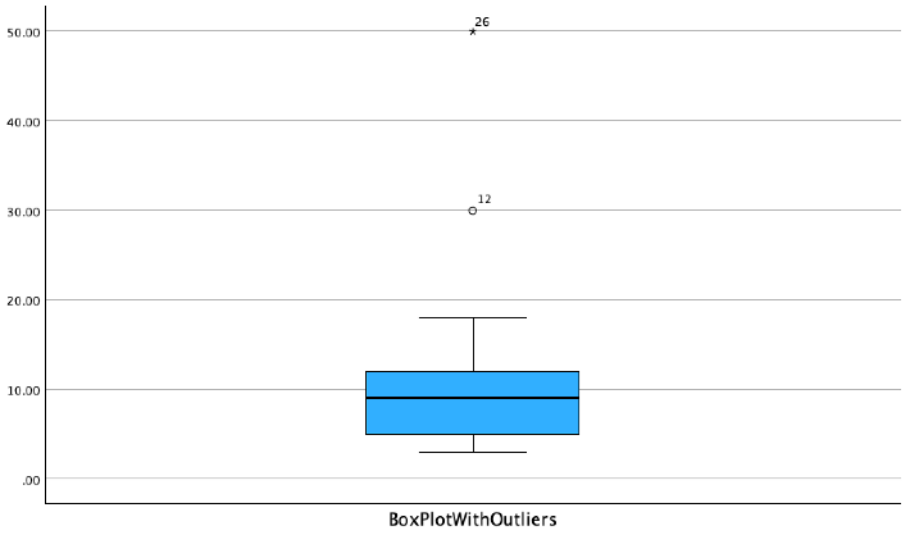


There are multiple ways to determine if a case/value is an outlier. You can evaluate if a case is qualitatively different (e.g., if everyone in your dataset has at least one sibling and one person is an only child, you might consider them an outlier, even though 0 is not that different from 1 numerically). You can eyeball it, meaning look to see if a particular value seems far away from others. Is it in the same ballpark as the rest of the data? There are also two systematic ways to evaluate if there are outliers. One way involves the interquartile range, the middle 50% of the data, which is the span of the box. An outlier is any number that is 150% of the IQR less than the lower quartile or more than the upper quartile, and an outlier that is 300% of the IQR less than the lower quartile or more than the upper quartile is considered an extreme outlier. In the box plot below the red line on the bottom show the span of the middle 50%, the blue line half of it, and those two combined for 150% of the IQR. The red and blue lines on the top show 1.5*IQR away from Q1 and Q3. The green and purple lines are the same span as the red line, so the line on top that is red, green, and purple shows 300% of the IQR from the upper quartile. You can see that the circle is an outlier (it is at least 1.5 IQRs above Q3) and the asterisk is an extreme outlier (it is at least 3 IQRs above Q3).



The other systematic way to determine outliers, which will sometimes give you different results, is discussed in Chapter 3. It is useful for box plots to show outliers, because within any box portion or whisker, one does not know where the data is concentrated, just that there is at least one case on either edge. By separating out outliers, you not only know about outlier values, but the whiskers better show the distribution of the outer quarters of the rest of the data.

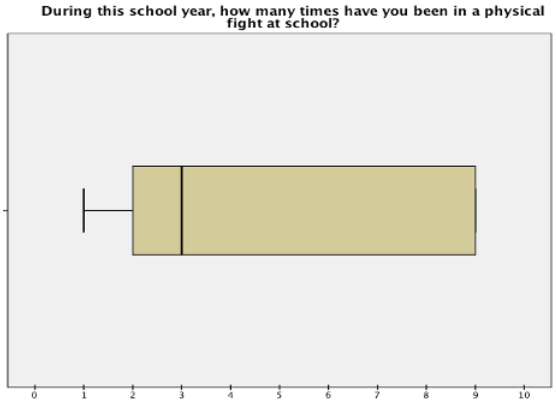
The figure below is the same box plot as above but how SPSS displays it without any editing.



The default for SPSS box plots are for them to be vertical, though you can choose to transpose them if you prefer them horizontal. The outliers here have labels attached with numbers. As you can see from the number line, these numbers are **not** the values of the cases. Instead they are the case numbers. If you go to the dataset's Data View in SPSS, you will see a column all the way on the left with numbers for each row/case (see image to right). These are the numbers that are referred to. This can let you look up the outliers in the dataset to see which cases they are. This is useful for knowing about these cases, but this can also be useful to doublecheck that the outliers are not there due to bad data entry such as a typo with a misplaced decimal point or extra zero. These case numbers will not mean anything to someone without your dataset, so are confusing to your audience. When you double-click the box plot to edit the graph, you can then click on these numbers and click delete/backspace, and they will disappear.



The box plot below, from the National Crime Victimization Survey's School Crime Supplement, shows the responses of the 3.7% of students who responded that they have been in at least one physical fight at school during the school year.



Do you see an upper tail? If you look closely, you can see the vertical line marking the edge of the whisker right on top of the upper edge of the box. This is because both the upper quartile and the maximum value are 9.

2A.3 Statistics in Action: Defining Democracy

Let's return to the question of how people define democracy, and whether it requires free elections. First, I need to generate a box plot in SPSS.

SPSS: Box plot

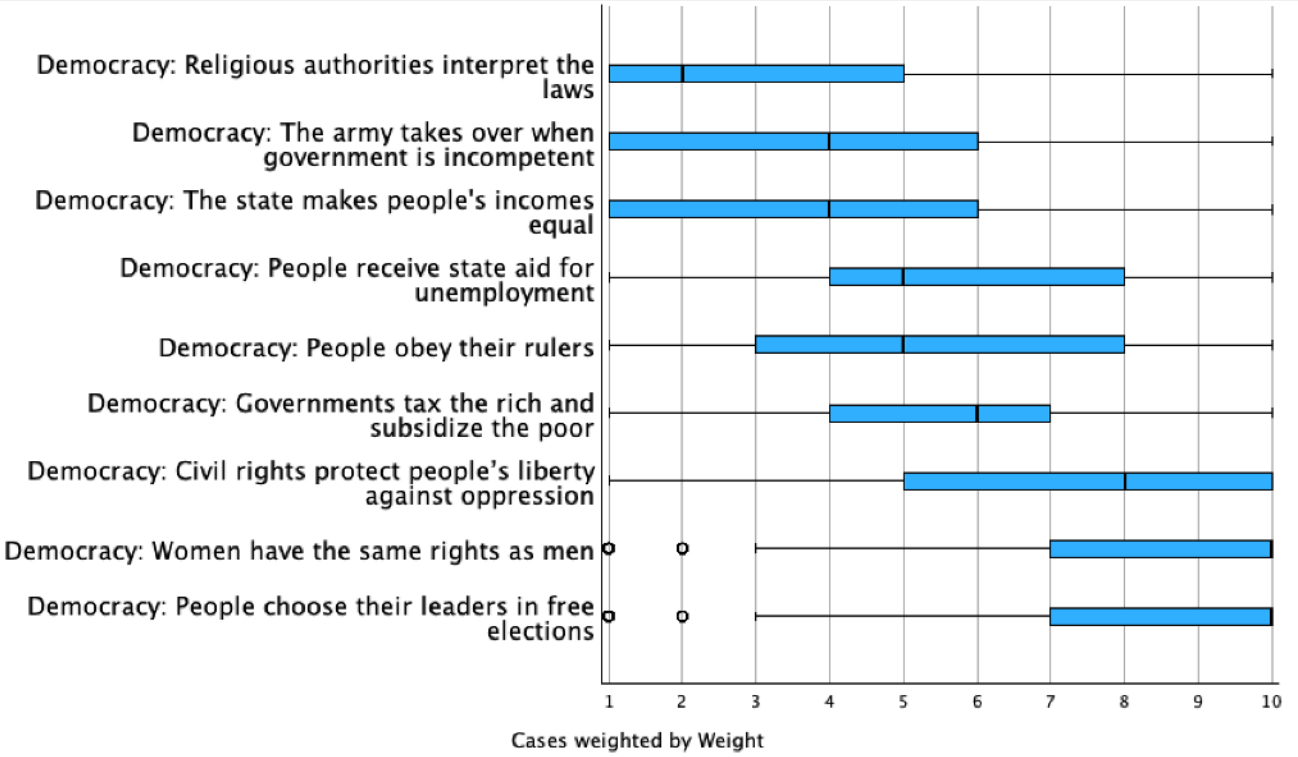
Menu:
Go to Graphs → Boxplot → Click Simple and click "Summaries of separate variables," unless you are trying to break it down by category and then click Define. Put your variable in the box/window labeled "Boxes Represent:" and then click OK.



I put all 9 of the defining democracy variables into Boxes Represent, and so have 9 side-by-side box plots. These are for U.S. respondents.

Syntax

```
*This is the Boxplot syntax for one or more variables.
*If you only have one variable then erase VariableName2.
*If you have multiple variables you can keep adding them with a space between each.
EXAMINE VARIABLES=VariableName1 VariableName2
/COMPARE VARIABLE
/PLOT=BOXPLOT
/STATISTICS=NONE
/NOTOTAL
/MISSING=LISTWISE.
```



While there is variation, we can see that the majority of U.S. respondents felt that people choosing their leaders in free elections is more an essential characteristic of democracy than not, as well as women having the same rights as men. We'll explore how this compares to other countries below. Overall, U.S. respondents especially disagreed that religious authorities interpreting the law and the army taking over when the government is incompetent are essential characteristics of democracy.

You can also make box plots in SPSS where your primary variable is broken down / categorized by another variable.

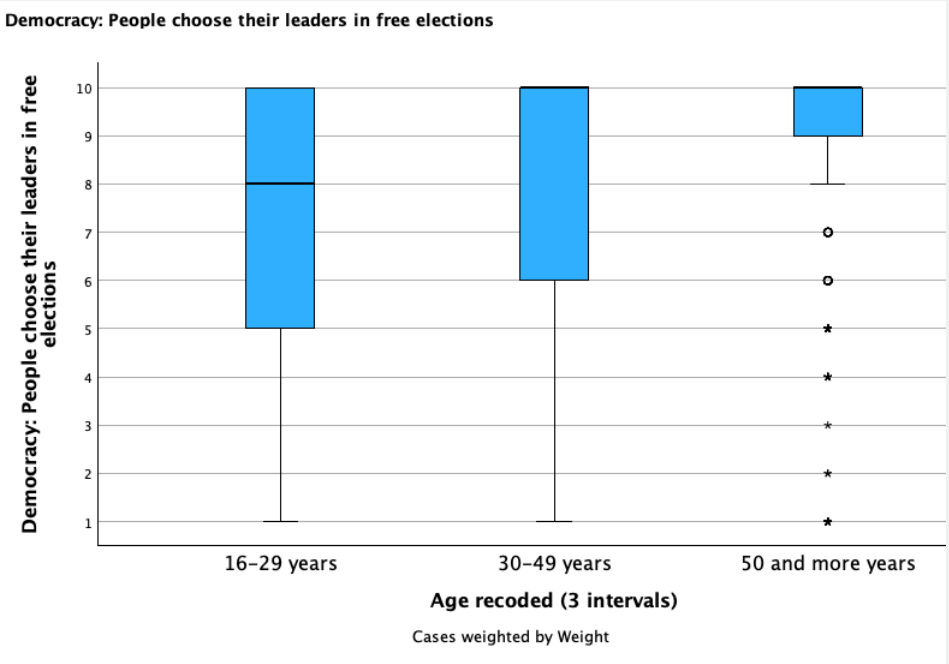
SPSS: Boxplots, broken down

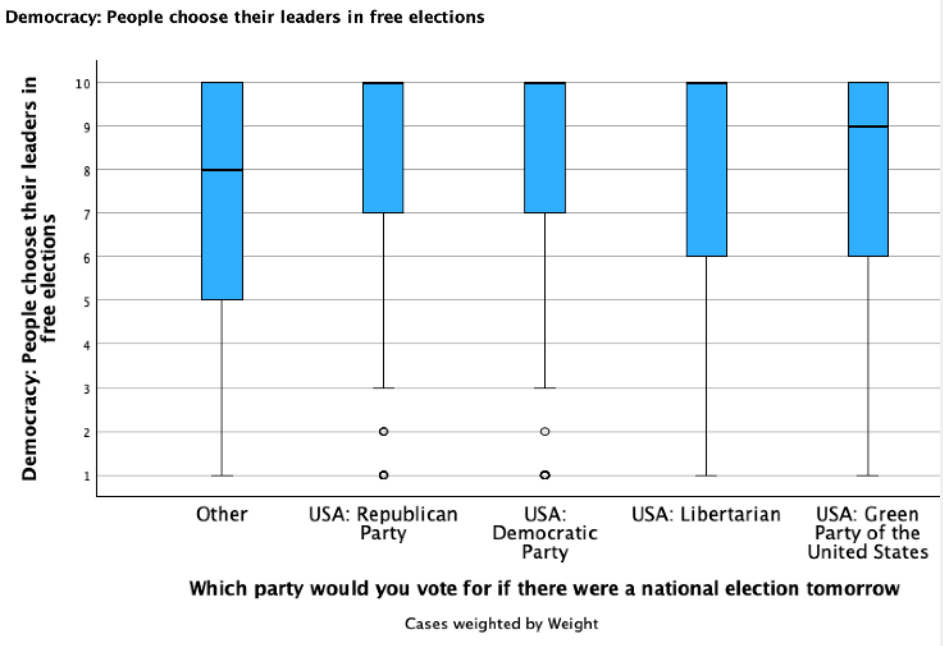
Menu:
Go to Graphs → Boxplot → Click Simple, but this time click "Summaries for separate groups of cases." Put your primary variable in the box/window labeled "Boxes Represent:" and the variable you want to break it down by in "Category Axis:" and then click OK.

Syntax

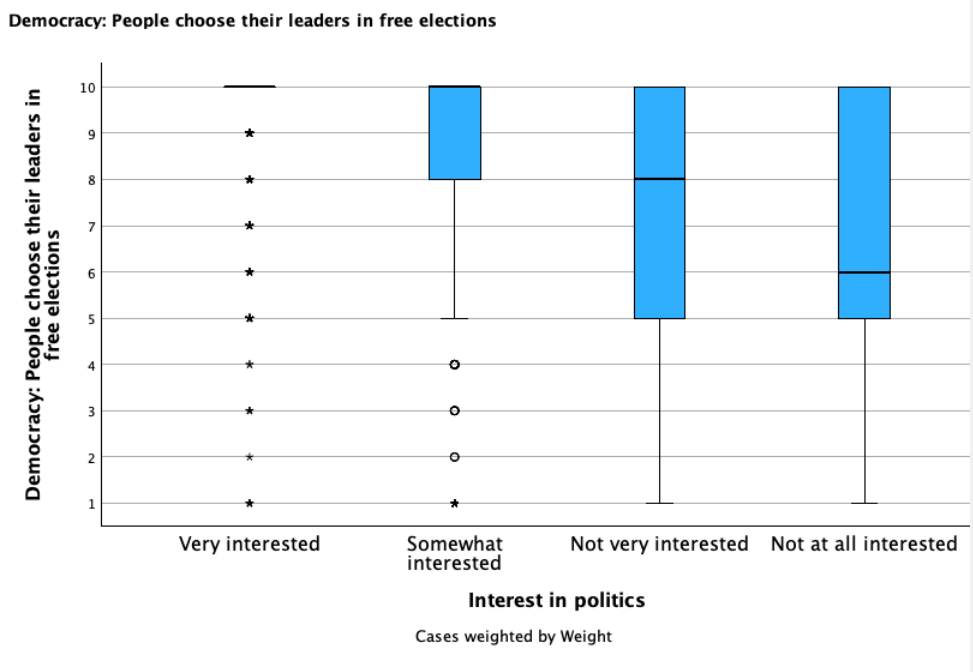
```
*This is the syntax for boxplots categorizing a variable by another one.
EXAMINE VARIABLES=PrimaryVariableName BY CategorizingVariableName
/PLOT=BOXPLOT
/STATISTICS=NONE
/NOTOTAL.
```

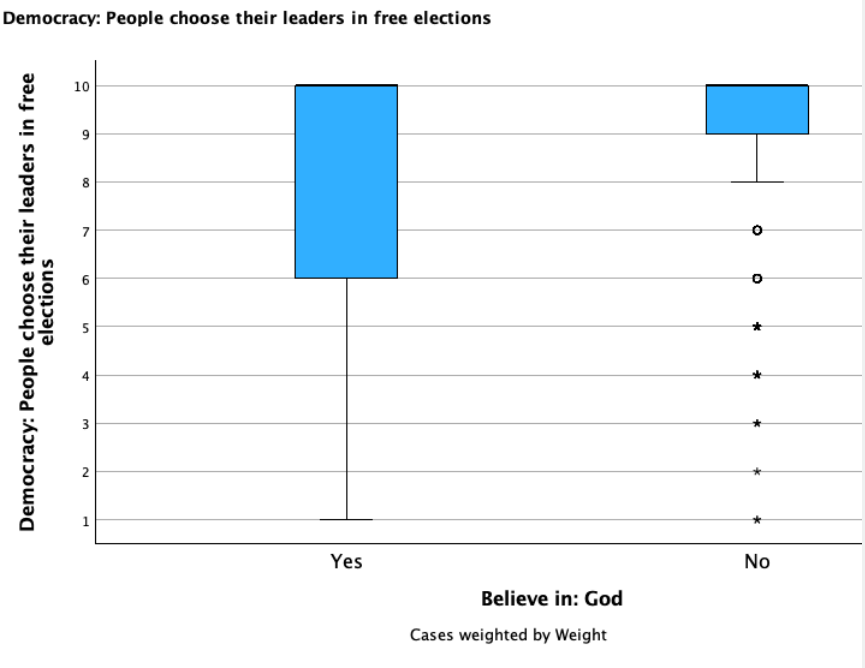
Check out the broken down box plots below and compare/contrast the groups:

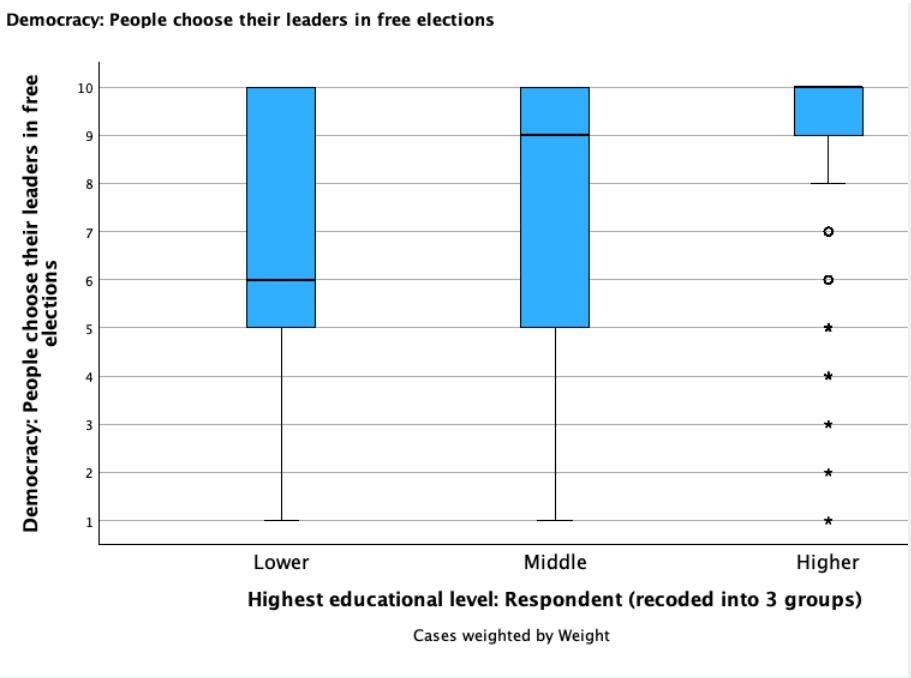


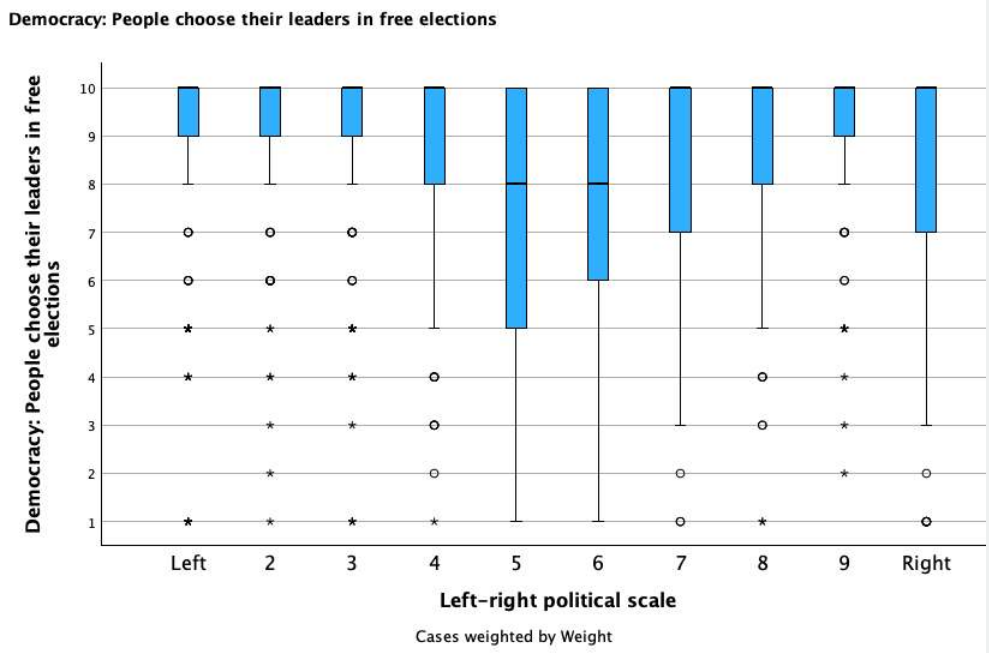


https://socialsci.libretexts.org/@go/page/186582


Democracy: People choose their leaders in free elections
Democracy: People choose their leaders in free elections
10
9
8
7
6
5
4
3
2
1
Very interested
Somewhat interested
Not very interested
Not at all interested
Interest in politics
Cases weighted by Weight


Democracy: People choose their leaders in free elections
Democracy: People choose their leaders in free elections
10
9
8
7
6
5
4
3
2
1
Yes
No
Believe in: God
Cases weighted by Weight





D. Five Point Summary

Let's map out each of the 5 points of the 5 point summary, what they are/mean, and how to interpret them.

The five points are:

- **Q_0: Minimum / Lower extreme**
- **Q_1: First / Lower quartile**
- **Q_2: Median**
- **Q_3: Third / Upper quartile**
- **Q_4: Maximum / Upper extreme**

Q_x means the value is greater than x quarters of the data.

1. Median

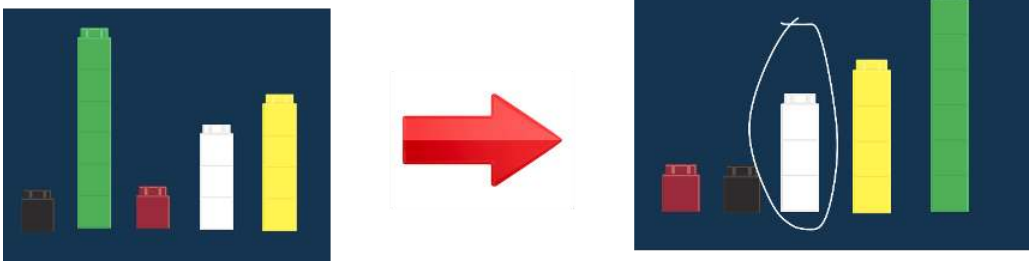
Median is a measure of center. It is one type of average (though colloquially average is often used synonymously with mean, see Chapter 3).

Definition: Median

Median: The middle number in a group of ordered numbers.

Median = Middle





The median is the 50th percentile. It is greater than 50% of the data and less than 50% of the data.
(As Q_2, the median is greater than two quarters, or one-half, or 50%, of the data.)

We'll revisit this next chapter when we compare the median to the mean, a different measure of central tendency. However, note that when you calculate the median, it does not matter if the minimum and maximum values are closer or further away from the middle. That does not change the middle. Because of this, the median is not sensitive to extreme values, which makes it a statistic often used for things like typical housing price or typical income earnings.

Interpretation: Median

General template: Half of respondents/cases had [median value] or fewer [label]; half of respondents/cases had [median value] or more [label].

For example, if we surveyed people and the median age was 53, we could write:
"Half of respondents were 53 years or younger; half of respondents were 53 years or older."

While this is a specific interpretation of the median, you could provide more context by bringing other data in like the minimum and maximum. Let's say the survey was only of adults, and that the minimum age was 18 and the maximum age was 89. Then you could write,
"Half of respondents were between 18 and 53 years old; the other half were between 53 and 89 years old."

If your median is the mean/midpoint of two middle values, then no cases actually have those values. Let's say that in the above example, no one was 53 years old, and that the two middle numbers were 51 and 55. You could provide a more contextual description, writing, "Half of respondents were between 18 and 51 years old; the other half were between 55 and 89 years old." If you wanted to ground this in the median, since you know no respondents were 53, you could also write, "Half of respondents were younger than 53; the other half were older than 53."

Writing: "The median is 53." tells your audience nothing. You need to give them context and tell the story of your statistic. Put it in real-life context.

2. Minimum, Maximum, and Range

Q_0, the minimum, is the lowest value. **Q_4, the maximum**, is the highest value.

Q_0 is greater than zero quarters of the data; it is the minimum value that has at least one case. Q_4 is greater than four quarters / all the rest of the data; it is the maximum value that has at least one case.

Interpretation: Minimum, Maximum

General Template:
Q_0 Minimum: The fewest/lowest/least [label] is [minimum value].
Q_4 Maximum: The fewest/lowest/least [label] is [minimum value].

Or, when the minimum is the lowest possible response and the maximum is the highest possible response:
Q_0 Minimum: At least one respondent/case has [minimum value] [label].
Q_4 Maximum: At least one respondent/case has [maximum value] [label].

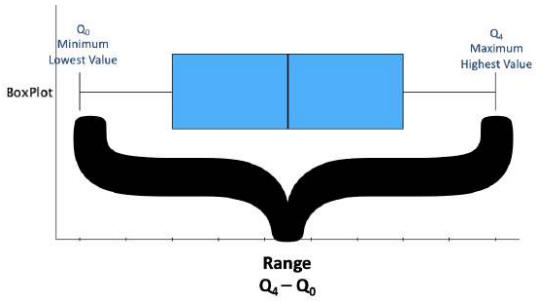
Examples:
Q_0 Minimum: The fewest number of pets someone has is 1 pet.
Q_4 Maximum: The greatest number of pets someone has is 9 pets.

Or,
Q_0 Minimum: At least one classmate has no pets.
Q_4 Maximum: At least one classmate has 5 or more pets.
At least one classmate has 9 pets. No one has more than 9 pets.

Range

Definition: Range

The range is the distance between the minimum and maximum values.



Interpretation: Range

General Template: Among our respondents/cases, there is a difference of [range value] between the respondent/case with the most [label] and the respondent/case with the least/fewest [label] (ranging from [minimum value] to [maximum value]).

Example: Among the survey takers, there is a difference of 8 pets from the person with the most pets to the person with the fewest pets (ranging from 1 pet to 9 pets).

While the range is only the difference of the minimum and maximum, providing those values as well gives context that

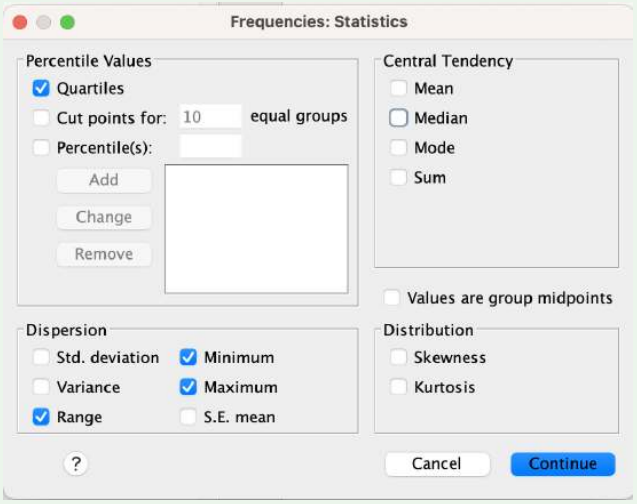
otherwise leaves your audience without knowing where to place the range. For example, a range of 8 pets could have been from a minimum and maximum of 0 to 8, 1 to 9, 25 to 33, or something else.

SPSS: Range

Finding the range with SPSS:

1. You can find the range when you run frequencies like you were doing for the five point summary.

Menu:
Make sure to also check the box for range in the bottom left of the statistics window.



Syntax

```
*Add RANGE to the STATISTICS line to include a calculation of range.
FREQUENCIES VARIABLES=VariableName
/NTILES=4
/STATISTICS=RANGE MINIMUM MAXIMUM
/ORDER=ANALYSIS.
```

Output:
The row labeled range gives you the range (e.g., here it is 9, 10-1=9):

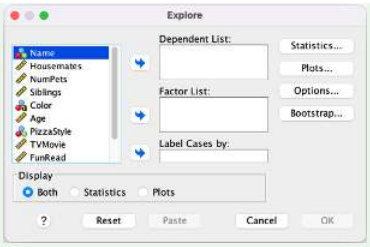
Statistics

Democracy: People choose their leaders in free elections

N	Valid	2534
	Missing	62
Range		9
Minimum		1
Maximum		10
Percentiles	25	7.00
	50	10.00
	75	10.00

2. You can find the range with Examine/Explore. However, with the menu drop-down it requires an extra step to include the lower or upper quartiles.

Menu:
Go to Analyze → Descriptive Statistics → Explore. Put your variable into the Dependent List window/box. At the bottom where it says "Display," make sure you either have "Both" or "Statistics" checked (Plots will only give you graphs/figures).



If you want to include the lower and upper quartiles, click on the Statistics button, check the Percentiles box, check the Quartiles box, then click Continue. Click OK to run the Explore command.

Syntax

```
*This will generate the range, five point summary, a box plot, and a histogram.
EXAMINE VARIABLES=VariableName
/PLOT BOXPLOT HISTOGRAM
/PERCENTILES(25,50,75) HAVERAGE
/STATISTICS DESCRIPTIVES
/CINTERVAL 95
/MISSING LISTWISE
/NOTOTAL.
```

3. Lower Quartile, Upper Quartile, and Interquartile Range

Definition: Q_1, the lower quartile

Q_1, the lower quartile, is the median of the lower half of the data. It is the 25th percentile, greater than 1/4 or 25% of the data (and less than 3/4 or 75% of the data).

Definition: Q_3, the upper quartile

Q_3, the upper quartile, is the median of the upper half of the data. It is the 75th percentile, greater than 3/4 or 75% of the data (and less than 1/4 or 25% of the data).

Interpretation: Lower Quartile, Upper Quartile

Template:
Q_1: 25% (1/4) of the respondents/cases had [quartile value] or fewer [label]. 75% (3/4) of the respondents/cases had [quartile value] or more [label].
Q_3: 75% (3/4) of the respondents/cases had [quartile value] or fewer [label]. 25% (1/4) of the respondents/cases had [quartile value] or more [label].

Example:
Q_1: 25% (1/4) of the parents had 4 or fewer pets. 75% (3/4) of the parents had 4 or more pets.
Q_3: 75% (3/4) of the parents had 6 or fewer pets. 25% (1/4) of the respondents/cases had 6 or more pets.

Again, you can be more specific by incorporating the minimum and maximum into your interpretation, even though it goes beyond the quartile statistic, e.g., "25% (1/4) of the parents had between 1 and 4 pets; 75% (3/4) had between 4 and 8 pets."

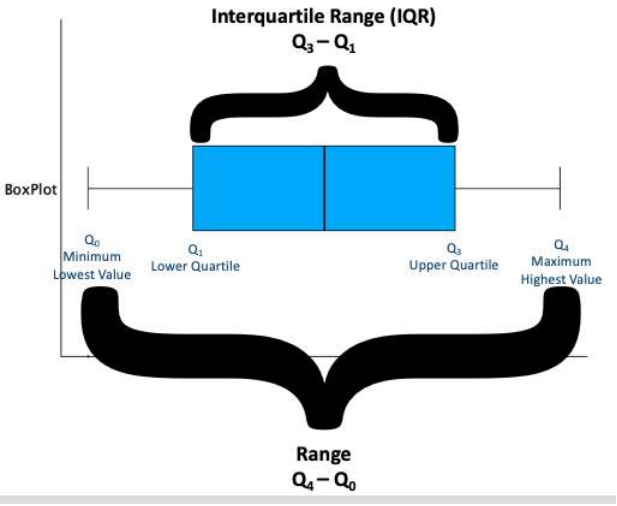
If Q1 is also the minimum or Q3 is also the maximum, you do not need to say [value] or [lower/more]. For example, if Q1 is 0 pets, you could say, "At least 1/4 of parents did not have any pets." You would not want to add, "3/4 of parents had 0 or more pets" since that does not add any useful information.

You can also look at your data to provide more accurate descriptions. What if Q3 is 4.75? Assuming no one has 4.75 pets, but that there are people with 4 pets and people with 5 pets, you could write, "75% (3/4) of the parents had between 1 and 4 pets; 25% (1/4) of the parents had between 5 and 8 pets." If you want to give the actual statistic, clarify in parentheses, e.g., "75% (3/4) of the parents had 4.75 or fewer (between 1 and 4) pets; 25% (1/4) of the parents had 4.75 or more (between 5 and 8) pets."

Interquartile Range (IQR)

Definition: Interquartile Range (IQR)

Interquartile range (IQR): The inter-quartile range (IQR) is the distance spanning the middle 50% of data, the distance between the lower and upper quartile values.



Interpretation: Interquartile Range

Template: The middle 50% of the data ranges from individuals/cases with [Q_1 value] [label] to individuals/cases with [Q_3 value] [label]. There is a range of [IQR value] [label] within the middle 50% of the data from the individual/case with the least to the individual/case with the most [label].

Example: The middle 50% of the data ranges from individuals with 1 pet to individuals with 4 pets. There is a range of 3 pets within the middle 50% of data from the person with the least to the person with the most pets.

SPSS: Interquartile range

Using SPSS, the frequencies command will not give you an IQR. However, if you run frequencies to get your lower and upper quartile, you can do your own math: Q_3-Q_1=IQR.

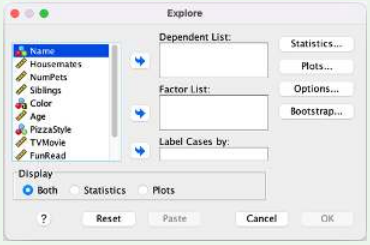
If you run Explore (menu) or Examine (syntax), as long as you have statistics checked / included, you will get a row with the Interquartile Range in your Descriptives output. You do not need to add the quartiles to the drop-down menu in order to generate an interquartile range.

SPSS: Using Explore/Examine to Examine Relationships

Just like we broke down box plots above by a second variable, you can do this using the Explore/Examine command, generating statistics (as well as a box plot) broken down by the categorical variable.

Menu:

When in the Explore window, put the categorical variable into the "Factor list:" box.



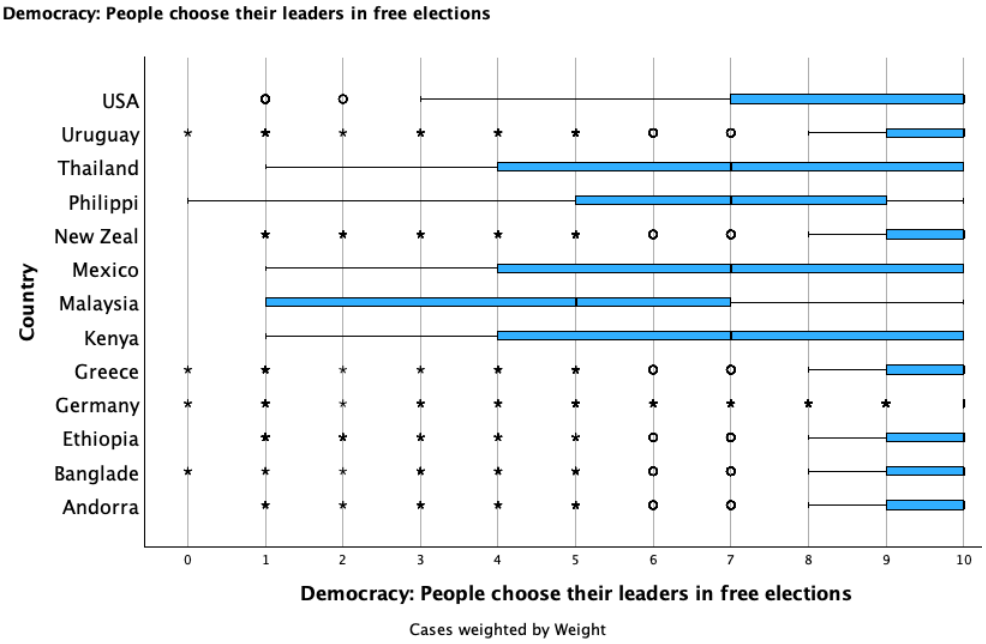
Syntax

```
EXAMINE VARIABLES=PrimaryVariableName BY CategoricalVariable2Name
/PLOT BOXPLOT
/COMPARE GROUPS
/PERCENTILES(25,50,75) HAVERAGE
/STATISTICS DESCRIPTIVES
/CINTERVAL 95
/MISSING LISTWISE
/NOTOTAL.
```

E. Synthesis Application: Cross-Country Comparison

My perception that free elections is a crucial component of democracy was similar to most U.S.-Americans. However, is this just my U.S. sensibilities? Or is this a univeral understanding of democracy? Let's compare this variable across different countries.

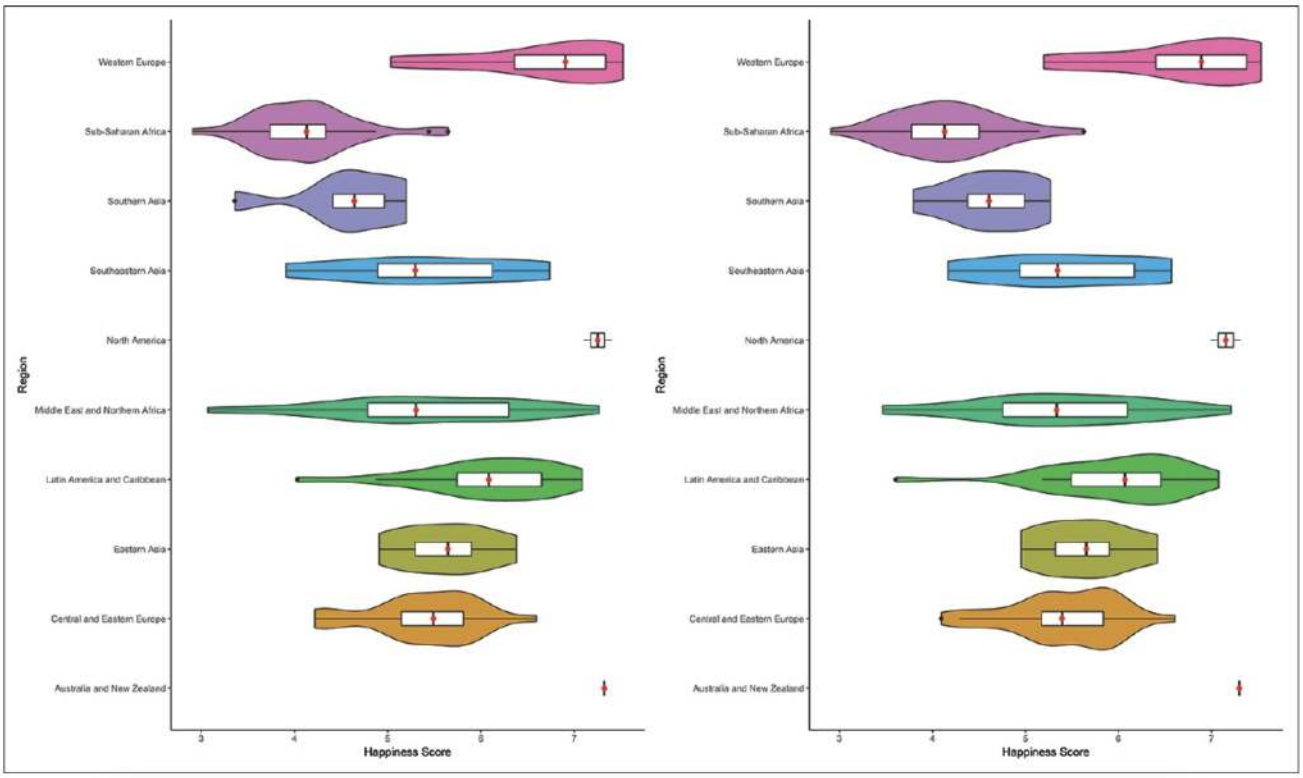
The median is 10 in 27 countries, 9 in 13 countries, 8.5 in 1 country, 8 in 17 countries, 7 in 5 countries, and 5 in 1 country. In all but one of the over 60 countries surveyed, a majority of respondents feel that free elections is an essential component or democracy or lean in that direction. However, not all countries are in as much agreement as the United States internally that it is an essential characteristic. There are also some countries where there is more agreement compared to the United States. The box plots below show the United States as well as a select group of other countries that have either comparatively high or low agreement that free elections are an essential component of democracy.



Example: Violin plots of happiness

Box and whisker plots are a great way to quickly get a summary of the distribution of data for a particular variable. Another plot, the violin plot, is similar to a box plot, but it also shows density distribution --- you get to see where the data is within each quartile.

Here is an example of a series of violin plots from a 2020 journal article by Vidushi Jaswal et al., in the Journal of Family Medicine and Primary Care. The left side is showing 2016 data and the right side is showing 2017 data. Each row has a violin plot for a different region. The violin plots show happiness scores from a Gallup World Poll.



This page titled 2: Levels of Measurement; Five Point Summaries is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Ezra Temko (Consortium of Academic and Research Libraries in Illinois (CARLI)) .

3: More Measures of Central Tendency and Spread- Mean, Standard Deviation, Z-scores

Chapter 2 focused on a measure of central tendency (median) and spread (the distribution and variation of data variation, through 5 point summaries and box plots). Chapter 3 focuses on another measure of central tendency: mean, and another measure of spread: standard deviation.

Statistics in Action

In order to show you how mean and standard deviation operates at different levels of measurement, this chapter will have examples around three questions. We will explore, within a U.S. context, (gender and racial) pay gaps, feelings towards journalists following the 2020 election, and Christian proselytization.

1. Mean

1.1. What is the mean?

Definition: Mean

Mean: The average[a] value; the value each case would have if all cases' values were redistributed equally among all the cases.

[a] *While both the mean and median are averages, the mean is the one colloquially used synonymously with average. When people make a claim like, "the average cost of a wedding in the United States is over $16,000" they are referring to mean.*

Notation

You will sometimes see mean notated with one of two symbols: $\bar{x}$ or μ

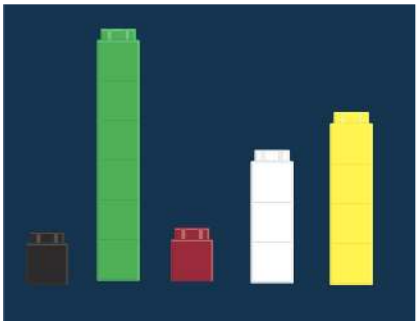
$\bar{x}$ Pronounced x-bar, this symbol is used for the mean of sample data (Samples are subsets of populations; we will learn about samples in the next chapter).

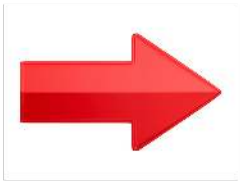
μ Pronounced myoo, the lowercase Greek letter mu is used for the mean of a population.

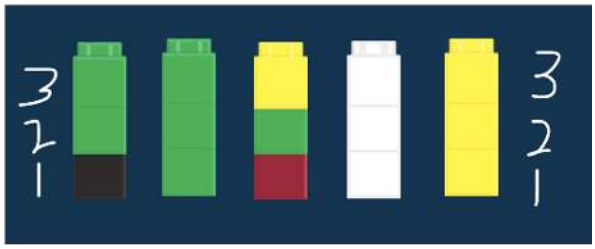
The mean is about equal distribution.
For example, if you took everyone's wealth and redistributed it so that everyone had an equal amount of wealth, that amount of wealth each person would have is the mean amount of wealth.

When you think of mean, think **"balancing out"**









1.2 Finding the Mean

Calculating the Mean

Let's say you and 9 of your friends were hosting a car wash. Some cars are dirtier than others, some of you are more efficient than others, etc. You don't all end up washing the same number of cars.



You worked your butt off and washed 10 cars. 2 of your friends washed 7 cars. 1 of your friends washed 6 cars. 4 of your friends washed 5 cars. 1 of your friends washed 4 cars. And 1 of your friends only washed 1 car.

Person	Cars Washed
You	
Friend 1	
Friend 2	
Friend 3	
Friend 4	
Friend 5	
Friend 6	
Friend 7	
Friend 8	
Friend 9	

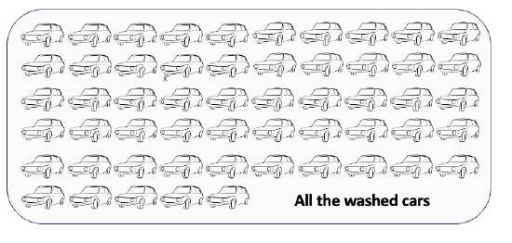
What is the average number of cars washed?

Well, we know you didn't wash a typical amount of cars among your friend group!

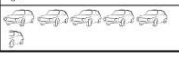
One way to measure this is the **median**. The median number of cars washed per person was 5 cars.
1 4 5 5 **5 5** 6 7 7 10 (Those are the number of cars washed in order. There are two middle numbers, both 5. Their midpoint or median is 5.)

Another way to measure this is the **mean**.

In total, you and your friends washed 55 cars.



How many cars would each of you have needed to wash to each wash the same number of cars? There were 10 of you. I can divide 55 cars up into 10 even groups. 55 ÷ 10 = 5.5. So each of you would have had to wash 5.5 cars. The mean is 5.5, another measure of the average number of cars washed per person.

Person	Cars Washed	Even Redistribution
You		
Friend 1		
Friend 2		
Friend 3		
Friend 4		
Friend 5		
Friend 6		
Friend 7		
Friend 8		
Friend 9		

How do I interpret the mean?

On average, each person washed 5 ½ cars. If you balance out the number of cars washed by each person, each person would have washed 5 ½ cars.

Comparing the mean and median: Is the average 5 or 5.5?

While the mean and median are both averages, they did not give you the exact same answer. This is because the mean took into account all the cars washed and redistributed them, while the median looked for the middle number of cars washed. You washed a lot of cars, which brought the mean up, since all 10 of those cars were included in the total number of cars washed. But whether you had washed 5 or 50 cars, the median would still be 5, as it did not change the number of cars washed by your friends who washed the middle number of cars. Means are influenced by extreme values and outliers, while medians are not.

Mean Calculation

Mean = total sum of all values ÷ number of items/cases

Sometimes you will see the formula written out with symbols, like this:

$$\bar{x} = \frac{\sum x}{n}$$

$$mean = \frac{sum\ of\ all\ the\ scores}{number\ of\ scores}$$

Remember that $\bar{x}$ is the sample mean. Σ, the Greek capital letter Sigma, means sum of. n is the sample size (number of cases). So this formula tells you that you find the sum of the values of all the cases and divided it by the number of cases.

Finding the Mean with SPSS

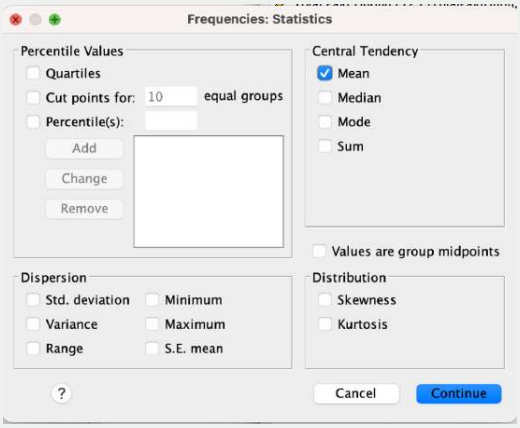
1. You can find the mean when you run frequencies like you were doing for the range and five point summary.

Menu:

In SPSS, go to Analyze → Descriptive Statistics → Frequencies.

Click on the Statistics button to the right. Then check the box for mean in the upper right of the statistics window and click

Continue.



Syntax:

*Include MEAN on the STATISTICS line to include a calculation of mean.
FREQUENCIES VARIABLES=VariableName
/FORMAT=NOTABLE
/STATISTICS=MEAN
/ORDER=ANALYSIS.

Output:

The row labeled mean gives you the mean (e.g., here it is 537.67)

Frequencies

Statistics

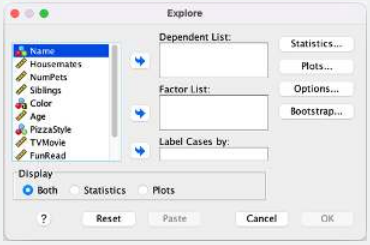
ID

N	Valid	1759
	Missing	1449
Mean		537.67

2. You can find the mean with Examine/Explore.

Menu:

Go to Analyze → Descriptive Statistics → Explore. Put your variable into the Dependent List window/box. At the bottom where it says "Display," make sure you either have "Both" or "Statistics" checked (Plots will only give you graphs/figures).



Click OK to run the Explore command.

Syntax:

*This will generate the mean along with other statistics and plots.
EXAMINE VARIABLES=VariableName
/PLOT BOXPLOT HISTOGRAM
/PERCENTILES(25,50,75) HAVERAGE
/STATISTICS DESCRIPTIVES
/CINTERVAL 95

```
/MISSING LISTWISE
/NOTOTAL.
```

1.3 The mean in action at different levels of measurement

Levels of Measurement

Means work best with ratio-level variables. They do not work with nominal-level variables. You can interpret means for ordinal-level ratio-like variables, with caution, and the mean for indicator variables has interpretable meaning, but tells a different story than for ratio-level variables.

Caution: No Means for Nominal Variables

Nominal-level variables do not have interpretable means!
(What would be a mean of people's names? The numbers assigned are arbitrary.) For nominal-level variables only the mode has meaning.

Example: In the SPSS output below, SPSS calculated a mean for respondent race. However, a mean of 1.37 does not have any actual meaning. The mode of 1 means that the value coded as 1 (in this case, White) was the most common/frequent response. That means that (in this sample) there were more white people than black people, and more white people than people of other races.

Statistics

RACE OF RESPONDENT

N	Valid	2538
	Missing	0
Mean		1.37
Median		1.00
Mode		1

RACE OF RESPONDENT

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	WHITE	1885	74.3	74.3	74.3
	BLACK	356	14.0	14.0	88.3
	OTHER	297	11.7	11.7	100.0
	Total	2538	100.0	100.0	

Indicator-level variables do have an interpretable mean, but with a different interpretation. For indicator variables, the mean is the proportion of respondents with the value coded as 1. An example below elaborates on this.

Statistics in Action: Pay gaps (ratio)

What do gender and race income gaps look like in the contemporary United States?

I analyzed U.S. Census data to find out. I used an IPUMS dataset (IPUMS originally stood for Integrated Public Use Microdata Series) of American Community Survey (ACS) 5-Year Data for 2017 to 2021. The ACS is intended to be representative of the U.S. population. I also used weighting to account for sampling error.

Pay gaps are earnings ratios. For example, a gender pay gap takes women's pay and divides it by men's pay; the resulting statistic is how much women on average earn for every dollar men earn.

Mean vs. Median

Comparing the mean and median tells us about the distribution of the data.

- When the mean is higher than the median, there are likely some extreme high values.
- When the mean is lower than the median, there are likely some extreme low values.
- When the mean and median are similar, the data is likely more balanced.

Think about this with housing prices. If you were looking for neighborhoods where you might be able to afford to buy or rent a house, you probably want to know the median price, as a few really expensive or inexpensive homes won't influence what the typical price is, and will give you a misrepresentation of a typical house. However, if you wanted to compare neighborhoods for what kinds of property taxes they brought in, or what kinds of economic capital the neighborhoods had, or to take into account that there are some much more or less expensive homes in the neighborhood, you may want to look at the mean.

Pay gaps

Conventionally pay gaps are based on median annual earnings. Why the median?

The median gives a better picture of the "typical" U.S. earner. It ignores the extremes (that are extreme), so doesn't take into account people with very high or very low incomes. This also means it doesn't take into account extreme income inequality, but because it's giving the "middle" earnings, they will be more "typical" for what the average U.S. earner would be making. The median might be a better measure of how much the typical person in one group is going to be making compared to the typical person in another group.

On the other hand, using the mean distorts pay gaps. It takes into account people with very high or very low incomes. Maybe this is less "typical," but it also better represents the distributions of incomes in our country and what it would look like if these earnings were redistributed (as a whole or among groups). Because there may be gender and race pay gaps among CEOs and other high-paying positions, ignoring these upper extremes only gives a partial picture of pay gaps. The mean might be a better measure of what gaps there are between groups and what that looks like per person.

So, which is a better measure of center? Of a typical U.S. adult? Of pay gaps?
There is no right answer (other than that mode is not the answer). Mean and median are both useful, and tell different stories.

Furthermore, conventionally the pay gap is calculated based solely on full-time year-round civilian workers, excluding institutionalized populations (e.g., people who live in nursing homes, military barracks, are incarcerated, etc.). The Pew Research Center also calculates pay gaps that include part-time workers. While we can look at industry specific pay gaps, and other pay gaps, who is working and how much they are working can also generate differences. Below I'll show you the mean and median both for the full-time civilian group and for everyone age 16 and older. The conventional group includes people who are at least 16 years old, report usually working at least 35 hours per week, worked 50 to 52 weeks in the past year, and are part of a civilian non-military non-institutionalized population. This included over 5 million cases. The second group includes everyone age 16+ regardless of institutionalization or work status. This included over 12 million cases.

To interpret the pay gaps in the table below, focus on the two groups in parentheses that are part of the ratio.
Template: [Group 1] earns [amount] for every $1 [Group 2] earns.
Example: First pay gap (Gender Pay Gap, Female:Male), using medians, among full-time, year-round, non-institutionalized U.S. workers ages 16+. The pay gap is $0.80. So, "Women earn 80¢ for every $1 men earn."
Example: Last row (Intersectional Pay Gaps, Hispanic Female:NH White Male), using medians, among all U.S. persons ages 16+. The pay gap is $0.34. So, "Hispanic women earn 34¢ for every $1 White men earn."

	Full-time, year-round civilian workers (age 16+)		*All age 16+*	
	Median	**Mean**	**Median**	**Mean**
Total	$53,948.00	$ 75,269.21	$28,575.00	$46,187.72
Gender Income & Pay Gap				
Male	$59,343.00	$ 84,676.98	$36,695.00	$57,281.92
Female	$47,690.00	$ 62,752.25	$22,000.00	$35,494.27
Pay Gap (Female:Male)	$ 0.80	$ 0.74	$ 0.60	$ 0.62
Race Income & Pay Gap				
Non-Hispanic (NH) White	$60,000.00	$ 83,482.70	$33,000.00	$52,193.78
NH Black	$43,105.00	$ 55,123.06	$21,579.00	$32,536.13
Pay Gap (NH Black:NH White)	$ 0.66	$ 0.72	$ 0.62	$ 0.65
NH Asian-American & Pacific Islander (AAPI)	$68,884.00	$ 94,846.17	$29,289.00	$55,050.03
Pay Gap (NH AAPI:NH White)	$ 1.14	$ 1.15	$ 1.05	$ 0.89
NH Other/Multiracial	$51,928.00	$ 70,806.07	$22,785.00	$39,551.42
Pay Gap (NH Other/Multi : NH White)	$ 0.85	$ 0.87	$ 0.76	$ 0.69
Hispanic	$40,000.00	$ 53,202.32	$21,562.00	$31,834.58
Pay Gap (Hispanic : NH White)	$ 0.64	$ 0.67	$ 0.61	$ 0.65
Intersectional Income & Pay Gaps				
NH White Male	$66,316.00	$ 95,104.46	$43,158.00	$65,848.25
NH White Female	$52,335.00	$ 67,497.56	$25,000.00	$38,936.95
Pay Gap (NH White Female: NH White Male)	$ 0.79	$ 0.71	$ 0.58	$ 0.59
NH Black Male	$45,000.00	$ 58,676.46	$22,658.00	$34,656.56
NH Black Female	$41,700.00	$ 51,774.50	$20,934.00	$30,634.57
Pay Gap (NH Black Female : NH Black Male)	$ 0.93	$ 0.88	$ 0.92	$ 0.88
Pay Gap (NH Black Female : NH White Male)	$ 0.63	$ 0.54	$ 0.49	$ 0.47
NH AAPI Male	$77,368.00	$105,927.69	$40,801.00	$69,939.40
NH AAPI Female	$60,789.00	$ 80,815.81	$20,934.00	$41,766.98
Pay Gap (NH AAPI Female : NH AAPI Male)	$ 0.79	$ 0.76	$ 0.51	$ 0.60
Pay Gap (NH AAPI Female : NH White Male)	$ 0.92	$ 0.85	$ 0.49	$ 0.63
NH Other/Mutliracial Male	$56,000.00	$ 78,563.38	$27,900.00	$46,869.77
NH Other/Mutliracial Female	$46,800.00	$ 61,642.17	$19,050.00	$32,618.75
Pay Gap (NH Other/Multi Female : NH Other/M	$ 0.84	$ 0.78	$ 0.68	$ 0.70
Pay Gap (NH Other/Multi Female : NH White M	$ 0.71	$ 0.65	$ 0.44	$ 0.50
Hispanic Male	$42,390.00	$ 57,155.41	$28,053.00	$39,216.22
Hispanic Female	$36,634.00	$ 47,070.36	$14,837.00	$24,312.07
Pay Gap (Hispanic Female : Hispanic Male)	$ 0.86	$ 0.82	$ 0.53	$ 0.62
Pay Gap (Hispanic Female : NH White Male)	$ 0.55	$ 0.49	$ 0.34	$ 0.37

Interpreting the mean

Template: The average [variable label] for [respondents/target population] is [mean value with label]. If you took everyone's [variable label] who is [respondents/target population] and redistributed it equally/evenly, each person would have [mean value with label].

This example: The average income for all U.S. respondents ages 16 or older is about $46,188. If you took everyone who is 16 or older's income and redistributed it so that everyone who is 16 or older earned the same amount, each person would have an income of just over $46,000.

Another example: The average age in the class is 8.5 years old. If you took all the years everyone in the class has been alive and distributed them equally, everyone in the class would be about 8 1/2 years old.

Statistics in Action: Feelings towards journalists (ordinal, ratio-like)

Following the 2020 elections, how did U.S. adults feel about journalists?

The American National Election Studies (ANES) study surveyed a representative sample of U.S. eligible voters. This analysis is based on a survey done in November and December following the 2020 elections.

ANES asks a series of questions using a feelings thermometer scale. Respondents rate their feelings towards political leaders, other people, and groups, on a scale from 0 to 100, with 0 being very cold feelings, 50 being neutral, and 100 being very warm feelings.



I ran (weighted) summary statistics for the question asking how respondents feel about journalists. Here were my results:

Statistics

POST: Feeling thermometer: journalists

N	Valid	2620
	Missing	23
Mean		50.33
Median		50.00
Mode		50
Std. Deviation		29.129
Minimum		0
Maximum		100

The mean is 50.33. The median is 50.

On average, U.S. eligible voters who took the survey feel neutral towards journalists. If you took everyone's responses, added all the temperatures together, and redistributed them evenly, everyone would have pretty middle/neutral feelings. We can also see from the median that half of respondents had neutral to warm feelings and half had neutral to cold feelings.

Did feelings differ based on whether or not people were Trump supporters?

I broke down the summary statistics by who respondents voted for. Here were my results:

Descriptives

	Did you vote for president? If so, who did you vote for for president?		Statistic
POST: Feeling thermometer: journalists	Joe Biden	Mean	69.45
		Median	70.00
		Std. Deviation	20.400
		Minimum	0
		Maximum	100
		Range	100
		Interquartile Range	30
	Donald Trump	Mean	28.25
		Median	30.00
		Std. Deviation	23.076
		Minimum	0
		Maximum	100
		Range	100
		Interquartile Range	40
	Other Candidate	Mean	39.08
		Median	40.00
		Std. Deviation	24.950
		Minimum	0
		Maximum	100
		Range	100
		Interquartile Range	45
			.063
			-.676
	Did not vote	Mean	59.50
		Median	63.07
		Std. Deviation	25.206
		Minimum	0
		Maximum	100
		Range	100
		Interquartile Range	20

Here we can see that, among respondents who voted for Trump, they on average felt fairly cold or unfavorable towards journalists (mean=28.25), whereas respondents who voted for Biden on average felt fairly warm or favorable towards journalists (mean=69.45). Note that this is imprecise, as we are adding numbers where people with the same feelings towards journalists may have reported different numbers, and we only have descriptors given to us for 9 benchmark temperatures. Sometimes we treat ordinal-level variables as ratio-like. It's not perfect, but it gives us some insights we otherwise may not have.

Statistics in Action: Proselytizing for Jesus (indicator)

I'm Jewish. When I moved from the Northeast to the Midwest, I experienced Christian proselytization permeating the culture in a way I had not experienced before. Are there regional differences that support my feelings? Or has this just been a result of who I have personally encountered?

The General Social Survey (GSS), a study that is representative of U.S. adults, regularly asks the question, "Have you ever tried to encourage someone to believe in Jesus Christ or to accept Jesus Christ as his or her savior?"

After re-coding the variable so that 0=no and 1=yes, I had an indicator variable, a dichotomous variable with 0 and 1 as its coding. Because it is an indicator variable, when the mean is calculated, all values being added up consist of 0s and 1s, so the sum ends up being the total number of cases with a value of 1. When the number of cases with a value of 1 is divided by the total number of cases, the quotient is **the proportion of cases with a value of 1.** Means for indicator variables are interpreted as the percent with a value of 1.

Here are my summary statistics and a frequency table for the GSS question, using 2022 data:

Statistics

Have you ever tried to encourage someone to believe in Jesus Christ or to accept Jesus Christ as his or her savior?

N	Valid	3479
	Missing	65
Mean		.38
Median		.00
Mode		0
Std. Deviation		.486
Range		1
Minimum		0
Maximum		1
Percentiles	25	.00
	50	.00
	75	1.00

Have you ever tried to encourage someone to believe in Jesus Christ or to accept Jesus Christ as his or her savior?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	No	2148	60.6	61.7	61.7
	Yes	1331	37.6	38.3	100.0
	Total	3479	98.2	100.0	
Missing	System	65	1.8		
Total		3544	100.0		

The mean is 0.38, which indicates that 38% of respondents had a 1 value (and the other 100%-38%=62% had a 0 value). Interpreted, 38% of respondents, almost 2/5, have engaged in Christian proselytization. You can see that this matches the frequency table below the output with the mean. The summary statistics also include a median and mode, both 0; with two possible values, these both simply tell you which value has a majority of cases.

Next I broke down this variable by region. I used GSS's region variable, which breaks down regions as follows:

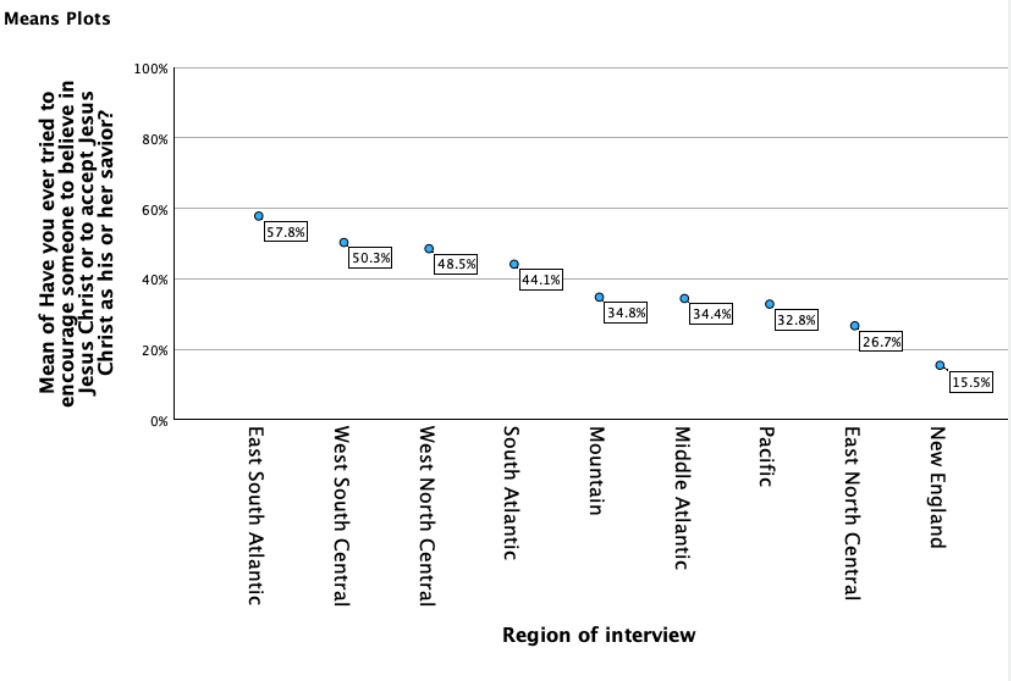


Here are my means, broken down by region:

Descriptives

	Region of interview		Statistic
Have you ever tried to encourage someone to believe in Jesus Christ or to accept Jesus Christ as his or her savior?	New England	Mean	.15
	Middle Atlantic	Mean	.34
	East North Central	Mean	.27
	West North Central	Mean	.49
	South Atlantic	Mean	.44
	East South Atlantic	Mean	.58
	West South Central	Mean	.50
	Mountain	Mean	.35
	Pacific	Mean	.33

To quickly interpret this, I made a graph using the one-way ANOVA means plot feature (we'll learn about one-way ANOVA in Chapter 7).



Here you can see that there are substantive differences in Christian proselytization around the country. While in New England about 2/13 of respondents proselytize for Jesus, and about 1/4 of those in the East North Central, it's closer to 1/3 in the Pacific, Middle Atlantic, and Mountain regions, above 2/5 in the South Atlantic, around 1/2 in West North Central and West South Central, and almost 3/5 of respondents in the East South Atlantic.

Interpretation (mean with an indicator variable)

Template:
Convert your mean into a percentage (e.g., 0.38 becomes 38%).
[mean]% of [cases] [label of 1].

Examples:
Mean=0.33
About 1/3 (33%) of U.S. adult respondents in the Pacific region have tried to encourage someone to believe in Jesus Christ or to accept Jesus Christ as their savior.

Mean=0.58.
58% of U.S. eligible voters voted in the 2020 election.

1.4 The Mode

Other Measures of Central Tendency: Mode

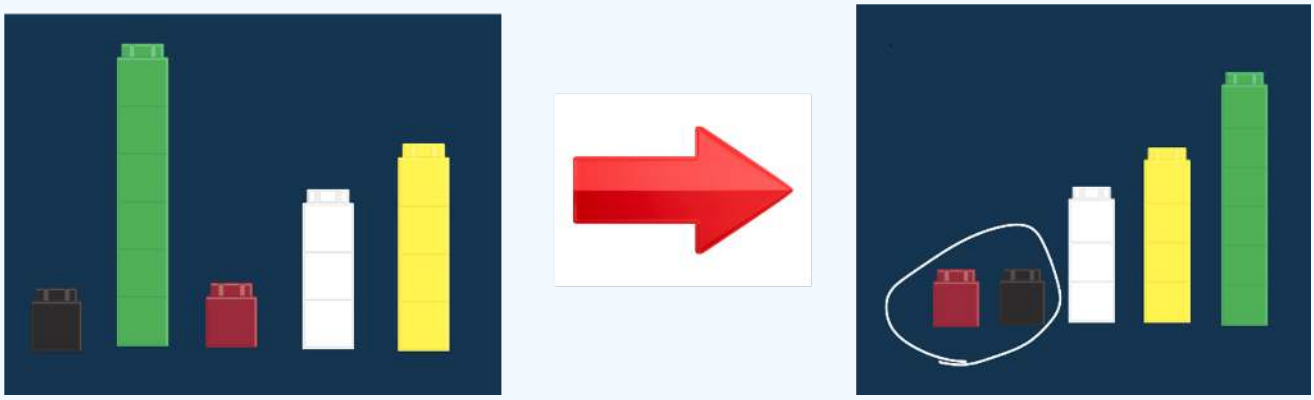
The mean and median are two averages. Another measure of central tendency is the **mode**.

Definition: Mode

The mode is the repeated value that appears most frequently

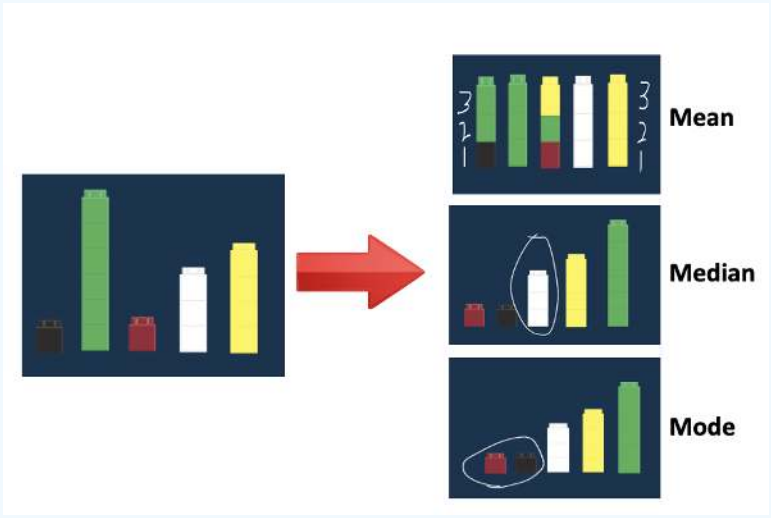
Mode = Most





Data can have no mode if no values repeat / each value is unique, called **amodal** (pronounced ay-mohd-l with a long a), e.g., if no one washed the same number of cars. Data can have multiple modes if more than one value is tied for most popular, called **bimodal** for two modes and **multimodal** for multiple modes, e.g., a mode of both 4 and 5 if 3 friends had washed 4 cars, 3 friends had washed 5 cars, and you didn't have any other number of cars that 3 or more people washed, meaning the most common number of cars washed was 4 and 5 cars.

The mode is generally useless. In some cases it can tell you the most **common** or **popular** item (e.g., What's your favorite ice cream flavor?), and in elections the mode, called the **plurality**, tells you who got the most votes, even if they did not get a majority. However, the mode can be very misleading. Let's say 3 people washed 2 cars each, and then among the remaining car-washers, 2 washed 19 cars, 2 washed 20 cars, 2 washed 21 cars, and 1 washed 22 cars. The mode is 2 cars, even though it does not seem to be the typical number of cars washed. Or let's say that 30 kids voted on which movie they wanted to watch, and 7 said Frozen, 6 said The Lego Batman Movie, 6 said The Lego Ninjago Movie, 5 said The Lego Movie, and 4 said The Lego Movie 2: The Second Part. The mode is Frozen, even though 3/4 of the kids said they want to watch some form of Lego movie. Finally, the more possible values (e.g., consider poverty rates, especially rounded to the hundredths place), the less likely the mode is to be useful, because multiple cases having the same value do not indicate there is any grouping or typicality around that value.



Measures of Central Tendency: Mean, Median, Mode

2. Standard Deviation

2.1 Standard Deviation

Definition: Standard Deviation

Standard deviation: An estimate of the average distance of cases' data values from their overall mean.

Variance: Another term you may hear is variance, which is the square of the standard deviation.

Notation

You will sometimes see standard deviation notated with one of two symbols: s or σ

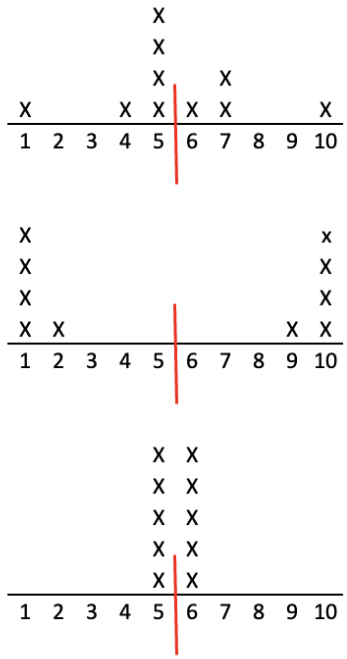
s A lower-case s is used for the standard deviation of sample data.

σ A lower-case Greek letter sigma is used for the standard deviation of a population.

What is standard deviation?

Standard deviation is a measure of how much variability there is in the data.

Returning to our car washing example, you and your friends washed a mean of 5.5 cars each. That's the top figure below, with the Xs representing each person and the red line indicating the mean. However, let's look at two other possibilities of how a group of 10 washed 55 cars. Look at the two charts below it. Both of those also have a mean of 5.5, but their distributions are very different. With the mean, we know one measure of central tendency, of how many cars each person washed on average, but as you can see in the middle chart, you can have a mean of 5.5 without anyone washing close to that number of cars, or in the bottom chart, you can have a mean of 5.5 and everyone washed close to that many cars. The mean does not tell us anything about the distribution or variation of how many cars people washed. It is solely a measure of central tendency.



Standard deviation is a statistic that measures how much variability there is in the data. It is an aggregate measure of how far away each number in the set of data is from their mean.

- The larger the standard deviation, the more spread out the data values are and the more variability they have. They tend to be further away from the mean.

- The smaller the standard deviation, the less spread out the data values are and the less variability they have. They tend to be closer to the mean.

? Exercise 3.1

The three charts above have standard deviations of 0.5, 2.3, and 4.6. Can you match these standard deviations to the correct charts?

Answer

The second chart has the most variability or spread; its standard deviation is 4.6. The last chart has the least variability or spread; it's standard deviation is 0.5. Our original data, the top chart? It has some variation from the mean; its standard deviation is 2.3.

2.2 Finding the standard deviation

✔ Calculating the standard deviation

Let's look at how the standard deviation is calculated. You will see it is basically an average distance of the values from the mean. Follow along with the calculation steps to get a sense of the what standard deviation means conceptually.

Standard deviation calculation

Sometimes you will see the formula written out with symbols, like this:

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Again, s means standard deviation, Σ means sum, x-bar is the mean, and n is the number of cases. Basically this formula tells you to follow the steps below:

1. **Calculate the mean.**
2. **Calculate the difference between each value and the mean.**
3. **Square each of the differences.**
4. **Add up the squared differences.**
5. **Divide by n-1**
6. **Take the square root of your answer.**

Let's return to the car washing example. We already have our raw data (how many cars each person washed).

Step 1. Calculate the mean

The standard deviation is a measure of average departure from the mean, so we need to start by determining the mean. We already did Step 1 of calculating our mean above. 55÷10=5.5

Step 1 *Mean*	Data Values
	10
	7
	7
	6
	5
	5
	5
	5
	4
	1
Sum:	55
Sample size:	10
Mean:	5.5

Step 2: Calculate the difference between each value and the mean.

Next we find out how different each case is from the mean. This generates a list of deviations from the mean.

Step 2 *Individual Deviations*	Data Values	Mean	Value - Mean
	10	5.5	4.5
	7	5.5	1.5
	7	5.5	1.5
	6	5.5	0.5
	5	5.5	-0.5
	5	5.5	-0.5
	5	5.5	-0.5
	5	5.5	-0.5
	4	5.5	-1.5
	1	5.5	-4.5
Sum:	55		
Sample size:	10		
Mean:	5.5		

Step 3. Square each of the differences.

Notice that some of the deviations are positive and some are negative. The final step is to find their average. However, adding them up would be problematic, as for example adding the case that is 1.5 above and 1.5 below the mean would give you a summed deviation of 0 instead of a summed deviation of 1.5. One way to account for this would be to take the absolute value of each deviation. If you did that and continued through the final steps, you would end up with a statistic called the mean absolute deviation. We use a different solution to make all the deviations positive---squaring them. This will give us the standard deviation that works with normal distributions and inferential statistics (see Chapter 4).

Step 3. *Squre each difference*	Data Values	Mean	Value - Mean	$(\text{Value - Mean})^2$
	10	5.5	4.5	20.25
	7	5.5	1.5	2.25
	7	5.5	1.5	2.25
	6	5.5	0.5	0.25
	5	5.5	-0.5	0.25
	5	5.5	-0.5	0.25
	5	5.5	-0.5	0.25
	5	5.5	-0.5	0.25
	4	5.5	-1.5	2.25
	1	5.5	-4.5	20.25
Sum:	55			
Sample size:	10			
Mean:	5.5			

Step 4. Add up the squared differences.

We are going to find an average of the differences. First we find the sum of our squared differences. This is the total of all the squared deviations.

Step 4 Sum of squared differences	Data Values	Mean	Value - Mean	(Value - Mean)^2
	10	5.5	4.5	20.25
	7	5.5	1.5	2.25
	7	5.5	1.5	2.25
	6	5.5	0.5	0.25
	5	5.5	-0.5	0.25
	5	5.5	-0.5	0.25
	5	5.5	-0.5	0.25
	5	5.5	-0.5	0.25
	4	5.5	-1.5	2.25
	1	5.5	-4.5	20.25
Sum:	55		Sum:	48.5
Sample size:	10			
Mean:	5.5			

Step 5. Divide by n-1

We divide by one less than the number of cases. This has to do with degrees of freedom---if we know the mean and we know the values for the other 9 cases, then we already have enough information to figure out the 10th case. Additionally this makes the standard deviation a little bigger, which when we are estimating population parameters from sample statistics, helps us account for sampling error.

Step 5 Divide by n-1	Data Values	Mean	Value - Mean	(Value - Mean)^2
	10	5.5	4.5	20.25
	7	5.5	1.5	2.25
	7	5.5	1.5	2.25
	6	5.5	0.5	0.25
	5	5.5	-0.5	0.25
	5	5.5	-0.5	0.25
	5	5.5	-0.5	0.25
	5	5.5	-0.5	0.25
	4	5.5	-1.5	2.25
	1	5.5	-4.5	20.25
Sum:	55		Sum:	48.5
Sample size:	10		n-1	9
Mean:	5.5		Sum/(n-1)	5.3889

This value, here 5.4, is called the **variance**.

Step 6. Take the square root of your answer.

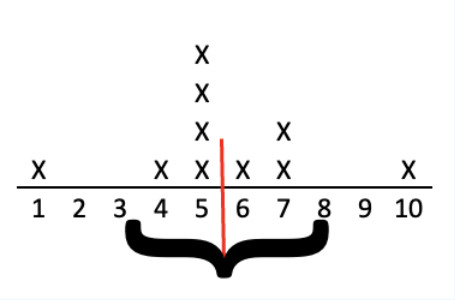
Recall that we squared our deviations. Therefore 5.4 is not a good estimate of average distance from the mean. We need to

unsquare our variance by taking its square root. This gives us the standard deviation.

Step 6 *Square root*	Data Values	Mean	Value - Mean	$(\text{Value - Mean})^2$
	10	5.5	4.5	20.25
	7	5.5	1.5	2.25
	7	5.5	1.5	2.25
	6	5.5	0.5	0.25
	5	5.5	-0.5	0.25
	5	5.5	-0.5	0.25
	5	5.5	-0.5	0.25
	5	5.5	-0.5	0.25
	4	5.5	-1.5	2.25
	1	5.5	-4.5	20.25
Sum:	55		Sum:	48.5
Sample size:	10		n-1	9
Mean:	5.5		Sum/(n-1)	5.3889
			Square root	2.3214

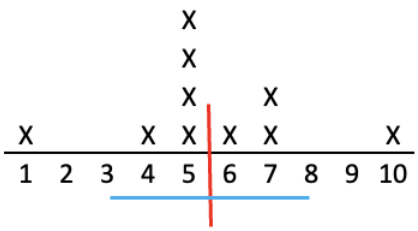
How do I interpret the standard deviation?

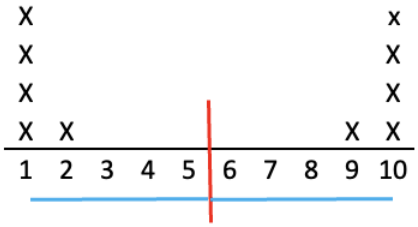
On average, each person washed 5 ½ cars. However, there was some variation from this. On average people washed 2.3 more or fewer cars than 5.5 (e.g., 3.2 and 7.8 cars).

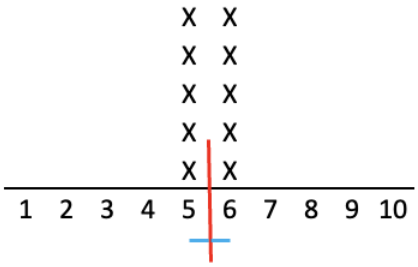


Note: We can see in this example that the "typical" car washer actually did not wash 3.2 or 7.8 cars. Most car washers are closer to the mean. However, with the extreme values of 1 car and 10 cars, this draws the standard deviation out considerably.

However, check out our other two examples of car washers below. You can see in these examples, the standard deviation shows us about where the typical cases are:



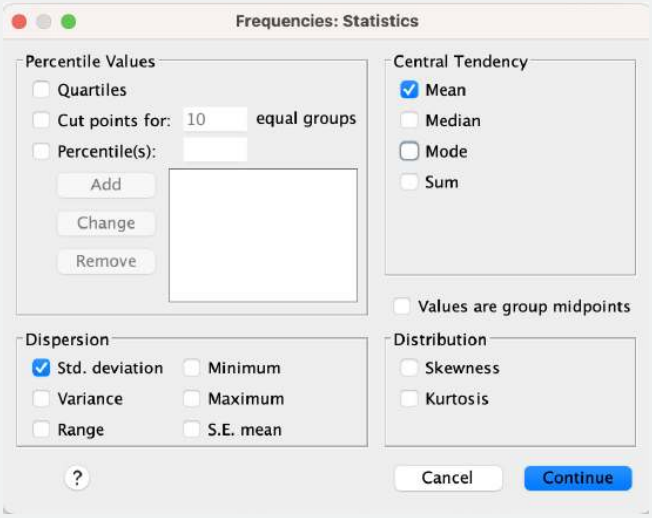




Finding the standard deviation with SPSS

1. You can find the standard deviation the same way you found the mean.

Menu:
In SPSS, go to Analyze → Descriptive Statistics → Frequencies.
Click on the Statistics button to the right. Then check the box for standard deviation in the bottom left corner of the statistics window and click Continue.



Syntax:
*Include standard deviation (stddev) on the STATISTICS line to include a calculation of standard deviation.

FREQUENCIES VARIABLES=VariableName
/FORMAT=NOTABLE
/STATISTICS=MEAN STDDEV
/ORDER=ANALYSIS.

Output:
The row labeled standard deviation gives you the standard deviation (here it is 17.93).

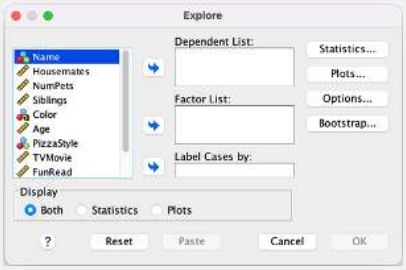
Statistics

Age of respondent

N	Valid	3360
	Missing	184
Mean		47.13
Std. Deviation		17.930

2. You can find the standard deviation with Examine/Explore.

Menu:
Go to Analyze → Descriptive Statistics → Explore. Put your variable into the Dependent List window/box. At the bottom where it says "Display," make sure you either have "Both" or "Statistics" checked (Plots will only give you graphs/figures).



Click OK to run the Explore command.

Syntax:
*This will generate the mean along with other statistics and plots.
EXAMINE VARIABLES=VariableName
/PLOT BOXPLOT HISTOGRAM
/PERCENTILES(25,50,75) HAVERAGE
/STATISTICS DESCRIPTIVES
/CINTERVAL 95
/MISSING LISTWISE
/NOTOTAL.

2.3 Interpreting the standard deviation

Statistics in action: Feelings towards journalists

Remember our summary statistics for feelings towards journalists?

Statistics

POST: Feeling thermometer: journalists

N	Valid	2620
	Missing	23
Mean		50.33
Median		50.00
Mode		50
Std. Deviation		29.129
Minimum		0
Maximum		100

While the mean and median were close to neutral, we saw with the breakdown by who people voted for for president that the typical repondent was not actually neutral.

Here we can see a standard deviation of 29.1. This means that on average, a respondent was about 29.1 degrees away from the mean of 50. This is about 20 or 80 (50-29.1 and 50+29.1, rounded). So we know that most respondents were not neutral, but felt warm or felt cold. Alternatively there could be a collection of neutral respondents and then a collection of respondents at 0 and 100, very cold and very warm. We don't know without seeing more of the distribution, since standard deviation is just one statistic that summarizes variation into one number. This is where box plots and other measures of variation can be helpful.

Next chapter we'll learn about normal distributions, and there will be more specific ways to be able to interpret standard deviations when they are part of normally distributed data. For now, just focus on whether there seems to be little variation, with cases close to the mean, some variation, or a lot of spread, with cases far away from the mean.

Levels of Measurement

Standard deviation is deviation from the MEAN. Therefore standard deviation only works with variables you can calculate and interpret a mean for. It is also useless for indicator variables (just give the percentage distribution for each of the two categories).

Interpreting standard deviation

The standard deviation can be a bit harder to access, so depending on your audience, it can be helpful to start your storytelling with a refresher/summary of what standard deviation is, e.g., "The standard deviation is a measure of how much individual cases vary away from the mean. Unlike the mean and median, which are measures of central tendency, this is a measure of spread."

Template: There was [not much, some, a good deal of] variation among [respondents/target population]. The average [respondent/target population] had an average deviation of about [value] from the mean of [mean value] [label].

Example: There was some variation among car washers. The average car washer had an average deviation of about 2 cars from the mean of 5 1/2 cars.

Example: There was a good deal of variation in how many shots were taken in each game. The average number of shots in a game had an average deviation of about 5½ shots from the mean of 11.7 shots per game.

2.4 Z-scores

What are “z-scores”?

Definition: Z-score

A z-score is the number of standard deviations a value is from its mean.

z = 0 is the mean (zero standard deviations away from the mean)
z = -2 is two standard deviations below the mean
z = +3 is three standard deviations above the mean
Etc.

For example, let's say you have a mean of 5 and a standard deviation of 1.5. Here would be the values at various z-scores.

Z-score	Value
z=-3	0.5
z=-2	2
z=-1	3.5
z=0	5
z=1	6.5
z=2	8
z=3	9.5

This may not seem very relevant yet. However, it will come in handy in the future.

2.5 Outliers

Outliers

Last chapter discussed one systematic way to determine if a case for a ratio-like variable is an outlier. You took 1.5 times the IQR and then went that far out below the lower quartile and above the upper quartile. Anything beyond those values was an outlier.

Another way that people systematically determine outliers is by using the threshold of three standard deviations away from the mean. If a value is below z=-3 or above z=+3, it can be considered an outlier.

For example, in the table for 2.4 above, any values below 0.5 or above 9.5 would be considered outliers.

These different methods of determining outliers can lead to different outcomes in terms of whether or not a value is considered an outlier.

2.6 Other measures of variation

Exercise

If half of your survey respondents' favorite color is green, one-quarter prefer orange, and one-quarter prefer black, what is the mean and standard deviation for your favorite color variable?

Answer

Favorite color is a nominal-level variable, so you cannot calculate and interpret a sensical mean or standard deviation.

Index of Qualitative Variation (IQV)

The IQV can be used to measure variation for nominal-level variables.

It ranges from 0 to 1.

- 0 = no diversity at all (all cases in one category)

- 1 = maximum diversity (equal number of cases in all categories)

For example, if in a class everyone got As, the IQV value would be 0. If 1/5 got As, 1/5 Bs, 1/5 Cs, 1/5 Ds, and 1/5 Fs, the IQV value would be 1.

Additional measures you may come across:

Gini Coefficient

The Gini Coefficient is used for income inequality. It also ranges from 0 to 1, with 0 representing perfect income equality (everyone has an equal share) to 1 for for perfect income inequality (1 person/group has all the income).

Index of Dissimilarity

The Index of Dissimilarity measures racial segregation of two groups in a geographic area. It ranges from 0 at maximum integration to 100 at maximum segregation.

This page titled 3: More Measures of Central Tendency and Spread- Mean, Standard Deviation, Z-scores is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Ezra Temko (Consortium of Academic and Research Libraries in Illinois (CARLI)) .

4: Histograms, Normal Distributions, and the Central Limit Theorem

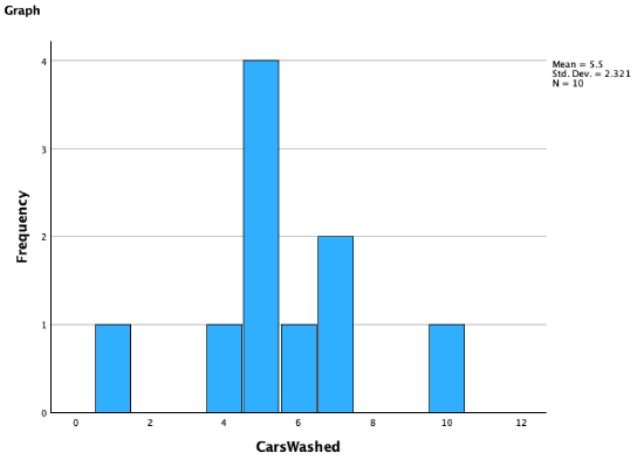
This chapter covers histograms, normal and skewed distributions, and introduces you to inferential statistics, including through the Central Limit Theorem and a discussion of weighting.

4a. Histograms

4a.1. Histograms

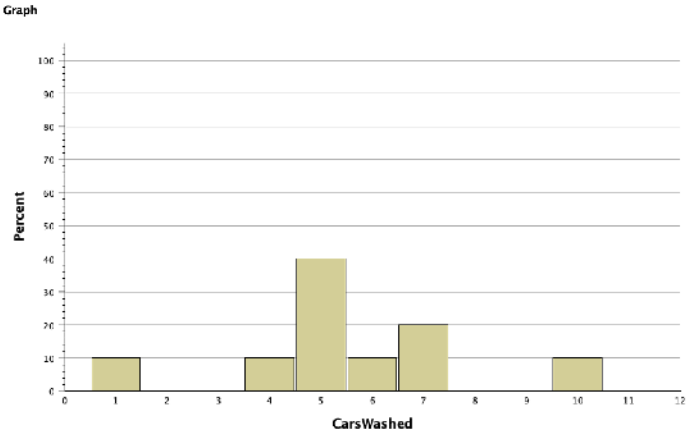
Definition: Histogram

Histogram: A frequency distribution graph for numeric data.



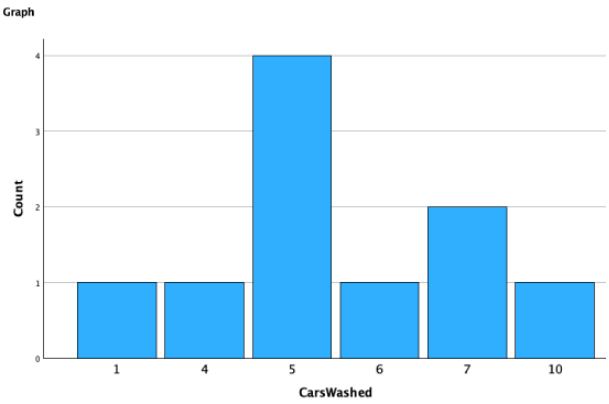
Returning to the example from Chapter 3, the histogram above shows the frequency of how many cars you and each of your nine friends washed.

While the graph above shows counts, another way to display a histogram is to show percentages.



The key difference between a histogram and a bar graph is that a bar graph is a frequency distribution table of categorical rather than quantitative data. Therefore the x-axis on a bar graph has category labels rather than a number line.

For example, here is what happens when I try to run the cars washed variable as a bar graph:



You can see that it includes the various numbers of cars washed as different categories. While generally a histogram is more appropriate for quantitative data, there are some exceptions. For example, if you have a ratio-like Likert scale variable, making a bar graph will give you labels (e.g., "strongly disagree") instead of numbers on the x-axis.

Traditionally the bars in a bar graph have space in between them, while the bars in a histogram are touching.

How to run a histogram in SPSS

1. *Syntax:*

The syntax language below will produce a histogram. Just replace "VariableName" with the name of your variable.

```
GRAPH
/HISTOGRAM=VariableName.
```

Alternatively, the Explore menu item and Examine syntax can also produce a histogram.

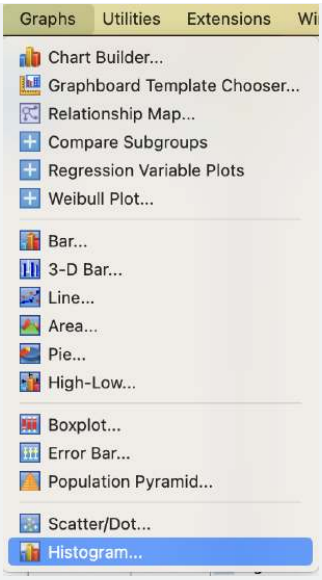
```
EXAMINE VARIABLES=VariableName
/PLOT HISTOGRAM
/PERCENTILES(25,50,75) HAVERAGE
/STATISTICS DESCRIPTIVES
/CINTERVAL 95
/MISSING LISTWISE
/NOTOTAL.
```

Another option is to use the Frequencies syntax to produce a histogram.

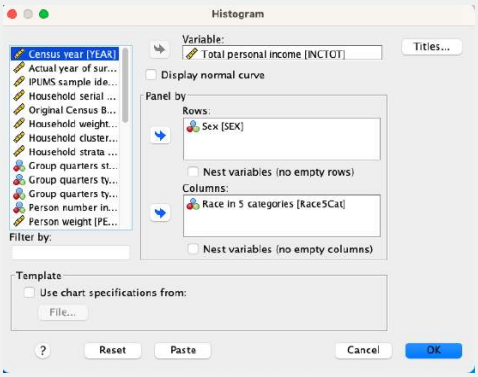
```
FREQUENCIES VARIABLES=VariableName
/FORMAT=NOTABLE
/HISTOGRAM
/ORDER=ANALYSIS.
```

2. *Menu:*

To produce a histogram using the drop-down menus, click on Graphs → Histogram.



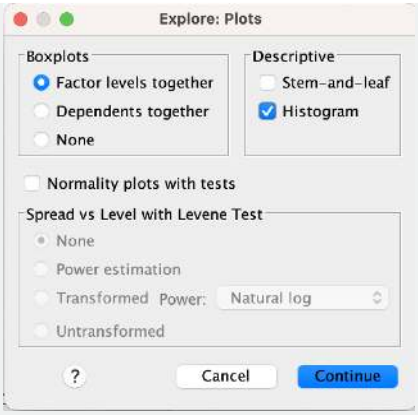
Put the variable you want a graph of in the variable box, and click OK.
Below you will see I added variables to the panel rows and columns boxes as well, which produced the series of partial histograms for income you will find below.



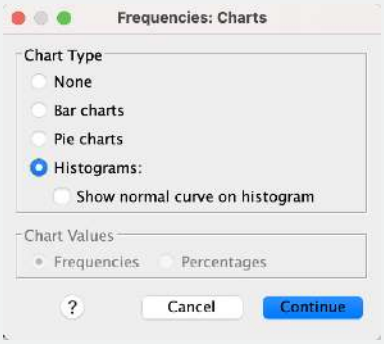
Tip for datasets with long lists of variables:

Some datasets have a lot of variables to look through to find the one you want! If you right-click on the list of variables in that left menu, you can choose to show the variable name instead of the label, and can sort them alphabetically. If I sorted the variable list alphabetically by name and then hit an "I" on the keyboard, SPSS would bring me down to the next variable that starts with an I.

Alternatively, you can produce a histogram as part of Explore. Go to Analyze → Descriptive Statistics → Explore, and put your variable into the Dependents list. Make sure you either have Plots checked on the bottom or Both, not just Statistics. Click the Plots button and make sure the "Histogram" option is checked. Then click Continue and OK.

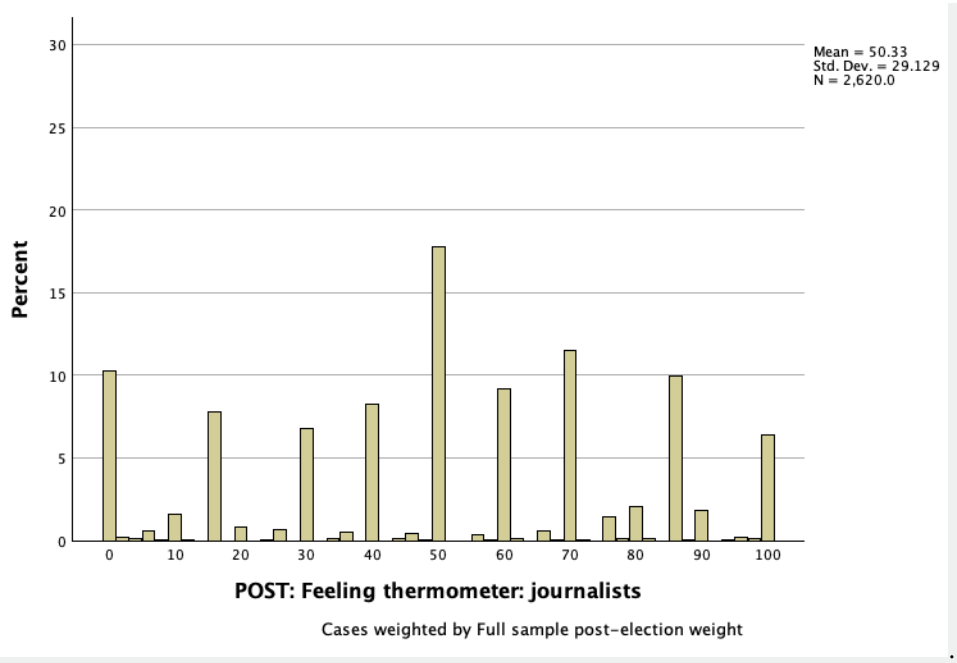


Or, you can produce a histogram as part of Frequencies. Go to Analyze → Descriptive Statistics → Frequencies, click the Charts button on the right, and check the Histograms button, then click Continue. Drag your variable into the variables box and click OK to generate your histogram.



Statistics in Action: Feelings towards journalists

Continuing with the Chapter 3 example, here is a histogram of U.S. eligible voter respondents' feelings towards journalists on the feelings thermometer following the 2020 election.





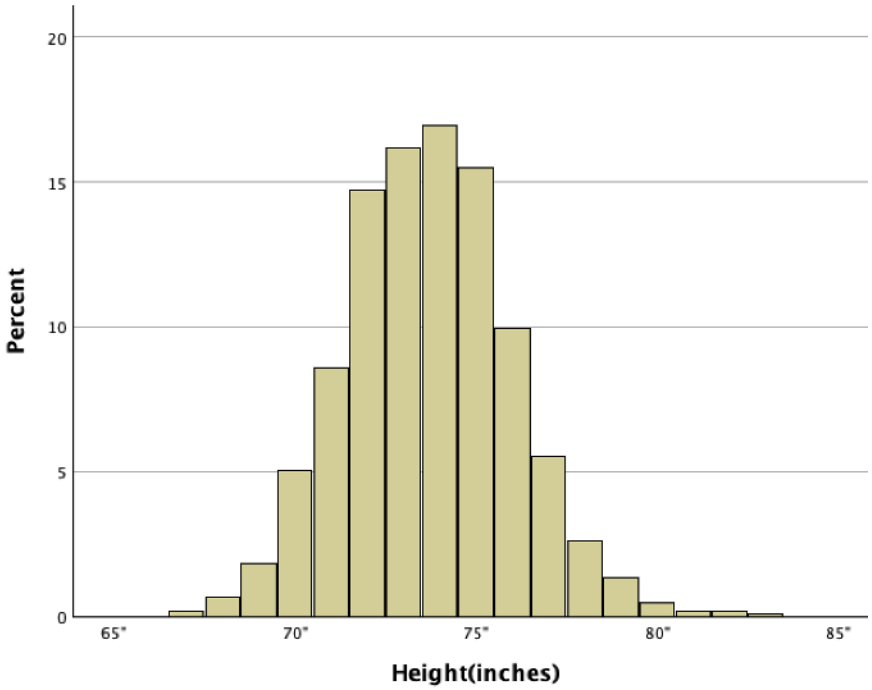
This histogram looks a little strange to me at first glance. The reason for this is because of the thermometer diagram respondents were shown to help them rate journalists. While respondents could give any number from 0 to 100, most gave numbers highlighted on the thermometer.

In this histogram, we can see that the most common response (over 1/6 of respondents) was neutral, but that there were respondents that rated journalists across the temperatures, and that there was still a sizable proportion of respondents who felt very cold (over 1/10 of them) or very warm (over 1/20).

4a.2 Normal and Skewed Distributions

Histograms come in different shapes based on data's distribution.

Pictured below is a histogram of the height of 1,035 Major League Baseball players.



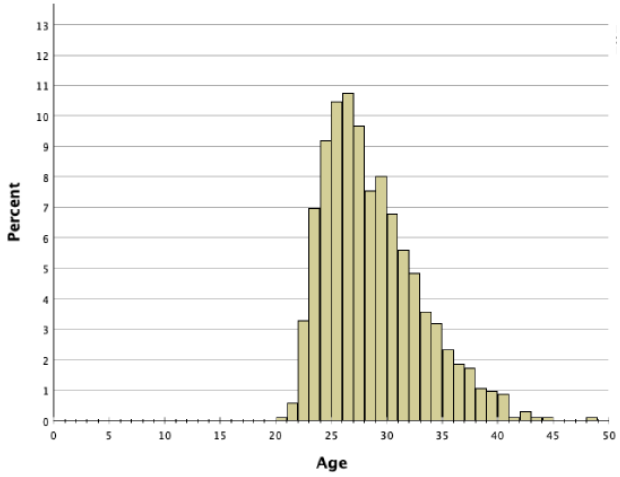
This histogram is in an important shaped called a "normal distribution." Notice how the bell-shaped histogram is relatively symmetric. At the peak are the mean (73.7"), median (74"), and mode (74"). There is more data (more players' heights) in this center, and then the frequency of players gradually decreases as you go out to both sides.

Definition: Normal Distribution

Normal distribution: A bell-shaped symmetrical histogram with data concentrated in the middle and evenly tapering off to both sides.

Normal distributions occur in real life, from height to weight to standardized tests to how long people cook their spaghetti for. All of these would be relatively symmetrical, with a clustering around the mean and median and then a declining number of cases as you go out in either direction.

In contrast, check out the histogram below of the same baseball players' ages.



You can clearly see the peak, and the data slowly tapering off to the right, but to the left of the peak the data quickly plummets. The

data's distribution is not symmetrical. And with good reason ---- all the baseball players are adults! If you saw a tail on the left comparable to the tail on the right, that would mean there were kids playing Major League Baseball.

When a histogram that is unimodal (one mode, meaning it has one peak) is not symmetrical, it is called "skewed." Data with a tail to the right, like the histogram of the age variable above, is positively skewed, and data with a histogram that has a tail to the left is negatively skewed.

Definition: Skewed distribution

Skewed distribution: An asymmetrical histogram.

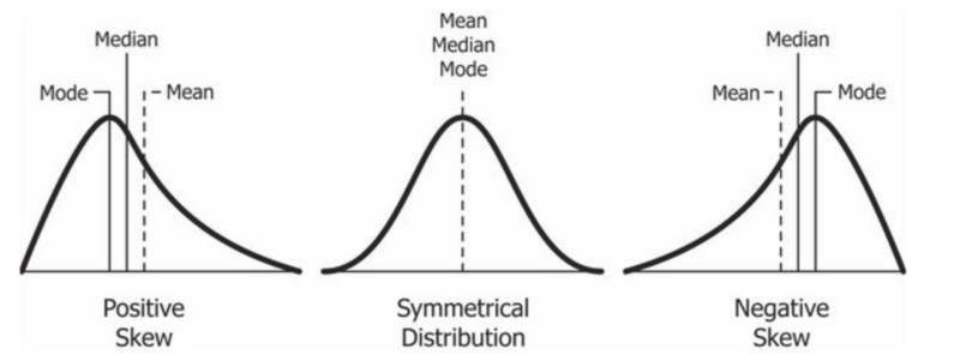
Histogram shapes, mean, and median

As we saw in the normal distribution above, measures of central tendency all conglomerate at the peak of the distribution curve.

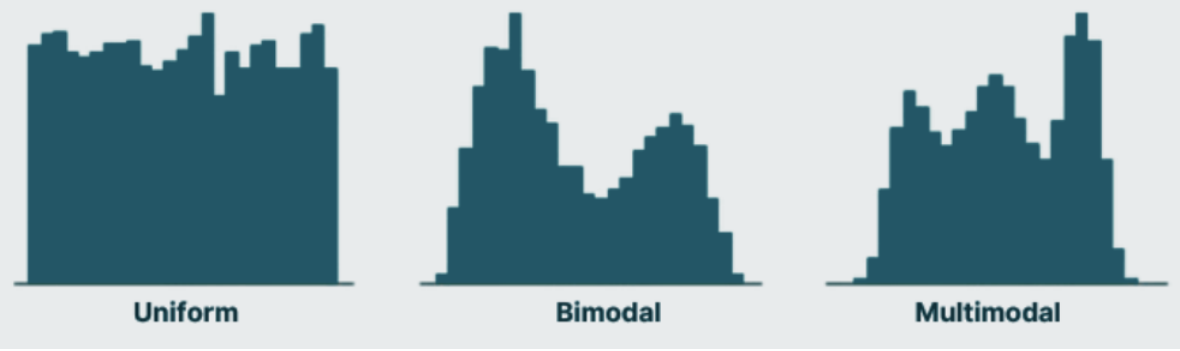
However, skewed distributions are different.

- In positively skewed distributions, the tail to the right represents some high extreme values that do not exist to the left. This pulls the mean up.
- In negatively skewed distributions, the tail to the left represents some low extreme values that do not exist to the right. This pulls the mean down.

The tail pulls the mean.



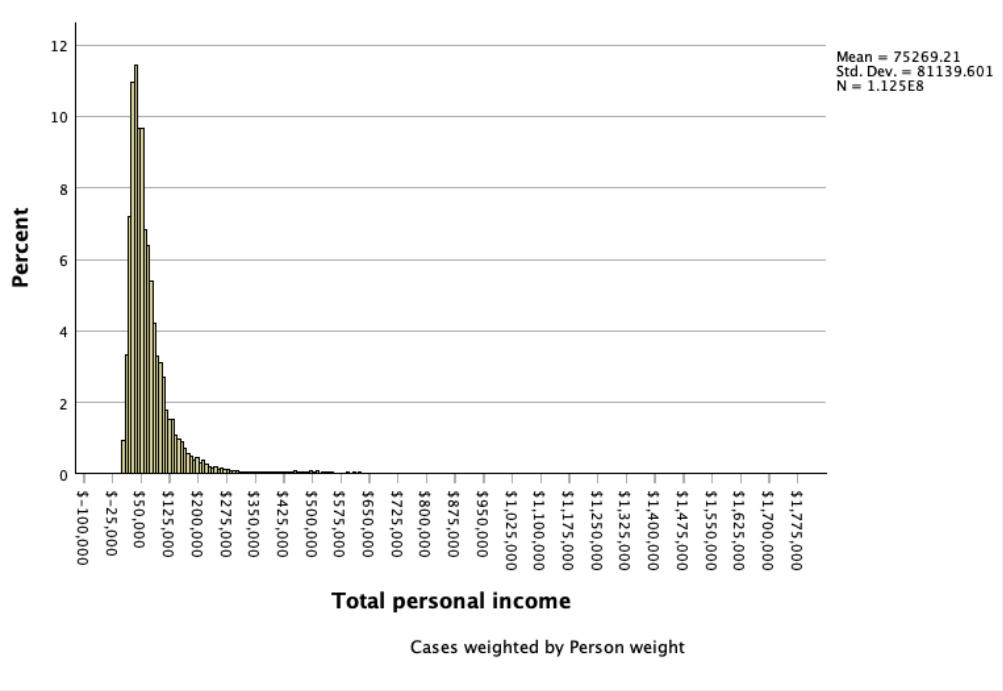
Histograms can come in other shapes as well. Below, the uniform distribution shows a relatively equal proportion of cases across values, while the bimodal and multimodal distributions show multiple peaks / concentrations of cases.

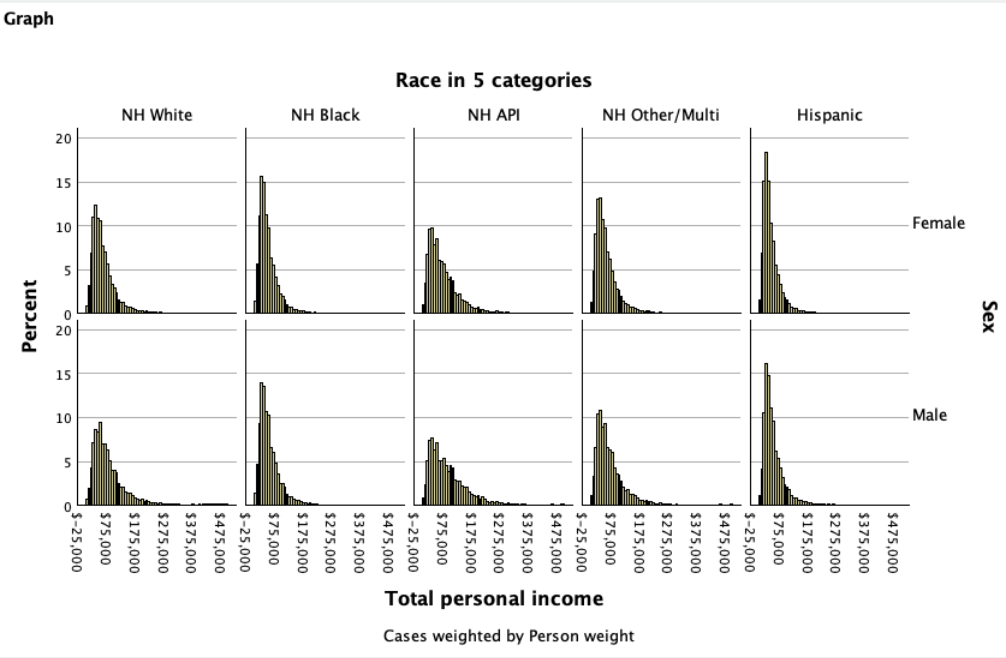


Statistics in Action: Income and pay gaps

Last chapter we looked at income and its distribution by gender and race. Below are histograms of some of this data ---- income among full-time, year-round, civilian workers ages 16 and older. Note that there were some outlier cases with really

high incomes that are not shown in the race/gender histograms. There were so few cases comparatively that they were not really visible, and limiting the x-axis scale lets you see the rest of the distribution a bit better.

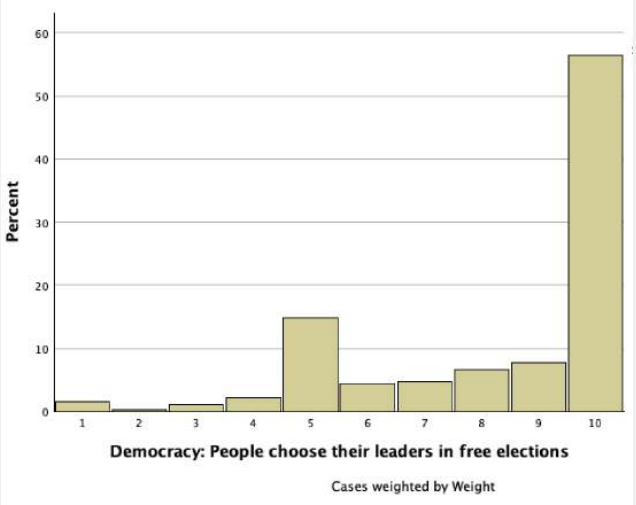




All of these histograms have a tail to the right, rather than being normally distributed and symmetrical. These high-income earners pull the tail to the right, which suggests that mean income is likely higher than median income.

Statistics in Action: Defining Democracy

Returning to the World Values Survey question about whether people choosing their leaders in free elections is an essential characteristic of democracy, the histogram below shows U.S. adult responses in 2017.



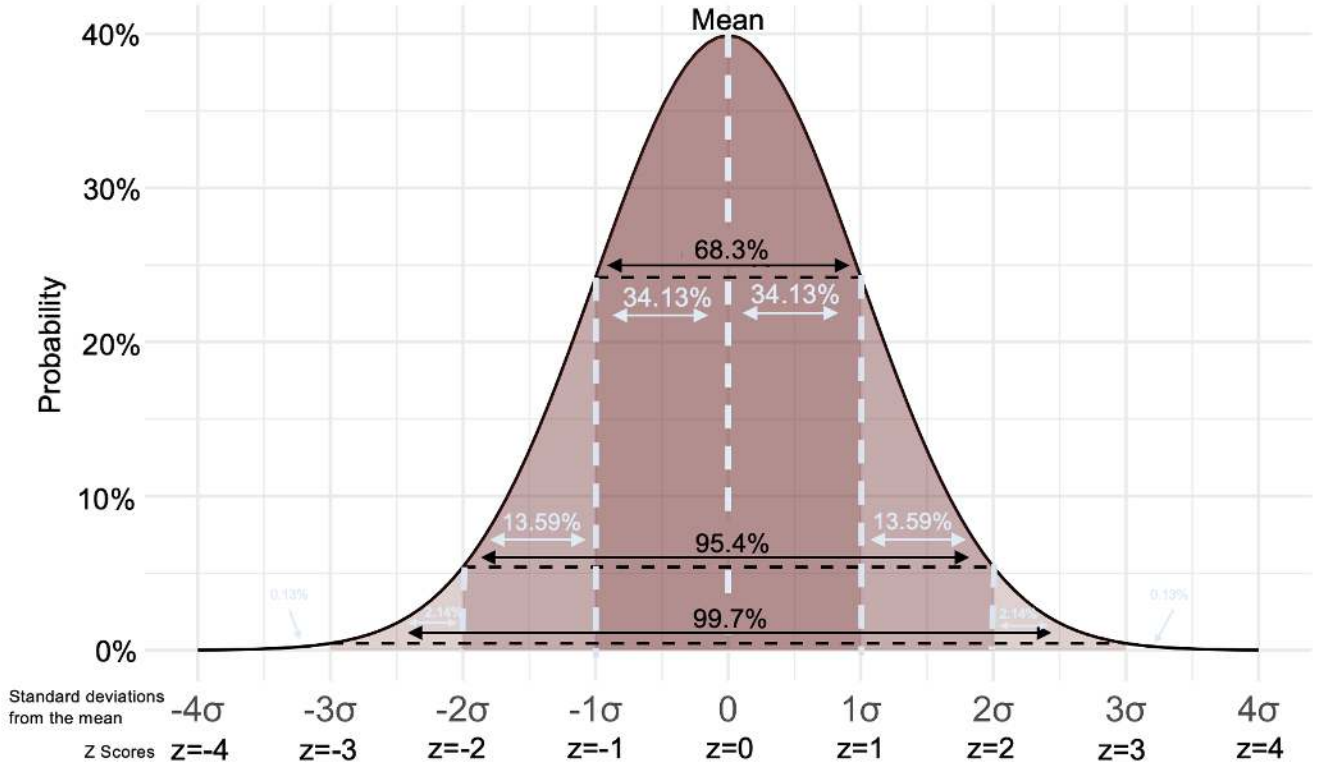
This distribution is also not symmetrical. The peak is all the way to the right at 10 (a majority of respondents felt people choosing their leaders in free elections is an essential characteristic of democracy) and then there is a tail to the left. This is a negatively skewed distribution, and the tail to the left, the respondents who felt free elections is not essential or not as essential, draws the mean lower. Based on this histogram the mean is likely less than the median.

4a.3 Standard Normal Distributions: Z-Scores and Probability under the Curve

Definition: Standard normal distribution

Standard normal distribution: Also known as a z-distribution, this is a special normal distribution standardized by z-scores.

The figure below shows a standard normal distribution with the proportion/probability of cases in various portions of/under the curve. For example, 34.13% of cases are between the mean and one standard deviation above the mean.



A histogram shows your data. The curve imposed on top of the bars therefore has all the data under it. 100% of the data (probability=1) is under the curve. In a standard normal distribution, just over 2/3 (about 68%) of data is within one standard deviation of the mean, about 95% of data is within 2 standard deviations of the mean (95% of the data is within 1.96 standard deviations), and about 99.7% of data is within 3 standard deviations of the mean. Data further than three standard deviations from the mean are considered outliers.

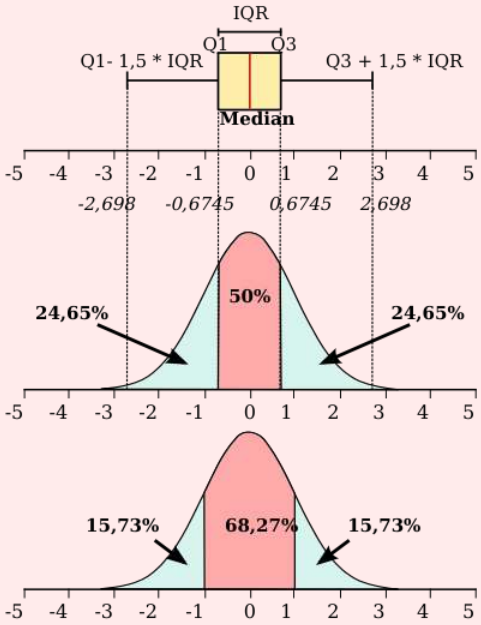
In the last chapter, we calculated values at various z-scores:

Z-score	Value
z=-3	0.5
z=-2	2
z=-1	3.5
z=0	5
z=1	6.5
z=2	8
z=3	9.5

If your data is normally distributed, you can now figure out what percentage of cases are expected to be between these various values. For example, about 68% of data would be between 3.5 and 6.5 (within one standard deviation of the mean) and about 95% of data would be between 2 and 8 (within two standard deviations of the mean).

Caution

Don't get confused between the standard normal distribution and the five point summary and box plots. Remember that while about 68% of data is within one standard deviation of the mean in a normal distribution, an interquartile range is where the middle 50% of data is, between the lower and upper quartile. These are two separate concepts and not interchangable.



4b.1 Inferential Statistics

When you conduct a research study or ask a research question, you have a **population** you are interested in. For example, if you want to know about how U.S. likely voters felt about a candidate debate, the population you are interested in consists of all U.S. likely voters.

Definition: Population

All individuals, groups, objects, or cases of interest.

However, there were over 150 million people who voted in the 2020 presidential election. You don't have the time or resources to efficiently survey them all, and if you tried and didn't do a great job (e.g., you missed a portion of them) your population parameters (measures describing populations, whereas statistics describe samples) may be inaccurate.

Instead, you would survey a **sample** of likely voters.

Definition: Sample

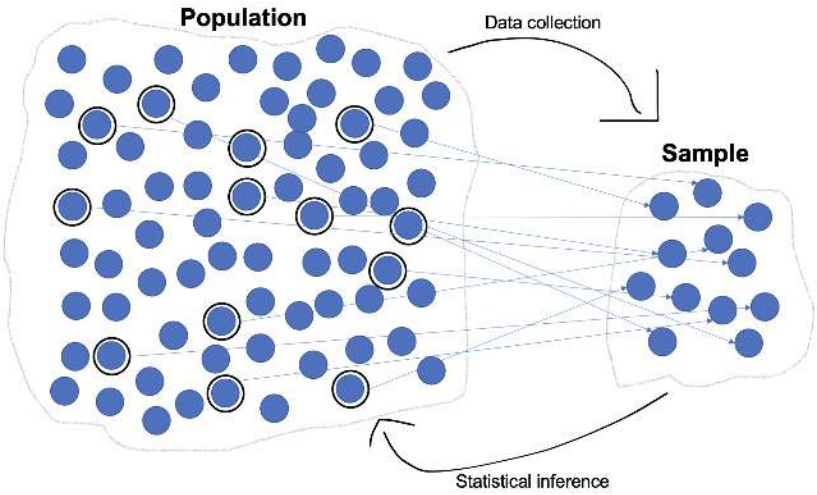
Sample: A subset of cases from the target population.

You then use the sample statistics to estimate population parameters.

Let's say you surveyed 200 people who watched a debate and you found out that 2/3 of them thought the most interesting part was how a fly landed on a candidate's head and stayed there for a really long time. This is a **descriptive statistic**. Descriptive statistics describe your sample data.

You know something about how those 200 debate watchers feel. But, if your target population is all likely voters, or everyone who watched the debate, then who cares about these particular 200 debate watchers? You want to know about your target population, not your sample.

In order to describe the population, I need to engage in **inferential statistics**. This is where I make claims about the target population based on my sample data. When we move to confidence intervals in the next chapter and statistical significance tests thereafter, we are engaging in inferential statistics.



Definition: Inferential Statistics

Inferential statistics: Statistical procedures where we make claims about a population based on corresponding sample data.

In order to make generalizations from your sample to your target population, the sample needs to be representative of that population. This is a requirement for engaging in inferential statistics. For example, if you only spoke to climate change activists after a candidate debate, you could not generalize their perceptions to U.S. voters as a whole. Or what if the 200 debate watchers interested in the fly were interested because they were interacting on social media during the debate, and most debate watchers did not even notice the fly? I can only make inferences about a population if I have a representative sample from that particular population. Sociologists are especially insistent on representative samples because of our understanding of the importance of social context. You will find that many psychology studies that engage in inferential statistics do not seem to have a representative sample

(often using college students, for example). Psychologists who do this may rationalize performing inferential statistics because they believe the processes they are studying are universal (e.g., human nature; hardwired reactions, etc.) rather than context-specific.

There are various sampling methods. Non-probability sampling methods like convenience, purposive, and snowball sampling do not generate representative samples. Probability sampling methods that use random sampling, giving each case in the population an equal chance of being selected, produce samples that, if you have enough cases, are representative. With random sampling, as your sample size increases, you control for various differences in your sample (e.g., gender, political leanings, etc.) such that as your sample gets bigger it will be more representative of and reflect the actual target population. Even with probabiilty sampling, your sample likely will not be an exact microcosm of your population. This is called **sampling error** and is a reason we estimate population parameters derived from a sample instead of having definitive conclusions about them.

Definition: Sampling Error

Sampling error: The difference between sample statistics and population parameters resulting from samples not being perfectly representative of the populations they are drawn from.

Population Studies

Not all research studies are based on samples. You can also survey the full population. For example, every 10 years the U.S. Census collects data on all U.S.-American residents. Or if you were to want to find out information about high school students at a particular high school and give them all a survey, you are surveying the population, not a subset of that population. You would not do inferential statistics with population data---- you don't need to make inferences about the population, as you already have population parameters! For example, if you survey all students at your local high school and find out 15% of them own their own car, you already know what percentage own their own car. There's no reason to build confidence intervals or generalize further.

4b.2 Weighting

Even with random sampling, sometimes you still end up with less than representative samples. One statistical technique to attempt to make our data even more representative is called weighting. Weighting basically adjusts your sample so that the sample better reflects the population.

For example, weighting can adjust for non-response. Lets say a survey you conduct has more women than men respond, and so 60% of your respondents are women, but only 52% of your target population are women. If you have men's responses count as 1.2 each and women's responses count as 0.867 each, you end up balancing out for gender. Statistical programs that have weight variables can do this to take into account things like gender, race, income, etc. so long as we know the actual population parameters for it.

Definition: Weighting

Weighting: A process of adjusting statistical calculations to compensate for sampling error.

Weighting is not always done just to balance out errors in your collection. Weighting can also correct for intentional oversampling by weighting each case inversely to its probability of selection. Sometimes you want to weight because you have oversampled a particular population.

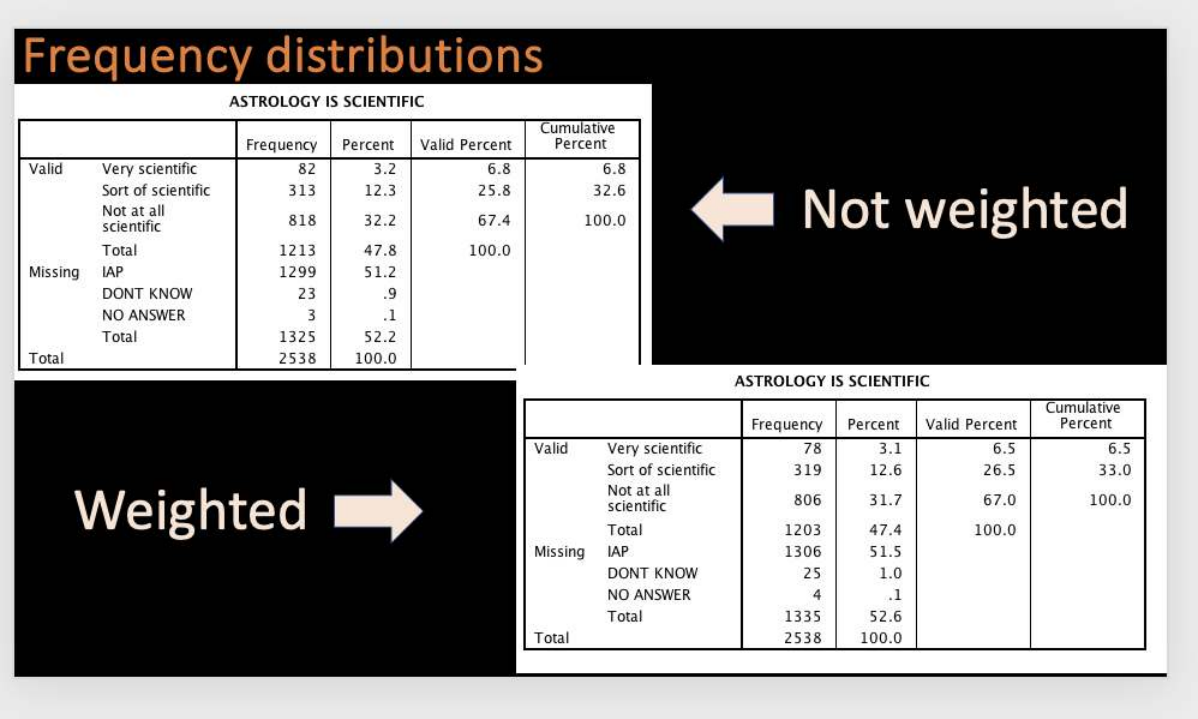
For example, when I was preparing a lecture for a Society & Mental Health course on sexuality, I analyzed data from the General Social Survey for information about LGBT identified individuals. While the data was interesting, when I looked at how many people were in the sample, it was quite clear I could not make any inferences about the truth of this data. Depending on one's research question, one may want to oversample populations like sexual and gender minorities, racial minorities, etc., so that there are enough people in the sample. If you were to do a random sample of 300 university students at a large university, you probably could not say much about how graduate students think, because you would be basing it off too few a number of people that would be included in that random sample. So, you can oversample, making the sample intentionally disproportionate. However, if you

want to make statements about the population, you don't want to still have more of a particular population proportionally---you want each group to reflect its actual proportion. Hence, you can balance this out using weighting.

People might also have different probabilities of selection built into the survey design. For example, if every household has an equal probability of being selected, but only one individual from each household can be selected, then people from larger households have a lower probability of being selected than people from smaller households or who live alone. These types of survey designs can be adjusted for at the analysis stage through weighting.

While it is important to use weighting when available, you want to report unweighted sample sizes, which are often (but not always) different from a weighted sample size. That way you can share the actual number of participants/cases. There also can be issues if you have small sample sizes, are using large weights, or if there are very few people from a subgroup that is being up-weighted (when respondents are being counted as multiple cases). In one poll, a 19-year-old black man in Illinois was being weighted as much as 30 times more than the average respondent. If your weighting and unweighted analyses are extremely different, that's a flag for concern.

Example: The table below is based on the survey question, "Would you say that astrology is very scientific, sort of scientific, or not at all scientific?" Notice in the frequency tables below that the weighting distribution and sample size is different, but not that different, from the unweighted distribution.



		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Very scientific	82	3.2	6.8	6.8
	Sort of scientific	313	12.3	25.8	32.6
	Not at all scientific	818	32.2	67.4	100.0
	Total	1213	47.8	100.0	
Missing	IAP	1299	51.2		
	DONT KNOW	23	.9		
	NO ANSWER	3	.1		
	Total	1325	52.2		
Total		2538	100.0		

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Very scientific	78	3.1	6.5	6.5
	Sort of scientific	319	12.6	26.5	33.0
	Not at all scientific	806	31.7	67.0	100.0
	Total	1203	47.4	100.0	
Missing	IAP	1306	51.5		
	DONT KNOW	25	1.0		
	NO ANSWER	4	.1		
	Total	1335	52.6		
Total		2538	100.0		

How to weight in SPSS

Syntax:
To turn on weighting, use the following syntax, replacing WeightVariableName with the name of the designated weight variable.

Weight by WeightVariableName.

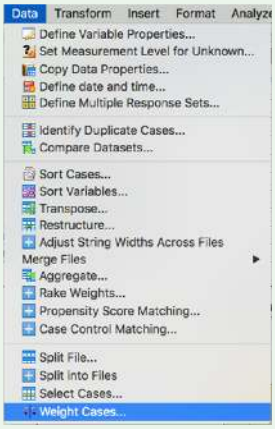
Once you turn on weighting, it will stay on for all the analyses you do with that dataset until you either close SPSS or turn weighting off.

To turn weighting off, use the following syntax:

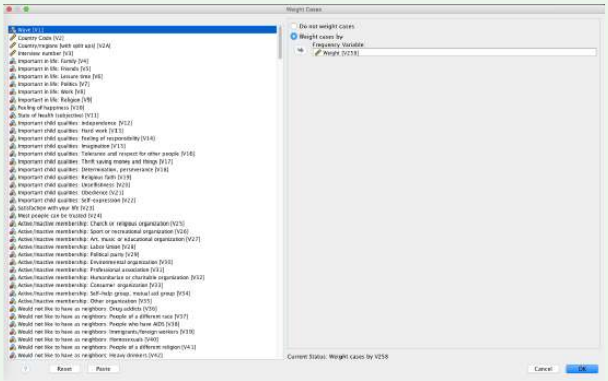
Weight Off.

Menu:

Go to Data → Weight Cases.



Select "Weight cases by." Select the designated weight variable and put it in the box where it says "Frequency Variable." Click OK.



Note that if you exit SPSS or close your dataset, weighting will turn off. The next time you open your dataset, if you want to use weighting, you will need to turn it on again.

To turn off weighting, go to Data → Weight Cases. Select "Do not weight cases" and click OK.

> Weight variables are carefully constructed to adjust cases in order to better reflect target populations. You must use the designated weight variable for weighting. Do not use any other variable, including variables that you are interested in learning about. If you mistakenly use a different variable, it will make all your results wrong. For example, if you use age, someone who is 80 years old will count 80 times, and someone who is 20 years old will count 20 times, etc. This will distort your data. Only use designated weight variables for weighting.

4b.3 Central Limit Theorem

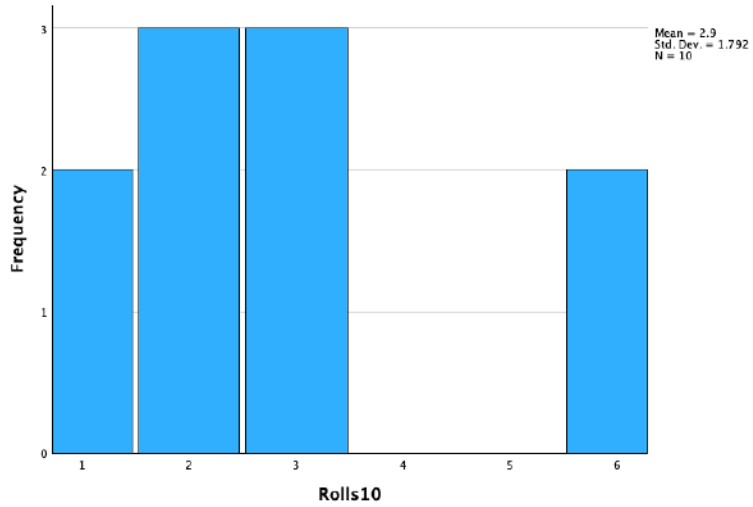
In this section we'll walk through an experiment that will demonstrate the Central Limit Theorem.

1. Population of Dice Rolls

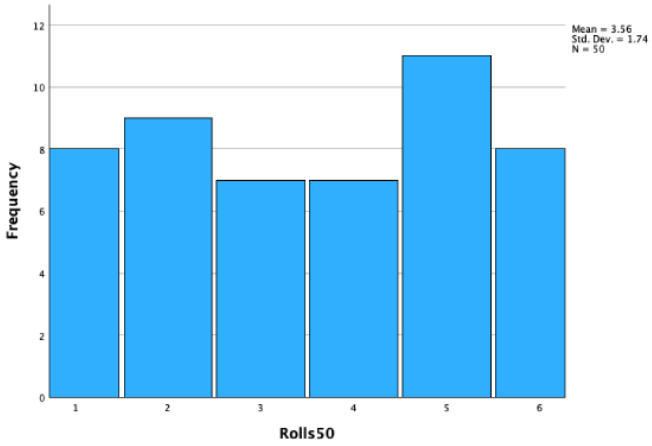
First, we're going to roll a die 1,000 times. This will constitute our population. What do you think the shape of our frequency distribution will look like?

Let's take look at cumulative histograms, where we take the first 10 rolls, 50 rolls, 100 rolls, 200 rolls, 500 rolls, and 1000 rolls. As we add more rolls, observe the shape of our distribution and if/how it changes as the sample size increases

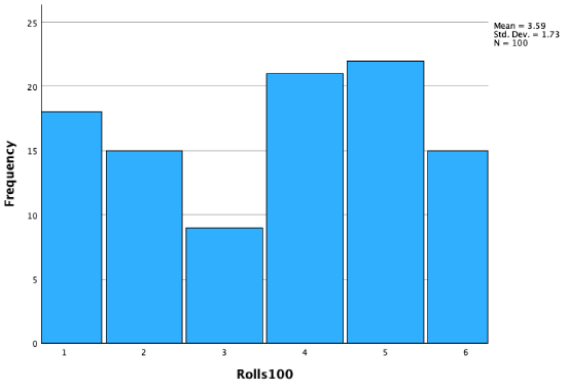
First 10 rolls:



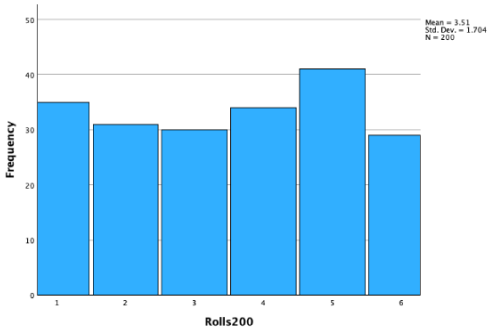
First 50 rolls:



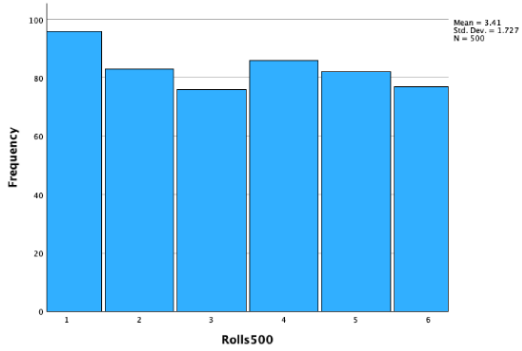
First 100 rolls:



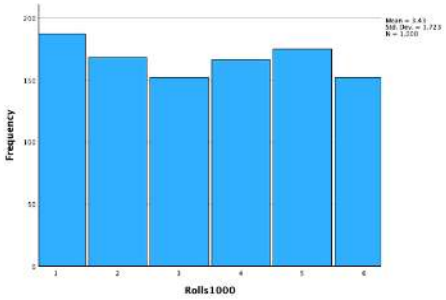
First 200 rolls:



First 500 rolls:



All 1000 rolls:



I hypothesized that it would look like a uniform distribution, because each of 1 through 6 has an equal 1/6 chance of being landed on. However, 1/6 is the probability for each die face each time, so if I roll a 1, the next roll does not increase my chances of rolling a 2 through 6 ---- 1 has just as much chance as 2 or 3 or 4 or 5 or 6 as getting rolled the next time. If you try rolling a die yourself (a real one, or using a random dice roll generator online, or random number generator with 1 through 6 as options), you'll notice that the outcomes look more patterned than you might expect! However, you'll notice that as the number of dice rolls increases, it begins to look more and more like the predicted theoretical uniform distribution.

The outcomes from our dice rolls constitute our population. While our theoretical population mean would be 3.5 (if you had an even number of 1s, 2s, 3s, 4s, 5s, and 6s), our population mean is 3.43.

2. Random Samples of Dice Rolls

Let's randomly sample our population of 1,000 dice rolls, calculate means of those samples, and see how our sample means compare to the population mean. As previously noted, sample means are (usually) not the same as the population mean ---- this is called sampling error. While samples are drawn from a population, they do not usually completely accurately reflect the population.

I am going to randomly select 10 of the 1,000 rolls. This is 1% of the population. It is random selection since each of the 1,000 rolls will have an equal chance (in this case 0.1%) of being selected. From that random selection of 10 rolls, I will calculate its mean. I will do this 1,000 times. While I could write down all the rolls and pick them out of a hat or write them on balls and spin them

around in a bingo cage, I used a random number generator to select cases (e.g., the 56th roll), then looked up what was rolled for that case (e.g., the 56th roll is a 3). There are plenty of random number generators online. I used Excel, with the formula =randbetween(1,1000) to randomly select a case. After I had 10 cases, I calclulated their mean by adding them up and dividing that total by 10. That gave me a sample mean. I repeated this 1,000 times.

Now that I have 1,000 sample means, I'm going to create a histogram of the sample means. This is called a sampling distribution.

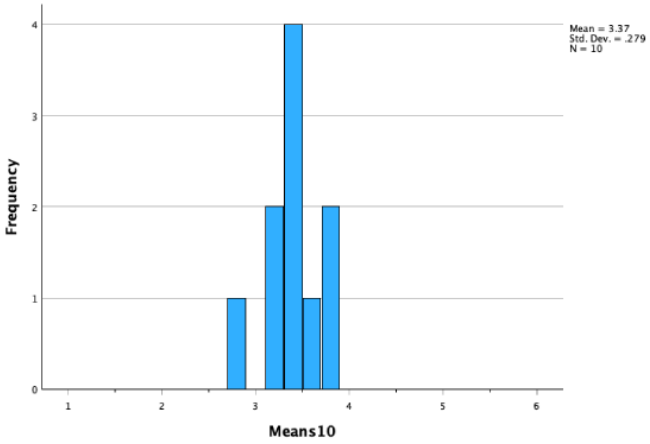
Definition: Sampling Distribution

A distribution of sample means from the same population.

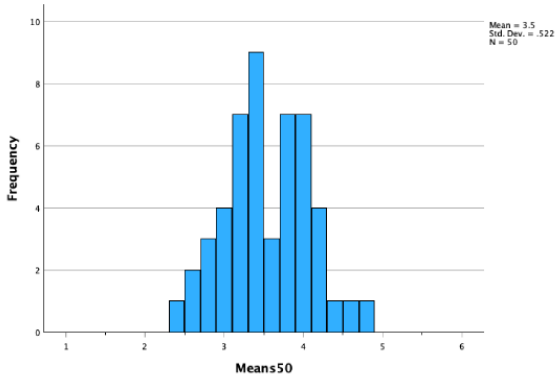
What do you think the shape of our frequency distribution will look like?

Let's look at cumulative histograms, first with the first 10 sample means, then the first 50 sample means, then 100, then 200, 500, and finally all 1,000 sample means. As more sample means are added, observe the shape of our distribution and if/how it changes as the number of samples change.

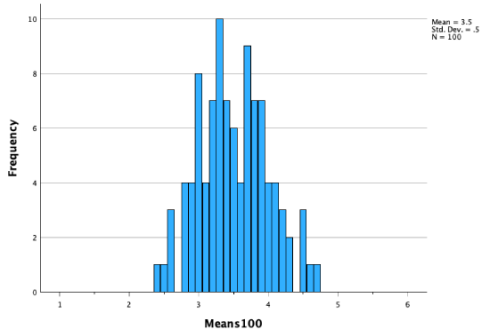
10 sample means:



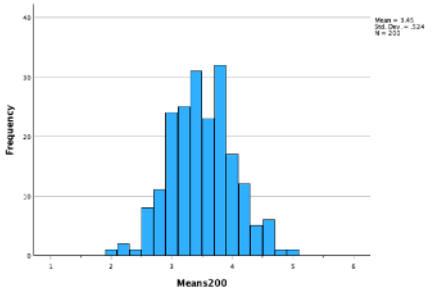
50 sample means:



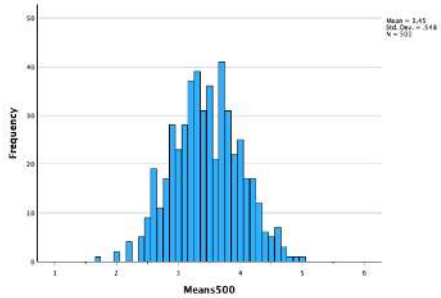
100 sample means:



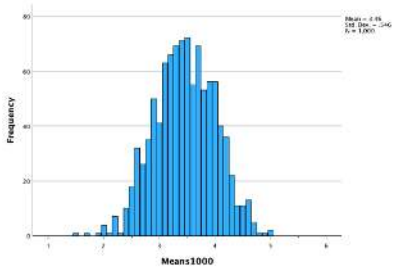
200 sample means:



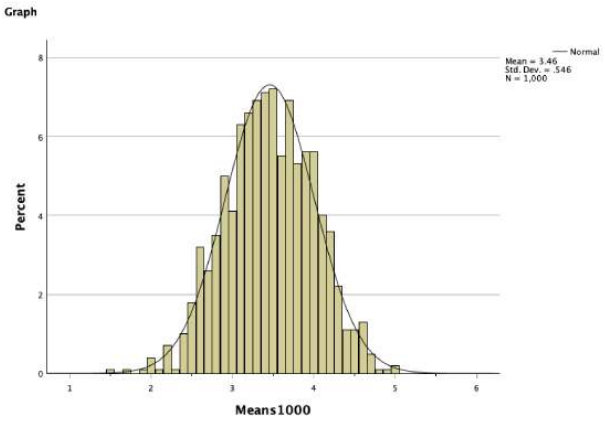
500 sample means:



All 1,000 sample means:



As our sampling distribution (our distribution of sample means) contains more and more sample means, we get a distribution that becomes closer and closer to a perfect bell curve, to a normal distribution.



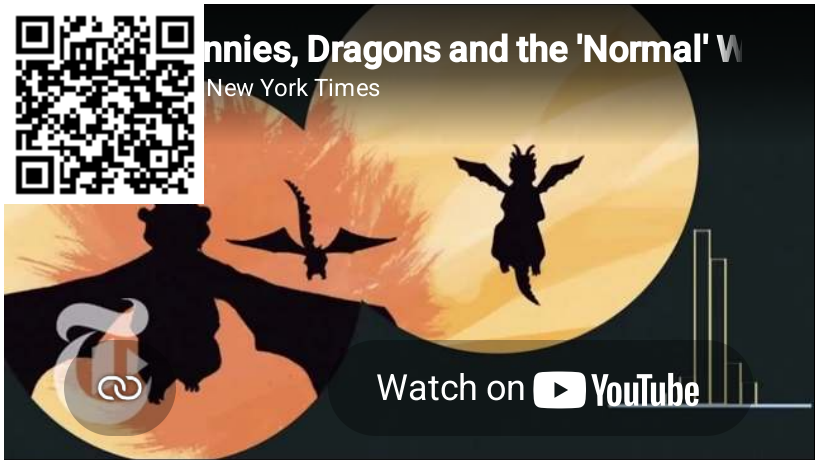
Remember that our population mean was 3.43. In the histogram above of all 1,000 sample means, you can see that there are more sample means close to 3.43, and fewer sample means as we get further out. Sample means further away from the population mean occurred less frequently. However, these still occurred. For example, there's a sample mean on there at 5. But you can see there was a low probability that we would get a sample mean of 5. It was more likely we would draw a sample mean closer to the actual population mean.

3. Central Limit Theorem

Definition: Central Limit Theorem

Central Limit Theorem: Given a random sample with enough cases, the distribution of sample means will be normally distributed.

Check out this animation that explains the Central Limit Theorem:

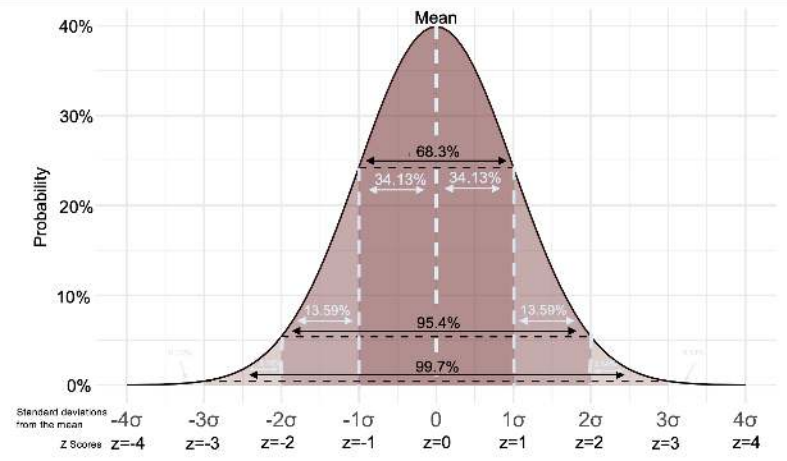


Sampling distributions have some useful terms and concepts embedded in them that are new, but are similar to concepts we've looked at already with frequency distribution histograms.

Standard Errors & T-Scores

Below, notice how standard error and t-scores are familiar in that they have almost parallel definitions to standard deviation and z-scores respectively, except they are about sampling distributions.

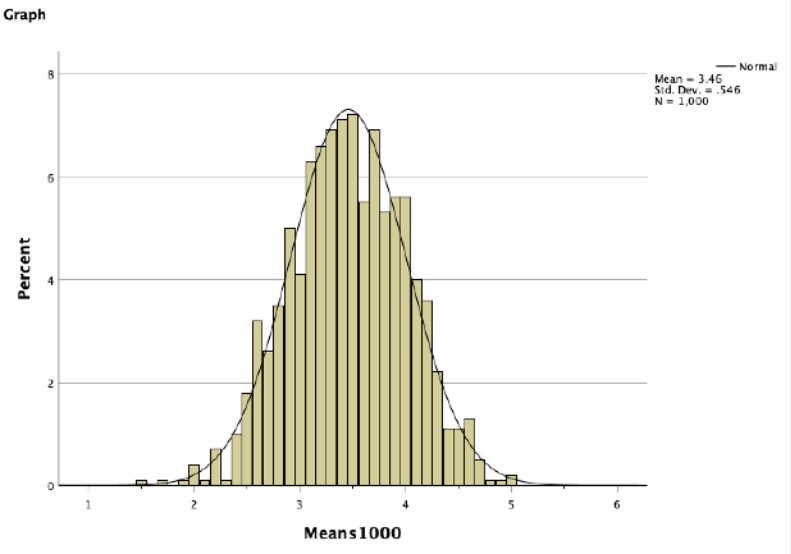
Looking back: This is a frequency distribution graph (histogram) that is normally distributed:



Standard deviation: Average distance of values from the sample mean

Z-score: How many standard deviations a value is away from the sample mean

Looking forward: This is a sampling distribution, also a normally distributed histogram:



Definition: Standard error

Standard error: The standard deviation for a sampling distribution. This is the average distance of sample means from the population mean

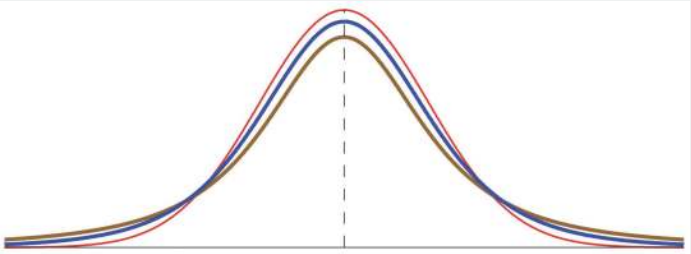
Definition: T-score

T-score: How many standard errors a value (sample mean) is away from the population mean.

A standard error does not communicate much intuitively, but it helps us calculate other useful statistics (e.g., to build confidence intervals, determine p-values, etc.). Standard errors are the standard deviation divided by the square root of the sample size.

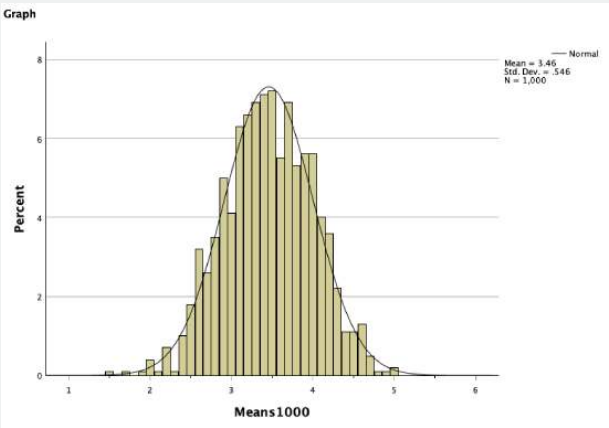
With a large enough sample size, t-scores and z-scores are identical. However, at smaller sample sizes t-scores will be a little different, as they won't have a standard normal distribution. The normal distributions below represent t-distributions at different sample sizes. Because of the differences in these curves, the probabilities under the curve are spread out differently.

We'll see how t-scores can differ and how to calculate them in the next chapter.

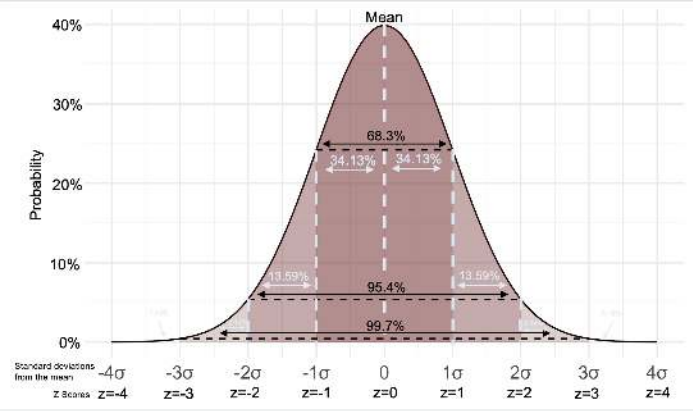


P-value

Remember that in our sampling distribution of sample means of dice rolls, we were more likely to pull samples with means closer to the actual population mean, and less likely to pull samples with means further away from the actual population mean:



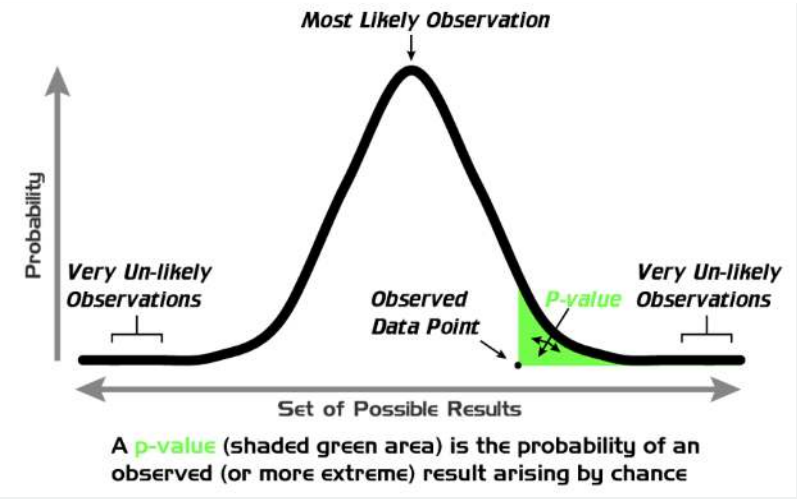
Given the characteristics of normal distributions, we can determine the probability of pulling a sample within various areas of this distribution.



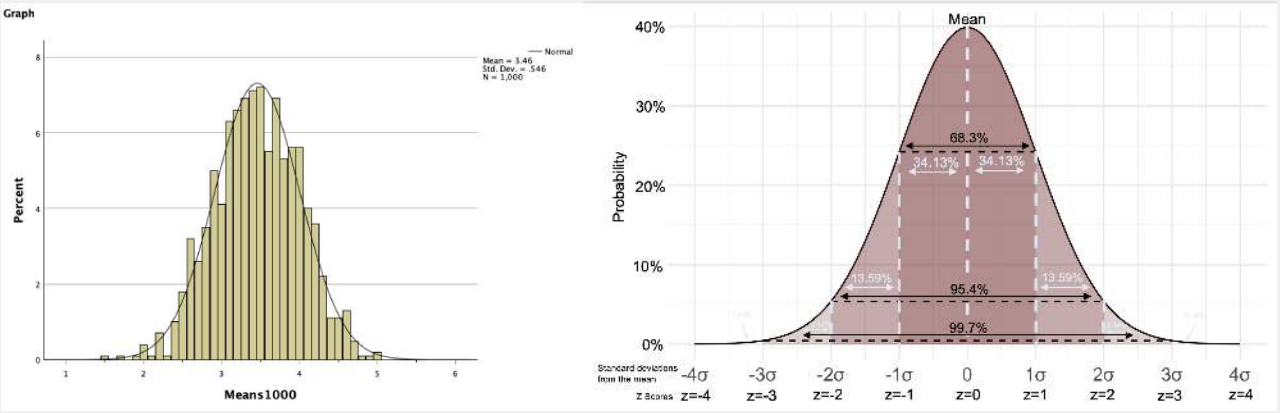
A **p-value**, or probability value, tells us how likely it is that we would get a particular sample mean (or one more extreme) given our actual population mean and distribution.

Definition: p-value

The probability (likelihood) of getting a particular sample statistic as extreme (or more extreme than) the outcome we observed, based on what we would expect from the sampling distribution.



Looking at our sampling distribution, we can thus calculate the probability of pulling a sample with a particular sample mean.



For example, 68% of sample means would be between 2.9 and 4.0 (in our empirical rather than theoretical sampling distribution, 72.1% of the sample means were between 2.9 and 4.0).

In our sampling distribution, 0.7% of sample means were 2.0 or less, and 0.2% of sample means were 5.0 or more.

How likely is it that we would get a sample mean of 5 or more if our population mean is actually 3.43? If the population mean is 3.43, there is only a 0.2% probability that we would draw a sample with a mean of 5 or more. If we draw a sample with a mean of 5 or more, there is only a 0.2% probability that the population mean is 3.43.

We'll come back to p-values in future chapters, but I wanted you to be introduced to the concept as you see a sampling distribution.

This page titled 4: Histograms, Normal Distributions, and the Central Limit Theorem is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Ezra Temko (Consortium of Academic and Research Libraries in Illinois (CARLI)) .

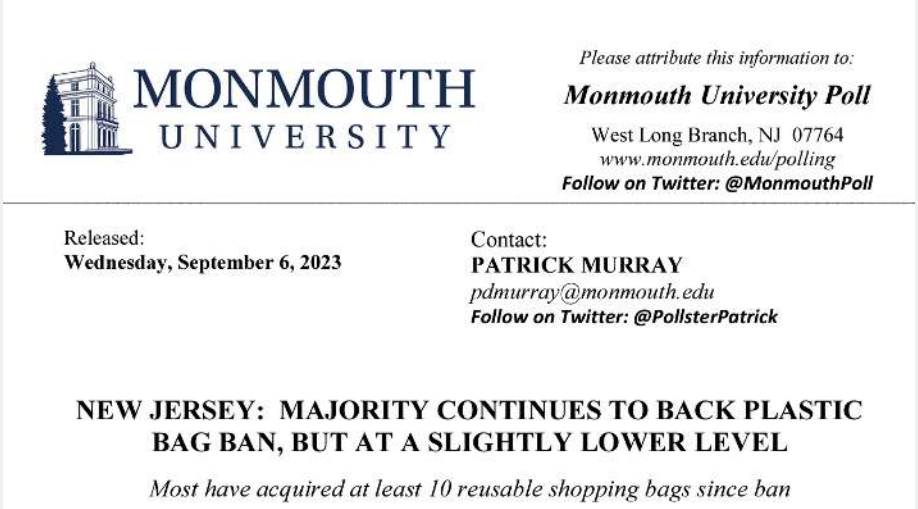
5: Confidence Intervals

This chapter is focused on confidence intervals. Next chapter we'll look at confidence intervals again, but with comparing subgroups.

Statistics in Action: Do New Jersey residents support the state's plastic bag ban?

We see polls in the news all the time, with claims about poll findings in headlines and media stories.

For example, when the Monmouth University Polling Institute released poll results, with this heading:



MONMOUTH UNIVERSITY

Please attribute this information to:
Monmouth University Poll
West Long Branch, NJ 07764
www.monmouth.edu/polling
Follow on Twitter: @MonmouthPoll

Released:
Wednesday, September 6, 2023

Contact:
PATRICK MURRAY
pdmurray@monmouth.edu
Follow on Twitter: @PollsterPatrick

NEW JERSEY: MAJORITY CONTINUES TO BACK PLASTIC BAG BAN, BUT AT A SLIGHTLY LOWER LEVEL

Most have acquired at least 10 reusable shopping bags since ban

it generated news stories with headlines like the following:

"Poll: Support for NJ's single-use plastic bag ban dips slightly" -NJBIZ

"Monmouth Poll: N.J. backs single-use plastic bag ban; oppose ban on paper bags at grocery stores" -New Jersey Globe

"Monmouth poll: One year later, New Jerseyans support plastic bag ban and have accumulated plenty of reusable bags" -ROI-NJ

"Poll: Support drops for NJ bag ban, 9% don't know it exists" -NJ 101.5

"New Poll Shows Majority of New Jersey Residents Oppose Ban on Grocery Paper Bags; Overall Support for Bag Ban Dips Slightly" -The Lakewood Scoop

What should we make of these headlines? Are they accurate? Do we need to be critical of such a poll?

> The Monmouth University Polling Institute's findings are based on **sample data**. When they write in their release,
>
> > *A majority of New Jerseyans continue to support a ban on single-use plastic bags (56%) although this is a few points lower than shortly before the ban went into effect (61% in April 2022).*
>
> they actually mean that 56% of **814 New Jersey adults** they surveyed answered "support," when asked, "Do you support or oppose banning single-use plastic bags in New Jersey, such as those given out at supermarkets or other stores?"
>
> But nobody cares about what 814 random survey respondents think. We want to know about the target population, how New Jersey adults overall think. This is where we can turn to inferential statistics, and specifically confidence intervals.

Confidence Intervals

Definition: Confidence Interval

Confidence interval: The values within which we are claiming a population mean or proportion falls, at a particular confidence level, based on a representative sample of that population.

As an example, in Chapter 3, I conducted a weighted analysis of an IPUMS dataset (IPUMS originally stood for Integrated Public Use Microdata Series) of American Community Survey (ACS) 5-Year Data for 2017 to 2021 and found that the **sample mean** annual income for full-time, year-round, non-institutionalized U.S. workers ages 16+ was **$75,269.21**.

At the 95% confidence level, the **confidence interval** for this mean is: **$75,254.22, $75,284.21**

The confidence interval consists of two numbers. I separate them with a comma because using a dash can be confusing (it also looks like a minus or negative sign).

The lower number, in this case $75,254.22, is called the **lower limit** or **lower bound**.

The greater number, in this case $75,284.21, is called the **upper limit** or **upper bound**.

We are 95% confident that the population mean is somewhere between the lower and upper bounds.

Interpretation; Confidence intervals

*For a confidence interval of a **mean**:*

General Template:
We are [confidence level, usually 95%] confident that, as of [date/time period sample data was collected], the mean [variable name/description] for [name of population] was somewhere between [lower limit with label] and [upper limit with label], meaning on average they were somewhere between [interpret lower limit and upper limit values in context].

Sometimes it flows better to name the population before the variable, e.g.,
[name of population]'s mean [variable name/description] (e..g, "U.S. adults' mean age" instead of "the mean age for U.S. adults").

Remember that to write the story of your confidence interval, you need to know information about your dataset (when was the data collected, who is the target population) and use your codebook to interpret the variable and values.

> Note that the word "respondents" is not in the template. Respondents are those who responded to a survey --- they constitute the sample. Confidence intervals are about the target population.

Examples:

Previous chapters presented a number of sample means and proportions. Here are examples of interpreting some of their confidence intervals:

We can be 95% confident that, from 2017 to 2021, the mean income for full-time, year-round, civilian U.S. workers ages 16 or older was somewhere between $75,254.22 and $75,284.21. We are 95% confident that these workers average annual income was just over $75,000.

We are 95% confident that, from 2017 to 2021, the mean income for U.S. residents ages 16 or older was somewhere between $46,179.57 and $46,195.87. We are 95% confident that these residents average income was just over $46,000.

We are 95% confident that, following the 2020 general election, U.S. eligible voters mean feelings thermometer scores towards journalists were somewhere between 49.21 and 51.45, meaning on average they were neutral or close to neutral in their

feelings towards journalists (the average is somewhere between just cooler or warmer than neutral).

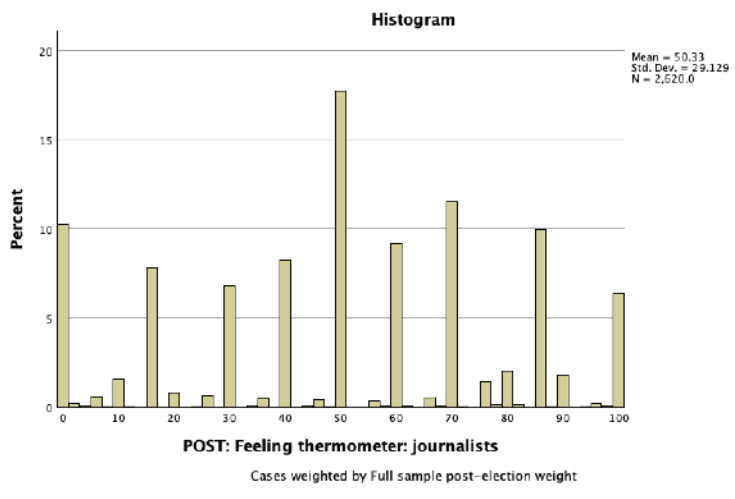
Let's say there was a feelings thermometer questions about raspberries from the same dataset and we had a CI with a lower limit of 30 and upper limit of 70 (this is made up). In this case:
We are 95% confident that, following the 2020 general election, U.S. eligible voters mean feelings thermometer scores towards raspberries were somewhere between 30 and 70, meaning on average they were somewhere between fairly cold/unfavorable and fairly warm/favorable. We are fairly confident that on average they do not love or hate raspberries, but we're not really sure whether on average they feel more cold, neutral, or warm towards raspberries.

Caution

The confidence interval is providing a range within which we are confident (at 95% confidence, or whatever level specified) the population **mean** falls. It is only a claim about the population mean.

When we were dealing with sample data and normal distributions, we made estimates like "95% of data falls between ___ and ____." This is **different** than a confidence interval, which is based on a sampling distribution of **means**.

Take a look at a histogram of our sample ANES data about feelings towards journalists:



Our 95% CI is 49.21, 51.45.

It is **wrong** to interpret the CI by claiming that that 95% of respondents (or U.S. eligible voters) feelings thermometer scores were between 49.21 and 51.45 degrees. Take a look at the histogram --- within our sample, 95% of respondents were between 0 and 100 (there are over 2.5% of respondents at both 0 and 100). 95% of the respondents were **not** between 49.2 and 51.5 (less than 20% of respondents were), and it would not make sense that this would then be the case for all U.S. eligible voters.

Instead, take a look at the sample mean of 50.33. This is a **sample** mean. Due to sampling error, we're not confident this accurately reflects the **population** mean. So 50.33 is the mean for the 2,647 respondents in the sample, but we're not confident it would be the mean for all U.S. eligible voters. The 95% confidence interval is saying that we are 95% confident that the population mean is somewhere between 49.2 and 51.5. It could be 50.33, it could be 49.5, it could be 51.3, we're not sure. But based on the logic of sampling distributions and their properties, given that we have a probability sample, and given our sample mean, the sample size, and the level of variation in our data (as measured by standard deviation), we can make a claim at the 95% confidence level that the population mean is somewhere close to that sample mean, somewhere between 49.5 and 51.3. This is about the population **mean**, a measure of central tendency, *not* the distribution/spread of data/responses/feelings.

*For a confidence interval of a **proportion**:*

General Template:
We are [confidence level, usually 95%] confident that, as of [date/time period sample data was collected], somewhere between

[lower limit] and [upper limit] of [name the population] [name the outcome].
For lower and upper limit, either use the percentages or convert these into benchmark fractions and say "about 1/5" or the like.
If you have already discussed when the sample was taken, you don't need to include its timing in your interpretation, but if this is new context, it's useful to know.

Example:
We are 95% confident that, as of 2022, somewhere between 36.7% and 39.9% (somewhere between about 1/3 and 2/5) of U.S. adults have tried to encourage someone to believe in Jesus Christ or to accept Jesus Christ at their savior.

Note that the CI was: 0.367, 0.399. Because this was an indicator variable, I interpret these means as proportions (the percent with the value coded as 1).

Statistic in Action: Do New Jersey residents support the state's plastic bag ban?

For the Monmouth poll discussed above:
We are 95% confident that, in mid-August 2023, somewhere between 50.6% and 61.4% of New Jersey adults support banning single-use plastic bags in New Jersey, such as those given out at supermarkets or other stores.

Confidence Levels

Confidence intervals can be constructed at different confidence levels. The most widely used confidence level is 95%. If the confidence level is not specified, you can generally assume it is a 95% confidence interval. However, 90% confidence levels are also commonly used, such as by the U.S. Census Bureau.

What does it mean to say, "We are 95% confident?"

Confidence levels come from p-values (probability values).
A 95% confidence level is based on a p-value of 5%, or 0.05.
A 90% confidence level is based on a p-value of 10%, or 0.10.

To find your confidence level from a p-value, subtract it from 1 (or 100%). For example, for a p-value of 0.05 or 5%: 1-0.05=0.95 or 100%-5%=95%

The U.S. Census explains confidence intervals for a p-value of 5% this way:

> *If we were to repeatedly make new estimates using exactly the same procedure (by drawing a new sample, conducting new interviews, calculating new estimates and new confidence intervals), the confidence intervals would contain the average of all the estimates 95% of the time. We have therefore produced a single estimate in a way that, if repeated indefinitely, would result in 95% of the confidence intervals formed containing the true value.*

Let's apply this using the confidence interval in the example shared above regarding feelings towards journalists.

> *We are 95% confident that, following the 2020 general election, U.S. eligible voters mean feelings thermometer scores towards journalists were somewhere between 49.21 and 51.45, meaning on average they were neutral or close to neutral in their feelings towards journalists (the average is somewhere between just cooler or warmer than neutral).*

Consider what you know about sampling distributions and the Central Limit Theorem to make sense of this.

In the ANES survey following the 2020 general election, 2,647 U.S. eligible voters responded to a question about how they felt about journalists, and the sample mean was 50.33, a smidgen warmer than neutral.

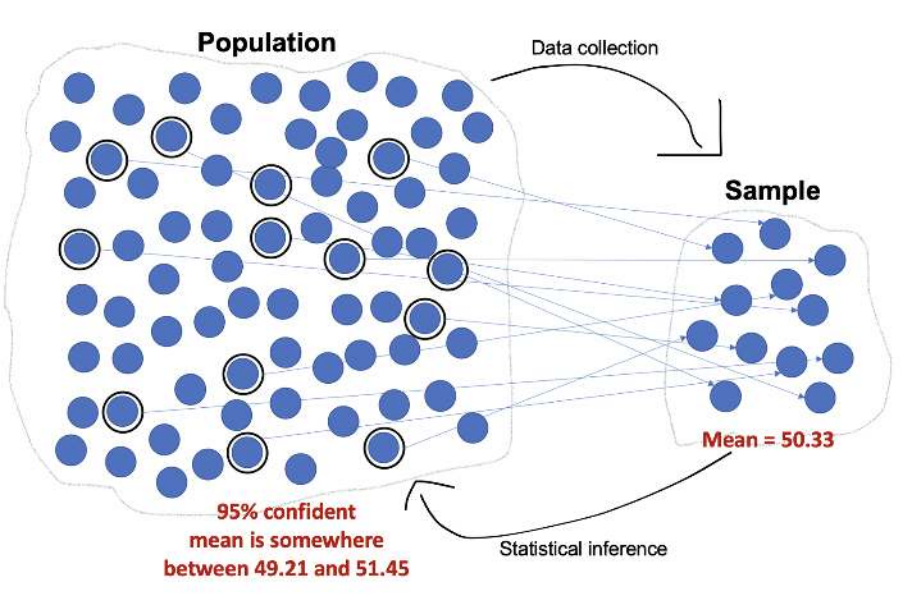
Let's turn to the 95% confidence interval: 49.21, 51.45

If the population mean were actually less than 49.21 or more than 51.45, there is only a 5% chance ANES would have ended up drawing a random sample with the sample mean, sample size, and standard deviation of the variable in the sample. Given that ANES used probability sampling (generating a sample representative of their target population, though with sampling error), and given the ANES sample data (the sample mean of 50.33, the sample size of 2,647, and the variability in the data with a standard deviation of 29.13), there is a 5% chance that the population mean is outside the parameters of our confidence interval (less than the lower bound of 49.21 or more than the upper bound of 51.45).

From this, we say we are 95% confident that the population mean is between 49.21 and 51.45.

Inferential Statistics: only for making population claims based on representative samples

The ANES target population was all U.S. eligible voters. ANES collected data from a random sample of U.S. eligible voters. The sample mean was 50.33. While we know the sample mean, due to sampling error we're not sure about the population mean. However, we can be 95% confident it is somewhere between 49.21 and 51.45.



Population data:

Inferential statistics are not used with population data. Confidence intervals are not a relevant concept. If you have population data and have a mean, it's the population mean.

- The average age of a U.S. Senator (as of 2023) is 63.5 years old. This is based on a census of senators. There is no inference to make. We already know (with 100% confidence!) the population mean.
- In Chapter 1, 30% of the top 20 kids films between 2015 and August 2023 had monarchy as a central form of government (mean=0.15). We know the population mean/proportion and don't construct a confidence interval.
- In Chapter 3, you and your 9 friends each washed a mean of 5.5 cars. This was calculated based on a census of collecting data on every person who washed the cars and how many cars they washed. You would not construct a confidence interval. You already know the mean.

Non-representative samples:

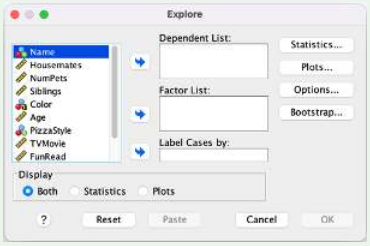
- If you don't have a representative sample (based on probability sampling), you should not engage in inferential statistics.

- If you wanted to know what proportion of kids movies between 2015 and August 2023 had monarchy as a central form of government, you cannot figure this out using the sample of top 20 kids films. The top kids films may be different from all films, so it is not a representative sample.
- If you wanted to know how many cars people wash on average during a car wash, you cannot figure this out with your convenience sample of you and your friends. You and your friends may not be typical of car washers more broadly.
- If every person in the target population you want to generalize to did not have an equal chance of being selected to be part of the sample, inferential statistics are inappropriate.

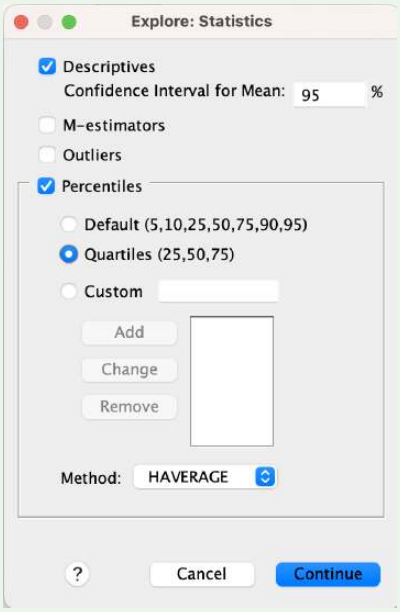
Constructing a confidence interval with SPSS

You can use Explore/Examine in SPSS to construct confidence intervals.

Menu:
Go to Analyze → Descriptive Statistics → Explore. Put your variable into the Dependent List window/box. At the bottom where it says "Display," make sure you either have "Both" or "Statistics" checked (Plots will only give you graphs/figures).



*If you want to construct a confidence interval at a confidence level other than 95%, click on the "Statistics" button, and where it says "Confidence Interval for Mean: 95%" replace the 95 with whatever confidence level you want to use.



Click OK to run the Explore command.

Syntax:
*This will generate a confidence interval along with other statistics.
EXAMINE VARIABLES=VariableName
/PLOT NONE
/PERCENTILES(25,50,75) HAVERAGE
/STATISTICS DESCRIPTIVES
/CINTERVAL 95

/MISSING LISTWISE
/NOTOTAL.

*If you want to construct a confidence interval at a confidence level other than 95%, replace the "95" on the /CINTERVAL line with the confidence level you want to use.

Reading the Output:

Here is what part of your output will look like:

Descriptives

			Statistic	Std. Error
POST: Feeling thermometer: journalists	Mean		50.33	.569
	95% Confidence Interval for Mean	Lower Bound	49.21	
		Upper Bound	51.45	
	5% Trimmed Mean		50.37	
	Median		50.00	
	Variance		848.477	
	Std. Deviation		29.129	
	Minimum		0	
	Maximum		100	
	Range		100	
	Interquartile Range		40	
	Skewness		-.171	.048
	Kurtosis		-.892	.096

In the Statistic column, the first row, "Mean," is your sample mean. The next two rows, "95% Confidence Interval for Mean," include your lower and upper bound for your confidence interval.

Remember that you can use this for proportions as well as long as your variable is set up as an indicator variable (dichotomous and coded 0 and 1).

If you are using population data, the first row (mean) is your population mean, and you would ignore the confidence interval rows.

Deconstructing Confidence Intervals

You can generate a confidence interval using SPSS, but where do the numbers come from?

Looking back, when data has a standard normal distribution, you can find where a certain percentage of the data falls as follows:

Sample mean ± distance from mean = range of values
where Distance = Z-score x Standard deviation

For example, if you wanted to know between what two values about the middle 95% of data falls and have a mean of 10 and a standard deviation of 3,

1. You know 95% of the data falls between approximately two standard deviation away from the mean, z=-2 to z=2
2. The distance away from the mean will be the z-score x standard deviation, so 2*3=6
3. The range of values will be the 10-6 and 10+6, so between 4 to 16

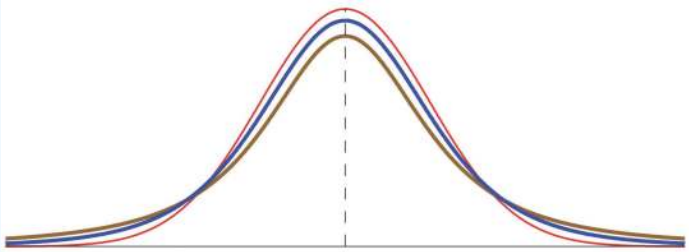
Confidence intervals work the same way, but using a normally distributed sampling distribution.

Mean ± distance from mean = confidence interval, or: Proportion ± distance from proportion = confidence interval
Distance = t-score x standard error

> Let's say we want to calculate a confidence interval with the same summary statistics, a mean of 10 and a standard deviation of 3.

Our mean is 10. Let's figure out the distance.

T-scores are based on t-distributions. A t-score for 95% of the area under the curve will be different depending on the curve, which depends on the sample size.



As the sample size gets large enough, the t-distribution will be a standard normal distribution, a z-distribution, and the t-score will equal the z-score, for a 95% confidence interval, 1.96.

There are t-tables you can use to look up approximate t-scores based on your selected confidence level and sample size, but you can do this more easily and precisely by using a t-score calculator. There are plenty online, such as this one.

When you use the online calculator, put your sample size into the box that says "degree of freedom." For significance level, put in the p-value. Remember this is 1 minus your confidence level, so for a confidence level of 95%, you would put in 0.05 (1-0.95=0.05). Click calculate. Use the two-tailed t-value.

Let's use a sample size of 100. That gives us a t-score of 1.98.

Standard Error, the standard deviation of the sampling distribution, is calculated based on the standard deviation and sample size.

$$\frac{s}{\sqrt{n}}$$

In our case:

$$\frac{3}{\sqrt{100}} = \frac{3}{10} = 0.3$$

You can also calculate standard error for a proportion, using this formula:

$$\sqrt{\frac{P(1-P)}{n}}$$

Distance = t-score x standard error

Distance = 1.98 x 0.3 = 0.6

Mean ± distance from mean = confidence interval

10 ± 0.6 = confidence interval
10 - 0.6 = 9.4, our lower bound

10 + 0.6 = 10.6, our upper bound
Confidence interval (95% confidence level): 9.4, 10.6

Interpretation: I am 95% confident the population mean is somewhere between 9.4 and 10.6.

Statistics in Action: Do New Jersey residents support the state's plastic bag ban?

You will see polls in the news frequently. Oftentimes they have misleading headlines and text, where they report out sample poll data as if they are describing the population.

For example, in the Monmouth poll release, the article states:

> *Overall, 45% of New Jerseyans say, based on their own experience, all stores in the state are following the plastic bag ban and another 45% say most stores are.*

However, 45% of New Jerseyans did not say this. 45% of the people who responded to their poll did.

Usually at the very end of articles like this you will find a mention of inferential statistics, but it is not incorporated into the article itself.

For example, in the Monmouth poll release, the last paragraph states:

> *The Monmouth University Poll was conducted by telephone from August 10 to 14, 2023 with 814 New Jersey adults. The question results in this release have a margin of error of +/- 5.4 percentage points for the full sample. The poll was conducted by the Monmouth University Polling Institute in West Long Branch, NJ.*

When polls use the term **margin of error** (abbreviated as MOE), this is the same as **distance from the mean/proportion** that we were using above.

Oftentimes they give one margin of error, a "maximum margin of error," constructed based on a proportion of 50%. Margins of error would differ for more specific proportions.

To calculate a confidence interval, I would use the same formula, taking the sample mean/proportion and adding/subtracting the distance (or in this case the MOE).

28. Do you support or oppose banning single-use plastic bags in New Jersey, such as those given out at supermarkets or other stores?

Trend:	Aug. 2023	April 2022*	Sept. 2019**
Support	**56%**	61%	65%
Oppose	**42%**	37%	29%
(VOL) Don't know	**2%**	2%	7%
(n)	***(814)***	*(802)*	*(713)*

** Question wording in 2022 was "Do you support or oppose a ban on..."*

*** Question wording in 2019 was "Would you support or oppose a ban on..."*

So for example, for this question:

Confidence Interval = Sample mean/proportion ± margin of error

56% ± 5.4% would give me a 95% confidence interval of 50.6%, 61.4%

We are 95% confident that, in mid-August 2023, somewhere between 50.6% and 61.4% of New Jersey adults support banning single-use plastic bags in New Jersey, such as those given out at supermarkets or other stores.

While our confidence interval does indicate that, at 95% confidence level, a majority of New Jersey adults support the state's plastic bag ban, recall that the headline included a second claim: "Majority continues to back plastic bag ban, but at a slightly lower level."

The provided sample statistics show that among those polled, support fell from 61% to 56% between April 2022 and August 2023. The April 2022 survey release indicated a margin of error of 3.5%. So, in April 2022, we can be 95% confident that somewhere between 57.5% and 64.5% of New Jersey adults supported the ban, and in August 2023, somewhere between 50.6% and 61.4% of New Jersey adults supported the ban. Yes, support could have fallen, but support could also be the same (it could have been 60% support both times), or it could have even increased (e.g., from 58% to 61%)! Be wary of discussions focused on changes in sample statistics that don't incorporate inferential statistics into their analyses. If I don't use the provided maximum margins of error and calculate my own for these proportions, I get 3.38% for 2022 and 3.42% for 2023. This makes the 95% confidence intervals for support 57.62% to 64.38% in 2022 and 52.58% to 59.42% in 2023. This still does not demonstrate, at 95% confidence, that there was a decrease in support.

Regardless, our confidence intervals are only as good as our question and data. Notice that the Monmouth poll question forced people to choose if they supported or opposed the ban. No one was allowed to be neutral on it, unless they volunteered that they don't know, even though they weren't given that option as part of the question.

It's also critical to be careful about subgroups. Their given margin of error is for all 814 respondents. But if you get into questions about subgroups (e.g., particular age groups, genders, political parties, religions, etc.), this decreases the sample size and increases the margin of error.

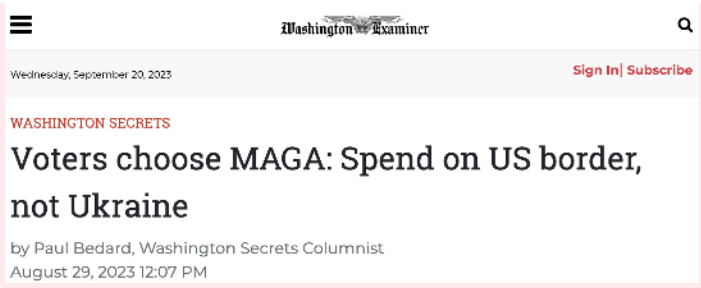
Caution

Even when you account for sampling error, you still can't fix bad data. Be careful of non-scientific push polls that are used to get out certain messages or to generate particular outcomes that can then be promoted.

For example, take a look at this headline/release:



and this corresponding news story:



The reported margin of error is 3%, so with survey findings that,

> *The latest Rasmussen Reports national telephone and online survey finds that 60% of Likely U.S. voters believe getting the migrant crisis at the U.S. border under control is more important for America's national security, while 30% say supporting Ukraine in its war against Russia is more important. Another 10% are not sure.*

we can be 95% confident that among U.S. likely voters, at least 23% more say migrant vs Ukraine.

But wait. Here are the first two survey questions:

> *1* Is the current situation with migrants at the U.S.-Mexico border a crisis?*
>
> *2* Which is more important for America's national security – getting the migrant crisis at the U.S. border under control, or supporting Ukraine in its war against Russia?*

The survey wording labels one option as a crisis and portrays the Russia-Ukraine War as an act of aggression on Ukraine's part. Different survey wordings can lead to very different answers.

What impacts confidence interval widths?

Confidence interval widths (the precision of your estimate) vary based on sample size, variation in sample data, and confidence levels.

Sample Size

- Larger samples have narrower confidence interval widths. You can make more precise predictions of the population mean.
- Smaller samples have wider confidence interval widths. With fewer cases, it's harder to be as sure about what the population mean is.

My university has about 13,000 students. If I were to randomly sample 100 students, it would be harder for me to make claims about the population with accuracy. If I were to randomly sample 1,000 students, I could do this with much more accuracy. If I were to randomly sample 12,000 students, my claims would be pretty precise.

Above, our confidence interval for the income of full-time, year-round, civilian U.S. workers ages 16. and older had a width of less than $30! A primary reason from this --- the sample had over 5 million cases! With that many randomly selected respondents from our target population, we can have much more certainty about the population mean.

Mathematically, a larger sample size will decrease the standard error and the t-score, which will make the distance/margin of error smaller.
That being said, at a certain point increasing the sample size does not affect the t-score much. At the 95% confidence level, a sample size of 120 has a t-score of 1.98, a sample size of 300 has a t-score of 1.97, a sample size of 500 has t-score of 1.96 (well, 1.96472), and a sample size of 1 million has a t-score of 1.96 (well, 1.959966). As the sample size increases, the t-distribution approaches a standard normal distribution (the z-distribution), and the t-score will equal the z-score.

Variation in Sample Data

- More variation (a larger standard deviation) widens confidence interval widths. With more variation, it's harder to be as sure what the population mean is.
- Less variation (a smaller standard deviation) narrows confidence interval widths. With more uniformity, you can make more precise predictions of the population mean.

If you randomly sample 200 people and they all have similar answers, you can be more comfortable estimating that the population mean is close to their mean. But if their answers have a lot of variation, there's more of a chance that sampling error might be impacting your mean.

Mathematically, more variation (a larger standard deviation) increases the standard error, which increases the distance/margin of error.

Confidence Levels

- Lower confidence levels (e.g., 80%) narrow confidence interval widths. If I don't have to be as confident the population mean is in my confidence interval, I can give a more precise estimate.
- Higher confidence levels (e.g., 99%) widen confidence interval widths. If I need to be more confident the population mean is in my confidence interval, I will give a broader range of possible means, increasing the chance the population mean is within that range.

Think about it. Using the feelings thermometer (0 cold to 100 warm) and a sample mean of 50, I can be 100% confident that a population mean will be somewhere between 0 and 100. I am not going to be as confident that the sample mean is between 49.9 and 50.1.

You could make a really narrow confidence interval if you constructed a 50% or even 5% confidence interval. But this is useless. Who cares about your claim if you are only half confident or 5% confident that your population mean is within your interval? It's just as likely or more likely to be outside your interval!

Mathematically, higher confidence levels increase t-scores, which increases the distance/margin of error.

This page titled 5: Confidence Intervals is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Ezra Temko (Consortium of Academic and Research Libraries in Illinois (CARLI)) .

6: Comparing Confidence Intervals

In Chapter 5 we explored confidence intervals and how we can make inferential claims about our target population based on representative sample data. In this chapter we'll again make inferential claims and explore confidence intervals, but this time we will compare confidence intervals, including among subgroups. When we have representative sample data, we can compare means and proportions at the population level. This chapter will also introduce language of significance and substantiveness, which will be followed up on in more depth in Chapter 7.

Statistics in Action: The 2024 Republican Presidential Primary

With our 24/7 news cycle, we are inundated with discussion of poll results, and small changes in polls often get blasted out as important and meaningful. Take this summer 2023 news release: "Reuters/Ipsos Survey: Despite indictments, Trump leads primary field as DeSantis loses support." The headline is supported by this claim: "Former President Donald Trump continues to lead the Republican primary field (47%). At the same time, Florida Governor Ron DeSantis' support has dropped to 13%, a six-point drop from mid-July."

Caution

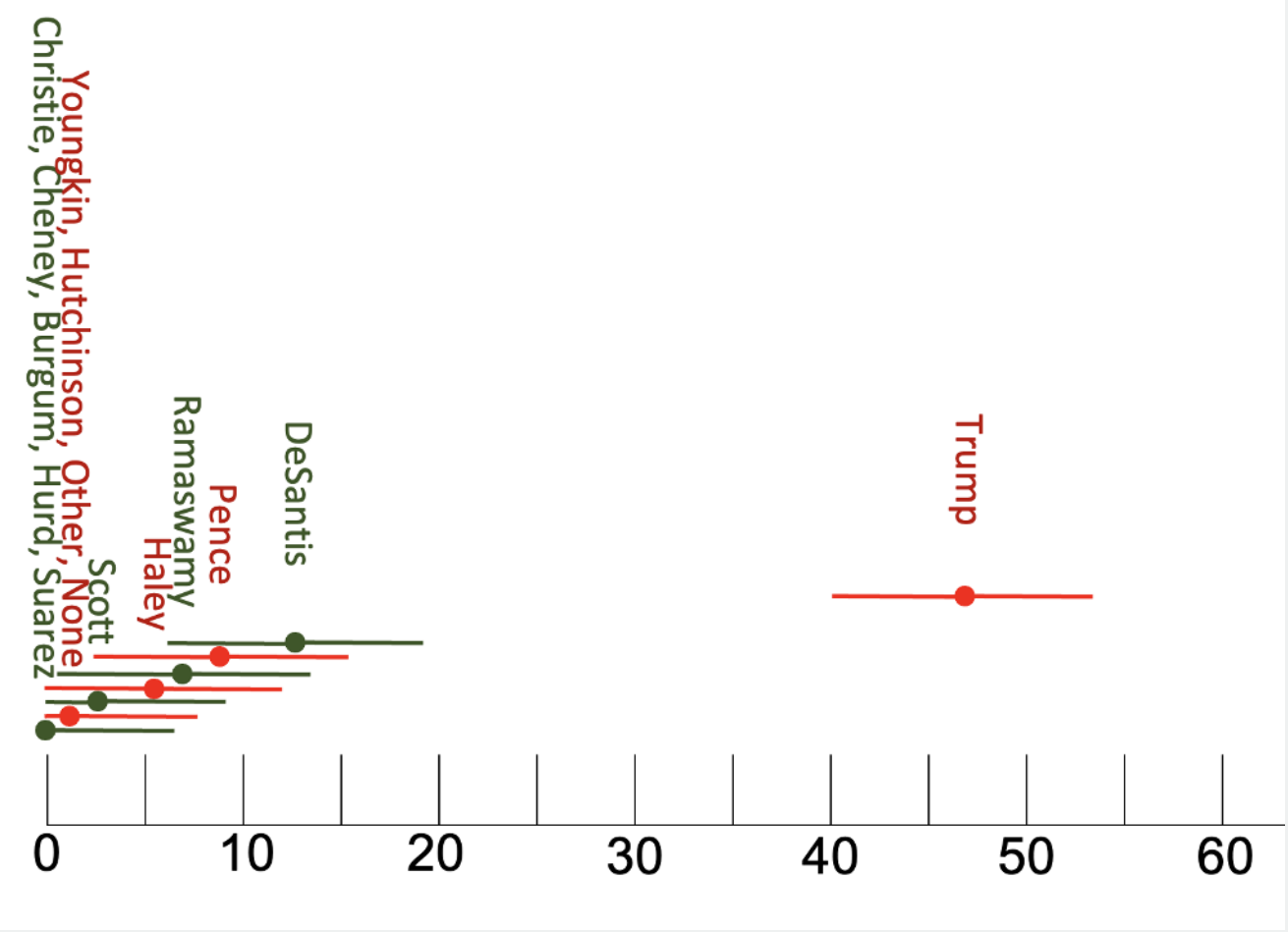
This poll uses a non-probability online sample, where respondents have been recruited into panels or click on online advertisements. While Reuters/Ipsos uses weighting, this cannot account for if the people in the sample are different from people who share their demographics that were not in the sample. Because it is not a probability sample, we would not do inferential statistics or build a confidence interval. Reuters/Ipsos, however, uses something called a "credibility interval." The American Association for Public Opinion Research states they are concerned with underlying biases associated with nonprobability online samples, though they also note traditional probability samples can also be problematic when they have areas of noncoverage and high levels of nonparticipation. For our purposes here, however, let's use the "credibility interval" as if it were a regular "margin of error" and look at the poll results and inferential statistics in kind.

Does Trump lead the primary field? Trump's sample proportion was 47%.
The confidence/credibility interval will be: 47% ± 6.4%, so among all Republicans, support for Trump was somewhere between 41.6% to 53.4%.

The next highest polling candidate, DeSantis, is at 13%. 13% ± 6.4% gives us a range of 6.6% to 19.4%.

The headline indicating that "Trump leads primary field" seems accurate.

The graph below shows the percentage each candidate got among the 365 Republican respondents in the sample, along with a line that represents the confidence interval for where we could reasonably expect the actual percentage of support to be among all U.S. Republicans.



We can see that, even taking into account the confidence intervals, Trump is leading the pack. If this were a traditional confidence interval, we would say that, at the 95% confidence level, the lowest support Trump might have is 41.6%, and the highest any other candidates might have is DeSantis at 19.4%. These confidence intervals do not touch or overlap, meaning that we are over 95% confident that, among all U.S. Republicans (at the time of the poll), there was more support for Trump than for DeSantis or other candidates. We refer to this as being "significantly" different.

Definition: Significant difference

If we are over 95% confident that two proportions (or means) are different, they are significantly different.

Statistical significance is a technical statistics term. It does not mean that the difference is meaningful or important, just that we are over 95% confident it exists among our target population. **If two 95% confidence intervals do not overlap, the proportions/means are significantly different.**

Is DeSantis ahead of Pence? DeSantis was 2nd place in the poll, at 13%, while Pence was 3rd place in the poll, at 8%. So 1/20 more Republican surveytakers preferred DeSantis to Pence. But again, we don't really care about the 365 Republicans who were surveyed. We want to know about all U.S. Republicans' preferences. If you look at the confidence intervals, they overlap.

If two 95% confidence intervals overlap, the proportions/means are not significantly different. They could be the same. At the 95% confidence level, DeSantis could be ahead of Pence, Pence could be ahead of DeSantis, or they could be tied. Thus, even though the sample proportions are different, the two are not significantly different. We cannot be over 95% confident that, among all U.S. Republicans, DeSantis is ahead of Pence or that DeSantis has a different level of support from Pence.

Example 6.1

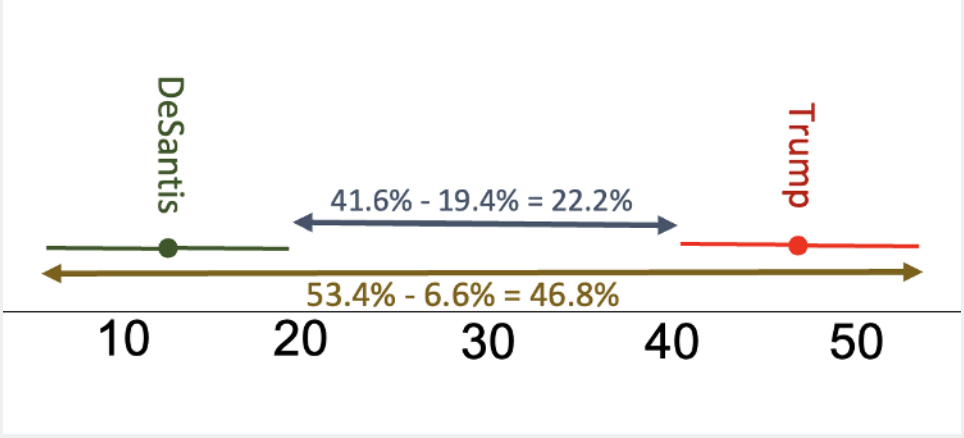
Sometimes you will see *significance* language in news stories. For example, check out the excerpts below from the article, "Poll finds dead heat in Mississippi governor's race." I bolded the language that refers to the concept of statistical significance. When the article says there is a statistical tie or an insignificant difference, it means that when accounting for sampling error, we cannot be over 95% confident that either candidate has the lead among our target population.

> *Two titans of Mississippi politics are **statistically tied** in the race for governor, a new survey shows, highlighting what will be one of the marquee matchups in what might otherwise be a quiet political year.*
>
> *The survey, conducted by the nonpartisan firm Mason-Dixon Polling and Strategy, finds Attorney General Jim Hood (D) holding a slim two-point lead over Lt. Gov. Tate Reeves (R), 44 percent to 42 percent, a **statistically insignificant edge** in a poll with a margin of error of plus or minus 4 percentage points.*
>
> *If retired Supreme Court Justice Bill Waller runs as an independent, Hood would cling to a similarly **insignificant** 40 percent to 38 percent **lead** over Reeves. Waller takes 9 percent of the vote in that scenario.*

While we determined that Trump has a **significant** lead over his primary opponents, it does not mean the difference is sizable. If we were over 95% confident that Trump had 0.0001% more support among U.S. Republicans than other candidates, he would still have a **significant** lead. It's not the dictionary definition of significant.

This then leads us to our next question, whether or not the difference is **substantive.**

In the poll sample, Trump was 34% ahead of DeSantis, 47% to 13%. However, with our 95% confidence intervals, Trump could be anywhere from 22.2% to 46.8% ahead of DeSantis. We are 95% confident that Trump is at least 22.2% ahead of DeSantis among U.S. Republicans.



To figure those numbers out, I looked at how close together the confidence intervals could be. I subtracted the upper bound of DeSantis' confidence interval from the lower bound of Trump's confidence interval. I am 95% confident that they are at least that far apart. To find out at most how far apart they might be at the 95% confidence level, I subtracted the lower bound of DeSantis' confidence interval from the upper bound of Trump's confidence interval.

Given that I am 95% confident that at over 1 in 5 more Republicans prefer Trump to DeSantis, this is a **substantive** difference.

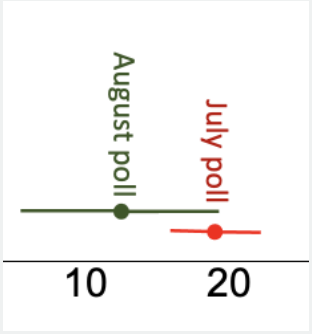
Definition: Substantive difference

A difference that is big, sizable, notable. A difference we care about. If possible, we evaluate this at the population rather than sample level.

For comparing support between DeSantis and Pence, if we are not even confident a relationship exists, then it's not going to be substantive either. Unless we are making a Type 2 error (a false negative, claiming there is not a significant difference when actually there is a difference among all U.S. Republicans between support for DeSantis and Pence), insignificant relationships are not substantive.

Going back to the article's original claims, the next part of the headline was that DeSantis lost support, with a supporting claim that he had lost 6% support over the past month. On the surface, this seems to likely not be justified, given that the supposed drop in support is less than the offered credibility interval. The comparison was to a poll the prior month of 1,640 Republicans with "a credibility interval of... plus or minus 3.0% for Republicans." Trump still came in at 47% in this poll, but DeSantis was at 19%. 19% ± 3% gives us a range of 16% to 22%.

So, we are claiming that among all Republicans, July 11-17 Desantis had somewhere between 16% and 22% of Republican support, and August 2-3 Desantis had somewhere between 6.6% and 19.4% of Republican support. While DeSantis's support among Republicans could have fallen from 22% to 6.6%, it could also have increased from 16% to 19.4%. The polls do not give us evidence to support the claim that DeSantis lost support among Republicans.

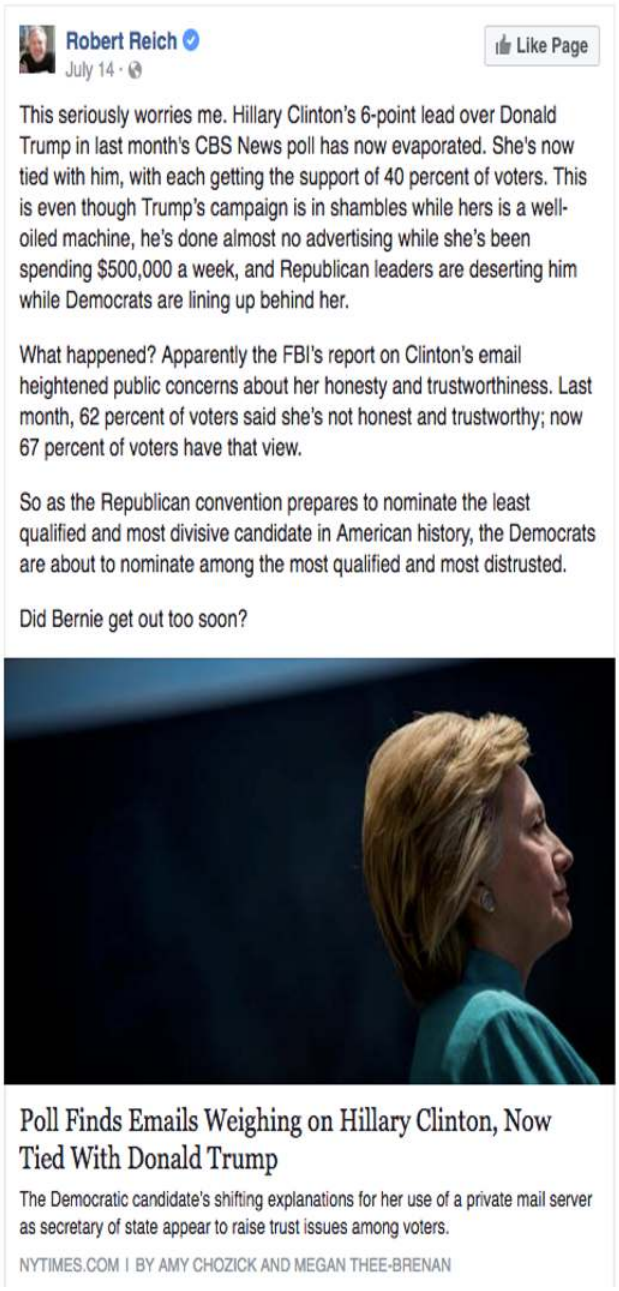


You can see in the figure above that the confidence intervals overlap. Yes, DeSantis could have fallen, but he could also be at the same level of support both times. We don't have support at the 95% confidence level to argue that DeSantis fell from July to August. DeSantis' July and August numbers are not **significantly** different.

Statistics in Action: A Non-Story About Hillary Clinton's 2016 bid for president

We see these kinds of claims across the media, and among social and political commentators, who get swept up in candidate polling. For example, I remember when former Secretary of Labor Bob Reich put out a Facebook post about a New York

Times article that sensationalized the changes from their June to July polls. Check out the post below on the left and then my discussion of it below on the right.



Reich claimed that Clinton went from having a lead among voters in June to being tied in July. But the change was not statistically significant.

- The claim was based on a New York Times article reflecting June 9-13 and July 8-12 CBS News/New York Times live phone polls.
- The June poll results showed that 43% of the 1,048 registered voters surveyed said they would vote for Clinton, compared to 37% for Trump. However, it had a 3% margin of error, so at the 95% confidence level, among all U.S. registered voters, Clinton had somewhere between 40% and 46% support and Trump had somewhere between 34% and 40% support. So Clinton may have had as much as a 12% lead... or may have been tied with Trump.
- The July poll results showed that 40% of the 1,358 registered voters surveyed said they would vote for Clinton, compared to 40% for Trump. This poll also had a 3% margin of error, so at the 95% confidence level, among all U.S. registered voters, both Clinton and Trump had somewhere between 37% and 43% support. So in July Clinton may have had a lead by as much as 6%, been tied, or Trump could have been ahead by as much as 6%.
- We can be 95% confident that from the time fo the June poll to the time of the July poll, Clinton had somewhere between 9% less support and 3% more support. We don't know whether Clinton went from having a lead to being tied. She could have went from having a lead to being behind. She could have been tied the whole time. Indeed, other polls at the time had different results.

Reich also claimed that in July, 5% more voters felt Clinton was not honest or trustworthy compared to June, and then argued that the FBI's report on Clinton's e-mail was the reason for this 5% shift. But this change was also not statistically significant.

- In the June poll, 62% of the 1,048 registered voters said they thought Clinton was not honest or trustworthy (note the question asked them, "Do you think Hillary Clinton is honest and trustworthy, or not?," trying to force them into a non-neutral opinion).
- In the July poll, 67% of the 1,358 registered voters surveyed said they thought Clinton was not honest or trustworthy.
- With the polls' 3% margins of error, at the 95% confidence level, these would reflect somewhere between 59% and 65% of U.S. registered voters at the time of the June poll and somewhere between 64% and 70% of U.S. registered voters at the time of the July poll.
- We can be 95% confident that, from the time fo the June poll to the time of the July poll, the percentage of U.S. registered voters who would evaluate Clinton as not honest or trustworthy changed somewhere between decreasing 1% to increasing 11% (including the proportion could have stayed

the same). And remember, this is with a 95% confidence level. This means that if the same poll had been conducted 100 times, we would expect the results to be within our margin of error (3%) of the actual percent of U.S. registered voters 95 of those 100 times. That also means that 1/20 of the time we would expect it not to be.

Sampling error takes into account that samples are often not exactly representative of their target populations. Taking into account margins of error helps us have a sense of how close we should expect the poll results to be to how the target population would respond to such a poll. Building these confidence intervals shows that there could have been real changes, or it could just be that the July poll's random sample of registered voters liked Clinton less than the sample from June, regardless of the month. Of course, taking into account the polls' margins of errors results in a much less sexy news story --- read the headline now: *We're not Sure if Anything Changed This Past Month*.

Statistics in Action: Financial Worries

Quinnipiac University Poll

March 30, 2023

Food Costs Top List Of Americans' Financial Concerns, Quinnipiac University National Poll Finds; 68% Concerned They Won't Have Enough Money To Live Comfortably In Retirement

A Quinnipiac University Poll RDD (random digit dialing) telephone survey of 1,795 U.S. adults, conducted March 9-13, 2013, had a reported margin of sampling error of "+/- 3.0 percentage points."

One poll question asked was, "What is your biggest personal financial concern right now: food costs, healthcare costs, mortgage or rent payments, energy bills, credit card or loan payments, college tuition, retirement savings, or loss of a job?"

According to the release,

> *Food costs rank as the most pressing financial worry, with 22 percent of Americans naming it as their biggest personal financial concern right now, followed by retirement savings (18 percent), healthcare costs (17 percent), mortgage or rent payments (13 percent), college tuition (8 percent), energy bills (7 percent), credit card or loan payments (5 percent), and loss of a job (4 percent), according to a Quinnipiac (KWIN-uh-pea-ack) University national poll of adults released today.*

Significant differences

Among all U.S. adults at the time of the survey, was food the biggest financial concern? Which concerns can we be confident

are significantly greater than/less than others among all U.S. adults?

While we know the survey results, we want to know about the concerns of all U.S. adults.

Here are the sample proportions along with confidence intervals. In the table, P is for the sample proportion, LB for lower bound of the confidence interval, and UB for upper bound of the confidence interval:

Financial Concerns	P	LB	UB
Food costs	22%	19%	25%
Retirement savings	18%	15%	21%
Healthcare costs	17%	14%	20%
Mortgage/Rent	13%	10%	16%
College tuition	8%	5%	11%
Energy bills	7%	4%	10%
Credit card/Loan	5%	2%	8%
Loss of a job	4%	1%	7%
DK/NA	4%	1%	7%
Something else (vol.)	3%	0%	6%

While the headline read that U.S.-Americans are most concerned about food costs, the proportion of U.S. adults concerned about food costs at the time of the survey is not significantly different from the proportion concerned about retirement savings or healthcare costs. We're not sure which of these three takes the top spot. However, we are confident a higher proportion of U.S. adults are most concerned about food costs compared to each of the other five options.

Substantive differences

While evaluating whether something is substantively different involves subjective judgment calls and evaluating differences in context (don't forget to use your codebook!), in general I look for differences that I am 95% confident are at least 5% different or differences where I am 95% confident that the results for two groups have qualitatively different meanings.

- We can be confident that substantively more U.S. adults are most concerned with food costs compared to college tuition, energy bills, credit cards/loans, or loss of a job. While we are confident that somewhere between 19% and 25% of U.S. adults are most concerned with food costs, these other areas are each a top concern for somewhere between 1% and 11% of U.S. adults.
- We can be confident that substantively more U.S. adults are most concerned with retirement savings compared to energy bills, credit cards/loans, or loss of a job. While we are confident that somewhere between 15% and 21% of U.S. adults are most concerned with retirement savings, these other areas are each a top concern for somewhere between 1% and 10% of U.S. adults.
- We can be confident that substantively more U.S. adults are most concerned with health care costs compared to credit cards/loans, or loss of a job. While we are confident that somewhere between 15% and 21% of U.S. adults are most concerned with retirement savings, these other areas are each a top concern for somewhere between 1% and 10% of U.S. adults.

Concerns by Age

The survey release continues,

> *There are differences among most listed groups, particularly when considering age.*
>
> *The top three personal financial concerns broken down by age:*

- *18 to 34 year olds: mortgage or rent payments (23 percent), college tuition (20 percent), and food costs (19 percent);*
- *35 to 49 year olds: food costs (24 percent), retirement savings (17 percent), and mortgage or rent payments (15 percent);*
- *50 to 64 year olds: retirement savings (25 percent), food costs (23 percent), and healthcare costs (19 percent);*
- *65 years of age and over: retirement savings (26 percent), food costs (22 percent), and healthcare costs (20 percent).*

These differences in the narrative were based on sample proportions. Upon request, I received margins of error for the age subgroups: 6.09% for ages 18-34, 5.46% for ages 35-49, 4.46% for ages 50-64, and 3.91% for ages 65+. The table below shows sample proportions and confidence intervals for each age group. Looking at each age group, the yellow highlights are the highest sample proportions for the age group, and the grey-shaded cells are the areas that, for that age group, are not significantly different than the yellow highlights.

	18-34 year olds			35-49 year olds			50-64 year olds			65+ year olds		
Financial Concerns	**P**	**LB**	**UB**	**P**	**LB**	**UB**	**P**	**LB**	**UB**	**P**	**LB**	**UB**
Food costs	19%	12.9%	25.1%	24%	18.5%	29.5%	23%	18.5%	27.5%	22%	18.1%	25.9%
Retirement savings	6%	-0.1%	12.1%	17%	11.5%	22.5%	25%	20.5%	29.5%	26%	22.1%	29.9%
Healthcare costs	13%	6.9%	19.1%	14%	8.5%	19.5%	19%	14.5%	23.5%	20%	16.1%	23.9%
Mortgage/Rent	23%	16.9%	29.1%	15%	9.5%	20.5%	10%	5.5%	14.5%	4%	0.1%	7.9%
College tuition	20%	13.9%	26.1%	6%	0.5%	11.5%	3%	0.0%	7.5%	1%	0.0%	4.9%
Energy bills	2%	0.0%	8.1%	7%	1.5%	12.5%	9%	4.5%	13.5%	10%	6.1%	13.9%
Credit card/Loan	8%	1.9%	14.1%	6%	0.5%	11.5%	4%	-0.5%	8.5%	3%	0.0%	6.9%
Loss of a job	4%	0.0%	10.1%	5%	-0.5%	10.5%	4%	-0.5%	8.5%	2%	0.0%	5.9%
DK/NA	2%	0.0%	8.1%	4%	0.0%	9.5%	1%	0.0%	5.5%	9%	5.1%	12.9%
Something else (vol.)	2%	0.0%	8.1%	3%	0.0%	8.5%	3%	0.0%	7.5%	4%	0.1%	7.9%

It could be that food costs were the top concern for each age group, though it could also be that food costs were not the top concern for any age group. What is most notable is that mortgage rent is substantially more likely to be a top concern for 18-34 year olds than 65+ year olds, college tuition is substantially more likely to be a top concern for 18-34 year olds than 50-64 and 65+ year olds, and retirement savings is substantially more likely to be the top concern for 50-64 and 65+ year olds compared to 18-34 year olds.

SPSS

Using Explore/Examine to Examine Relationships

Generating confidence intervals for subgroups is very similar to generating confidence intervals like we did in Chapter 5, but with one extra step (in bold below). This is the same process you used in Chapter 2 to generate comparative statistics and side-by-side box plots for analyzing relationships.

Menu:

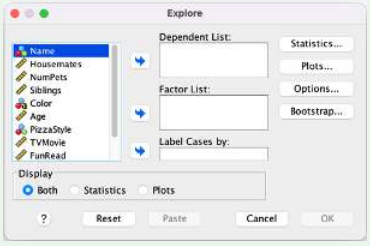
Go to: Analyze → Descriptive Statistics → Explore.

Put your variable into the "Dependent List:" window/box.

Put your categorical variable into the "Factor list:" box.

At the bottom where it says "Display," make sure you either have "Both" or "Statistics" checked. (Plots will only give you

graphs/figures.)



Syntax:

EXAMINE VARIABLES=PrimaryVariableName **BY CategoricalVariableName**
/PLOT BOXPLOT
/COMPARE GROUPS
/PERCENTILES(25,50,75) HAVERAGE
/STATISTICS DESCRIPTIVES
/CINTERVAL 95
/MISSING LISTWISE
/NOTOTAL.

Statistics in Action: feelings towards journalists

In Chapter 5, we found that, at the 95% confidence level, following the 2020 general election, U.S. eligible voters mean feelings thermometer scores towards journalists were somewhere between 49.21° and 51.45°, meaning on average they were neutral or close to neutral in their feelings towards journalists (the average is somewhere between just cooler or warmer than neutral).

Biden vs. Trump

Let's compare feelings towards journalists among those who voted for Biden vs. those who voted for Trump.

Case Processing Summary

	Did you vote for president? If so, who did you vote for for president?	Valid N	Valid Percent	Missing N	Missing Percent	Total N	Total Percent
POST: Feeling thermometer: journalists	Joe Biden	1016	99.5%	5	0.5%	1021	100.0%
	Donald Trump	826	99.3%	6	0.7%	832	100.0%
	Other Candidate	52	99.7%	0	0.3%	52	100.0%
	Did not vote	51	100.0%	0	0.0%	51	100.0%

Descriptives

	Did you vote for president? If so, who did you vote for for president?			Statistic
POST: Feeling thermometer: journalists	Joe Biden	Mean		69.45
		95% Confidence Interval for Mean	Lower Bound	68.20
			Upper Bound	70.71
		5% Trimmed Mean		70.49
		Median		70.00
		Std. Deviation		20.400
		Minimum		0
		Maximum		100
		Range		100
		Interquartile Range		30
	Donald Trump	Mean		28.25
		95% Confidence Interval for Mean	Lower Bound	26.68
			Upper Bound	29.83
		5% Trimmed Mean		26.89
		Median		30.00
		Std. Deviation		23.076
		Minimum		0
		Maximum		100
		Range		100
		Interquartile Range		40



Looking at the sample means, we can see that, among survey respondents who voted for Biden, the mean feeling towards journalists was 69.45 degrees, or about fairly warm/favorable. Among survey respondents who voted for Trump, the mean feeling towards journalists was 28.25 degrees, or about fairly cold/unfavorable. These are 41.2 degrees apart.

As noted, we don't really care about how the survey participants feel, however. We want to know about how all Biden voters and how all Trump voters feel. Here are their confidence intervals: Biden 95% CI: 68.20°, 70.71°; Trump 95% CI: 26.68°, 29.83°

1. Are they significantly different?
Yes! At the 95% confidence level, the warmest Trump voters might be, on average, is 29.83°, and the coldest Biden voters might be, on average, is 68.20°. These do not overlap. We are over 95% confidence that Trump and Biden voters had different mean feelings towards journalists following the 2020 elections.

2. Are they substantively different?
We are 95% confident that Trump and Biden supporters' mean feelings towards journalists are at least 38.37 degrees apart (and could be up to 44.03 degrees apart). (I got these numbers by finding the difference of the closest bounds from each CI to each other, 68.20-29.83, and the further bounds from each CI to each other, 70.71-26.68.) Given how far apart they are, and that we are confident Trump supporters average cold feelings and that Biden supporters average warm feelings, they are substantively different.

3. Do they have a causal relationship?
Next chapter we will discuss causality. Do you think that being a Biden vs. Trump supporter might have an impact on how one feels about journalists? Given Trump's repeated repeated claims that the media is promoting "fake news", I would say it's likely!

Definition: Causality/Causation

Causality / Causation: There is a causal relationship when one variable has an impact or effect on variation of another variable.

Gender
Let's compare feelings towards journalists among men vs. women eligible voters.

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	What is your sex?	N	Percent	N	Percent	N	Percent
POST: Feeling thermometer: journalists	Female	1337	99.0%	14	1.0%	1350	100.0%
	Male	1262	99.3%	8	0.7%	1270	100.0%

Descriptives

	What is your sex?			Statistic
POST: Feeling thermometer: journalists	Female	Mean		53.61
		95% Confidence Interval for Mean	Lower Bound	52.09
			Upper Bound	55.13
		Median		50.00
		Std. Deviation		28.291
		Minimum		0
		Maximum		100
		Range		100
		Interquartile Range		30
	Male	Mean		46.84
		95% Confidence Interval for Mean	Lower Bound	45.21
			Upper Bound	48.48
		Median		50.00
		Std. Deviation		29.626
		Minimum		0
		Maximum		100
		Range		100
		Interquartile Range		50



Looking at the sample means, we can see that, among survey respondents, women's mean feelings towards journalists were 53.61°, a tad warmer than neutral, and men's mean feelings towards journalists were 46.84°, a tad colder than neutral.

We don't really care about how the survey participants feel, however. We want to know about all U.S. men and women eligible voters feel.
Here are their confidence intervals: Women 95% CI: 52.09°, 55.13°; Men 95% CI: 45.21°, 48.48°

1. Are they significantly different?
Yes! At the 95% confidence level, the warmest men could be, on average, is 48.48°, and the coldest women could be, on average, is 52.09°. These do not overlap. We are over 95% confidence that men and women eligible voters had different mean feelings towards journalists following the 2020 elections.

2. Are they substantively different?

We are 95% confident that men and women, on average, are at least 3.61 degrees different (and at most 9.92 degrees different). (I got these numbers by finding the difference of the closest bounds from each CI to each other, 52.09-48.48, and the further bounds from each CI to each other, 55.13-45.21.) While these numbers are not that far apart on the 0° to 100° thermometer, given that we are 95% confident women's average leans warm (at least 52.09°) and men's average leans cold (at most 48.48°), this seems substantive to me.

3. Do they have a causal relationship?

Does gender impact variation in how people feel towards journalists? Following the 2020 election, this could make sense, if for no other reason than that, compared to men, women were more supportive of Biden and less supportive of Trump.

Age

Let's compare feelings towards journalists by age.

Case Processing Summary

	Age (in 3 categories)	Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
POST: Feeling thermometer: journalists	18-39	826	99.1%	8	0.9%	834	100.0%
	40-64	1181	99.3%	8	0.7%	1189	100.0%
	65+	543	99.1%	5	0.9%	548	100.0%

Descriptives

	Age (in 3 categories)			Statistic
POST: Feeling thermometer: journalists	18-39	Mean		52.90
		95% Confidence Interval for Mean	Lower Bound	51.03
			Upper Bound	54.76
		Median		50.00
		Std. Deviation		27.365
		Minimum		0
		Maximum		100
		Range		100
		Interquartile Range		30
	40-64	Mean		49.56
		95% Confidence Interval for Mean	Lower Bound	47.89
			Upper Bound	51.24
		Median		50.00
		Std. Deviation		29.376
		Minimum		0
		Maximum		100
		Range		100
		Interquartile Range		40
	65+	Mean		49.48
		95% Confidence Interval for Mean	Lower Bound	46.87
			Upper Bound	52.09
		Median		50.00
		Std. Deviation		30.913
		Minimum		0
		Maximum		100
		Range		100
		Interquartile Range		50



Looking at the sample means, we can see that survey respondents ages 18-39's mean feeling towards journalists was 52.90°, a tad warmer than neutral, survey respondents ages 40-64's mean feeling towards journalists was 49.56°, just a smidgen colder than neutral, and survey respondents ages 65 or older's mean feelings towards journalists was 49.48°, also a smidgen colder than neutral.

We don't really care about how the survey participants feel, however. We want to know about all U.S. eligible voters of these age groups feel on average.
Here are their confidence intervals:

- Ages 18-39: 51.03°, 54.76°
- Ages 40-64: 47.89°, 51.24°
- Ages 65+: 46.87°, 52.09°

1. Are they significantly different?
No! While the youngest age range (18-39) had average warmer feelings in the sample, at the 95% confidence level all the confidence intervals overlap. For example, they could all be at 51 degrees, a smidgen warmer than neutral, in their feelings towards journalists.

2. Are they substantively different?
First, let's look at how different each age group's means might be from one another at the 95% confidence level.
If you're not sure how to solve this, sketch out the confidence intervals above a number line, and then figure out which numbers to subtract.
Ages 18-39 & Ages 40-64: Ages 18-39 could be anywhere from 6.87 degrees warmer (54.76-47.89) to the same (they overlap from 51.03 to 51.24) to 0.21 degrees colder (51.24-51.03) than ages 40-64.

Ages 18-39 & Ages 65+: Ages 18-39 could be anywhere from 7.89 degrees warmer (54.76-46.87) to the same (they overlap from 51.03 to 52.09) to 1.06 degrees colder (52.09-51.03) than ages 65+.

Ages 40-64 & Ages 65+: Ages 40-64 could be anywhere from 4.37 degrees warmer (51.24-46.76) to the same (they overlap from 47.89 to 51.24) to 4.2 degrees colder (52.09-47.89) than ages 65+.

So, are they substantively different?

They could be, e.g., at 95% confidence ages 18-39 could have an average that is 7.89 degrees warmer than ages 65+ (54.76-46.87), with the younger group having warm feelings on average and the older group on average cold feelings. However, we don't have evidence to show this is the case. At the 95% confidence level, ages 65+ could have slightly warmer average feelings than ages 18-39 (e.g., 52.05° vs. 51.05°) or they could have the same average feelings (e.g., both could have a mean of 52°). All three confidence intervals overlap (e.g., all three age groups could have a mean of 51.2°), so at the 95% confidence level, we are not confident they are different at all, let alone by a sizable amount. We will not claim they are substantively different.

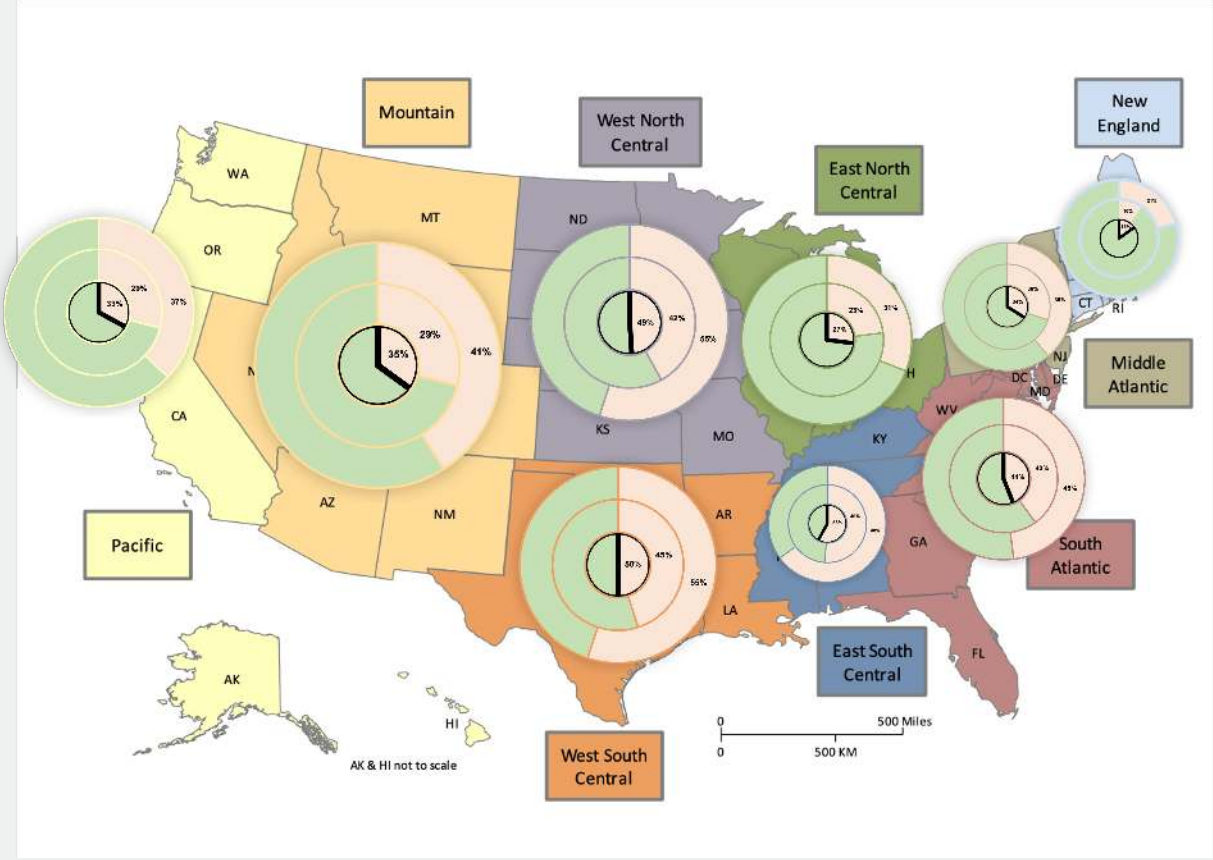
3. Are they causal?
Does age impact variation in how people feel towards journalists? Given that no relationship was identified, there cannot be a causal relationship. If age was affecting feelings towards journalists, we would expect to see different age groups have different average feelings towards journalists (barring any suppression effects, but don't worry about that right now).

Statistics in Action: Christian Prosyletization

For our final example, let's look back at the 2022 General Social Survey question asked of U.S. adults, "Have you ever tried to encourage someone to believe in Jesus Christ or to accept Jesus Christ as his or her savior?" This was recoded as an indicator variable (0=no, 1=yes). While the confidence intervals could be interpreted the same as the other examples with interpreting means (e.g., the sample mean in the Mountain region is 0.35; the 95% confidence interval for the Mountain region is 0.29 to 0.41), it makes more sense to interpret them as proportions (e.g., the sample proportion in the Mountain region is 35%, the 95% confidence interval for the Mountain region is 29% to 41%). Rather than comparing means, we're comparing the actual percentages of respondents (sample) or of the target population (in this case U.S. adults) that have the value of 1 (in this case, having engaged in Christian proselytization).

Region

The map below shows, for each region, sample proportions (inner circle) and confidence intervals at the 95% confidence level (lower bounds in middle circle, upper bounds in outer circle), with the peach-colored sectors with the labeled percentages the proportions who answered yes, that they have tried to encourage someone to believe in Jesus Christ or to accept Jesus Christ as their savior.



1. Are the regions significantly different?

https://socialsci.libretexts.org/@go/page/188425

- Significantly fewer New England adults proselytize for Jesus compared to any other regions, and vice versa. At the 95% confidence level, the maximum proportion of New England adults that have proselytized for Jesus is 21%, whereas all other regions are 23% or higher.
- The Middle Atlantic, East North Central, Pacific, and Mountain regions are not significantly different from one another (e.g., at the 95% confidence level, there could be 29% to 30% of adults in each of these regions that have proselytized for Jesus). They are, however, all significantly higher than New England and significantly lower than the remaining four regions.
- The South Atlantic, West North Central, and West South Central regions are not significantly different from one another (e.g., at the 95% confidence level, there could be 45% to 48% of adults in each of these regions that have proselytized for Jesus).
- The South Atlantic region is significantly lower than the East South Central/Atlantic region, and vice versa (at the 95% confidence level, at most 48% of South Atlantic adults have proselytized for Jesus and at least 51% of East South Central/Atlantic adults have).
- The West North Central, West South Central, and East South Central/Atlantic regions are not significantly different from one another (e.g., at the 95% confidence level, there could be 51% to 55% of adults in each of these regions who have proselytized for Jesus).

2. Are the regions substantively different (at the 95% confidence level)?

Since we are not confident about differences existing between certain regions, we also will not claim those regions have substantively different percentages of their adult residents who have proselytized for Jesus. This means that we do not have evidence that the Middle Atlantic, East North Central, Pacific, or Mountain regions are substantively different from one another, that the South Atlantic, West North Central, and West South Central regions are substantively different from one another, or that the West North Central, West South Central, and East South Central/Atlantic regions are substantively different from one another. Among the regions that are significantly different from one another, we need to evaluate whether those differences are sizable. New England is substantively different from the other seven regions; we are 95% confident New England is at least 8% lower than them (the highest New England could be is 21% and the lowest any of the other seven could be is 29%).

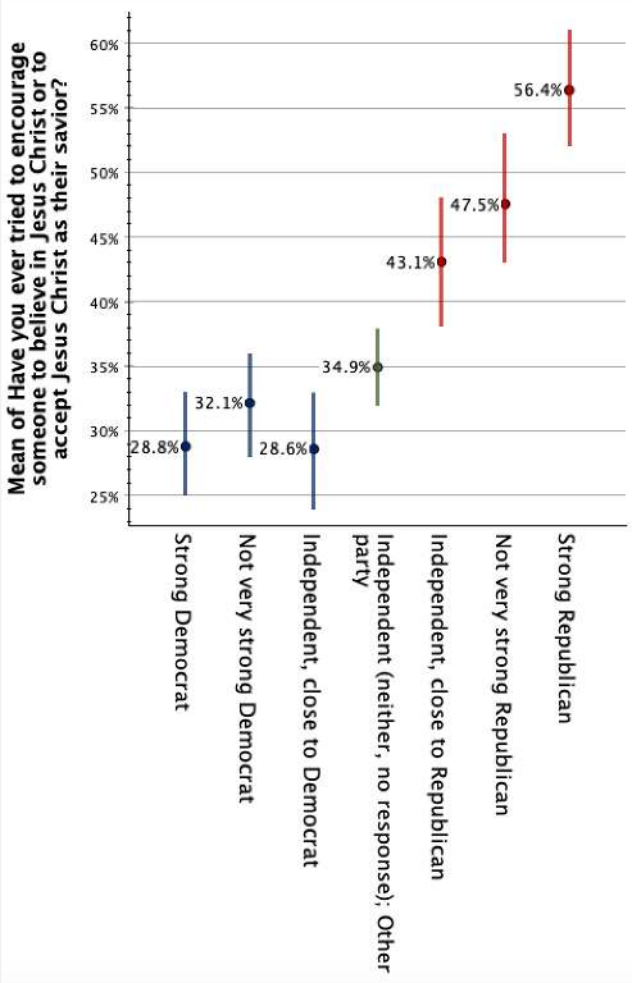
- The East North Central and Pacific regions are both substantively different from the South Atlantic, West North Central, West South Central, and East South Central/Atlantic regions. At the 95% confidence level, the East North Central region is at maximum 31% and Pacific region at maximum 33%, while the other four named regions are all at least 40% or higher.
- The Mountain region and Middle Atlantic region are substantively less than the East South Central/Atlantic region. At the 95% confidence level, the Mountain region is at most 41% and the Middle Atlantic region at most 39%, and these are meaningfully less than the East South Central/Atlantic region which is at least 51%.
- The Middle Atlantic region is also substantively less than the West South Central region, as they are at least 5% different, so at least 1/20 more adults in West South Central have proselytized compared to the Middle Atlantic region.
- We cannot be confident, at the 95% confidence level, that any other regions are 5% or more different from one another. For example, while New England is significantly lower than East North Central, and it might be substantively different (e.g., New England could be at 10% and East North Central at 31%), we are only 95% confident that New England and East North Central are at least 2% different (New England could be at 21% and East North Central at 23%). We do not have evidence to claim they are substantively different.

3. Is there a causal relationship?

The regions with more proselytizing are more religious and have a higher proportion of evangelical Christian residents. These factors could impact proselytizing rates.

Political Party

The graph below shows sample proportions along with 95% confidence intervals of those who said they have engaged in Christian proselytization, broken down by self-reported political party identification.



The graph is based on these confidence intervals:

Descriptives

	Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or what?			Statistic
Have you ever tried to encourage someone to believe in Jesus Christ or to accept Jesus Christ as their savior?	Strong Democrat	Mean		.2880
		95% Confidence Interval for Mean	Lower Bound	.2484
			Upper Bound	.3276
	Not very strong Democrat	Mean		.3215
		95% Confidence Interval for Mean	Lower Bound	.2786
			Upper Bound	.3643
	Independent, close to Democrat	Mean		.2860
		95% Confidence Interval for Mean	Lower Bound	.2386
			Upper Bound	.3335
	Independent (neither, no response); Other party	Mean		.3493
		95% Confidence Interval for Mean	Lower Bound	.3176
			Upper Bound	.3809
	Independent, close to Republican	Mean		.4307
		95% Confidence Interval for Mean	Lower Bound	.3818
			Upper Bound	.4796
	Not very strong Republican	Mean		.4754
		95% Confidence Interval for Mean	Lower Bound	.4250
			Upper Bound	.5257
	Strong Republican	Mean		.5636
		95% Confidence Interval for Mean	Lower Bound	.5175
			Upper Bound	.6096

1. Are the parties significantly different?

Looking at the graph above, you can see which confidence intervals do and do not overlap. (The one exception is

Independents/Other Party and Independents close to Republicans, as they are too close to tell, but they do not overlap --- they are 0.09% apart.)

- Strong Democrats, not very strong Democrats, Independents close to Democrats, and Independents are all not significantly different from one another. At the 95% confidence level, they could all be anywhere between 31.76% and 32.76%, and the first three Democratic or Democratic leaning groups could all be between 27.87% and 32.76%.
- These four groups are all significantly different from Independents close to Republicans, not very strong Republicans, and strong Republicans. At the 95% confidence level, the most any of the first three groups could be is 36.43%, the most the Independents group could be is 38.09%, and the least any of the last three Republican or Republican leaning groups could be is 38.18%.
- Independents close to Republicans are not significantly different from not very strong Republicans. At the 95% confidence level, they both could be between 42.5% and 47.96%.
- Independents close to Republicans are significantly different from strong Republicans. At the 95% confidence level, Independents close to Republicans are at most 47.96%, while strong Republicans are at least 51.75%.
- Not very strong Republicans and strong Republicans are not significantly different from each other. At the 95% confidence level, they could both be between 51.75% and 52.57%.

2. Are the parties substantively different?

We cannot claim that any groups that are not significantly different are substantively different. They could be, e.g., at the 95% confidence level, it could be that less than 1/4 of strong Democrats proselytize (24.84%) while over 1/3 of not very strong Democrats proselytize (36.43%), but these two groups could be identical, e.g., it could be that 3/10 of both groups proselytize.

Among the groups that are significantly different, let's explore whether we can be confident at the 95% confidence level that they are substantively different.

- Strong Democrats are substantively different from Independents close to Republicans (at least 5.4% lower), not very strong Republicans (at least 9.7% lower), and strong Republicans (at least 19.0% lower).
- Not very strong Democrats are substantively different from not very strong Republicans (at least 6.1% lower) and strong Republicans (at least 15.3% lower).
- Independents closer to Democrats are also substantively different from not very strong Republicans (at least 9.2% lower) and strong Republicans (at least 18.4% lower).
- Independents are substantively different from strong Republicans (at least 13.7% lower).
- All other groups could be less than 5% different, at the 95% confidence level. While other groups might be substantively different from one another (e.g., Independents and not very strong Republicans could be 20.81% different), we do not not have sufficient evidence to claim that they are.

3. Does party have a causal effect on Christian proselytizing?

Perhaps there are differences by party with respect to respecting difference, inclusion, and ethnocentrism. However, it also could be that party is not impacting proselytization, but different religious beliefs and traditions (atheist/agnostic, non-Christian, Christian but not evangelical, Christian evangelical) impact both political party and proselytization in the same direction.

Collapsing Categories

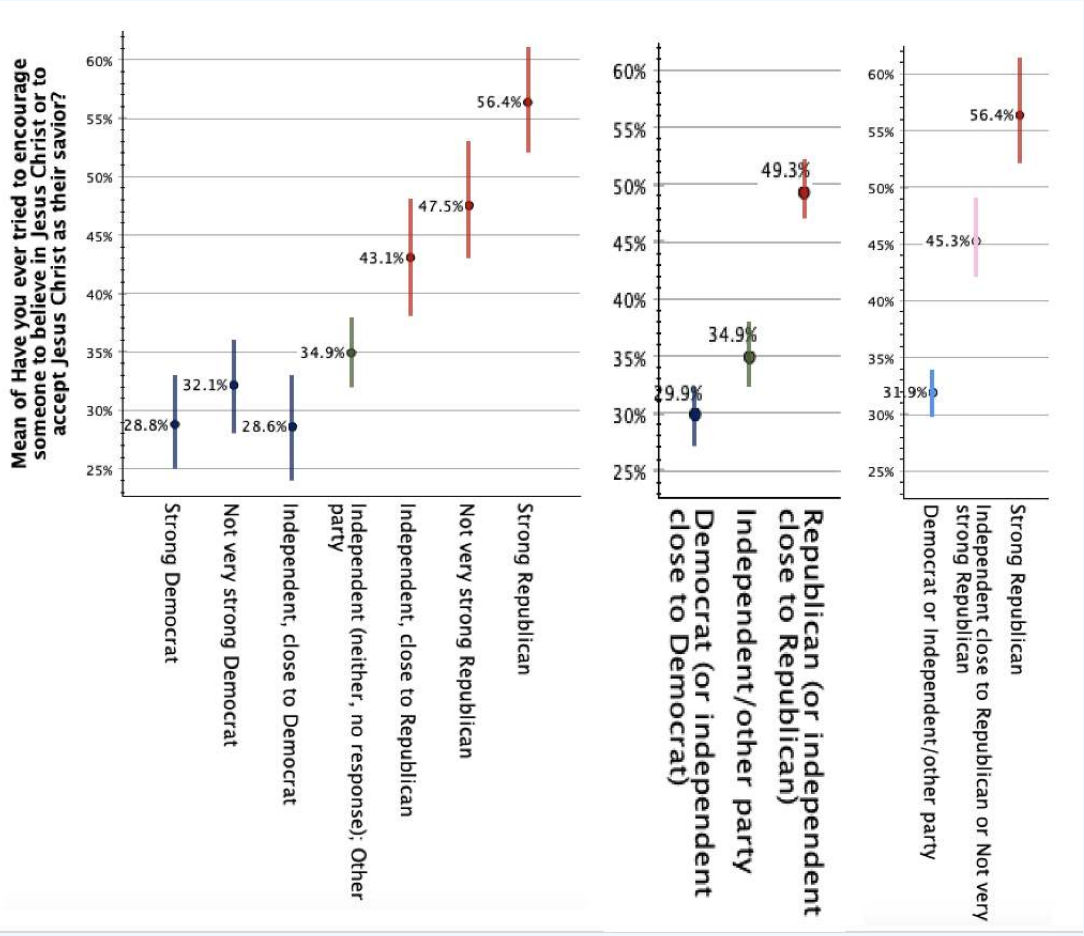
Remember that confidence intervals' widths vary based on sample size, as well as the extent of uniformity/variability within the data for that particular category. Here are confidence intervals of the same data, but this time: 1) party collapsed into three categories, Democrats, Independents, and Republicans, and 2) party collapsed into three categories based on levels of proselytization, Democrats/Independents, weak Republicans, and strong Republicans.

Descriptives

	Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or what?			Statistic
Have you ever tried to encourage someone to believe in Jesus Christ or to accept Jesus Christ as their savior?	Democrat or Independent/other party	Mean		.32
		95% Confidence Interval for Mean	Lower Bound	.30
			Upper Bound	.34
	Independent close to Republican or Not very strong Republican	Mean		.45
		95% Confidence Interval for Mean	Lower Bound	.42
			Upper Bound	.49
	Strong Republican	Mean		.56
		95% Confidence Interval for Mean	Lower Bound	.52
			Upper Bound	.61

Descriptives

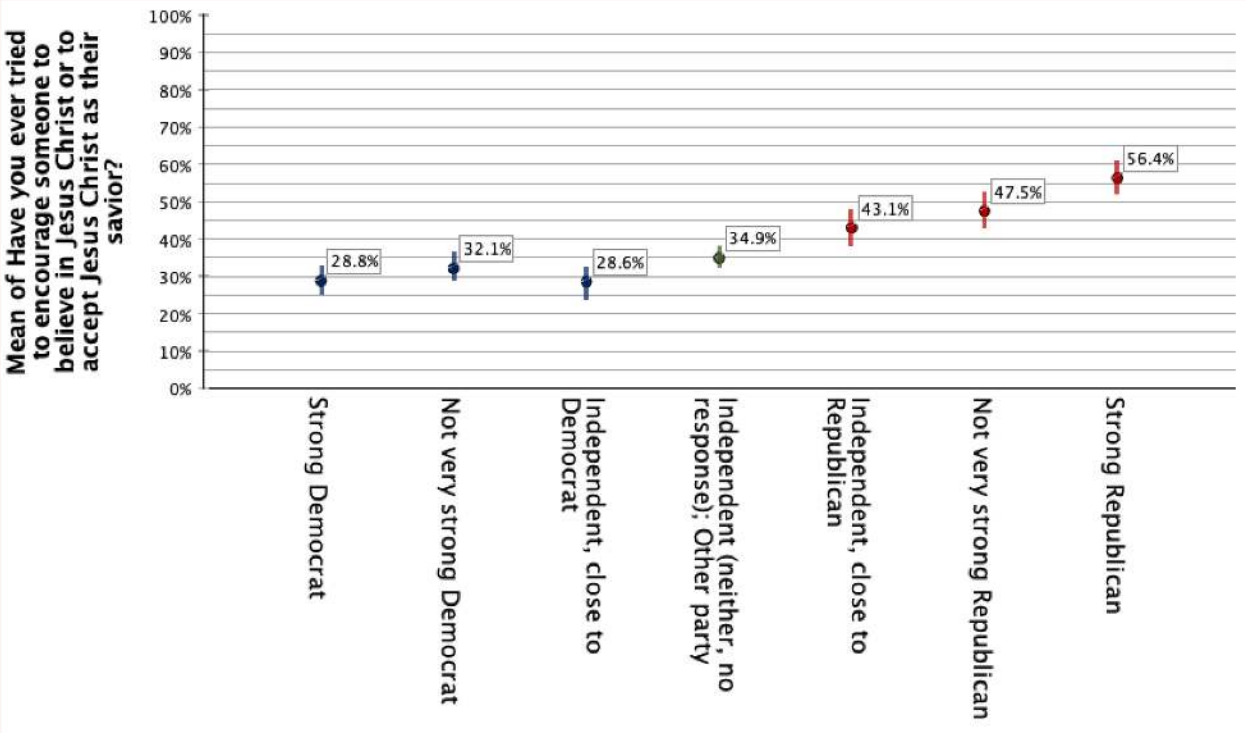
	Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or what?			Statistic
Have you ever tried to encourage someone to believe in Jesus Christ or to accept Jesus Christ as their savior?	Democrat (or independent close to Democrat)	Mean		.30
		95% Confidence Interval for Mean	Lower Bound	.27
			Upper Bound	.32
	Independent/other party	Mean		.35
		95% Confidence Interval for Mean	Lower Bound	.32
			Upper Bound	.38
	Republican (or independent close to Republican)	Mean		.49
		95% Confidence Interval for Mean	Lower Bound	.47
			Upper Bound	.52



This all comes from the same data, but what do you notice when you compare the graphs? What stories do they tell? Do they tell the same story? Are there differences in what is or is not substantively different, or the extent to which it appears different? If you were to communicate the story of this data, which graph would you choose?

Caution

The graph above exaggerates differences, because the y-axis scale goes from 23% to 62%. While zooming in allows you to see differences more closely, it also makes them appear more pronounced than they actually are. Here's the same data, graphed with a y-axis scale from 0% to 100%.



On the one hand, we don't want to exaggerate differences. Are Independent close to Republican and not very strong Republicans really that different from strong Republicans? Their sample means are 11% apart, and at the 95% confidence level they could be 19% apart, but also at the 95% confidence level, they could be only 3% apart (49% and 52%), with about half of people in each group having engaged in Christian proselytization. If these were the numbers, for every 100 people, you would have 49 who did proselytize in the both groups, 48 who did not in both groups, and then the difference would be among the final 3 out of 100 people. However, you also do not want to dismiss these differences. When we are dealing with sample proportions and means, differences can appear smaller than they are. 3% of strong Republicans constitute over 1 million U.S. adults and 3% of independents close to Republican or not very strong Republicans constitute over 1½ million adults. That is a lot more/fewer people who have/have not proselytized.

This page titled 6: Comparing Confidence Intervals is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Ezra Temko (Consortium of Academic and Research Libraries in Illinois (CARLI)) .

7: Hypothesis Tests, T-test, & ANOVA

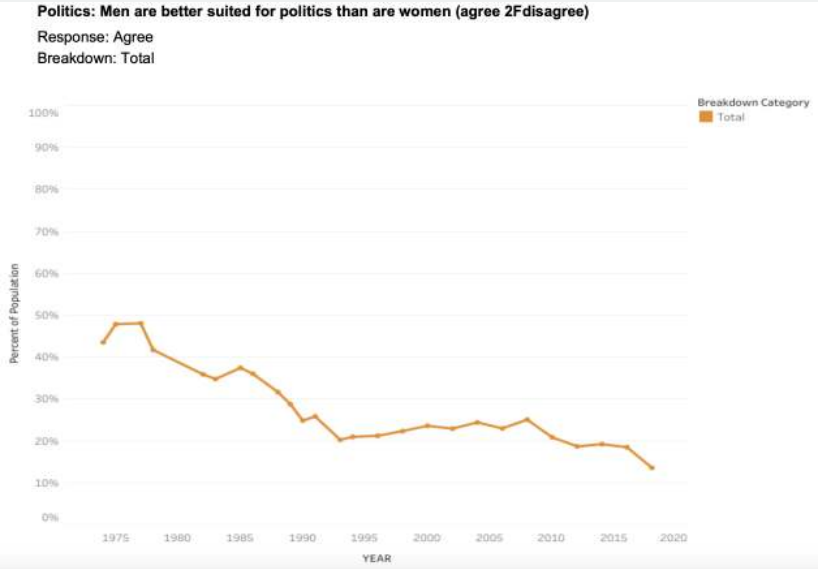
In the last chapter, we compared confidence intervals. This chapter will also involve inferential statistics that compare claims of population means, but through a slightly different approach: hypothesis testing. After reviewing one-sample t-tests, which are most similar to the logic of confidence intervals, we'll explore two-sample t-tests and one-way ANOVA. For these latter two tests, we'll also explore criteria for causality claims. These tests are about comparing means, so they will only work when your primary variable of interest is either ratio-like or an indicator variable.

1. Using a sample mean to judge the legitimacy of claims about a population: One-sample t-test

Statistics in Action: Gender Bias & Political Candidate Evaluation

The General Social Survey, which is representative of U.S. adults, frequently includes the following survey question: "Tell me if you agree or disagree with this statement: Most men are better suited emotionally for politics than are most women."

Here are the historical results of the percentage that agreed:



Most recently, in 2022, 15.23% of respondents said they agree with the statement.

Less than one decade ago, this proportion was closer to 20%, or 1 in 5 respondents. Could it still be that 20% of U.S.-American adults would agree that women are generally less suitably emotionally for politics than men?

In Chapter 6, to answer this we would construct a 95% confidence interval to see if it does or does not overlap with 20%. If the confidence interval's upper bound was 20% or greater, that would mean that if the population proportion were really 20%, there is less than a 5% probability I would randomly draw this sample (given the sample proportion of 15.23% among 1,142 respondents).

Here is the confidence interval output from SPSS:

Explore

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Most men are better suited emotionally for politics than are most women.	1326	37.4%	2218	62.6%	3544	100.0%

Descriptives

			Statistic	Std. Error
Most men are better suited emotionally for politics than are most women.	Mean		.1523	.00987
	95% Confidence Interval for Mean	Lower Bound	.1330	
		Upper Bound	.1717	

We are 95% confident that, as of 2022, somewhere between 13.3% and 17.2% of U.S. adults would agree that most men are better suited emotionally for politics than most women. This means we are also 95% confident that less than 20% of U.S. adults would agree.

While SPSS gives us a confidence interval, remember that behind the scenes it had to calculate the confidence interval bounds.

Here was our confidence interval formula:
T-score x Standard Error = Distance (or margin of error)

The t-score here assumes a 95% confidence level, which corresponds with that less than 5% probability that the population proportion is actually 20% (or any value outside the CI). This reflects back to sampling distributions and a normal t-distribution / bell curve, with the 5% on the tail ends. With a sample size over 1,000, our t-score for a 95% confidence level is 1.96.

SPSS gave us a standard error of 0.987%. We could calculate this as well, using the standard error formula for proportions or means: $\sqrt{\frac{P(1-P)}{n}} = \sqrt{\frac{0.1523(1-0.1523)}{1142}} = 0.0106$ $\frac{s}{\sqrt{n}} = \frac{0.3595}{\sqrt{1142}} = 0.0106$

Both give me 0.987% (like SPSS) if I use the weighted sample size instead.

T-score x Standard Error = Distance (or Margin of Error)
1.96 x 0.00987 = 0.0193452 or 1.93%
15.23% + or - 1.93
CI: 13.30% to 17.16%

Okay, so we're over 95% confident that less than 20% of U.S. adults feel women are generally less suitably emotionally for politics than men.

Are we over 99% confident?

I could answer this again by making another confidence interval:

Explore

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Most men are better suited emotionally for politics than are most women.	1326	37.4%	2218	62.6%	3544	100.0%

Descriptives

			Statistic	Std. Error
Most men are better suited emotionally for politics than are most women.	Mean		.1523	.00987
	99% Confidence Interval for Mean	Lower Bound	.1269	
		Upper Bound	.1778	

We are 99% confident that, as of 2022, somewhere between 12.7% and 17.8% of U.S. adults would agree that most men are better suited emotionally for politics than most women. This means we are also 99% confident that less than 20% of U.S. adults would agree.

Are we over 99.9% confident?

I could continue to test out different confidence intervals to find out how confident I can be. But there's another way to do this ---- called a **one-sample t-test** --- to find out how confident I can be that it's not actually 20% of U.S. adults that would agree that most men are better suited emotionally for politics than most women.

Definition: One-sample t-test

One-sample t-test: A statistical test used to evaluate claims about whether a population mean is different from a particular hypothesized value.

Do you remember fact families? For example, 2 groups of 3 will give me 6, the same product as 3 groups of 2. 6 divided into 2 groups gives me 3, and 6 divided into 2 groups will give me 3.

Fact Family	
2x3=6	6÷3=2
3x2=6	6÷2=3

We can do the same thing with the confidence interval formula:
T-score x Standard Error = Distance
Distance / Standard Error = T-score

When we take the distance between two values, and divide them up by the standard deviation of the normal distribution, we can determine how many standard deviations/errors away the two values are from one another (our t-score). From there, we

can figure out the probability and confidence level based on areas under the curve. For confidence intervals, we had already put in a particular p-value (e.g., p=0.05) and corresponding confidence level (e.g., 95%), which we used along with sample size to determine the t-score, and eventually the distance. But if we already know distance, we can instead solve for the t-score, and then find out a particular probability and confidence level.

For a one-sample t-test, we want to figure out the probability that we would pull this random sample if the population proportion is actually 20% (or if the population mean is actually 0.2). We start with our distance (in this case the distance of the sample mean, 15.23%, from our test value for the population mean, in this case 20%), and use that and the standard error to calculate the t-score, and then from the t-score we can determine the p-value.

Let's try it:
Could the population mean be 20%?
(If the population mean really were 20%, what is the probability that we would randomly pull this sample?)

Distance ÷ Standard Error = T-score
Distance = sample mean – test value = 15.23% – 20.00% = –4.77%
Standard Error = 0.00987 (calculated above)
Distance ÷ Standard Error = -4.77 ÷ 0.00987 = –483.28

The t-score, –483.28, represents how many standard errors 15.23% is away from 20.00%. We took the distance between the two means and made their units standard errors. This means that our sample mean is 483 standard errors less than / to the left of the test value of 20%.

You can use your t-score, with your sample size, to determine the p-value. Remember that different sample sizes determine what the normal t-distribution looks like and thus what the areas under the curve are. There are tables you can use to look up p-values based on a t-score and particular sample size, but instead we can just use an online calculator (such as this one), which will be easier and even more precise. If you use the linked online calculator, use the box "P from t", put the t-score in the box next to the t, and then where it says "DF," which stands for degrees of freedom, put in one less than the sample size. Click Compute P and we find that p<0.0001.

This means there is less than a 0.01% probability that we would randomly pull this sample if the population mean really where 20%.

To go from a p-value to a confidence level, calculate 1-p, or if you have converted p into a percentage already, 100%-p%. For example, if p=0.05, 1-0.05=0.95 or 100%-5%=95%.

Here, p<0.0001, so 1-0.0001=0.9999, or 99.99% (or p<0.01%, so 100% - 0.01% = 99.99%)
We are over 99.99% confident that the population mean is not 20%.

That matches --- here's a confidence interval at the 99.99% confidence level. You can see the CI does not overlap 0.2:

Descriptives

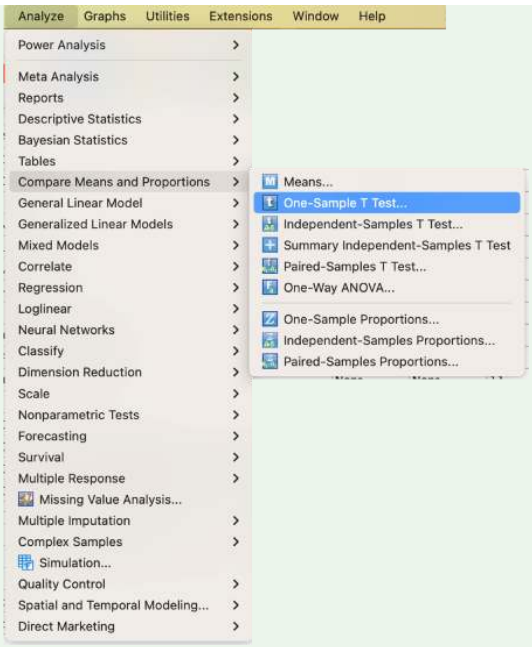
			Statistic	Std. Error
Most men are better suited emotionally for politics than are most women.	Mean		.1523	.00987
	99.99% Confidence Interval for Mean	Lower Bound	.1138	
		Upper Bound	.1909	

You probably don't want to have to do those calculations each time. Here's how to conduct a one-sample t-test in SPSS:

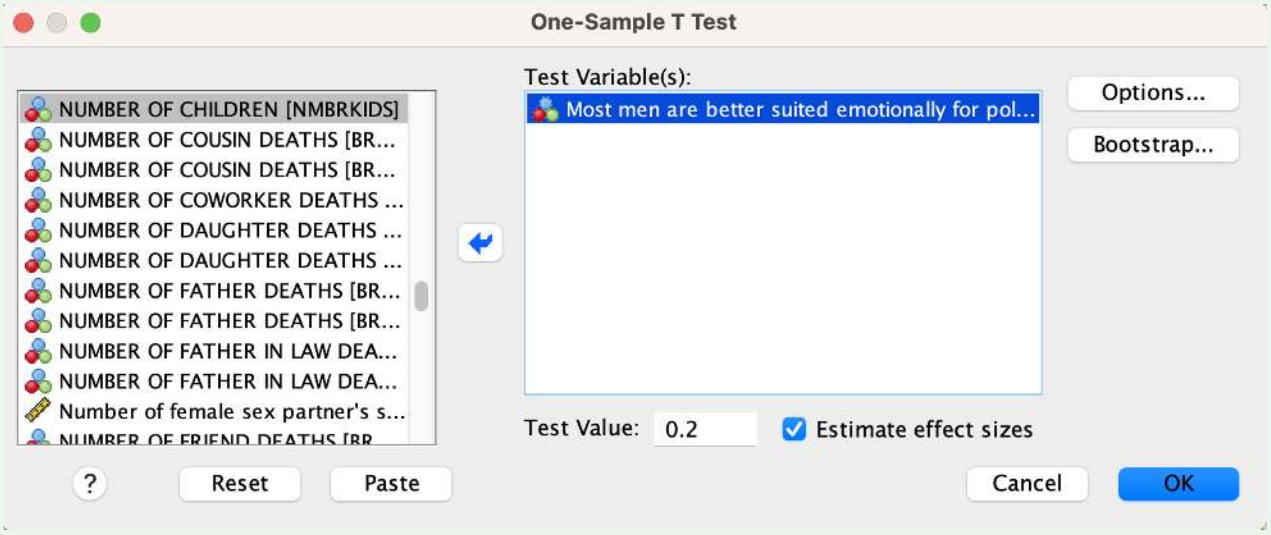
SPSS: One-sample t-test

How to:

Menu option:
Analyze → Compare Means and Proportions → One-Sample T-test



Select your variable. Write your hypothesized population mean in the test value box and click OK.



Syntax option:

T-TEST
/TESTVAL=TestValue#
/MISSING=ANALYSIS
/VARIABLES=VariableName
/CRITERIA=CI(.95).

Replace TestValue# with the test value (in this case 0.2) and VariableName with the name of your variable of interest.

Output:

Here is what your output will look like:

T-Test

One-Sample Statistics

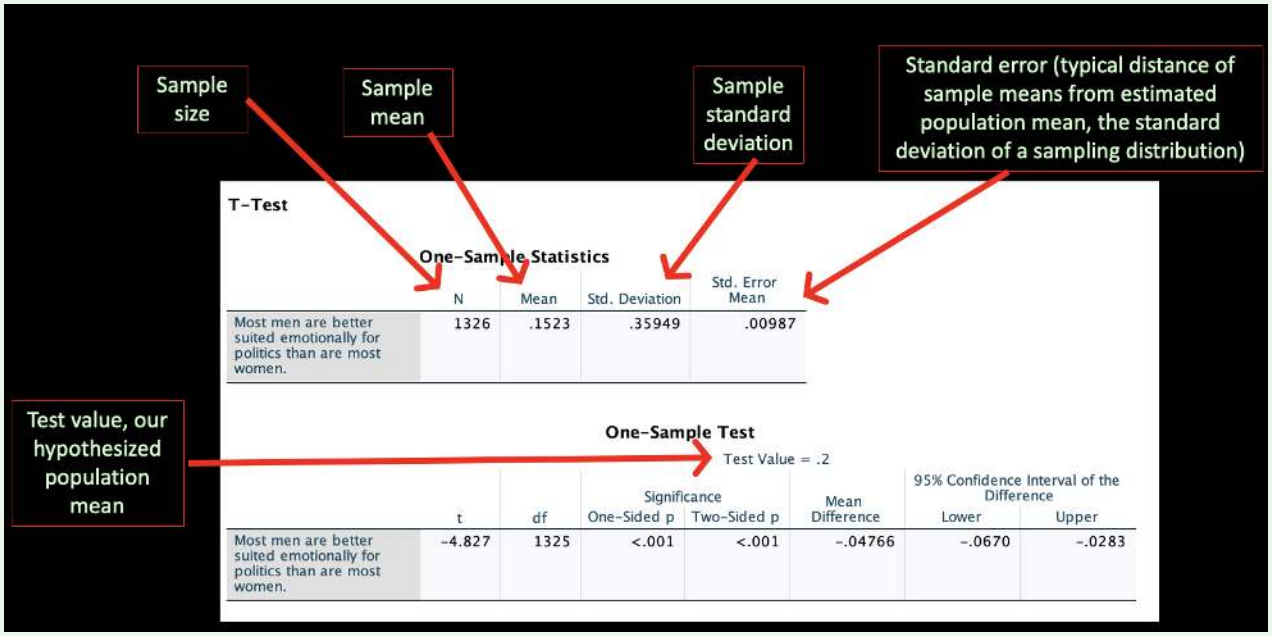
	N	Mean	Std. Deviation	Std. Error Mean
Most men are better suited emotionally for politics than are most women.	1326	.1523	.35949	.00987

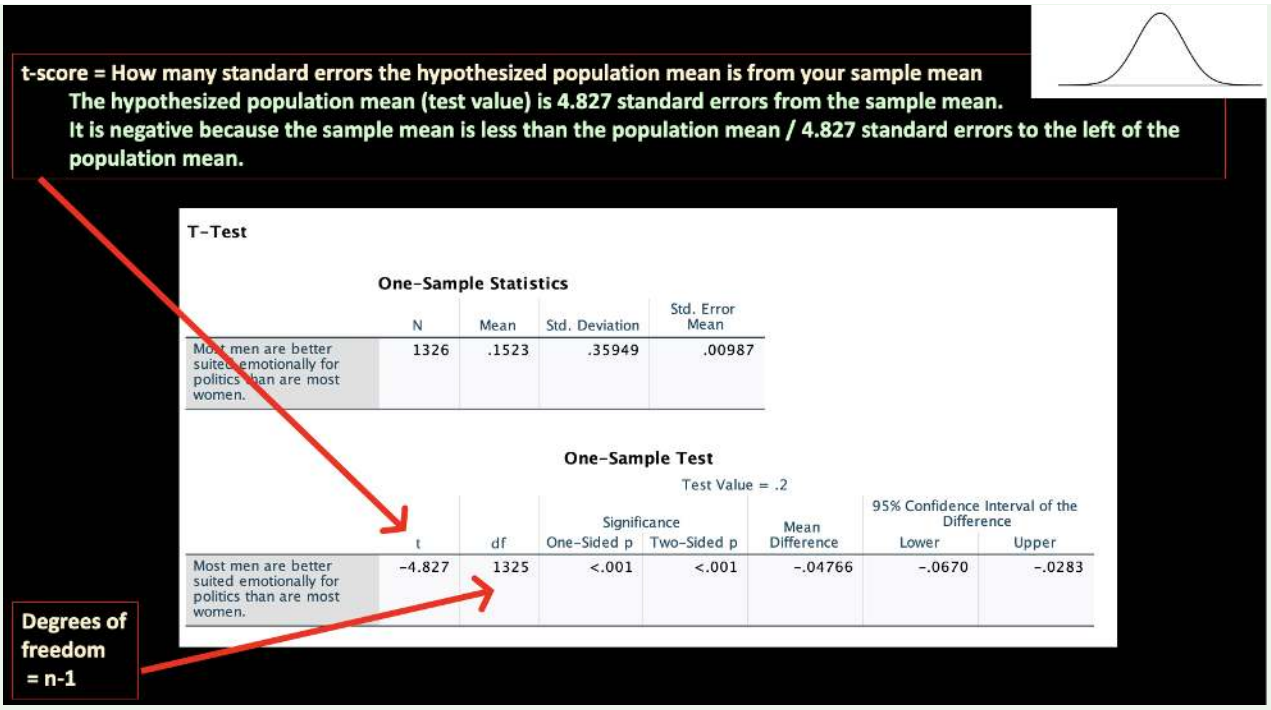
One-Sample Test

Test Value = .2

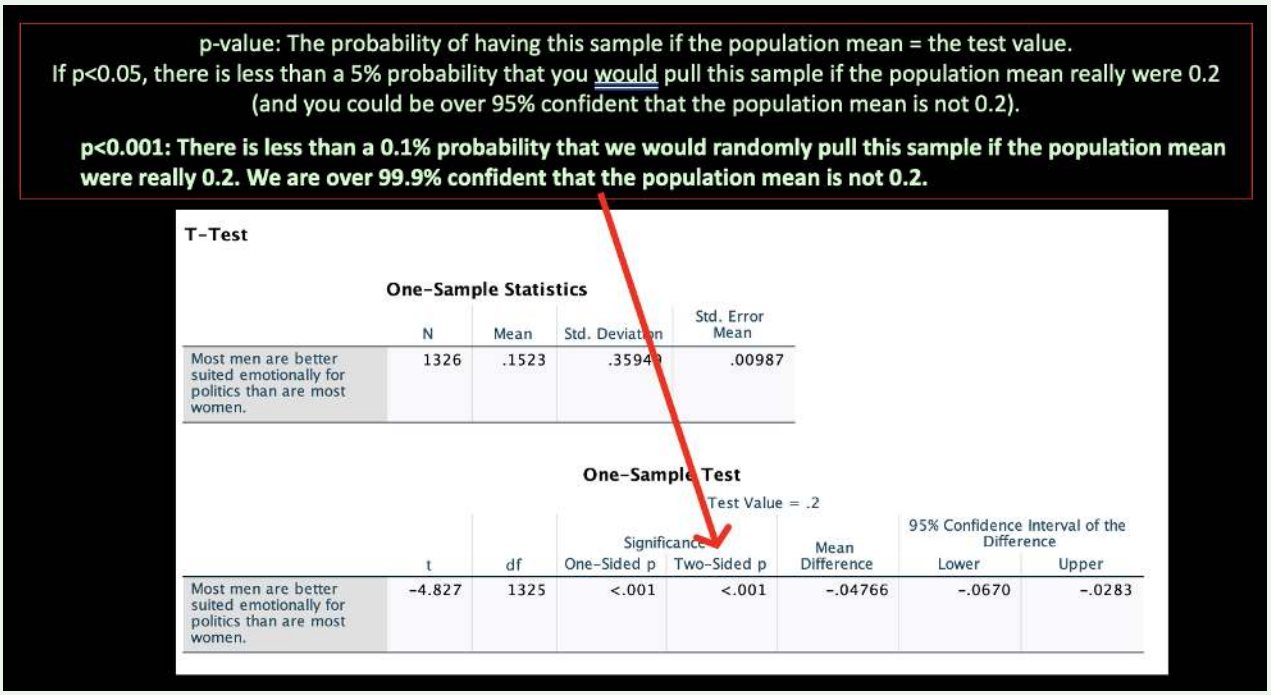
	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
Most men are better suited emotionally for politics than are most women.	-4.827	1325	<.001	<.001	-.04766	-.0670	-.0283

Here's how to interpret what's in the output:

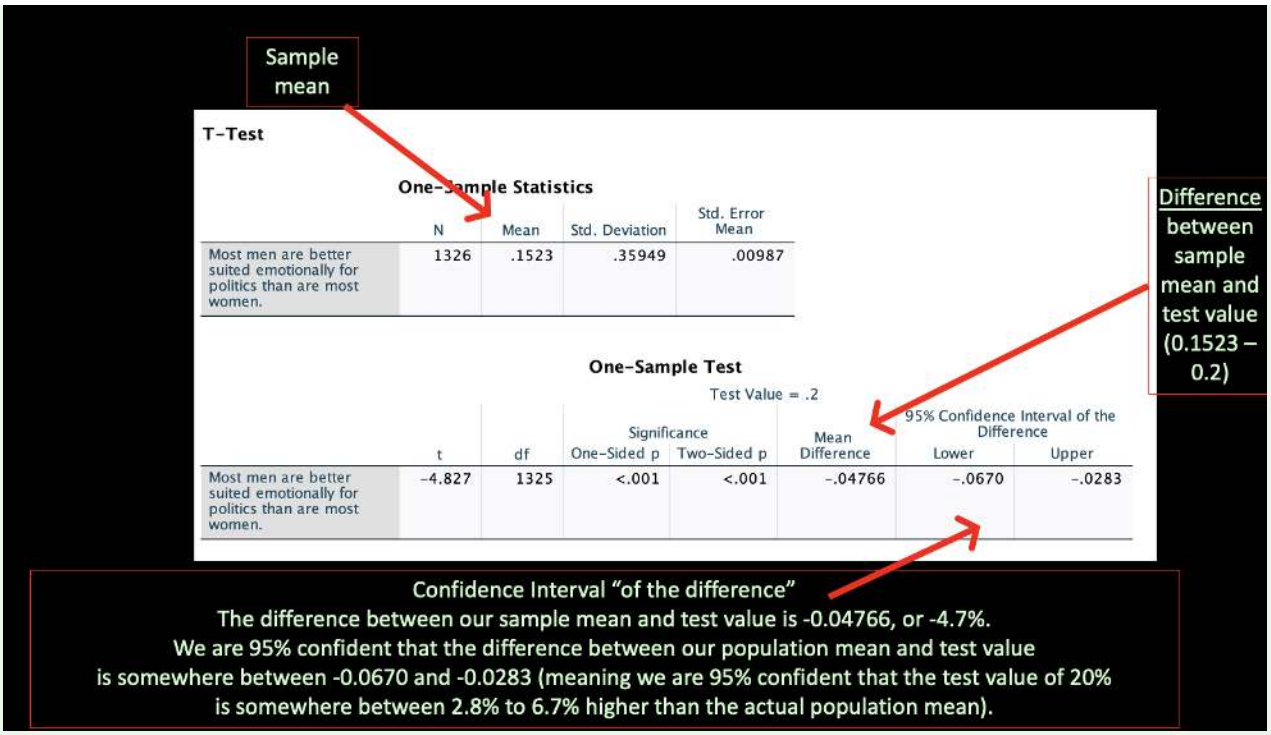




The more standard errors the sample mean and test population mean are from each other, the bigger the t-score.



Use the "Two-sided p" under Significance. The one-sided p-value is based on a one-tailed test, which assumes a particular direction. If you think about the sampling distribution, the two-sided p reflects the area under the curve on both sides, at both tails.



Don't get confused between "confidence intervals" like we covered in Chapters 5 and 6 and the "95% Confidence Interval of the Difference." Do you remember how we calculated how close and far apart two population means or proportions might be from one another based on confidence intervals? This is more similar to those numbers. It's a confidence interval for how different the sample mean is from the test value. We know that the actual difference between our sample mean of 0.1523 and our test value of 0.2 is -0.04766. But due to sampling error, that might not be the actual difference among our target population. We are building a confidence interval around the difference between the two. If this confidence interval overlaps 0, then we know the population mean could be the test value (that they might be the same, at the 95% confidence level).

Hypothesis Testing

Definition: Hypothesis test

Hypothesis test (also known as "statistical test"): A statistical procedure that tests an inferential claim about a population using representative sample data.

We conduct “statistical tests” to test claims about a population based on sample data.

Hypothesis tests always start with a null and alternative hypothesis.

A hypothesis is a claim about a population based on statistics from a sample of this population.

You can have your own hypotheses, but for statistical tests we will always have the same two:

$H_{\varnothing}$: Null hypothesis
The null hypothesis is always that there is no difference or no relationship (among your target population).

H_a: Alternative hypothesis
The alternative hypothesis is always that there is a difference or a relationship (among your target population).

Definitions: Hypothesis, null and alternative

Hypothesis: An inferential claim about a population.

- **Null hypothesis ($H_{\varnothing}$):** There is no difference/relationship (among the target population).
- **Alternative hypothesis (H_a):** There is a difference/relationship (among the target population).

Your statistical test will generate a "test statistic" (for one-sample t-tests, a t-value), which can be used to determine a p-value.

Conventionally, if $p<0.05$, you reject your null and accept your alternative hypothesis. This is a statistically "significant" difference or relationship. This means there is less than a 5% probability you would pull your sample if the null hypothesis were true.

Caution

Using $p<0.05$ as a cut-off point is arbitrary --- it's just based on convention. Why $p<0.05$ rather than $p<0.01$ or $p<0.10$, or something else? Does it really make sense to reject your null hypothesis if $p=0.049$ but fail to reject it if $p=0.051$? There are also different real-life situations where it might make more sense to lean towards having Type 1 errors (false positives, rejecting your null hypothesis when actually there is no difference/relationship) or Type 2 errors (false negatives, failing to reject your null hypothesis when actually there is a difference/relationship). Remember also that the probability-value is influenced by things like sample size. A p-value of <0.05 also does not tell you that a relationship matters or is important. **A p-value is not a substitute for having to make judgments.**

To this end, going forward I will be reporting out $p<0.05$ as significant, but also reporting out $p<0.10$ as marginally significant and $p<0.15$ as approaching significance, and using the following key to reflect various p-values:
^ $p<0.15$ ~ $p<0.1$ * $p<0.05$ ** $p<0.01$ *** $p<0.001$

Statistics in Action: Gender Bias & Political Candidate Evaluation, continued

Let's do a one-sample t-test, set up as a hypothesis test.

We are confident that it's not 20% (1/5) of U.S. adults who agree that "most men are better suited emotionally for politics than are most women." But could it be that 1/6 (16.67%) of U.S. adults agree? We already know that our 95% confidence interval overlapped 1/6, so we know $p>0.05$. We are not over 95% confident that the population mean is not 1/6. Let's find out what the probability is that we would pull this sample if the population proportion were actually 1/6.

Hypothesis Test

$H_{\varnothing}$: Null hypothesis: $\mu=16.67\%$ (The population mean is 1/6)
(Our sample proportion was 15.23%. We're actually saying here that the sample mean of 15.23% could equal a population mean of 16.67%, that there's no significant difference.)

H_a: Alternative hypothesis: $\mu\neq16.67\%$ (The population mean is not 1/6)
(Our sample proportion was 15.23%. We're actually saying here that the sample mean of 15.23% does not equal a population mean of 16.67%, that there is a significant difference.)

One-sample t-test:

One-Sample Test

Test Value = .166667

	t	df	Significance One-Sided p	Significance Two-Sided p	Mean Difference	95% Confidence Interval of the Difference Lower	95% Confidence Interval of the Difference Upper
Most men are better suited emotionally for politics than are most women.	-1.451	1325	.074	.147	-.01432	-.0337	.0050

One-Sample Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval Lower	95% Confidence Interval Upper
Most men are better suited emotionally for politics than are most women.	Cohen's d	.35949	-.040	-.094	.014
	Hedges' correction	.35969	-.040	-.094	.014

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation.
Hedges' correction uses the sample standard deviation, plus a correction factor.

Test Statistic: t=-1.451 (Weighted) Sample size=1,326

P-value: 0.147 **Confidence Level:** 1-0.147=100%-14.7%=85.3% (or 1-0.147=0.853=85.3%)

Determination: Not significant (approaching significance)

Because p is not <0.05, we cannot reject our null or accept our alternative hypothesis. We are not over 95% confident that the population proportion is not 1/6. It could be. There is a 14.7% probability we would pull our sample if the population proportion were 1/6.

Notice how the 95% confidence interval of the difference overlaps 0. This means that we cannot be 95% confident that the population proportion is not the test value.

While not part of hypothesis tests, the next question we ask is whether there is **substantive significance**. Can we be over 95% confident that there is a substantive or meaningful difference? Recall that we are 95% confident that, as of 2022, somewhere between 13.3% and 17.2% of U.S. adults would agree that most men are better suited emotionally for politics than most women. That's not substantively different from 16.67%, as it might be 16.67%. You could argue either way about whether it's substantively different from 20%. On the one hand, at the 95% confidence level, we're only confident that there's at least a 2.8% difference. On the other hand, that difference comes out to be over 9 million people, and elections can be pretty close.

You will likely not come across one-sample t-tests in sociology scholarship. One practical application for them would be if you have a variable about change that occurred and you want to compare that mean to a test value of zero to evaluate whether you can be confident there was any change. Regardless, let's move on to a statistical test you will see more often in the literature, the two-sample t-test.

2. Comparing two population means: two-sample t-test

While the one-sample t-test was more similar to, in Chapter 5, when we constructed one confidence interval and saw what values were outside or inside its bounds, a two-sample t-test is somewhat similar to, in Chapter 6, when we compared two confidence intervals to see whether or not they overlapped. Just like with the one-sample t-test, instead of starting with a confidence level and determining the distance, we'll start with a distance and determine the confidence level.

Definition: Two-sample t-test

Two-sample t-test: A statistical test used to evaluate claims about whether two population means are different.

One-sample vs. two-sample t-tests

Remember that our formula for a one-sample t-test was:
t-score* = *distance* ÷ *standard error

It's the same for a two-sample t-test.

A t-score represents how many standard errors one population's mean is from the other. We take the distance and make its units standard errors, as reflected in this formula:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{standard\ error}$$

For a one-sample t-test:
Distance = sample mean – hypothesized population mean (test value)

$$t = \frac{distance\ (between\ your\ sample\ mean\ and\ hypothesized\ population\ mean)}{standard\ error\ (from\ your\ sample)}$$

For a two-sample t-test:
Distance = first sample mean – second sample mean

$$t = \frac{distance\ (between\ your\ first\ sample\ mean\ and\ second\ sample\ mean)}{standard\ errors\ (from\ both\ samples)}$$

Rather than comparing a sample mean to a hypothesized population mean, we are comparing two sample means.
Rather than making an inference about the sample mean in terms of what population mean it could be and comparing that to a test value, we are making inferences about our sample means and comparing the two estimated population means.

Standard error has a similar formula, but it has to be modified because we now have two different samples, each with their own mean and standard error.

One-sample t-test:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\frac{s}{\sqrt{n}}}$$

Two-sample t-test:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

For the two-sample test, n_1 is the sample size for the first group/subsample, and n_2 is the sample size for the second group/subsample.

The further apart the two sample means are, the bigger the t-score.

The smaller the standard error, the bigger the t-score.

Hypothesis testing: Two-sample t-test

$H_\varnothing$: Null hypothesis: $\mu_1=\mu_2$ The population means are equal.
The null hypothesis is always that there is no difference or no relationship (among your target population).

H_a: Alternative hypothesis: $\mu_1\neq\mu_2$ The population means are not equal.
The alternative hypothesis is always that there is a difference or a relationship (among your target population).

Your statistical test will generate a "test statistic" (for two-sample t-tests, a t-value), which can be used to determine a p-value.

Conventionally, if $p<0.05$, you reject your null and accept your alternative hypothesis. This is a statistically "significant" difference or relationship. This means there is less than a 5% probability you would pull your sample if the null hypothesis were true.

Statistics in Action: Gender pay gaps

In Chapter 3 we looked at gender and race pay gaps. However, we were only looking at sample data. In this example we'll return to these pay gaps but evaluate them using statistical tests. Again, we're using 2017 to 2021 American Community Survey data. Our target population below will be full-time, year-round, non-institutionalized U.S. workers ages 16+.

A two-sample t-test only works for comparing two groups. In this case, we'll compare men and women's mean earnings.

$H_\varnothing$: Null hypothesis: $\mu_M=\mu_W$ On average, men and women have the same incomes.
(There is no relationship between gender and income.)

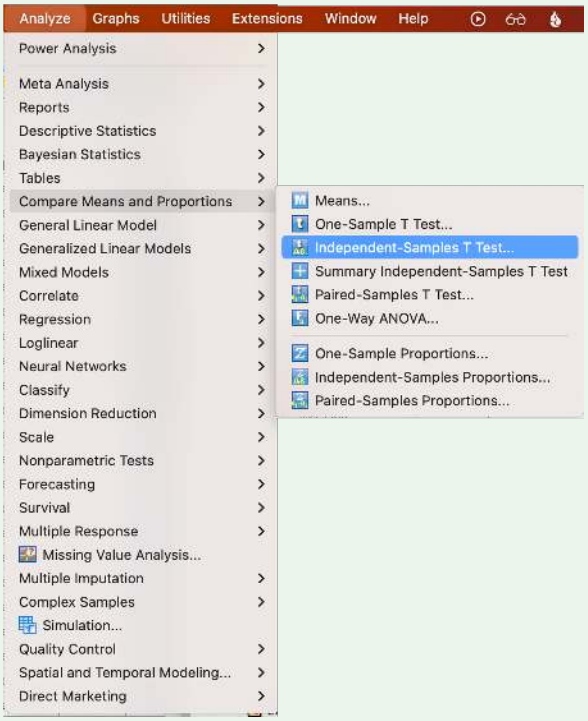
H_a: Alternative hypothesis: $\mu_M\neq\mu_W$ On average, men and women have different incomes.
(There is a relationship between gender and income)

You might have your own hypothesis as well (e.g., men have higher average incomes than women), but our hypothesis test will only be evaluating the above.

SPSS: Two-sample t-test

How to:

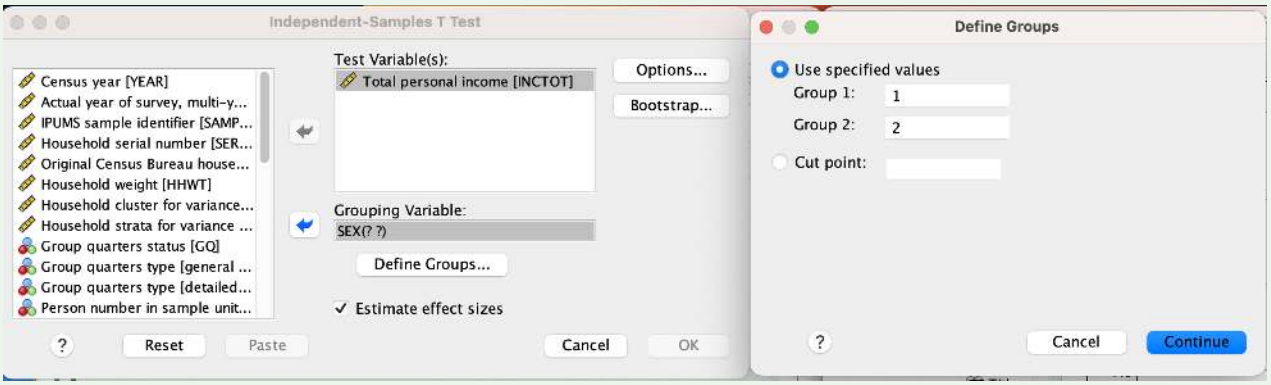
Menu option:
Analyze → Compare Means and Proportions → Independent-Samples T Test



Select your test variable (the one you are interested in the means of, here the income variable).

Select your grouping variable (here gender).
Click "Define Groups" and then use the codes for the two groups you want to compare the means for. You have to look up the coding for the groups you have. Here it was 1 and 2. It could have been 0 and 8 or any other two numbers. You can do two-sample t-tests with a categorical variable that has more than two categories, as long as you only select two for

comparison. After putting in your specified values, Click continue and then click OK.



Syntax option:
T-TEST GROUPS=GroupVariableName(# #)
/MISSING=ANALYSIS
/VARIABLES=TestVariableName
/CRITERIA=CI(.95).

Replace TestVariableName with the name of the variable you are interested in the means of (in this example, the income variable).
Replace GroupVariableName with the name of your grouping variable (in this example, the gender variable).
Replace the #s with the codes for the categories you are using.

My syntax was:
T-TEST GROUPS=Sex(1 2)
/MISSING=ANALYSIS
/VARIABLES=IncTot
/CRITERIA=CI(.95).

Output:

T-Test

Group Statistics

	Sex	N	Mean	Std. Deviation	Std. Error Mean
Total personal income	Male	64226720	84676.98	92459.765	11.537
	Female	48272914	62752.25	60782.868	8.748

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means							
						Significance				95% Confidence Interval of the Difference	
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
Total personal income	Equal variances assumed	2194787.322	<.001	1431.379	112499632	<.001	<.001	21924.729	15.317	21894.707	21954.750
	Equal variances not assumed			1514.254	110648749	<.001	<.001	21924.729	14.479	21896.350	21953.107

The descriptive statistics at the top should look familiar to you.

Now to the two-sample t-test output.

You have two different rows, but don't let that confuse you. In order for statistical tests to work, there are a lot of assumptions that need to be met. For the two sample t-test, you either need samples of similar sizes or for the two samples to have equal variances. If you have similar sample sizes for the two groups or if the "Sig." p-value in the first row under "Levene's Test for Equality of Variances" is 0.05 or greater, you can use the first row, the pooled two-sample t-test. If you do not have similar sample sizes for the two groups and the Sig. p-value under the Levene's test is <0.05, we need to use

the second row, the unpooled two-sample t-test. The second row is more conservative as it is, and so if you'd prefer to just use that row to be on the safe side that is okay as well. The most important thing is to not confuse the Sig. in the second column, which is Levene's test, as being a p-value for the t-test, or to try to compare the first and second rows instead of just focusing on the row you are interpreting. If it's helpful, you can delete the first row (double-click on the output, highlight the first row, and right click and click delete).

Let's take a look at the second row of the t-test. Here is what that output looks like if I delete the first row:

Independent Samples Test

		t-test for Equality of Means							
				Significance				95% Confidence Interval of the Difference	
		t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
Total personal income	Equal variances not assumed	1514.254	110648749	<.001	<.001	21924.729	14.479	21896.350	21953.107

You can see a t-score (1,514.254) and a p-value (use the "Two-sided p") of <0.001.

Is there a significant difference?
Because p<0.05, we can reject our null and accept our alternative hypothesis. There is less than a 0.1% probability that we would have pulled this sample if, between 2017 and 2021, men and women (who are full-time, year-round, non-institutionalized U.S. workers ages 16+) had the same mean earnings.

p<0.001 1-0.001=0.999=99.9% OR p<0.1% 100%-0.1%=99.9%
We are over 99.9% confident that men and women have different mean earnings. They are significantly different.
(Note: I said over because p<0.001. If p=0.001 I would have said "We are 99.9% confident.")

All the way to the right you will notice a "95% Confidence Interval of the Difference." If this overlapped 0, we would know that men and women's mean earnings could be the same at the 95% confidence level.

Is it substantive?
Continuing to look at this output, the "mean difference" represents the first sample mean minus the second sample mean. Men's mean income was $84,676.98. Women's mean income was $62,752.25. The difference was $21,924.73. That's for the sample. Due to sampling error, we can only estimate how different men and women's mean incomes are. At the very right of the output we get a "95% Confidence Interval of the Difference." This is a confidence interval around the $21,924.73. We are 95% confident that, among our target population, men's mean income is somewhere between $21,953.11 and $21,896.35 more than women's mean income. Given that this is over $20,000, this definitely seems substantive.

Is it causal?
This is asking whether gender has an impact or effect on variation in income. I would argue it is causal. We have a significant and substantive relationship. The causal order makes sense: while gender might cause differences in pay, it would not make sense for how much one earns to change someone's gender. Gender is a relatively stable social location and patriarchy could impact the kinds of jobs and pay men vs. women earn. Theoretically, it makes sense that gender impacts mean earnings. There is a robust body of sociological scholarship explaining various reasons for the gender pay gap (e.g., the motherhood penalty and fatherhood bonus, pinkcollar jobs/workforce segregation and differential valuation, outright discrimination, etc.). There are no other preceding variables that would explain away this relationship.

Evaluating Causality

Causality / Causation: There is a causal relationship when one variable has an impact or effect on variation of another variable.

When we are evaluating relationships between variables, we label one variable the "independent variable" (IV) and one variable the "dependent variable" (DV).

Definition: Independent and Dependent Variables

Dependent Variable: The outcome; the variable potentially being impacted by the independent variable.

Independent Variable: The determinant; the variable potentially impacting the dependent variable.

IV → DV

In the gender pay gap example, gender is the independent variable and income is the dependent variable. Gender is having an effect on income variation.

To determine causality, we need to conduct a controlled study and have validity and reliability. We would conduct an experiment and hold everything else constant other than the one thing we change, and simultaneously not change that factor for another group so we can make a comparison. But... human beings are not "guinea pigs." And we can't hold conditions constant while we conduct an experiment on ourselves. Also, these types of controlled studies would cost a ton of money. So, we do our best to make causal claims using the data and theory available to us.

There are five criteria that need to be met to claim causality. One can demonstrate causation is highly likely, but in social sciences it cannot be “proven.”

1. **Significance**
2. **Substantiveness**
3. **Correct causal order**
4. **Theoretical sense/link/explanation**
5. **Nonspuriousness**

1. Relationship is significant

For a relationship to be causal, it must exist. Unless you are already looking at population data, this means that your inferential statistics test should show that the relationship is statistically significant. When we are testing for this relationship we are seeing whether the two variables co-vary, or vary together, also called covariance. This is often called association for qualitative measures and correlation for linear relationships.

2. Relationship is substantive

For us to claim that a relationship is causal, the relationship needs to be more than negligible. This may not be technically true --- if we had found that there was at least a $5 annual income mean difference between men and women, it could still be gender and patriarchy causing that difference --- but not enough for us to care. We want to look at relationships where the IV is having a substantial, meaningful impact.

3. Causal order

We might have a significant and substantive relationship (substantive significance), but before we can claim the IV is impacting the DV, we have to make sure we correctly identified which variable is the IV and which is the DV. Does X cause Y or does Y cause X? Think about time order and what types of constructs are more stable. There are certain relationships that have a particular causal order and cannot be the other way around. For example, while race (through racism) can impact variation in birth weight, birth weight would not impact someone's racial classification. Or while region could have an impact on college graduation rates, college graduation rates are not going to impact geographic region.

4. **Theory**

The relationship has to have a sensical explanation. How does it make theoretical sense? We have to be able to provide a rationale for how the IV is impacting the DV.

5. Nonspuriousness

The relationship cannot be spurious. A spurious relationship is when a variable that precedes the IV explains away the entire relationship. While you can test this with multivariate analyses, this relies on theory to figure out what variables to test for (and you

cannot test for it if you don't have the measures in your dataset). If a preceding variable only partially explains the relationship between the IV and DV, then the IV is still having its own causal impact, just a smaller one. We'll look at some examples of spurious relationships below, which will help make sense of this.

If there is a variable that comes *after* the first IV and before the second DV, it is okay for it to explain the relationship. This does not make the relationship spurious and it does not challenge the causality of the relationship. It is an explanation of how the IV impacts the DV, of its mechanism for doing so. For example, race impacts differentiation in political party registration. Obviously one's race cannot automatically cause you to register with one party or the other. We can explain the relationship (social, economic, and racial equity policies of the two parties; racial representation in the parties; union relationships with the parties; parent/family/community political party allegiances, etc.). These explanations are just explaining what makes the relationship causal, not negating its causality.

Caution

There are lots of examples in the news, in social media, in politics, etc., where people claim causality when they shouldn't. Just because a relationship exists does not mean it is a causal relationship.

Consider the following **existing** relationships. What might be going on in these situations other than a **causal** relationship? Try to brainstorm your thoughts on each, then continue to scroll for the answers.

1. As the number of fire trucks sent to a fire increases, the amount of damage from the fire also increases. Does sending more fire trucks to a fire cause more damage to the place on fire?

2. Children who get tutored get worse grades than children who do not get tutored. Do children getting tutored cause them to get worse grades than children who do not get tutored?

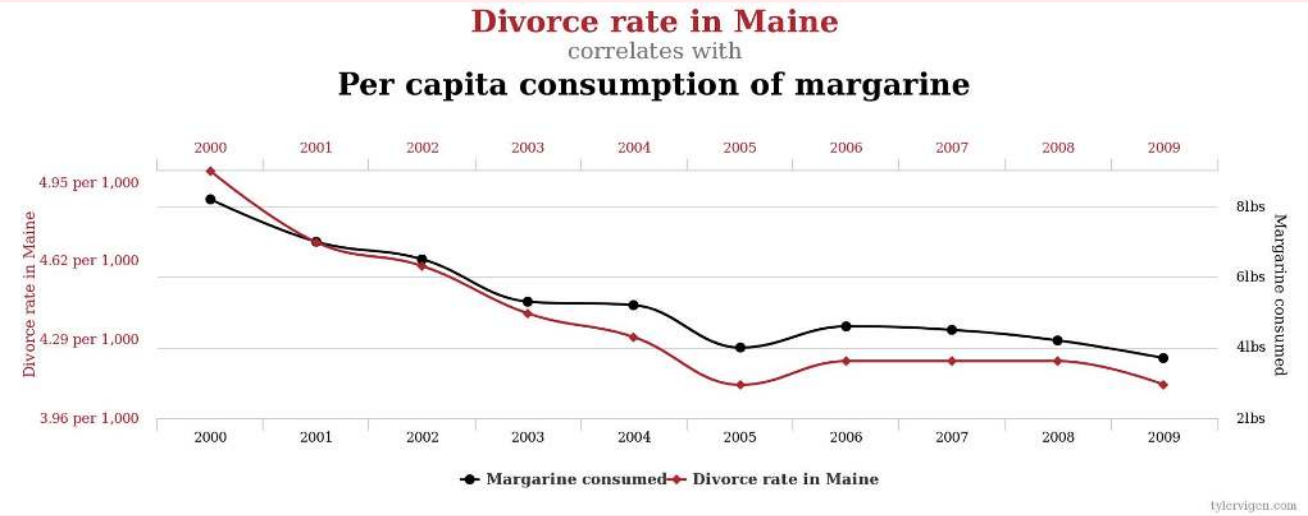
3. When there are spikes in ice cream sales, there are also spikes in children drowning. Does ice cream cause kids to drown?

4. The more police in a community, the more crimes in said community. Does having more police in a community cause people to commit more crimes in that community?

5. Kids who sleep with the light on are more likely to become short-sighted adults. Does sleeping with the light on as a young kid is will cause you to be short-sighted later in life?

6. On average, the bigger your palm, the shorter your lifespan. Does having a bigger palm cause you to have a shorter lifespan?

7. Do changes in the divorce rate in Maine cause U.S.-Americans to consume more/less margarine? They covary...



PAUSE HERE AND THINK THROUGH THE QUESTIONS ABOVE. WHEN YOU'RE READY, SCROLL DOWN FOR THE ANSWERS.

Answers

Q1. As the number of fire trucks sent to a fire increases, the amount of damage from the fire also increases. Does sending more fire trucks to a fire cause more damage to the place on fire?
A1: This is a spurious relationship. The preceding variable is the size of the fire. Bigger fires lead to both more fire trucks being sent out and to fires that have more damage.

Q2. Children who get tutored get worse grades than children who do not get tutored. Do children getting tutored cause them to get worse grades than children who do not get tutored?
A2: The causal order is wrong. Children who are getting worse grades are more likely to seek out tutoring.

Q3. When there are spikes in ice cream sales, there are also spikes in children drowning. Does ice cream cause kids to drown?
A3: This is spurious. The preceding variable is weather. Hotter weather = more swimming (and thus more drowning) and more ice cream. Colder weather = less of both.

Q4. The greater the number of police in a community, the greater the number of crimes in said community. Does having more police in a community cause people to commit more crimes in that community?
A4: This is a spurious relationship. It's all about the population size of the community. There are going to be more police and more crimes in a community that has a million people vs. a community with 300 people.

Q5. Kids who sleep with the light on are more likely to become short-sighted adults. Does sleeping with the light on as a young kid is will cause you to be short-sighted later in life?
A5: This is spurious. Being short-sighted is genetic! So parents who are short-sighted are more likely to put on a night light and more likely to have short-sighted kids. The parents being or not being short-sighted is the preceding variable that explains away the relationship.

Q6. On average, the bigger your palm, the shorter your lifespan. Does having a bigger palm cause you to have a shorter lifespan?
A6: This is spurious. The preceding variables are sex and gender. Females on average have smaller palms than males, and females and women on average have longer lifespans than males and men.

Q7. Do changes in the divorce rate in Maine cause U.S.-Americans to consume more/less margarine? They covary...
A7: I have no explanation for this one, so it fails to meet the theory criterion for causality. Strange things co-vary. If you look for patterns, you'll find them. You can find more of these at: http://www.tylervigen.com/spurious-correlations

Caution

There are usually multiple variables impacting a dependent variable. Just because other things also contribute, and maybe even more so than the independent variable you are looking at, does not mean you don't have a causal relationship. Causal just means the IV is impacting the DV, not that it is the only or main thing causing differentiation in the DV.

Example: Race, gender, and class all have causal impacts on who people vote for for president. There are many other things that influence who people vote for for president (e.g., political party, age, etc.). But none of these precede the other variables and explain away the entire relationship. While there are other determinants of who people vote, race, gender, and class still all have causal relationships with who people vote for.

3. Comparing more than two population means: one-way ANOVA

There are other types of t-tests besides one-sample and two-sample t-tests. For example, paired t-tests are used to compare before and after results. However, we're going to move to one final statistical test for this chapter, and it's not a t-test. Instead, let's turn to one-way ANOVA.

While a two-sample t-test allows us to evaluate, among our target population, whether the mean of a variable is different for two groups, what if you want to compare more than two groups?

There are various forms of ANOVA, but the one we're doing here, called a "one-way ANOVA," is for comparing means among more than two groups.

ANOVA stands for **Analysis of Variance**

Definition: One-way ANOVA

One-way ANOVA: A statistical test used to evaluate claims about whether three or more population means are different.

ANOVA is similar to a two-sample t-test in that we are comparing means and then dividing this by a measure of variation for our samples.

Two-sample t-test: **ANOVA:**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad F = \frac{BGSS/(\#of\ groups - 1)}{WGSS/(\#of\ cases - \#of\ groups)}$$

For ANOVA, we compare group means to the overall mean. If the group means are equal, they would each be the same as the population mean.

In the ANOVA equation above, BGSS stands for "Between Groups Sum of Squares." It is a measure of deviation between group means. The more deviation between group means, the bigger the BGSS, and the more likely we will find that not all groups have the same means.

To calculate BGSS:

- Calculate the overall mean for all data
- Calculate the mean for each group
- For each group, find the difference between the group mean and the overall mean, and square it.
- For each group, multiply this by the number of cases (we are treating each case as if it had the group mean; we'll account for within group variation later with WGSS)
- Add up these squared differences.

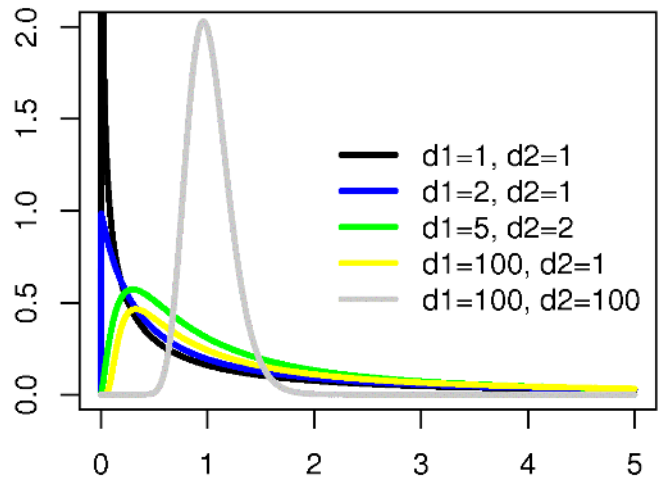
In the ANOVA equation above, WGSS stands for "Within Groups Sum of Squares." It is a measure of deviation within the groups. The more within group variation, the bigger the WGSS, and the less likely we will find that not all groups have the same means.

Calculating WGSS is very similar to calculating variance (which we do when we calculate standard deviation; variance is standard deviation squared). Here are the steps:

- For each group:
  - Calculate the mean.
  - Find the difference between each value and the mean.
  - Square the differences.
  - Add up the squared differences.
- Then add up each group's summed squared differences.

While confidence intervals and two-sample t-tests used normal t-distributions, ANOVA is different. It uses an F-distribution and the statistical test generates an F-statistic. F is the ratio of two variances (between group and within group). It is not a normal distribution (e.g., you can't have F=0 or F<0). While the t-distribution curves were based on sample size, the F-distribution is based both on sample size and on the number of groups you are comparing. However, it's otherwise the same idea. The area under the curve adds up to 1 (100%), the probability of observing any F value. The figure below shows various F-distribution curves. The legend shows "d" for degrees of freedom. These are based on number of groups and number of cases. Sometimes instead of seeing d1 and d2 or df1 and df2, you will see them referred to as the degrees of freedom for the numerator and denominator, because they are the formulas in parentheses in the F-statistic formula (df 1 = # of groups - 1; df2 = # of cases - # of groups).

$$F = \frac{BGSS/(\#of\ groups - 1)}{WGSS/(\#of\ cases - \#of\ groups)}$$



• Once we have our F-value, we are going to have to use both the number of cases and number of groups to determine the right F-distribution. You can use the same webpage as for two-sample t-tests, but would fill out the P from F information. There it says DFn and DFd to refer to degrees of freedom from the numerator and degrees of freedom from the denominator.

Hypothesis testing: one-way ANOVA

$H_\varnothing$: Null hypothesis: $\mu_1 = \mu_2 = \mu 3 \ldots$ All the population means are equal
The null hypothesis is always that there is no difference or no relationship (among your target population).

H_a: Alternative hypothesis: The population means are not all equal. At least one population mean is different from at least one other population mean.
The alternative hypothesis is always that there is a difference or a relationship (among your target population).

Your statistical test will generate a "test statistic" (for one-way ANOVA, an F-statistic), which can be used to determine a p-value.

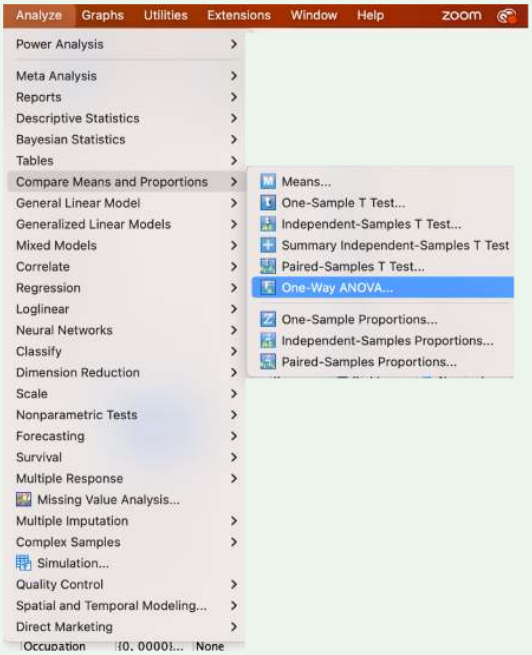
Conventionally, if $p<0.05$, you reject your null and accept your alternative hypothesis. This is a statistically "significant" difference or relationship. This means there is less than a 5% probability you would pull your sample if the null hypothesis were true.

If you reject your null hypothesis, you still do not know how many groups have different means, which groups are different, or the strength or direction of the difference.

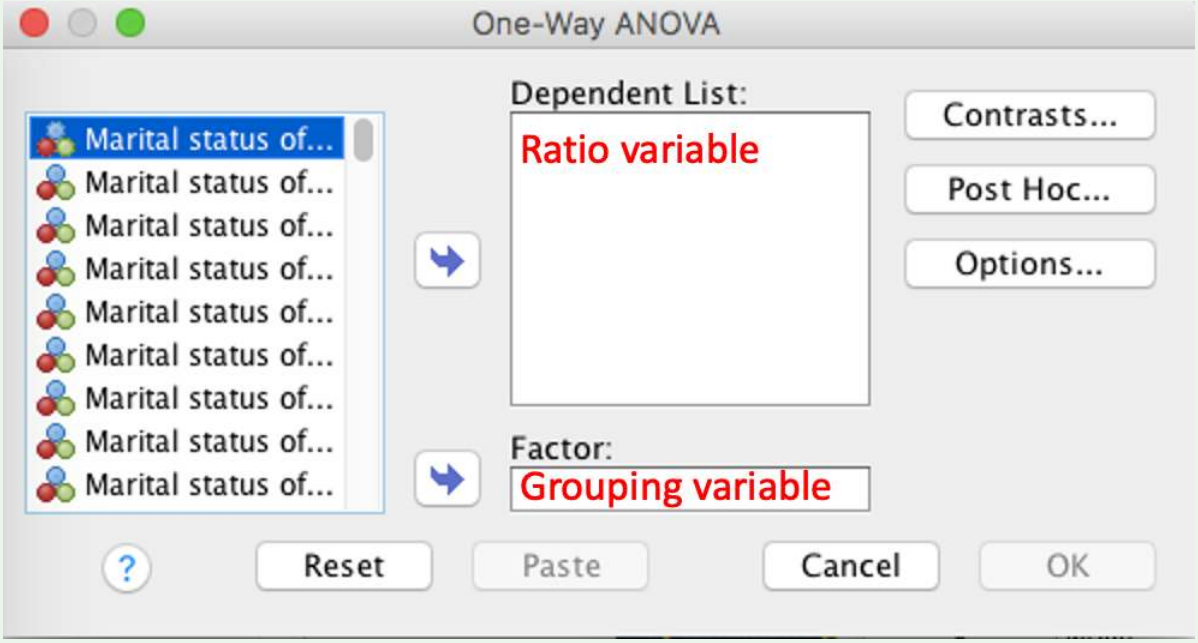
Using SPSS: one-way ANOVA

How to:

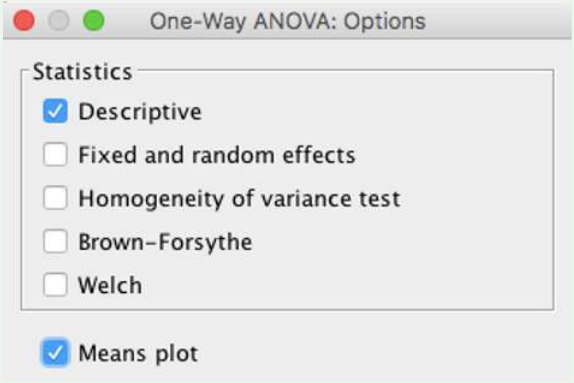
Menu option:
Analyze → Compare Means and Proportions → One-way ANOVA



Put the variable you want means for in the dependent list box and the variable with your group categories in the factor box.



Click on Options and select Descriptive and Means plot.



Click Continue and then OK.

Syntax option:

```
ONEWAY RatioVariableName BY GroupingVariableName
/STATISTICS DESCRIPTIVES
/PLOT MEANS
/MISSING ANALYSIS.
```

Caution

ANOVA will compare every group/category in your grouping variable. If you want to only compare some groups, collapse groups together, or otherwise change your categories, you need to create a new variable with the groups/categories you want and use that for your grouping variable.

Statistics in Action: Race pay gaps

Let's return to our 2017 to 2021 American Community Survey data. Our target population below will be full-time, year-round, non-institutionalized U.S. workers ages 16+. We're using ANOVA to compare average income by racial group because we have more than two racial groups. A two-sample t-test only works for comparing two groups. Because we want to compare more than two means, we're using ANOVA.

$H_{\varnothing}$: Null hypothesis: $\mu_W=\mu_B=\mu_A=\mu_O=\mu_H$
On average, NH (non-Hispanic) White, NH Black, NH API (Asian & Pacific Islander), NH Other/Multiracial, and Hispanic U.S. workers all have the same incomes.
(There is no relationship between race and income.)

H_a: Alternative hypothesis: Not all five racial groups have the same mean income. At least one racial group has a different mean income than at least one other racial group.
(There is a relationship between race and income)

SPSS Output:

ANOVA

Total personal income

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2.227E+16	4	5.567E+15	871863.591	<.001
Within Groups	7.184E+17	112499629	6385680919		
Total	7.407E+17	112499633			

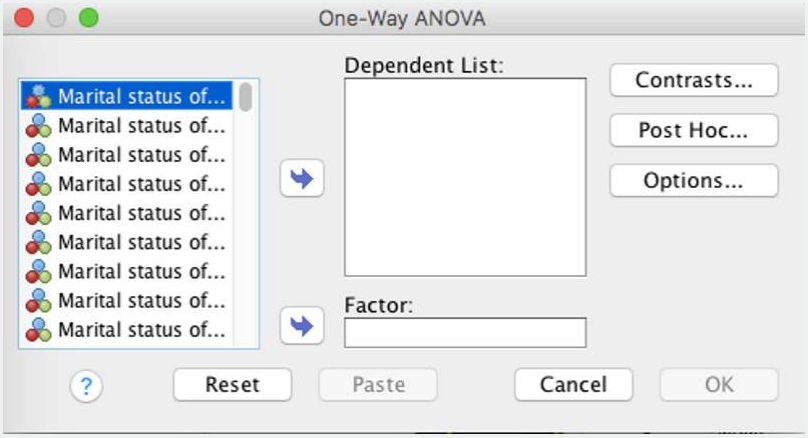
This is the ANOVA part. The E+15, E+16, and E+17 are for scientific notation (e.g., 2.227E+16 is for 2.227×10^{16}). Because these are sums of squares, they tend to be big numbers. Our F-statistic is 871,863.59. Based on the F-statistic, the number of groups, and the sample size, our p-value is <0.001.

1 - 0.001 = 0.999 = 99.9% *or* 100% - 0.1% = 99.9%

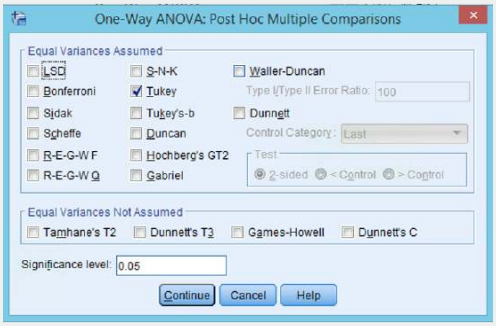
That means we can be over 99.9% confident that we can reject our null and accept our alternative hypothesis. We are over 99% confident that not all racial groups have the same mean income.

But that doesn't tell us much!

There are "post hoc" tests that can be used to test for which groups have significantly different means from which groups, with a lot of options. Tukey is the most common. You could click on the post hoc button in the One-Way ANOVA box:



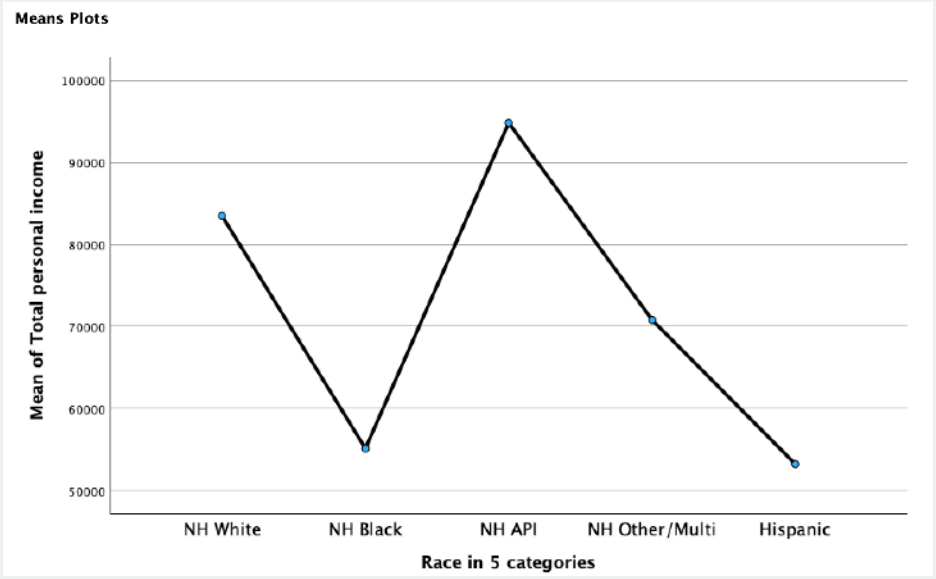
and then select the method you are going to use:



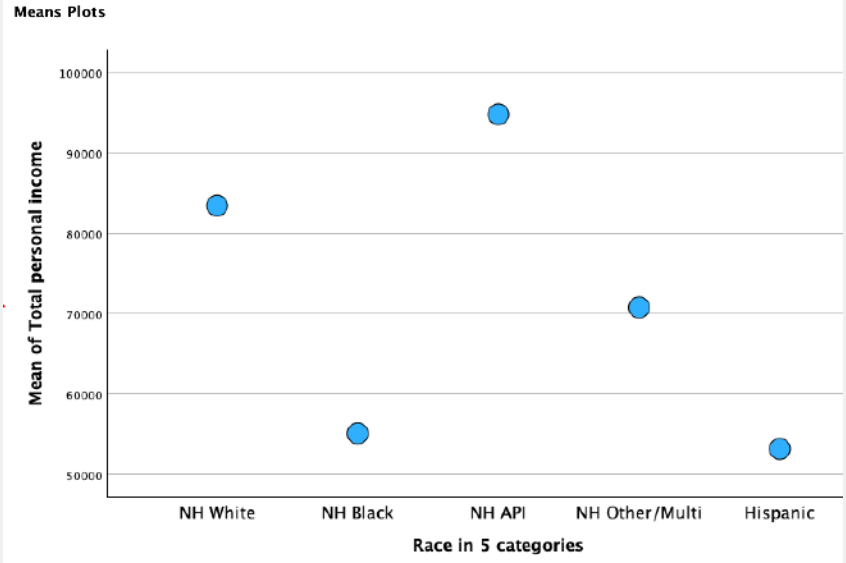
or you could add this to the last line of your ANOVA syntax:
/POSTHOC=TUKEY ALPHA(0.05).

However, when we ran the ANOVA, we also ran descriptive statistics, including confidence intervals. Rather than use a post-hoc test, we're just going to turn back to our skills from Chapter 6 and compare the racial groups based on their confidence intervals.

I generated a means plot with the ANOVA output. Here's what it looks like:



This tracks sample means. I find the lines distracting and inappropriately showing movement between categories, so I usually delete the lines and make the points bigger.



This is a quick visual way to get a sense of where the sample means are for each category in relation to one another.

Here is the rest of the descriptives output:

Oneway

Descriptives

Total personal income

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
NH White	69261799	83482.70	88647.450	10.652	83461.82	83503.58	-12084	1709394
NH Black	12832882	55123.06	50151.712	14.000	55095.62	55150.50	-9174	1428000
NH API	7141780	94846.17	97152.626	36.354	94774.92	94917.43	-10466	1726351
NH Other/Multi	3589218	70806.07	74604.958	39.379	70728.89	70883.25	-9279	1391860
Hispanic	19673955	53202.32	52602.763	11.859	53179.08	53225.57	-10466	1519697
Total	112499634	75269.21	81139.601	7.650	75254.22	75284.21	-12084	1726351

Because this is such a large dataset, we have very narrow confidence intervals, and so none of the confidence intervals overlap. Each racial group has a significantly different mean than each other racial group. We are over 95% confident that all the population means are different.

While all the means were different here, it doesn't have to be for the p-value in our ANOVA statistical test to be <0.05. If p<0.05, it just means that we are confident at least ONE group mean is different from at least ONE other group mean. The rest could all be the same.

Now that we've determined that they are significantly different, let's look at whether or not they are substantively different.

At the 95% confidence level:

- Hispanic upper bound $53,226; NH Black lower bound: $55,096. At least $1,870 apart.
- NH Black upper bound: $55,151; NH Other/Multi lower bound: $70,729. Over $10,000 apart. (Hispanic more)
- NH Other/Multi upper bound: $70,883; NH White lower bound: $83,462. Over $10,000 apart. (Hispanic and NH Black more)
- NH White upper bound: $83,504; NH API lower bound: $94,775. Over $10,000 apart. (Hispanic, NH Black, and NH Other/Multi more)

You could argue either way about the difference between the mean income for Hispanic and NH Black workers, but all the other group pairings are substantively different from one another.

Is it causal?

This is asking whether race has an impact or effect on variation in income. I would argue it is causal. We have a significant and substantive relationship. The causal order makes sense: while race might cause differences in pay, it would not make sense for how much one earns to change someone's race. Race is a relatively stable social location and racism could impact the kinds of jobs and pay workers of different racial groups earn. Theoretically, it makes sense that race impacts mean earnings. There is a robust body of sociological scholarship explaining various reasons for the racial pay gap (e.g., outright discrimination, wealth and opportunity gaps, segregation, etc.). There are no other preceding variables that would explain away this relationship.

Caution

Not all groups need to be significantly and substantively different from one another for the relationship to be causal.

For example, arguing that there is a causal relationship between race and income means you are claiming that race impacts variation in income. It does not mean that you are arguing every racial category difference has an impact on income.

This page titled 7: Hypothesis Tests, T-test, & ANOVA is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Ezra Temko (Consortium of Academic and Research Libraries in Illinois (CARLI)) .

8: Crosstabs and Chi-square

In this chapter we will evaluate relationships between two categorical variables. We will use crosstabulation tables for descriptive statistics and chi-square for inferential statistics.

Statistics in Action: Women Directors Across Top-Grossing Films

Introduction to Crosstabs

This section will introduce you to crosstabulations.

In Chapter 1, you saw an example of frequency tables from a USC Annenberg Inclusion Initiative report by Stacy L. Smith, Katherine Pieper, and Sam Wheeler with an intersectional analysis of gender and race for directors of the 100 top-grossing domestic fictional films in North America from 2007 to 2022 (fewer films were included in 2020 and 2021 due to pandemic-related performance).

Let's take a look at their report on "Inclusion in the Director's Chair." As part of the data collection and analysis, the institute collected a list of directors for each of the 1,488 films, and classified them by gender:



They also classified directors by studio:



What if you wanted to know if there is a relationship between studio and the proportion of women directors? Do some studios do a better or worse job when it comes to women's representation in directing their top films?

Here are the frequency tables we have for both variables:

Directors by Gender

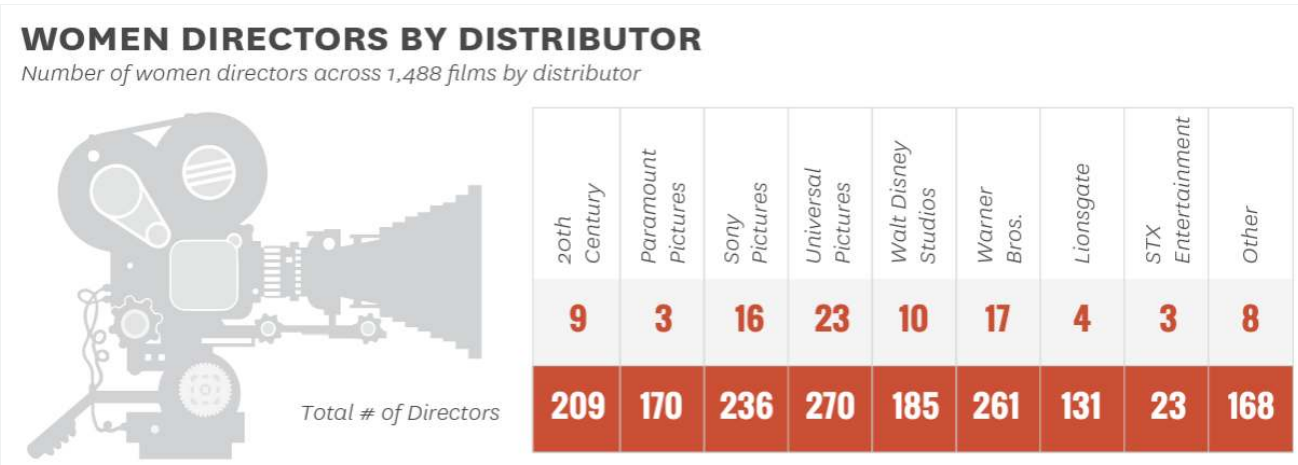
Gender	Frequency	Percent	Cumulative percent
Woman	93	5.63%	5.63%
Not woman	1560	94.37%	100.00%
Total	**1653**	100.00%	

Directors by Studio

Studio	Frequency	Percent	Cumulative percent
20th Century	209	12.64%	12.64%
Paramount Pictures	170	10.28%	22.93%
Sony Pictures	236	14.28%	37.21%
Universal Pictures	270	16.33%	53.54%
Walt Disney Studios	185	11.19%	64.73%
Warner Bros.	261	15.79%	80.52%
Lionsgate	131	7.92%	88.45%
STX Extertainment	23	1.39%	89.84%
Other	168	10.16%	100.00%
Total	**1653**	100.00%	

So, is there a relationship? We can't answer this question with the data above. To find out the relationship, we need to be able to go back to the raw data so that we can determine gender breakdown by studio. If there was no relationship / differences, we would expect all the studios to have the same proportion of women directors---we would expect 5.63% of directors in each studio to be women (just over 1/20). However, that's not the case.

Going back to the raw data, we can make a new table of directors that crosses gender and studio. For example, Jennifer Lee and Chris Buck were co-directors for Frozen II in 2019, which was a Walt Disney Studios production. Here you can see the USC Annenberg Inclusion Initiative's table of studios again, but this time they have counted how many of directors within each studio are women.



So, are some studios doing better when it comes to women's representation? Universal Pictures has the most women directors, but it also has the most films. Paramount Pictures and STX Entertainment have the fewer women directors, but over 13% (over 1/10) of STX Entertainment directors were women, and 1.8% (under 1/50) of Paramount Pictures directors were women. It's hard to compare studios using frequencies, because the studios do not have an equal number of films.

To make comparisons, we can calculate the percent of directors that are women for each studio.

Directors by Gender & Studio		Studio:	20th Century	Paramount Pictures	Sony Pictures	Universal Pictures	Walt Disney Studios	Warner Bros.	Lionsgate	STX Entertain.	Other	Total
Gender:	Women	Frequency	9	3	16	23	10	17	4	3	8	93
		Percent	**4.3%**	**1.8%**	**6.8%**	**8.5%**	**5.4%**	**6.5%**	**3.1%**	**13.0%**	**4.8%**	**5.6%**
	Not Women	Frequency	200	167	220	247	175	244	127	20	160	1,560
		Percent	**95.7%**	**98.2%**	**93.2%**	**91.5%**	**94.6%**	**93.5%**	**96.9%**	**87.0%**	**95.2%**	**94.4%**
	Total	Frequency	209	170	236	270	185	261	131	23	168	1,653
		Percent	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

This table is called a crosstabulation, or crosstab for short. It can also be called a contingency table. It tabulates (count up systematically and arranges in table form) our data across two variables.

Definition: Crosstabulation

A crosstabulation, or crosstab, is a table that shows subgroup distribution of data, enabling examination of potential relationships between variables.

Title: Crosstabulation of DV by IV

		IV Label			
		IV Category A	IV Category B	IV Category C	Total
DV Label	DV Category A	Frequency Column %	Frequency Column %	Frequency Column %	Frequency Column %
	DV Category B	Frequency Column %	Frequency Column %	Frequency Column %	Frequency Column %
	Total	Frequency 100%	Frequency 100%	Frequency 100%	Frequency 100%

Source: abc

When I look at the percent of women by studio, I can make comparisons, going across the row and comparing how similar or different the percentages are. Here we can see that STX Entertainment actually has the highest proportion of women directors, at 13%, and Paramount Pictures the lowest proportion, at 1.8%. Universal Pictures has the second highest proportion of women directors, at 8.5% (over 1/12), and Lionsgate has the second lowest proportion, at 3.1% (about 3/100).

Conversely I could compare the percent of directors that are men, though with a dichotomous variable (there were no non-binary directors), this will just be the inverse of the above, e.g., 87.0% of STX directors were men compared to Paramount with 98.2% men directors. Still, before I give an award to STX Entertainment, I might want to be cognizant that they did not have that many films, comparatively (1.4% of all directors). Because their total number of directors is 23, if just one woman director had instead been a man, they would have dropped to below 9% (though still the highest proportion!), or if just one man director had instead been a woman, they would have jumped up to over 17%. Percentages can fluctuate a lot more with smaller sample sizes, something to be cognizant of when you are comparing percentages.

Later in this chapter we'll discuss chi-square, a statistical test related to crosstabs. For this data, we would not conduct any inferential statistics. This is already population data, containing all of the top-100 grossing films from each year. It's also not a representative sample of all films --- the top-grossing films might be different from other films in terms of who directs them. To consider this a representative sample, each film out of all the films we are interested in would have had to have an equal chance of being included.

Note that the percentages in the table above are column percentages. I wanted to look at whether studio has an effect on women's representation. Studio is my independent variable and gender is my dependent variable. I needed to look at the percent of women directors within each studio. My independent variable is represented in the columns and my dependent variable in the rows.

This is what the table would have looked like with row percentages:

Directors by Gender & Studio		Studio:	20th Century	Paramount Pictures	Sony Pictures	Universal Pictures	Walt Disney Studios	Warner Bros.	Lionsgate	STX Entertain.	Other	Total
Gender:	Women	Frequency	9	3	16	23	10	17	4	3	8	93
		Percent	**9.7%**	**3.2%**	**17.2%**	**24.7%**	**10.8%**	**18.3%**	**4.3%**	**3.2%**	**8.6%**	**100%**
	Not Women	Frequency	200	167	220	247	175	244	127	20	160	1,560
		Percent	**12.8%**	**10.7%**	**14.1%**	**15.8%**	**11.2%**	**15.6%**	**8.1%**	**1.3%**	**10.3%**	**100%**
	Total	Frequency	209	170	236	270	185	261	131	23	168	1,653
		Percent	13%	10%	14%	16%	11%	16%	8%	1%	10%	100%

If I had instead used row percentages and compared them, I would have mistakenly thought Universal Studios had the highest proportion of women directors out of any studio! They don't, but because they have the most movies (and the second highest proportion of women directors), they have the highest proportion of women directors out of all women directors. Almost 1/4 of women directors from these top 1,488 films were from Universal Studios.

You will see crosstabs in a variety of formats. Oftentimes in journal articles multiple crosstabs will be combined into one table, and they may not give redundant information (e.g., if they give the percent of women directors, they might not give the percent of directors who are not women). In the infographic above from the report, only frequencies and not percentages were given, and the number of non-women directors was left out.

Here is another table from the USC Annenberg Inclusion Initiative's report, this one focused on representation of historically marginalized racial and ethnic groups.

UNDERREPRESENTED DIRECTORS BY DISTRIBUTOR

Number and percentage of underrepresented directors across 1,488 films by distributor

DISTRIBUTION COMPANY	# OF UR DIRECTORS	% OF UR DIRECTORS	TOTAL # OF DIRECTORS
20TH CENTURY	29	13.9	209
PARAMOUNT PICTURES	17	10	170
SONY PICTURES	41	17.4	236
UNIVERSAL PICTURES	53	19.6	270
WALT DISNEY STUDIOS	18	9.7	185
WARNER BROS.	31	11.9	261
LIONSGATE	28	21.4	131
STX ENTERTAINMENT	1	4.4	23
OTHER	34	20.2	168
TOTAL	252	15.2	1,653

Here there are frequencies and percentages for underrepresented directors, no frequencies or percentages for white directors, and then frequencies for the total number of directors. This table uses row percentages, with the independent variable represented in the rows and the dependent variable in the columns.

While crosstabs can be displayed in a variety of ways, we will use a standard convention in setting ours up, which will allow us to have the information we need in the table and have a consistent method to analyze them. For our tables, the independent variable will be on top, displayed in the columns, and the dependent variable will be on the side, displayed in the rows. We will include both counts and column percentages. Column percentages are the percent of the count for that cell out of the total for that column (for that independent variable category/subgroup).

For example, if we wanted to look at how women directors among these top films are distributed by studio, and our independent variable was gender and dependent variable studio, instead of the table above with row percentages, we would put gender on the top, studio on the side, and use column percentages. The table below shows this, with percentages out of gender. Out of all women directors, how are they distributed by studio?

Studio	Gender					
	Women		Not Women		Total	
20th Century	9	9.7%	200	12.8%	209	12.6%
Paramount Pictures	3	3.2%	167	10.7%	170	10.3%
Sony Pictures	16	17.2%	220	14.1%	236	14.3%
Universal Pictures	23	24.7%	247	15.8%	270	16.3%
Walt Disney Studios	10	10.8%	175	11.2%	185	11.2%
Warner Bros.	17	18.3%	244	15.6%	261	15.8%
Lionsgate	4	4.3%	127	8.1%	131	7.9%
STX Extertainment	3	3.2%	20	1.3%	23	1.4%
Other	8	8.6%	160	10.3%	168	10.2%
Total	**93**	100.0%	**1,560**	100.0%	**1,653**	100.0%

Here we can view the percent of total women directors that come from each studio. Comparing the column percentages as we go across the row, we can also see that some studios are making relatively similar proportional contributions to the totals for women directors and men directors (e.g., Walt Disney Studios has about 11% of all women directors and 11% of all men directors), while others are making differential contributions (e.g., Paramount Pictures had 3.2% of all women directors but 10.7% of all men directors).

Crosstabs are important for us to be able to analyze relationships. When you read poll results or analyze summary statistics by variable, you cannot compare relationships unless you are also given this information already broken down for you or have the raw data and can analyze it yourself. This chapter will teach you how to make crosstabs from a dataset.

When to use crosstabs

Crosstabs is useful when you want to compare two **categorical** variables. Note that in the analysis above, there were 9 studios and 2 genders. The 9th studio was actually a conglomerate catch-all, "other." You want to use crosstabs when you have a limited number of categories. What if there were 100 different studios included? That would have been a really long table and make comparisons difficult. You can use variables with a lot of categories if you are making a reference table (e.g., if you want to compare all 50 states and D.C., or all countries in the world), but in general you will want variables with a small number of categories. For example, if you wanted to use age and have data with adults, having every year from 18 to 89+ is going to be a big table that will also obscure patterns. Instead, you would want to use a variable that has age in categories (e.g., 18-35, 40-64, 65+). If you are using a ratio-level variable like number of siblings, you would likely want to have an upper cut-off like how the above report used "other" for studios, for example "5 or more siblings." If a table looks overwhelming or you have a number of column totals close to 0 (categories with a very low sample size), you may want to consider collapsing variables into fewer categories.

Statistics in Action: Schools, Race, & Security Measures

In this section you'll learn how to run a crosstab and chi-square using SPSS, and also learn about what chi-square is.

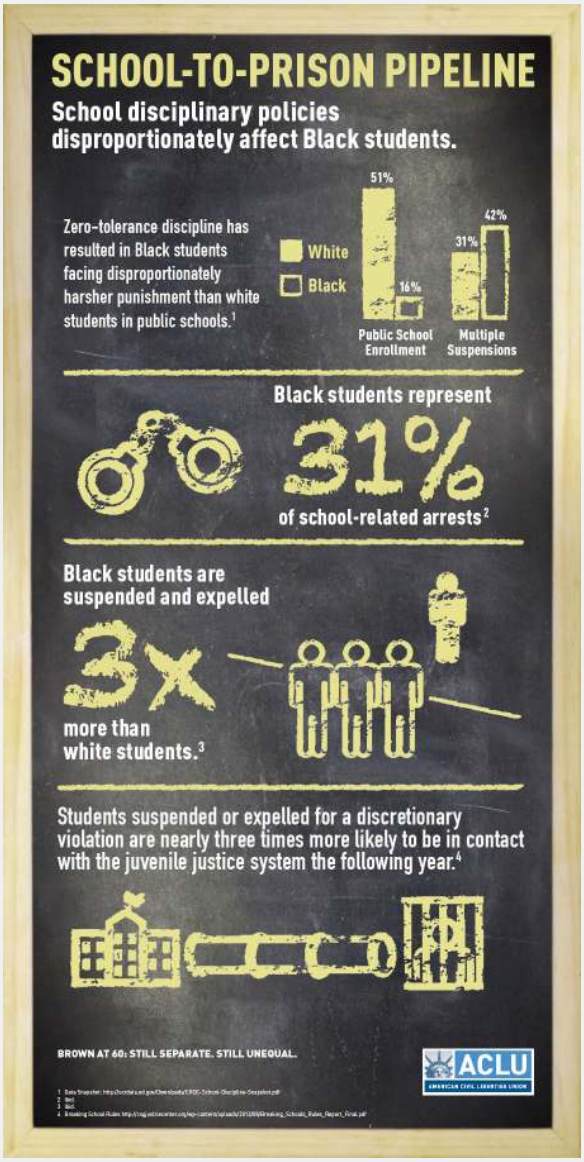
This section uses data from the National Crime Victimization Survey's School Crime Supplement, collected by the U.S. Census Bureau in 2019. The sample is representative of its target population, U.S. primary or secondary education (K-12) students ages 12 to 18. Students in private schools were included, but home-schooled children were not. The analyses below are adjusted using provided weighting, which adjusts the sample "to produce estimates of the number of persons ages 12 to 18."

Below you can see frequency tables for race. NH means non-Hispanic. The unweighted sample sizes are the actual number of student respondents. The weighted sample sizes reflect estimates of their actual proportion, e.g., at the time there were just over 16 million non-Hispanic white students ages 12 to 18 in the United States.

	Unweighted		Weighted	
Race	Frequency	Percent	Frequency	Percent
NH White only	8,856	62.0%	16,061,624	54.9%
NH Black only	1,550	10.9%	3,867,483	13.2%
NH AAPI only	812	5.7%	1,949,552	6.7%
NH U.S. indigenous only	74	0.5%	135,732	0.5%
Hispanic & White only	2,713	19.0%	6,658,166	22.7%
Other Multiracial	268	1.9%	607,863	2.1%
Total	**14,273**	100.0%	**29,280,419**	100.0%

The school-to-prison pipeline is a term describing how "children are funneled out of public schools and into the juvenile and criminal justice systems," disproportionately children of color.

Here is an infographic from the American Civil Liberties Union about the school-to-prison pipeline:



In the School Crime Supplement, students are asked whether their school has security guards or assigned police officers, and whether their school has metal detectors, including wands. Race is a categorical variable (in this case I separated race into 6 categories), and so is whether or not students have these in their schools.

First, we'll generate SPSS outputs for these variables. Then we'll analyze the crosstabs. Finally we'll turn to chi-square. SPSS below will give us cross-tabs as well as run chi-square tests.

Using SPSS

First, let's take a look at security guards. We are wondering about whether race has an effect on whether or not a school has security guards or assigned police officers. That means that race is the independent variable and guards/officers the dependent variable.

How to:

Syntax:
CROSSTABS
/TABLES=DVName BY IVName

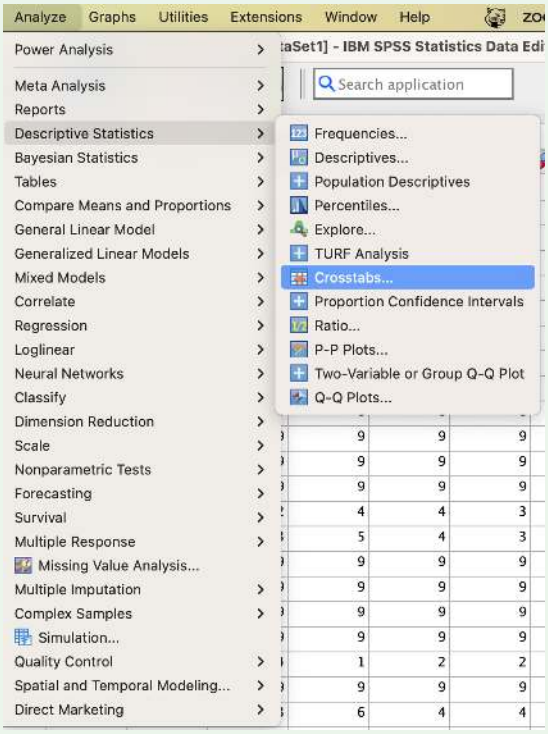
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ
/CELLS=COUNT COLUMN
/COUNT ROUND CELL.

I substitute the names of my variables where it says DVName (the name of my dependent variable) and IVName (the name of my independent variable) and then click play.

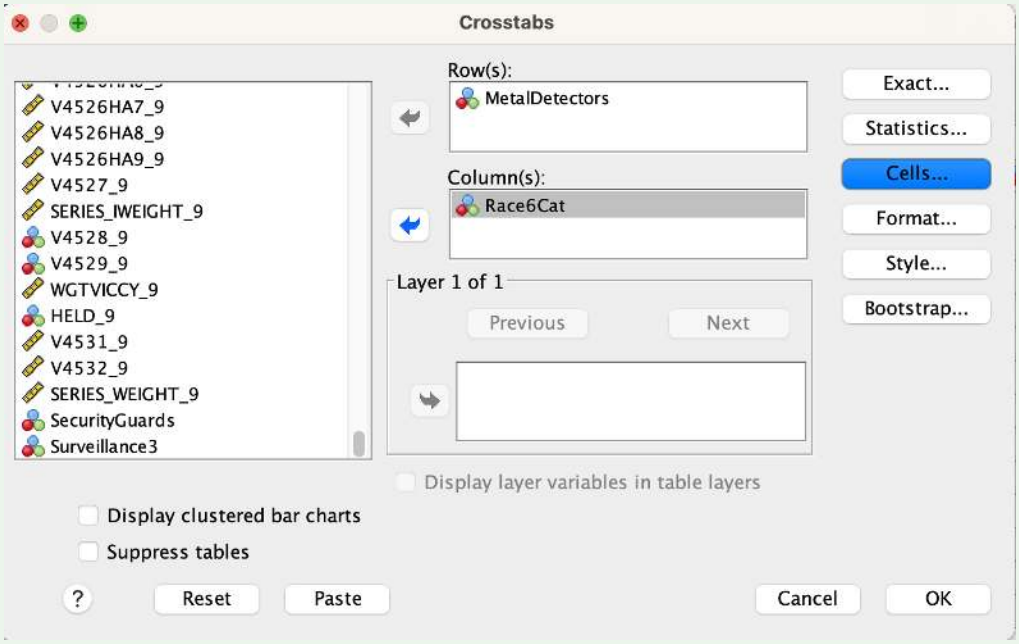
CROSSTABS
/TABLES=SecurityGuards BY Race6Cat
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ
/CELLS=COUNT COLUMN
/COUNT ROUND CELL.

Next, let's take a look at metal detectors and wands. We are wondering about whether race has an effect on whether or not a school has metal detectors, including wands. That means that race is the independent variable and metal detectors is the dependent variable.

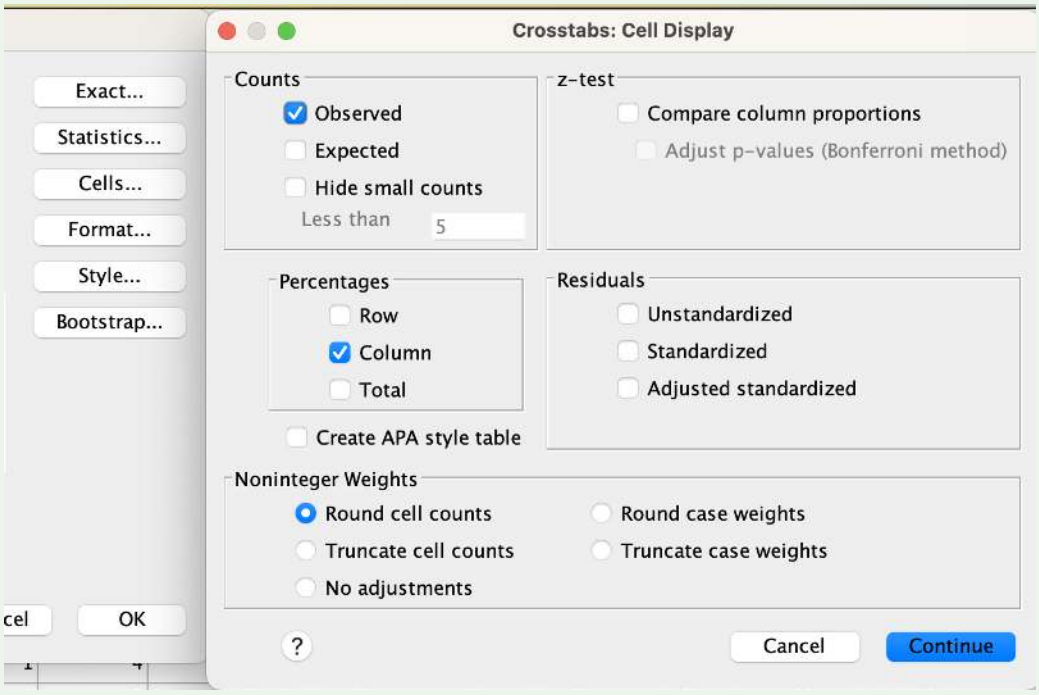
Menu:
First, go to Analyze → Descriptive Statistics → Crosstabs.



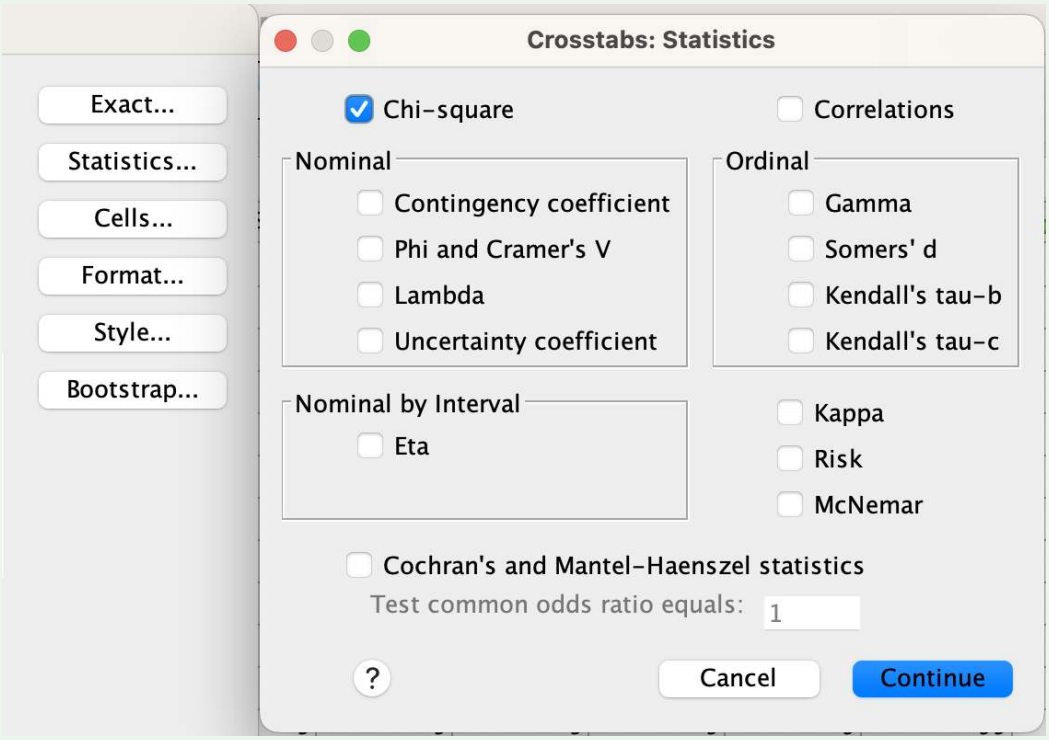
Next, put your independent variable into the "Column(s)" box and your dependent variable into the "Row(s)" box.



After that, click the "Cells" box on the right side. In the window that pops up, make sure that "Observed" is checked under "Counts." Under "Percentages," check the box for "Column." Don't check any other boxes. Click the "Continue" button.



After that, go to the "Statistics" box on the right side. Check the box for "Chi-square" and click "Continue."



Finally, click "OK."

Caution

It's vital that your **independent variable goes in the columns box** and your dependent variable in the rows box, and not the other way around. Otherwise your column percentages will not match what you intend them to and you won't be able to make comparisons the same way.

Make sure to turn on **column percentages**. Without them, you can't make comparisons between categories given the differences in sample sizes among categories.

Don't check other boxes (like expected counts, or row percentages). These will add additional values to each cell that will be confusing.

Gamma

When you clicked in the "Statistics" box, another box you could check, under the "Ordinal" column, is "Gamma." Gamma is a descriptive statistic that measures strength and direction. It requires two ordinal-level variables, though sometimes indicator variables are used. Gamma ranges from -1 to +1. If the Gamma value is positive, that means that an increase in the independent variable is associated with an increase in the dependent variable. If the Gamma value is negative, that means that an increase in the independent variable is associated with a decrease in the dependent variable. The number measures strength by comparing positive to negative trends. A Gamma of zero indicates no relationship. The further away from zero, the stronger the relationship. Generally a Gamma value that is at least -0.3 or +0.3 indicates at least a moderate association, at least -0.5 or +0.5 indicates a substantial association, and at least -0.7 or +0.7 indicates a very strong association.

Here is our SPSS output for race and security guards:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Security guards or assigned police officers (Does your school take any measures to make sure students are safe?) * Race in 6 categories	21145084	72.2%	8135334.797	27.8%	29280418.8	100.0%

Security guards or assigned police officers (Does your school take any measures to make sure students are safe?) * Race in 6 categories Crosstabulation

			Race in 6 categories						
			NH White only	NH Black only	NH AAPI only	NH U.S. indigenous only	Hispanic & White only	Other Multiracial	Total
Security guards or assigned police officers (Does your school take any measures to make sure students are safe?)	No	Count	3019402	436765	268821	23897	848996	90486	4688367
		% within Race in 6 categories	25.9%	15.4%	19.8%	21.7%	17.9%	21.1%	22.2%
	Yes	Count	8646732	2408131	1086868	86221	3890389	338376	16456717
		% within Race in 6 categories	74.1%	84.6%	80.2%	78.3%	82.1%	78.9%	77.8%
Total		Count	11666134	2844896	1355689	110118	4739385	428862	21145084
		% within Race in 6 categories	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	224125.707[a]	5	<.001
Likelihood Ratio	230209.033	5	<.001
Linear-by-Linear Association	112092.360	1	<.001
N of Valid Cases	21145084		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 24415.77.

Here is our SPSS output for race and metal detectors:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Metal detectors, including wands (Does your school take any measures to make sure students are safe?) * Race in 6 categories	19795894	67.6%	9484524.797	32.4%	29280418.8	100.0%

Metal detectors, including wands (Does your school take any measures to make sure students are safe?) * Race in 6 categories Crosstabulation

			Race in 6 categories						
			NH White only	NH Black only	NH AAPI only	NH U.S. indigenous only	Hispanic & White only	Other Multiracial	Total
Metal detectors, including wands (Does your school take any measures to make sure students are safe?)	No	Count	10157031	1896204	1028494	97208	3664440	390447	17233824
		% within Race in 6 categories	91.7%	71.6%	84.9%	96.2%	84.4%	92.4%	87.1%
	Yes	Count	913376	750875	182737	3790	679376	31916	2562070
		% within Race in 6 categories	8.3%	28.4%	15.1%	3.8%	15.6%	7.6%	12.9%
Total		Count	11070407	2647079	1211231	100998	4343816	422363	19795894
		% within Race in 6 categories	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	826613.904[a]	5	<.001
Likelihood Ratio	736673.047	5	<.001
Linear-by-Linear Association	96119.312	1	<.001
N of Valid Cases	19795894		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 13071.60.

Interpreting Crosstabs

Security guards or assigned police officers (Does your school take any measures to make sure students are safe?) * Race in 6 categories Crosstabulation

			Race in 6 categories						
			NH White only	NH Black only	NH AAPI only	NH U.S. indigenous only	Hispanic & White only	Other Multiracial	Total
Security guards or assigned police officers (Does your school take any measures to make sure students are safe?)	No	Count	3019402	436765	268821	23897	848996	90486	4688367
		% within Race in 6 categories	25.9%	15.4%	19.8%	21.7%	17.9%	21.1%	22.2%
	Yes	Count	8646732	2408131	1086868	86221	3890389	338376	16456717
		% within Race in 6 categories	74.1%	84.6%	80.2%	78.3%	82.1%	78.9%	77.8%
Total		Count	11666134	2844896	1355689	110118	4739385	428862	21145084
		% within Race in 6 categories	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Looking at this crosstab, we can see six different races, as well as a total that includes everyone in the sample. For each racial category, it shows the distribution of whether respondents said they do or do not have a security guard or assigned police officer in their school. With the particular weighting used here, the frequencies are estimates of the target population. So for example, we would estimate that 3,019,402 non-Hispanic white only U.S. students ages 12 to 18 do not have a security guard or assigned police officer in their school (25.9% of respondents, just over 1/4), and that 8,646,732 U.S. students ages 12 to 18 do have a security guard or assigned police officer in their school (71.4% of respondents, just under 3/4).

How to look for and analyze relationships

The table's story includes answers to three questions:

1. **Is there a relationship?**
2. **If so, how substantive it it?**

3. **If so, what is the relationship? (E.g., what is its direction?)**

1. Is there a relationship?

To analyze whether or not there is a relationship, look across rows and see if the percentages are the same or if they are different. Overall, 77.8% of student respondents report having security guards or police officers in their school. If there was no relationship between race and having security guards/police officers in one's school, then we would expect each racial category to have about 77.8% of student respondents reporting this. Look across and see what you notice.

Overall, 77.8% of students ages 12 to 18 are estimated to have a security guard or police officer in their school. That's over 3/4 of students. Each racial group also has somewhere between 74.1% and 84.6% of their respondents who reported having a security guard or police officer in their school. But don't make a common mistake of getting fooled into thinking there is no relationship just because most people have a security guard or police officer in their school, regardless of / across racial categories. We are trying to analyze variation. When I look across, I see that the percentage of students varies by race.

2. How substantive is the relationship?

Small percentage changes indicate a weak relationship. Substantial percentage changes indicate a strong relationship.

The biggest difference I see is that NH Black only respondents are over 10% more likely to report having a security guard or police officer in their school compared to NH White only respondents (84.6% vs. 74.1%). While what constitutes a substantive relationship is subjective, a good rule of thumb is to definitely characterize a relationship as being substantial anytime there is at least a 5% difference.

3. What is the relationship (e.g., what is its direction)?

If these were ordinal variables (e.g., age and income), we could classify the relationship by direction (e.g., as age increases, income increases). However, race is nominal so we cannot do that.

Looking at the crosstab, I am looking for which racial groups are more likely to have security guards or police officers in their school and which are less likely. I see that over 4/5 of NH Black only, Hispanic & White only, and NH AAPI only student respondents have security guards or assigned police officers in their school, between 3/4 and 4/5 of other multiracial and NH U.S. indigenous students do, and just under 3/4 of NH White only students do.

Try again with the crosstab below. Answer the three questions:

1. **Is there a relationship?** (Do column percentages vary as you move across rows?)
2. **If so, how substantive it it?** (Do they vary a lot?)
3. **If so, what is the relationship? (E.g., what is its direction?)** (Which IV categories have higher and which have lower column percentages?)

Metal detectors, including wands (Does your school take any measures to make sure students are safe?) * Race in 6 categories Crosstabulation

			Race in 6 categories						
			NH White only	NH Black only	NH AAPI only	NH U.S. indigenous only	Hispanic & White only	Other Multiracial	Total
Metal detectors, including wands (Does your school take any measures to make sure students are safe?)	No	Count	10157031	1896204	1028494	97208	3664440	390447	17233824
		% within Race in 6 categories	91.7%	71.6%	84.9%	96.2%	84.4%	92.4%	87.1%
	Yes	Count	913376	750875	182737	3790	679376	31916	2562070
		% within Race in 6 categories	8.3%	28.4%	15.1%	3.8%	15.6%	7.6%	12.9%
Total		Count	11070407	2647079	1211231	100998	4343816	422363	19795894
		% within Race in 6 categories	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

https://socialsci.libretexts.org/@go/page/189028

Overall, just over 1/10 of student respondents report having metal detectors in their school (about 2.6 million children). When I compare the different column percentages for different racial groups, I see differences, with column percentages ranging from 3.8% to 28.4%. With a difference of over 25% between two of the racial groups, there appears to be a relationship, and it is a substantial one.

- 15.1% of NH AAPI only respondents and 15.6% of Hispanic & White only respondents reported having metal detectors in their schools. These are pretty similar. However, just because there is no substantive difference between two racial groups does not mean there is not an overall relationship between race and having metal detectors in one's school. It just means that the relationship only applies to differences between certain racial groups.

NH Black only students (respondents) are most likely to report having metal detectors in their school. Over 1/4 of these respondents reported this. Over 10% lower are Hispanic & White only and NH AAPI students, where over 1/7 of students reported this. Less than 1/10 of NH White only, other multiracial, and NH U.S. indigenous only students reported this. There are substantive differences. Indeed, while there are over 8 million more NH White only students ages 12-18 in the U.S. than NH Black students, only about 160,000 more NH White only students than NH Black only students are estimated to have metal detectors in their schools. NH Black students are over 20% more likely than NH White students to report having metal detectors in their schools. That's over 1 student out of every 5.

Chi-Square

We know there is a relationship between race and whether or not a 12-16 year old U.S. student who took the School Crime Supplement survey has a police officer/security guard and/or a metal detector in their school. But we are not interested in those 14,273 respondents. The survey was conducted to allow us to generalize to all 12-16 year old U.S. students.

Chi-square ($\chi 2$) is an inferential statistic. It lets us evaluate population claims for categorical variables from representative data. If you have a cross-tab, chi-square is usually the corresponding statistical test you use for hypothesis testing.

Chi (χ) is a Greek letter. While it rhymes with chai, it starts with a k sound, not a ch sound. The inferential statistic is called chi-square (notated as χ^2) because we square values when we are calculating it.

Hypothesis Testing

$H_\varnothing$: Null hypothesis
The null hypothesis is that there is no difference or no relationship (among your target population). In Chapter 7, the null hypothesis was also that there is no difference between the two variables in terms of their means, that they're equal. For chi-square, it's the same--- that the frequencies in our crosstab that we would expect if there was no relationship are equal to the frequencies we actually have from our sample.

H_a: Alternative hypothesis
The alternative hypothesis is that there is a relationship (among your target population). Chi-square tests to see whether the column percentages we were looking at before are different enough, taking into account sample size and the number of categories both variables have, such that we can claim they are likely different among our target population.

Significance and p-values:
From our chi-square test, we can use a chi-square table, or an online calculator like this one (for df, the degrees of freedom, use the total number of categories for both variables minus 2) to calculate our p-value. The more categories your variables have, the higher your chi-square value will need to be to pass the **critical value** corresponding to p=0.05. The p-value is the probability we would draw our random sample with the column percentages we did if there really was no relationship between the two variables.

For the p-value, look at the "Chi-Square Tests" output. The p-value is in the cell at the intersection of the "Pearson Chi-Square" row and the "Asymptomatic Significance (2-sided)" column.

Remember, our conventional standard is that if $p<0.05$, we reject our null, and claim there is a significant relationship. We are over 95% confident the two variables are related among our target population.

Statistics in Action: Schools, Race, & Security Measures

Security Guards/Police Officers

$H_\varnothing$: Null hypothesis There is no relationship among U.S. students ages 12-18 between race and whether or not there is a security guard or police officer assigned to a school.

H_a: Alternative hypothesis There is a relationship among U.S. students ages 12-18 between race and whether or not there is a security guard or police officer assigned to a school.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	224125.707[a]	5	<.001
Likelihood Ratio	230209.033	5	<.001
Linear-by-Linear Association	112092.360	1	<.001
N of Valid Cases	21145084		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 24415.77.

For the p-value, look at the cell at the intersection of the "Pearson Chi-Square" row and the "Asymptomatic Significance (2-sided)" column. Here, p<0.001. That is less than 0.05, so we reject our null hypothesis that there is no relationship. We are over 99.9% confident (1-0.001=99.9%) that there is a relationship, among all U.S. 12-18 year old students, between race and whether or not there is a security guard or police officer assigned to a school. (Note: Given the large sample size with weighting, I also ran chi-square with the unweighted data. The chi-square value was much lower, 52.3, but the p-value was still <0.001.)

Metal Detectors/Wands

$H_\varnothing$: Null hypothesis There is no relationship among U.S. students ages 12-18 between race and whether or not there are metal detectors, including wands, in the student's school.

H_a: Alternative hypothesis There is a relationship among U.S. students ages 12-18 between race and whether or not there are metal detectors, including wands, in the student's school.

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	826613.904[a]	5	<.001
Likelihood Ratio	736673.047	5	<.001
Linear-by-Linear Association	96119.312	1	<.001
N of Valid Cases	19795894		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 13071.60.

For the p-value, look at the cell at the intersection of the "Pearson Chi-Square" row and the "Asymptomatic Significance (2-sided)" column. Here, p<0.001. That is less than 0.05, so we reject our null hypothesis that there is no relationship. We are over 99.9% confident (1-0.001=99.9%) that there is a relationship, among all U.S. 12-18 year old students, between race and whether or not there are metal detectors, including wands, in the student's school. (Note: Given the large sample size with weighting, I also ran chi-square with the unweighted data. The chi-square value was much lower, 222.3, but the p-value was still <0.001.)

Behind the Scenes: Calculating Chi-Square

Among all students surveyed, 12.9% had metal detectors in their schools.

Metal detectors, including wands (Does your school take any measures to make sure students are safe?) * Race in 6 categories Crosstabulation

			Race in 6 categories						
			NH White only	NH Black only	NH AAPI only	NH U.S. indigenous only	Hispanic & White only	Other Multiracial	Total
Metal detectors, including wands (Does your school take any measures to make sure students are safe?)	No	Expected Count							17233824.0
		% within Race in 6 categories							87.1%
	Yes	Expected Count							2562070.0
		% within Race in 6 categories							12.9%
Total		Expected Count	11070407.0	2647079.0	1211231.0	100998.0	4343816.0	422363.0	19795894.0
		% within Race in 6 categories	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

If there was no relationship between race and whether or not students have metal detectors in their schools, then we would **expect** that percentage to be the same across racial groups. 12.9% of NH White only students would have metal detectors in their schools, 12.9% of NH Black only students would have metal detectors in their schools, 12.9% of NH AAPI only students, etc.

We would end up with an **expected crosstab** that looks like this:

Metal detectors, including wands (Does your school take any measures to make sure students are safe?) * Race in 6 categories Crosstabulation

			Race in 6 categories						
			NH White only	NH Black only	NH AAPI only	NH U.S. indigenous only	Hispanic & White only	Other Multiracial	Total
Metal detectors, including wands (Does your school take any measures to make sure students are safe?)	No	Expected Count	9637627.2	2304482.6	1054468.3	87926.4	3781620.6	367699.0	17233824.0
		% within Race in 6 categories	87.1%	87.1%	87.1%	87.1%	87.1%	87.1%	87.1%
	Yes	Expected Count	1432779.8	342596.4	156762.7	13071.6	562195.4	54664.0	2562070.0
		% within Race in 6 categories	12.9%	12.9%	12.9%	12.9%	12.9%	12.9%	12.9%
Total		Expected Count	11070407.0	2647079.0	1211231.0	100998.0	4343816.0	422363.0	19795894.0
		% within Race in 6 categories	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Instead, our actual sample, our **observed crosstab,** looks like this:

Metal detectors, including wands (Does your school take any measures to make sure students are safe?) * Race in 6 categories Crosstabulation

			Race in 6 categories						
			NH White only	NH Black only	NH AAPI only	NH U.S. indigenous only	Hispanic & White only	Other Multiracial	Total
Metal detectors, including wands (Does your school take any measures to make sure students are safe?)	No	Count	10157031	1896204	1028494	97208	3664440	390447	17233824
		% within Race in 6 categories	91.7%	71.6%	84.9%	96.2%	84.4%	92.4%	87.1%
	Yes	Count	913376	750875	182737	3790	679376	31916	2562070
		% within Race in 6 categories	8.3%	28.4%	15.1%	3.8%	15.6%	7.6%	12.9%
Total		Count	11070407	2647079	1211231	100998	4343816	422363	19795894
		% within Race in 6 categories	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Our null hypothesis is that the observed and expected frequencies are equal, that these two tables match.

For each cell, we use the formula: $(f_o-f_e)^2 \div f_e$ where f is for frequency, o is for observed, and e is for expected. For example, for NH White only students saying "no" to having metal detectors, the expected frequency (if there was no relationship, if 87.1% of NH White only respondents had said no) is 9,637,627.2. The observed frequency, the actual number who said "no" (weighted) was 91.7% of the sample, 10,157,031.

$(f_o-f_e)^2 \div f_e = (10{,}157{,}031-9{,}637{,}627.2)^2 \div 9{,}637{,}627.2 = 27{,}992.4$

I would then do this for each of the other cells (NH White only Yes, NH Black only No, NH Black only yes, etc.). Then I would find the sum of these numbers. The bigger the differences between the actual data and expected data, the bigger the chi-square value. The bigger the chi-square value, the smaller the p-value.

This page titled 8: Crosstabs and Chi-square is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Ezra Temko (Consortium of Academic and Research Libraries in Illinois (CARLI)) .

8.1: Crosstabs for Research Methods

Statistics in Action: Women Directors Across Top-Grossing Films

Introduction to Crosstabs

Let's take a look at the USC Annenberg Inclusion Initiative's report on "Inclusion in the Director's Chair," which analyzed the gender for directors of the100 top-grossing domestic fictional films in North America from 2007 to 2022 (fewer films were included in 2020 and 2021 due to pandemic-related performance).

As part of the data collection and analysis, the institute collected a list of directors for each of the 1,488 films, and classified them by gender:



They also classified directors by studio:



What if you wanted to know if there is a relationship between studio and the proportion of women directors? Do some studios do a better or worse job when it comes to women's representation in directing their top films?

Here are the frequency tables we have for both variables:

Directors by Gender

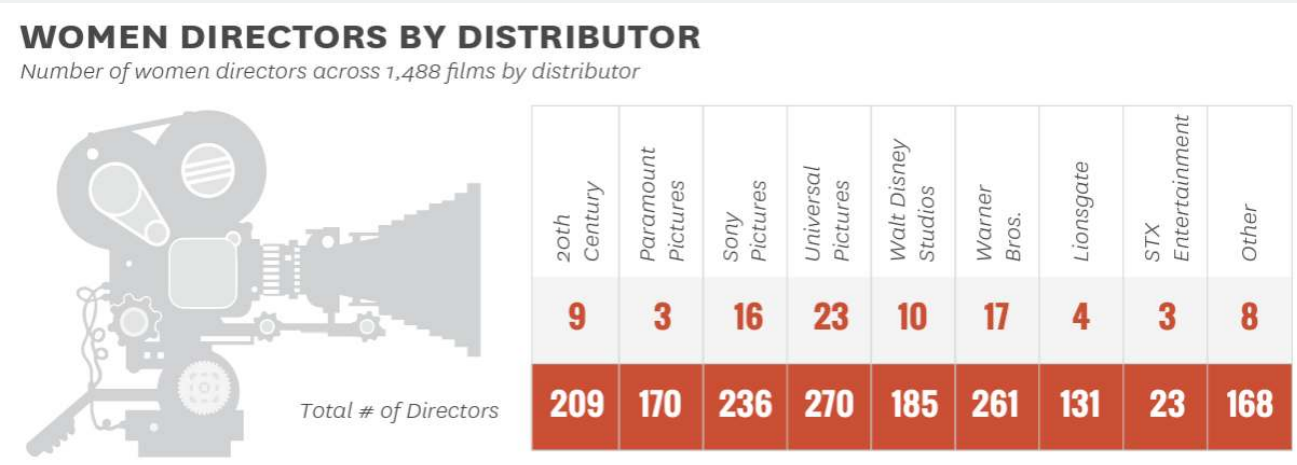
Gender	Frequency	Percent	Cumulative percent
Woman	93	5.63%	5.63%
Not woman	1560	94.37%	100.00%
Total	**1653**	100.00%	

Directors by Studio

Studio	Frequency	Percent	Cumulative percent
20th Century	209	12.64%	12.64%
Paramount Pictures	170	10.28%	22.93%
Sony Pictures	236	14.28%	37.21%
Universal Pictures	270	16.33%	53.54%
Walt Disney Studios	185	11.19%	64.73%
Warner Bros.	261	15.79%	80.52%
Lionsgate	131	7.92%	88.45%
STX Extertainment	23	1.39%	89.84%
Other	168	10.16%	100.00%
Total	**1653**	100.00%	

So, is there a relationship? We can't answer this question with the data above. We don't know, using the data above, because we don't know the gender breakdown by studio. If there was no relationship / differences, we would expect all the studios to have the same proportion of women directors---we would expect 5.63% of directors in each studio to be women (just over 1/20). However, that's not the case.

Going back to the raw data, we can make a new table of directors that crosses gender and studio. For example, Jennifer Lee and Chris Buck were co-directors for Frozen II in 2019, which was a Walt Disney Studios production. Here you can see the USC Annenberg Inclusion Initiative's table of studios again, but this time they have counted how many of directors within each studio are women.



So, are some studios doing better when it comes to women's representation? Universal Pictures has the most women directors, but it also has the most films. Paramount Pictures and STX Entertainment have the fewer women directors, but over 13% (over 1/10) of STX Entertainment directors were women, and 1.8% (under 1/50) of Paramount Pictures directors were women. It's hard to compare studios using frequencies, because the studios do not have an equal number of films.

To make comparisons, we can calculate the percent of directors that are women for each studio.

Directors by Gender & Studio		Studio:	20th Century	Paramount Pictures	Sony Pictures	Universal Pictures	Walt Disney Studios	Warner Bros.	Lionsgate	STX Entertain.	Other	Total
Gender:	Women	Frequency	9	3	16	23	10	17	4	3	8	93
		Percent	**4.3%**	**1.8%**	**6.8%**	**8.5%**	**5.4%**	**6.5%**	**3.1%**	**13.0%**	**4.8%**	**5.6%**
	Not Women	Frequency	200	167	220	247	175	244	127	20	160	1,560
		Percent	**95.7%**	**98.2%**	**93.2%**	**91.5%**	**94.6%**	**93.5%**	**96.9%**	**87.0%**	**95.2%**	**94.4%**
	Total	Frequency	209	170	236	270	185	261	131	23	168	1,653
		Percent	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

This table is called a crosstabulation, or crosstab for short. It can also be called a contingency table. It tabulates (count up systematically and arranges in table form) our data across two variables.

Definition: Crosstabulation

A crosstabulation, or crosstab, is a table that shows subgroup distribution of data, enabling examination of potential relationships between variables.

Title: Crosstabulation of DV by IV

		IV Label			
		IV Category A	**IV Category B**	**IV Category C**	**Total**
DV Label	**DV Category A**	Frequency Column %	Frequency Column %	Frequency Column %	Frequency Column %
	DV Category B	Frequency Column %	Frequency Column %	Frequency Column %	Frequency Column %
	Total	Frequency 100%	Frequency 100%	Frequency 100%	Frequency 100%

Source: abc

When I look at the percent of women by studio, I can make comparisons, going across the row and comparing how similar or different the percentages are. Here we can see that STX Entertainment actually has the highest proportion of women directors, at 13%, and Paramount Pictures the lowest proportion, at 1.8%. Universal Pictures has the second highest proportion of women directors, at 8.5% (over 1/12), and Lionsgate has the second lowest proportion, at 3.1% (about 3/100).

(Conversely I could compare the percent of directors that are men, though with a dichotomous variable (there were no non-binary directors), this will just be the inverse of the above, e.g., 87.0% of STX directors were men compared to Paramount with 98.2% men directors. Still, before I give an award to STX Entertainment, I might want to be cognizant that they did not have that many films, comparatively (1.4% of all directors). Because their total number of directors is 23, if just one woman director had instead been a man, they would have dropped to below 9% (though still the highest proportion!), or if just one man director had instead been a woman, they would have jumped up to over 17%. Percentages can fluctuate a lot more with smaller sample sizes, something to be cognizant of when you are comparing percentages.

Note that the percentages in the table above are column percentages. I wanted to look at whether studio has an effect on women's representation. Studio is my independent variable and gender is my dependent variable. I needed to look at the percent of women directors within each studio. My independent variable is represented in the columns and my dependent variable in the rows.

This is what the table would have looked like with row percentages:

Directors by Gender & Studio		Studio:	20th Century	Paramount Pictures	Sony Pictures	Universal Pictures	Walt Disney Studios	Warner Bros.	Lionsgate	STX Entertain.	Other	Total
Gender:	Women	Frequency	9	3	16	23	10	17	4	3	8	93
		Percent	**9.7%**	**3.2%**	**17.2%**	**24.7%**	**10.8%**	**18.3%**	**4.3%**	**3.2%**	**8.6%**	**100%**
	Not Women	Frequency	200	167	220	247	175	244	127	20	160	1,560
		Percent	**12.8%**	**10.7%**	**14.1%**	**15.8%**	**11.2%**	**15.6%**	**8.1%**	**1.3%**	**10.3%**	**100%**
	Total	Frequency	209	170	236	270	185	261	131	23	168	1,653
		Percent	13%	10%	14%	16%	11%	16%	8%	1%	10%	100%

If I had instead used row percentages and compared them, I would have mistakenly thought Universal Studios had the highest proportion of women directors out of any studio! They don't, but because they have the most movies (and the second highest proportion of women directors), they have the highest proportion of women directors out of all women directors. Almost 1/4 of women directors from these top 1,488 films were from Universal Studios.

You will see crosstabs in a variety of formats. Oftentimes in journal articles multiple crosstabs will be combined into one table, and they may not give redundant information (e.g., if they give the percent of women directors, they might not give the percent of directors who are not women). In the infographic above from the report, only frequencies and not percentages were given, and the number of non-women directors was left out.

Here is another table from the USC Annenberg Inclusion Initiative's report, this one focused on representation of historically marginalized racial and ethnic groups.

UNDERREPRESENTED DIRECTORS BY DISTRIBUTOR

Number and percentage of underrepresented directors across 1,488 films by distributor

DISTRIBUTION COMPANY	# OF UR DIRECTORS	% OF UR DIRECTORS	TOTAL # OF DIRECTORS
20TH CENTURY	29	13.9	209
PARAMOUNT PICTURES	17	10	170
SONY PICTURES	41	17.4	236
UNIVERSAL PICTURES	53	19.6	270
WALT DISNEY STUDIOS	18	9.7	185
WARNER BROS.	31	11.9	261
LIONSGATE	28	21.4	131
STX ENTERTAINMENT	1	4.4	23
OTHER	34	20.2	168
TOTAL	252	15.2	1,653

Here there are frequencies and percentages for underrepresented directors, no frequencies or percentages for white directors, and then frequencies for the total number of directors. This table uses row percentages, with the independent variable represented the rows and the dependent variable in the columns.

While crosstabs can be displayed in a variety of ways, we will use a standard convention in setting ours up, which will allow us to have the information we need in the table and have a consistent method to analyze them. For our tables, the independent variable will be on top, displayed in the columns, and the dependent variable will be on the side, displayed in the rows. We will include both counts and column percentages. Column percentages are the percent of the count for that cell out of the total for that column (for that independent variable category/subgroup).

For example, if we did want to look at how women directors among these top films are distributed by studio, and our independent variable was gender and dependent variable studio, instead of the table above with row percentages, we would put gender on the top, studio on the side, and use column percentages. The table below shows this, with percentages out of gender. Out of all women directors, how are they distributed by studio?

Studio	Gender					
	Women		Not Women		Total	
20th Century	9	9.7%	200	12.8%	209	12.6%
Paramount Pictures	3	3.2%	167	10.7%	170	10.3%
Sony Pictures	16	17.2%	220	14.1%	236	14.3%
Universal Pictures	23	24.7%	247	15.8%	270	16.3%
Walt Disney Studios	10	10.8%	175	11.2%	185	11.2%
Warner Bros.	17	18.3%	244	15.6%	261	15.8%
Lionsgate	4	4.3%	127	8.1%	131	7.9%
STX Extertainment	3	3.2%	20	1.3%	23	1.4%
Other	8	8.6%	160	10.3%	168	10.2%
Total	**93**	100.0%	**1,560**	100.0%	**1,653**	100.0%

Here we can view the percent of total women directors that come from each studio. Comparing the column percentages as we go across the row, we can also see that some studios are making relatively similar proportional contributions to the totals for women directors and men directors (e.g., Walt Disney Studios has about 11% of all women directors and 11% of all men directors), while others are making differential contributions (e.g., Paramount Pictures had 3.2% of all women directors but 10.7% of all men directors).

Crosstabs are important for us to be able to analyze relationships. When you read poll results or analyze summary statistics by variable, you cannot compare relationships unless you are also given this information already broken down for you or have the raw data and can analyze it yourself. This chapter will teach you how to make crosstabs from a dataset.

When to use crosstabs

Crosstabs is useful when you want to compare two **categorical** variables. Note that in the analysis above, there were 9 studios and 2 genders. The 9th studio was actually a conglomerate catch-all, "other." You want to use crosstabs when you have a limited number of categories. What if there were 100 different studios included? That would have been a really long table and make comparisons difficult. You can use variables with a lot of categories if you are making a reference table (e.g., if you want to compare all 50 states and D.C., or all countries in the world), but in general you will want variables with a small number of categories. For example, if you wanted to use age and have data with adults, having every year from 18 to 89+ is going to be a big table that will also obscure patterns. Instead, you would want to use a variable that has age in categories (e.g., 18-35, 40-64, 65+). If you are using a ratio-level variable like number of siblings, you would likely want to have an upper cut-off like how the above report used "other" for studios, for example "5 or more siblings." If a table looks overwhelming or you have a number of column totals close to 0 (categories with a very low sample size), you may want to consider collapsing variables into fewer categories.

Statistics in Action: Schools, Race, & Security Measures

In this section you'll learn how to run a crosstab and chi-square using SPSS, and also learn about what chi-square is.

This section uses data from the National Crime Victimization Survey's School Crime Supplement, collected by the U.S. Census Bureau in 2019. The sample is representative of its target population, U.S. primary or secondary education (K-12) students ages 12 to 18. Students in private schools were included, but home-schooled children were not. The analyses below are adjusted using provided weighting, which adjusts the sample "to produce estimates of the number of persons ages 12 to 18."

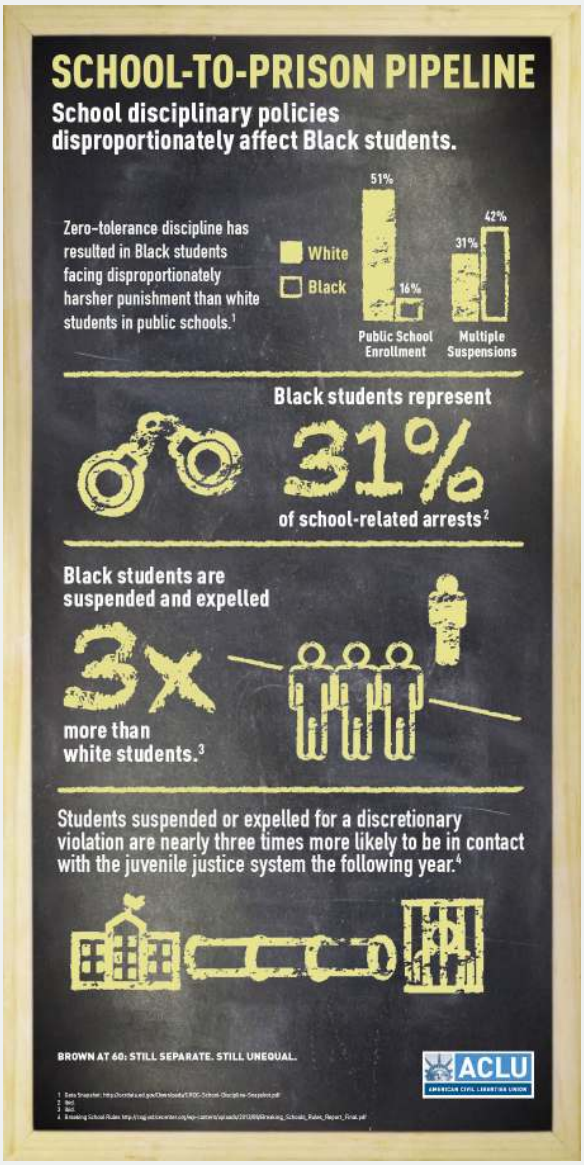
Below you can see frequency tables for race. NH means non-Hispanic. The unweighted sample sizes are the actual number of student respondents. The weighted sample sizes reflect estimates of their actual proportion, e.g., there are just over 16 million

non-Hispanic white students ages 12 to 18 in the United States.

Race	Unweighted		Weighted	
	Frequency	Percent	Frequency	Percent
NH White only	8,856	62.0%	16,061,624	54.9%
NH Black only	1,550	10.9%	3,867,483	13.2%
NH AAPI only	812	5.7%	1,949,552	6.7%
NH U.S. indigenous only	74	0.5%	135,732	0.5%
Hispanic & White only	2,713	19.0%	6,658,166	22.7%
Other Multiracial	268	1.9%	607,863	2.1%
Total	**14,273**	100.0%	**29,280,419**	100.0%

The school-to-prison pipeline is a term describing how "children are funneled out of public schools and into the juvenile and criminal justice systems," disproportionately children of color.

Here is a infographic from the American Civil Liberties Union about the school-to-prison pipeline:



In the School Crime Supplement, students are asked whether their school has security guards or assigned police officers, and whether their school has metal detectors, including wands. Race is a categorical variable (in this case I separated race into 6 categories), and so is whether or not students have these in their schools.

Interpreting Crosstabs

Security guards or assigned police officers (Does your school take any measures to make sure students are safe?) * Race in 6 categories Crosstabulation

			Race in 6 categories						
			NH White only	NH Black only	NH AAPI only	NH U.S. indigenous only	Hispanic & White only	Other Multiracial	Total
Security guards or assigned police officers (Does your school take any measures to make sure students are safe?)	No	Count	3019402	436765	268821	23897	848996	90486	4688367
		% within Race in 6 categories	25.9%	15.4%	19.8%	21.7%	17.9%	21.1%	22.2%
	Yes	Count	8646732	2408131	1086868	86221	3890389	338376	16456717
		% within Race in 6 categories	74.1%	84.6%	80.2%	78.3%	82.1%	78.9%	77.8%
Total		Count	11666134	2844896	1355689	110118	4739385	428862	21145084
		% within Race in 6 categories	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Looking at this crosstab, we can see six different races, as well as a total that includes everyone in the sample. For each racial category, it shows the distribution of whether respondents said they do or do not have a security guard or assigned police officer in their school. With the particular weighting used here, the frequencies are estimates of the target population. So for example, we would estimate that 3,019,402 non-Hispanic white only U.S. students ages 12 to 18 do not have a security guard or assigned police officer in their school (25.9% of respondents, just over 1/4), and that 8,646,732 U.S. students ages 12 to 18 do have a security guard or assigned police officer in their school (71.4% of respondents, just under 3/4).

How to look for and analyze relationships

The table's story includes answers to three questions:

1. **Is there a relationship?**
2. **If so, how substantive it it?**
3. **If so, what is the relationship? (E.g., what is its direction?)**

1. Is there a relationship?

To analyze whether or not there is a relationship, look across rows and see if the percentages are the same or if they are different. Overall, 77.8% of student respondents report having security guards or police officers in their school. If there was no relationship between race and having security guards/police officers in one's school, then we would expect each racial category to have about 77.8% of student respondents reporting this. Look across and see what you notice.

Overall, 77.8% of students ages 12 to 18 are estimated to have a security guard or police officer in their school. That's over 3/4 of students. Each racial group also has somewhere between 74.1% and 84.6% of their respondents who reported having a security guard or police officer in their school. But don't make a common mistake of getting fooled into thinking there is no relationship just because most people have a security guard or police officer in their school, regardless of / across racial categories. We are trying to analyze variation. When I look across, I see that the percentage of students varies by race.

2. How substantive is the relationship?

Small percentage changes indicate a weak relationship. Substantial percentage changes indicate a strong relationship.

The biggest difference I see is that NH Black only respondents are over 10% more likely to report having a security guard or police officer in their school compared to NH White only respondents (84.6% vs. 74.1%). While what constitutes a substantive relationship is subjective, a good rule of thumb is that anytime there is at least a 5% difference to definitely characterize it as being substantial.

3. What is the relationship (e.g., what is its direction)?

If these were ordinal variables (e.g., age and income), we could classify the relationship by direction (e.g., as age increases, income increases). However, race is nominal so we cannot do that.

Looking at the crosstab, I am looking for which racial groups are more likely to have security guards or police officers in their school and which are les likely. I see that over 4/5 of NH Black only, Hispanic & White only, and NH AAPI only student respondents have security guards or assigned police officers in their school, between 3/4 and 4/5 of other multiracial and NH U.S. indigenous students do, and just under 3/4 of NH White only students do.

Try again with the crosstab below. Answer the three questions:

1. **Is there a relationship?** (Do column percentages vary as you move across rows?)
2. **If so, how substantive it it?** (Do they vary a lot?)
3. **If so, what is the relationship? (E.g., what is its direction?)** (Which IV categories have higher and which have lower column percentages?)

Metal detectors, including wands (Does your school take any measures to make sure students are safe?) * Race in 6 categories Crosstabulation

			Race in 6 categories						
			NH White only	NH Black only	NH AAPI only	NH U.S. indigenous only	Hispanic & White only	Other Multiracial	Total
Metal detectors, including wands (Does your school take any measures to make sure students are safe?)	No	Count	10157031	1896204	1028494	97208	3664440	390447	17233824
		% within Race in 6 categories	91.7%	71.6%	84.9%	96.2%	84.4%	92.4%	87.1%
	Yes	Count	913376	750875	182737	3790	679376	31916	2562070
		% within Race in 6 categories	8.3%	28.4%	15.1%	3.8%	15.6%	7.6%	12.9%
Total		Count	11070407	2647079	1211231	100998	4343816	422363	19795894
		% within Race in 6 categories	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Overall, just over 1/10 of student respondents report having metal detectors in their school (about 2.6 million children). When I compare the different column percentages for different racial groups, I see differences, with column percentages ranging from 3.8% to 28.4%. With a difference of over 25% between two of the racial groups, there appears to be a relationship, and it is a substantial one.

- 15.1% of NH AAPI only respondents and 15.6% of Hispanic & White only respondents reported having metal detectors in their schools. These are pretty similar. However, just because there is no substantive difference between two racial groups does not mean there is not an overall relationship between race and having metal detectors in one's school. It just means that the relationship only applies to differences between certain racial groups.

NH Black only students (respondents) are most likely to report having metal detectors in their school. Over 1/4 of these respondents reported this. Over 10% lower are Hispanic & White only and NH AAPI students, where over 1/7 of students reported this. Less than 1/10 of NH White only, other multiracial, and NH U.S. indigenous only students reported this. There are substantive differences. Indeed, while there are over 8 million more NH White only students ages 12-18 in the U.S. than NH Black students, only about 160,000 more NH White only students than NH Black only students are estimated to have metal detectors in their schools. NH Black students are over 20% more likely than NH White students to report having metal detectors in their schools. That's over 1 student out of every 5.

This page titled 8.1: Crosstabs for Research Methods is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Ezra Temko (Consortium of Academic and Research Libraries in Illinois (CARLI)) .

9: Elaborated Crosstabs

Chapter 9 introduces multivariate analysis. We will build on Chapter 8's focus on crosstabs and chi-square by breaking down our crosstabs by a third variable.

Statistics in Action: Schools, Race, & Security Measures

In Chapter 8, we found that, in the United States, as of 2019, non-Hispanic black students were much more likely than non-Hispanic white students to have metal detectors (including wands) in their schools.

Is this a causal relationship? Many U.S.-Americans think we live in a post-racial society where racism no longer exists. Let's say someone claims that the relationship is actually spurious --- that social class is a preceding variable that explains away the relationship. Let's put aside how racism contributes towards a higher proportion of black students being lower-income than white students. We can test whether the black-white metal detector gap is simply due to income differences, or whether race is a contributing factor. The way we do this is to **control** for income. We will hold income constant in order to isolate the effect of race on having a metal detector in school from the effect of income on having a metal detector in school.

Definition: Control variable

Control variable: An additional variable added to a model and held constant in order to improve our evaluation of relationships with the dependent variable.

Research Question for Elaborated Crosstabs

- For our bivariate (two variable) crosstab, we asked, "Is there a relationship between a student's race and whether or not they have metal detectors in their school?"
- When we add a control variable, we add a second question, "**Does this relationship differ by [the control variable]?**". In this case, does the relationship differ by social class?

Chapter 8 involved **bivariate statistics**, because we were using two variables. Here in Chapter 9, we are adding a third variable, the control variable, so we are now engaging in **multivariate statistics**, because we are using multiple, more than two, variables.

When we control for income using crosstabs, we generate a series of **partial crosstabs** of the same race-metal detector relationship. We can evaluate whether the relationship exists and what it looks like among lower-income students, among middle-income students, and among higher-income students. If income explains away the race-metal detector relationship, once we control for income, there should not be a relationship between race and having metal detectors in one's school in any of the partial crosstabs. So for example, holding income constant, if race is not a factor, low-income students should have the same rates of metal detectors in their schools regardless of race / across racial groups.

Definition: Partial (or elaborated) crosstab

Partial crosstabs, also called elaborated crosstabs, are crosstabulations that add a control variable to the original bivariate crosstab, producing separate bivariate crosstabs for each subgroup.

Here we will make a series of crosstabs among income subgroups.

Total: Race (NH black vs. NH white) & Metal Detectors in Schools

First, let's look at our bivariate crosstab and chi-square again. Sometimes you will see a smaller sample size when you look at the overall bivariate crosstab using an elaborated/partial crosstab, because cases are only included if there is useful data for all three variables. In this case, we have income data for all 3,942 students that we had race and metal detector data on, so the sample size is the same.

Metal detectors, including wands (Does your school take any measures to make sure students are safe?) * Race (0=NH White, 1=NH Black) * Household Income Crosstabulation

Household Income				Race (0=NH White, 1=NH Black) NH White	NH Black	Total
Total	Metal detectors, including wands (Does your school take any measures to make sure students are safe?)	No	Count	11,807,300	2,236,513	14,043,813
			% within Race (0=NH White, 1=NH Black)	91.7%	71.4%	87.7%
		Yes	Count	1,064,356	896,775	1,961,131
			% within Race (0=NH White, 1=NH Black)	8.3%	28.6%	12.3%
	Total		Count	12,871,656	3,133,288	16,004,944
			% within Race (0=NH White, 1=NH Black)	100.0%	100.0%	100.0%

Chi-Square Tests

Household Income		Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Total	Pearson Chi-Square	218.512[a]	1	<.001		
	Continuity Correction[b]	216.343	1	<.001		
	Likelihood Ratio	170.260	1	<.001		
	Fisher's Exact Test				<.001	<.001
	Linear-by-Linear Association	218.457	1	<.001		
	N of Valid Cases	3942				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 60.49.
b. Computed only for a 2x2 table
c. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 29.85.
d. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 21.48.
e. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 17.79.
f. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.83.

I am over 99.9% confident there is a relationship among all U.S. students 12 to 18 years old between being NH black vs. NH white and having metal detectors in schools. In the sample, NH black students are over 20% more likely to have metal detectors in their schools compared to NH white students.

Because of the large sample size generated from weighting (since it calculated estimates of all 12 to 18 year old students, over 16 million cases), almost all chi-square tests are going to give me significant results (here it was $p<0.001$, and this continues for all the ones below). Instead, for the chi-square test above (and the ones I show below) I ran chi-square with weighting off to reflect the actual number of respondents.

Household Income <$30,000: Race (NH black vs. NH white) & Metal Detectors in Schools

Now let's look at the same relationship, but just among students whose household income is less than $30,000.

Metal detectors, including wands (Does your school take any measures to make sure students are safe?) * Race (0=NH White, 1=NH Black) * Household Income Crosstabulation

Household Income				Race (0=NH White, 1=NH Black) NH White	NH Black	Total
Less than $30,000	Metal detectors, including wands (Does your school take any measures to make sure students are safe?)	No	Count	1,300,180	675,576	1,975,756
			% within Race (0=NH White, 1=NH Black)	91.6%	66.5%	81.1%
		Yes	Count	119,053	340,183	459,236
			% within Race (0=NH White, 1=NH Black)	8.4%	33.5%	18.9%
	Total		Count	1,419,233	1,015,759	2,434,992
			% within Race (0=NH White, 1=NH Black)	100.0%	100.0%	100.0%

Chi-Square Tests

Household Income		Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Less than $30,000	Pearson Chi-Square	61.524[c]	1	<.001		
	Continuity Correction[b]	59.625	1	<.001		
	Likelihood Ratio	57.774	1	<.001		
	Fisher's Exact Test				<.001	<.001
	Linear-by-Linear Association	61.414	1	<.001		
	N of Valid Cases	561				

For this partial crosstab, our target population is not all U.S. students 12-18 years old. It is a subset of this population: all U.S. students 12-18 years old who have household incomes less than $30,000. We are over 99.9% confident that, among U.S. students 12-18 years old who have household incomes less than $30,000, there is a relationship between being NH white vs. black and whether or not one has metal detectors in their school. In the sample, NH black students (from households with incomes <$30K) are over 25% more likely than NH white students (from households with incomes <$30K) to have metal detectors in their schools. We now know that the race gap for metal detectors in schools is not simply a matter of income. Among lower income households, there is still a race gap between NH white and NH black students. In fact, it got wider, not smaller.

We can also compare our first and second crosstab, and observe that, compared to all students in the sample, NH white students from households with less than $30,000 have about the same likelihood of having metal detectors in their school (8.4% vs 8.3%). However, in the sample 7% more poor NH black students have metal detectors in their schools compared to NH black students as a whole (33.5% vs. 28.6%). The race gap actually got wider when we looked at poor students. It didn't go away.

Household Income $30,000 to $49,999: Race (NH black vs. NH white) & Metal Detectors in Schools

Now let's look at the same relationship, but just among students whose household income is at least $30K but less than $50K.

Metal detectors, including wands (Does your school take any measures to make sure students are safe?) * Race (0=NH White, 1=NH Black) * Household Income Crosstabulation

Household Income				Race (0=NH White, 1=NH Black) NH White	NH Black	Total
$30,000 to $49,999	Metal detectors, including wands (Does your school take any measures to make sure students are safe?)	No	Count	1,802,867	517,663	2,320,530
			% within Race (0=NH White, 1=NH Black)	90.6%	63.2%	82.6%
		Yes	Count	186,563	300,994	487,557
			% within Race (0=NH White, 1=NH Black)	9.4%	36.8%	17.4%
	Total		Count	1,989,430	818,657	2,808,087
			% within Race (0=NH White, 1=NH Black)	100.0%	100.0%	100.0%

Chi-Square Tests

Household Income		Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
$30,000 to $49,999	Pearson Chi-Square	61.343[d]	1	<.001		
	Continuity Correction[b]	59.283	1	<.001		
	Likelihood Ratio	52.481	1	<.001		
	Fisher's Exact Test				<.001	<.001
	Linear-by-Linear Association	61.249	1	<.001		
	N of Valid Cases	648				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 60.49.
b. Computed only for a 2x2 table
c. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 29.85.
d. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 21.48.
e. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 17.79.
f. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.83.

For this partial crosstab, our target population is not all U.S. students 12-18 years old. It is a subset of this population: all U.S. students 12-18 years old who have household incomes in the $30Ks and $40Ks. We are over 99.9% confident that, among U.S. students 12-18 years old who have household incomes between $30,000 and $49,999, there is a relationship between being NH white vs. black and whether or not one has metal detectors in their school. In the sample, NH black students (from households with incomes $30K to <$50K) are over 25% more likely than NH white students (from households with incomes $30K to <$50K) to have metal detectors in their schools. We again find here the race gap for metal detectors in schools is not simply a matter of income. Among households making between $30,000 and $49,999, there is still a race gap between NH white and NH black students.

Household Income $50,000 to $99,999: Race (NH black vs. NH white) & Metal Detectors in Schools

Now let's look at the same relationship, but just among students whose household income is at least $50K but less than $100K.

Metal detectors, including wands (Does your school take any measures to make sure students are safe?) * Race (0=NH White, 1=NH Black) * Household Income Crosstabulation

Household Income				Race (0=NH White, 1=NH Black) NH White	NH Black	Total
$50,000 to $99,999	Metal detectors, including wands (Does your school take any measures to make sure students are safe?)	No	Count	3,989,094	701,459	4,690,553
			% within Race (0=NH White, 1=NH Black)	91.5%	76.2%	88.8%
		Yes	Count	370,906	218,748	589,654
			% within Race (0=NH White, 1=NH Black)	8.5%	23.8%	11.2%
	Total		Count	4,360,000	920,207	5,280,207
			% within Race (0=NH White, 1=NH Black)	100.0%	100.0%	100.0%

Chi-Square Tests

Household Income		Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
$50,000 to $99,999	Pearson Chi-Square	39.011[e]	1	<.001		
	Continuity Correction[b]	37.348	1	<.001		
	Likelihood Ratio	31.325	1	<.001		
	Fisher's Exact Test				<.001	<.001
	Linear-by-Linear Association	38.980	1	<.001		
	N of Valid Cases	1281				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 60.49.
b. Computed only for a 2x2 table
c. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 29.85.
d. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 21.48.
e. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 17.79.
f. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.83.

For this partial crosstab, our target population is not all U.S. students 12-18 years old. It is a subset of this population: all U.S. students 12-18 years old who have household incomes between $50,000 and $99,999. We are over 99.9% confident that, among U.S. students 12-18 years old who have household incomes between $50,000 and $99,999, there is a relationship between being NH white vs. black and whether or not one has metal detectors in their school. In the sample, NH black students (from households with incomes $50K to <$100K) are over 15% more likely than NH white students (from households with incomes $50K to <$100K) to have metal detectors in their schools. We again find here the race gap for metal detectors in schools is not simply a matter of income. Among households making between $50,000 and $99,999, there is still a race gap between NH white and NH black students.

Household Income $100,000+: Race (NH black vs. NH white) & Metal Detectors in Schools

Now let's look at the same relationship, but just among students whose household income is at least $100K.

Metal detectors, including wands (Does your school take any measures to make sure students are safe?) * Race (0=NH White, 1=NH Black) * Household Income Crosstabulation

Household Income				Race (0=NH White, 1=NH Black) NH White	NH Black	Total
$100,000+	Metal detectors, including wands (Does your school take any measures to make sure students are safe?)	No	Count	4,715,159	341,815	5,056,974
			% within Race (0=NH White, 1=NH Black)	92.4%	90.3%	92.3%
		Yes	Count	387,834	36,850	424,684
			% within Race (0=NH White, 1=NH Black)	7.6%	9.7%	7.7%
	Total		Count	5,102,993	378,665	5,481,658
			% within Race (0=NH White, 1=NH Black)	100.0%	100.0%	100.0%

Chi-Square Tests

Household Income		Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
$100,000+	Pearson Chi-Square	1.104[f]	1	.293		
	Continuity Correction[b]	.654	1	.419		
	Likelihood Ratio	.984	1	.321		
	Fisher's Exact Test				.327	.202
	Linear-by-Linear Association	1.103	1	.294		
	N of Valid Cases	1452				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 60.49.
b. Computed only for a 2x2 table
c. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 29.85.
d. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 21.48.
e. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 17.79.
f. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.83.

For this partial crosstab, our target population is not all U.S. students 12-18 years old. It is a subset of this population: all U.S. students 12-18 years old who have household incomes $100,000 or higher. Our p-value (unweighted) is p=0.293. We are not over 95% confident that, among U.S. students 12-18 years old who have household incomes $100,000 or higher, there is a relationship between being NH white vs. black and whether or not one has metal detectors in their school. We could be making a Type 2 error (false negative), as there are only 73 NH black students in our sample who have household incomes $100,000 or higher (1,379 NH white students). However, we cannot claim, based on this data, that there is a NH black-NH white race gap among students from households making $100,000 or higher. In the sample, NH black students (from households with incomes $100K+) are just over 2% more likely than NH white students (from households with incomes $100K+) to have metal detectors in their schools.

Putting It Together

Income did not make the relationship between race and having a metal detector spurious. It did, however, moderate it. Overall in our sample, about 1 in 5 (20.3%) more NH black students had metal detectors in their schools compared to NH white students. However, for those from households with incomes less than $30,000 or $30,000 to $49,999, this was about 1 in 20 students higher (about 25%), for students from $50,000 to $99,999, about 1 in 20 students fewer (about 15%), and among students $100,000, the sample difference was about 1 in 50 (2.1%) students, and we were not confident that there is any difference for this subgroup. It may be that wealthy students, regardless of race, are less likely to have metal detectors in their schools, but that among lower-income and middle-income students, there is a sizeable race gap.

Elaborated Crosstab:

- ***IV: Race (NH black vs. NH white)***
- ***DV: Metal Detectors in Schools***
- ***Control Variable: Household income***

The SPSS outputs above started with an overall crosstab between the IV and DV and then showed partial crosstabs, with the same IV and DV but broken down among subsamples based on household income. When you run an elaborated crosstab in SPSS, all this information comes out in one table, all together.

Here are the series of crosstabs, for each income group and then the "total."

Metal detectors, including wands (Does your school take any measures to make sure students are safe?) * Race (0=NH White, 1=NH Black) * Household Income Crosstabulation

Household Income				Race (0=NH White, 1=NH Black) NH White	NH Black	Total
Less than $30,000	Metal detectors, including wands (Does your school take any measures to make sure students are safe?)	No	Count	1300180	675576	1975756
			% within Race (0=NH White, 1=NH Black)	91.6%	66.5%	81.1%
		Yes	Count	119053	340183	459236
			% within Race (0=NH White, 1=NH Black)	8.4%	33.5%	18.9%
	Total		Count	1419233	1015759	2434992
			% within Race (0=NH White, 1=NH Black)	100.0%	100.0%	100.0%
$30,000 to $49,999	Metal detectors, including wands (Does your school take any measures to make sure students are safe?)	No	Count	1802867	517663	2320530
			% within Race (0=NH White, 1=NH Black)	90.6%	63.2%	82.6%
		Yes	Count	186563	300994	487557
			% within Race (0=NH White, 1=NH Black)	9.4%	36.8%	17.4%
	Total		Count	1989430	818657	2808087
			% within Race (0=NH White, 1=NH Black)	100.0%	100.0%	100.0%
$50,000 to $99,999	Metal detectors, including wands (Does your school take any measures to make sure students are safe?)	No	Count	3989094	701459	4690553
			% within Race (0=NH White, 1=NH Black)	91.5%	76.2%	88.8%
		Yes	Count	370906	218748	589654
			% within Race (0=NH White, 1=NH Black)	8.5%	23.8%	11.2%
	Total		Count	4360000	920207	5280207
			% within Race (0=NH White, 1=NH Black)	100.0%	100.0%	100.0%
$100,000+	Metal detectors, including wands (Does your school take any measures to make sure students are safe?)	No	Count	4715159	341815	5056974
			% within Race (0=NH White, 1=NH Black)	92.4%	90.3%	92.3%
		Yes	Count	387834	36850	424684
			% within Race (0=NH White, 1=NH Black)	7.6%	9.7%	7.7%
	Total		Count	5102993	378665	5481658
			% within Race (0=NH White, 1=NH Black)	100.0%	100.0%	100.0%
Total	Metal detectors, including wands (Does your school take any measures to make sure students are safe?)	No	Count	11807300	2236513	14043813
			% within Race (0=NH White, 1=NH Black)	91.7%	71.4%	87.7%
		Yes	Count	1064356	896775	1961131
			% within Race (0=NH White, 1=NH Black)	8.3%	28.6%	12.3%
	Total		Count	12871656	3133288	16004944
			% within Race (0=NH White, 1=NH Black)	100.0%	100.0%	100.0%

Here are the chi-square outputs. SPSS runs a separate chi-square test for each of the (five above) crosstabs.

Chi-Square Tests

Household Income		Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Less than $30,000	Pearson Chi-Square	61.524[c]	1	<.001		
	Continuity Correction[b]	59.625	1	<.001		
	Likelihood Ratio	57.774	1	<.001		
	Fisher's Exact Test				<.001	<.001
	Linear-by-Linear Association	61.414	1	<.001		
	N of Valid Cases	561				
$30,000 to $49,999	Pearson Chi-Square	61.343[d]	1	<.001		
	Continuity Correction[b]	59.283	1	<.001		
	Likelihood Ratio	52.481	1	<.001		
	Fisher's Exact Test				<.001	<.001
	Linear-by-Linear Association	61.249	1	<.001		
	N of Valid Cases	648				
$50,000 to $99,999	Pearson Chi-Square	39.011[e]	1	<.001		
	Continuity Correction[b]	37.348	1	<.001		
	Likelihood Ratio	31.325	1	<.001		
	Fisher's Exact Test				<.001	<.001
	Linear-by-Linear Association	38.980	1	<.001		
	N of Valid Cases	1281				
$100,000+	Pearson Chi-Square	1.104[f]	1	.293		
	Continuity Correction[b]	.654	1	.419		
	Likelihood Ratio	.984	1	.321		
	Fisher's Exact Test				.327	.202
	Linear-by-Linear Association	1.103	1	.294		
	N of Valid Cases	1452				
Total	Pearson Chi-Square	218.512[a]	1	<.001		
	Continuity Correction[b]	216.343	1	<.001		
	Likelihood Ratio	170.260	1	<.001		
	Fisher's Exact Test				<.001	<.001
	Linear-by-Linear Association	218.457	1	<.001		
	N of Valid Cases	3942				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 60.49.
b. Computed only for a 2x2 table
c. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 29.85.
d. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 21.48.
e. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 17.79.
f. 1 cells (25.0%) have expected count less than 5. The minimum expected count is 4.83.

Given the length of the chi-square outputs, you may want to delete extra information so it is easier to look at and make comparisons. Remember that in SPSS you can double-click on outputs and then highlight and delete rows, columns, etc., that you do not want to appear. Here is a simplified version of the chi-square output, where I can focus on the p-values for each chi-square test.

Chi-Square Tests

Household Income		Value	df	Asymptotic Significance (2-sided)
Less than $30,000	Pearson Chi-Square	61.524	1	<.001
	N of Valid Cases	561		
$30,000 to $49,999	Pearson Chi-Square	61.343	1	<.001
	N of Valid Cases	648		
$50,000 to $99,999	Pearson Chi-Square	39.011	1	<.001
	N of Valid Cases	1281		
$100,000+	Pearson Chi-Square	1.104	1	.293
	N of Valid Cases	1452		
Total	Pearson Chi-Square	218.512	1	<.001
	N of Valid Cases	3942		

SPSS: Running elaborated crosstabs and chi-square in SPSS

Follow the same process you did in Chapter 8, with one additional step (highlighted in yellow below).

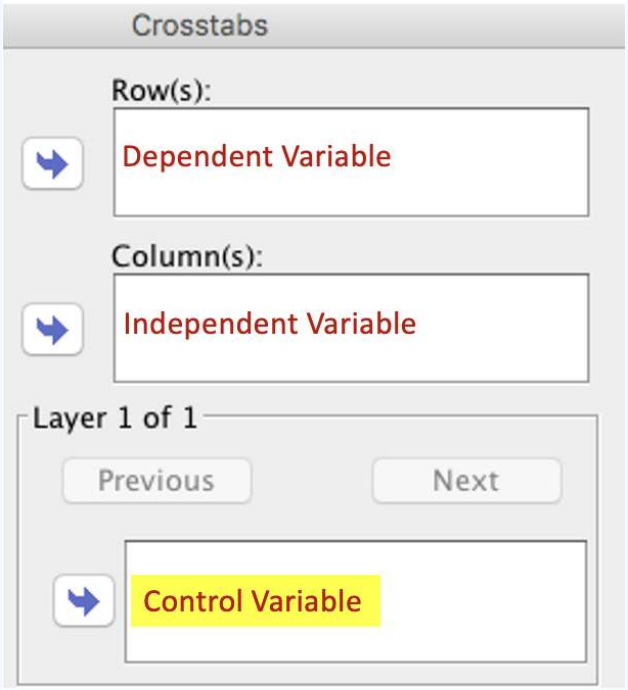
Syntax:

```
CROSSTABS
/TABLES=DependentVariableName BY IndependentVariableName BY ControlVariableName
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ
/CELLS=COUNT COLUMN
/COUNT ROUND CELL.
```

Menu:

First, go to Analyze → Descriptive Statistics → Crosstabs.

Next, put your independent variable into the "Column(s)" box, your dependent variable into the "Row(s)" box, and your control variable into the "Layer 1 of 1" box.



After that, click the "Cells" box on the right side. In the window that pops up, make sure that "Observed" is checked under "Counts." Under "Percentages," check the box for "Column." Don't check any other boxes. Click the "Continue" button.

After that, go to the "Statistics" box on the right side. Check the box for "Chi-square" and click "Continue."

Finally, click "OK."

Extension: Impact of control variable on initial bivariate crosstab

In the example above, we added a control variable, analyzed our initial bivariate crosstab and chi-square test, our partial crosstabs and chi-square tests, and then compared the bivariate model to the multivariate model. From this we figured out what happened to our original relationship after we added the control variable.

While you can analyze what happens when you add a control variable this way, there is a more systematic way to do so.

When you add a control variable, there are five potential outcomes:

- the initial relationship stays the same
- the initial relationship goes away
- the initial relationship diminishes
- the initial relationship varies
- the initial relationship increases (or there was no relationship, and now a relationship appears)

Interpreting the effects of adding a control variable can further depend on whether the control variable is an **antecedent** variable or **intervening** variable.

Definition: Antecedent variable

Antecedent variables are variables that precede, that come before, both the independent and dependent variables.

Antecedent → Independent → Dependent
Example: Degree → Occupation → Salary

Definition: Intervening variable

Intervening variables are variables that come in between the independent and dependent variables.

Independent → Intervening → Dependent
Example: Going to a restaurant → Eating → Satiety

Let's take a made-up example with made-up data of a bivariate relationship between finding a four-leaf clover and winning raffles. In our bivariate model, people who found a 4-leaf clover in the past year were also more likely to win a raffle in the past year.

		Found 4-leaf clover in past year?	
		Yes	No
Won raffle in past year?	Yes	60%	40%
	No	40%	60%
	Total	100%	100%

We'll consider two different control variables.

- The first is an antecedent variable, luck. Maybe lucky people are more likely to find 4-leaf clovers and win the raffle, and unlucky people are less likely to do either?
- The second is an intervening variable, praying to win a raffle. Maybe people who find a 4-leaf clover decide to put their luck to use, pray to win a raffle, and through doing so end up more likely to win a raffle.

The relationship stays the same → **Replication**
If the bivariate relationship stays the same, then the relationship still exists, even after controlling for the additional variable.

			Found 4-leaf clover in past	
			Yes	No
Lucky	Won raffle in past	Yes	60%	40%
		No	40%	60%
		Total	100%	100%
Unlucky	Won raffle in past	Yes	60%	40%
		No	40%	60%
		Total	100%	100%
Total	Won raffle in past year?	Yes	60%	40%
		No	40%	60%
		Total	100%	100%

There was no change after controlling for luck. Luck does not explain the relationship.

			Found 4-leaf clover in past	
			Yes	No
Prayed	Won raffle in past	Yes	60%	40%
		No	40%	60%
		Total	100%	100%
Did not pray	Won raffle in past	Yes	60%	40%
		No	40%	60%
		Total	100%	100%
Total	Won raffle in past year?	Yes	60%	40%
		No	40%	60%
		Total	100%	100%

There was no change after controlling for praying. Praying does not affect the relationship.

The relationship goes away

- ***Antecedent*** **= Spurious**

With an antecedent variable, if the bivariate relationship goes away, then the relationship is explained by the control (also called *confounding*) variable.

			Found 4-leaf clover in past	
			Yes	No
Lucky	Won raffle in past	Yes	51%	49%
		No	49%	51%
		Total	100%	100%
Unlucky	Won raffle in past	Yes	50%	50%
		No	50%	50%
		Total	100%	100%
Total	Won raffle in past	Yes	60%	40%
		No	40%	60%
		Total	100%	100%

The relationship went away. The relationship between finding a 4-leaf clover and winning a raffle is spurious. Luck explains the relationship.

- ***Intervening*** **= Mediation**

If the bivariate relationship goes away, then the relationship's mechanism is explained by the control variable.

			Found 4-leaf clover in past	
			Yes	No
Prayed	Won raffle in past	Yes	51%	49%
		No	49%	51%
		Total	100%	100%
Did not pray	Won raffle in past	Yes	50%	50%
		No	50%	50%
		Total	100%	100%
Total	Won raffle in past	Yes	60%	40%
		No	40%	60%
		Total	100%	100%

The relationship went away. Praying *mediates* (explains) the relationship. Praying is the mechanism through which finding a 4-leaf clover turns into winning a raffle. People who find 4 leaf clovers then pray to win raffles and then (perhaps enter more raffles and then) win raffles. The relationship is not spurious, because praying did not come before finding the 4-leaf clover.

The relationship decreases → Partially explained

In these examples, the relationship still exists, but it is less substantive than it was prior to introducing the control variable. In these cases, the control variable partially explains the relationships.

Antecedent:

			Found 4-leaf clover in past	
			Yes	No
Lucky	Won raffle in past	Yes	55%	50%
		No	45%	50%
		Total	100%	100%
Unlucky	Won raffle in past	Yes	53%	48%
		No	47%	52%
		Total	100%	100%
Total	Won raffle in past	Yes	60%	40%
		No	40%	60%
		Total	100%	100%

The relationship is not as substantive. The relationship is partially explained by the control variable, but the IV still has its own independent effect. Luck partially explains the relationship.

Intervening:

			Found 4-leaf clover in past	
			Yes	No
Prayed	Won raffle in past	Yes	55%	50%
		No	45%	50%
		Total	100%	100%
Did not pray	Won raffle in past	Yes	53%	48%
		No	47%	52%
		Total	100%	100%
Total	Won raffle in past	Yes	60%	40%
		No	40%	60%
		Total	100%	100%

The relationship is not as substantive. The way that finding a four-leaf clover translates into winning a raffle is partially explained by praying, but there are other mechanisms at play as well.

The relationship varies → **Moderation**

If the bivariate relationship varies / is different among subgroups, then the relationship is moderated by the control variable, meaning the control variable affects the relationship. For different control variable values, the relationship may be smaller, larger, etc.

			Found 4-leaf clover in past	
			Yes	No
Lucky	Won raffle in past	Yes	75%	25%
		No	25%	75%
		Total	100%	100%
Unlucky	Won raffle in past	Yes	58%	42%
		No	42%	58%
		Total	100%	100%
Total	Won raffle in past	Yes	60%	40%
		No	40%	60%
		Total	100%	100%

The relationship changes depending on whether someone is lucky or not. Luck *moderates* the relationship. The relationship looks different for folks who are lucky (where finding a 4 leaf clover is even more substantially associated with winning a raffle) and folks who are not lucky (finding a 4-leaf clover helps, but not at all as much).

			Found 4-leaf clover in past	
			Yes	No
Prayed	Won raffle in past	Yes	75%	25%
		No	25%	75%
		Total	100%	100%
Did not pray	Won raffle in past	Yes	58%	42%
		No	42%	58%
		Total	100%	100%
Total	Won raffle in past	Yes	60%	40%
		No	40%	60%
		Total	100%	100%

The relationship changes depending on whether someone prays to win a raffle or not. Praying *moderates* the relationship. The relationship looks different for folks who do and do not pray to win.

Relationship increases or appears

- ***Relationship increases***

			Found 4-leaf clover in past	
			Yes	No
Lucky	Won raffle in past	Yes	80%	20%
		No	20%	80%
		Total	100%	100%
Unlucky	Won raffle in past	Yes	80%	20%
		No	20%	80%
		Total	100%	100%
Total	Won raffle in past	Yes	60%	40%
		No	40%	60%
		Total	100%	100%

The effect of the finding clovers increased when controlling for luck, because luck un-supresses some of the power of finding a four leaf clover. Finding a clover especially makes a difference among people who are lucky, and also among people who are unlucky. After controlling, the relationship grows stronger/larger.

			Found 4-leaf clover in past	
			Yes	No
Prayed	Won raffle in past	Yes	80%	20%
		No	20%	80%
		Total	100%	100%
Did not pray	Won raffle in past	Yes	80%	20%
		No	20%	80%
		Total	100%	100%
Total	Won raffle in past	Yes	60%	40%
		No	40%	60%
		Total	100%	100%

The effect of the finding clovers increasing when controlling, because prayer un-supresses some of the power of finding a four leaf clover. Finding a clover especially makes a difference among people who prayed, and also among people who did not pray. After controlling, the relationship grows stronger/larger.

- ***Relationship appears***
  What if we had initially found no relationship between finding a 4-leaf clover and winning a raffle? Then, after adding our control variable, the relationship appeared?

			Found 4-leaf clover in past	
			Yes	No
Lucky	Won raffle in past	Yes	60%	40%
		No	40%	60%
		Total	100%	100%
Unlucky	Won raffle in past	Yes	60%	40%
		No	40%	60%
		Total	100%	100%
Total	Won raffle in past	Yes	51%	49%
		No	49%	51%
		Total	100%	100%

The effect of the finding clovers was concealed, but controlling for luck revealed its effect, un-supressing the underlying pattern. The relationship has emerged, in context. Finding a 4-leaf clover makes a difference among those who are lucky and among those who are not lucky.

			Found 4-leaf clover in past	
			Yes	No
Prayed	Won raffle in past	Yes	60%	40%
		No	40%	60%
		Total	100%	100%
Did not pray	Won raffle in past	Yes	60%	40%
		No	40%	60%
		Total	100%	100%
Total	Won raffle in past	Yes	51%	49%
		No	49%	51%
		Total	100%	100%

The effect of the finding clovers was concealed, but controlling for prayer revealed its effect/un-supressed the underlying pattern. The relationship has emerged, in context. Finding a 4-leaf clover makes a difference among those who pray and among those who don‘t.

Suppression effects do actually happen. For example, when I was analyzing support for public/employer sourcing of childcare funding using the 2012 General Social Survey, in my initial bivariate analysis, there was not a significant relationship between gender and support for public or employer sourcing of childcare funding. However, once I controlled for family income and political ideology, there was a relationship --- and men were more supportive! This relationship between gender and support for public or employer sourcing of childcare funding existed --- among the control variable groups. For example, if you hold

family income and political ideology constant, say liberals with low family incomes, you'll find men are slightly more supportive. However, when I looked at the bivariate relationship, this is not what I found. That's because women tend to be more liberal and less conservative, and tend to have lower family incomes. So when I'm comparing gender among all U.S. adults, women's decreased support that I found after controlling for the other variables was balanced out by there being more women who are liberal and low-income who have increased support. Here income and political ideology are suppressor variables.

Summary of Control Variable Effects

- The initial relationship stays the same
  - Replication: relationship still exists
- The initial relationship goes away
  - If control variable is antecedent
    - Spurious: Relationship explained by control (confounding) variable
  - If control variable is intervening
    - Relationship's mechanism explained by control (mediating) variable
- The initial relationship diminishes
  - If control variable is antecedent
    - Relationship is partially explained by control variable, but original independent variable still has its own independent effect
  - If control variable is intervening
    - Relationship's mechanism is partially but not fully explained by control variable
- The initial relationship varies
  - Moderation: relationship differs/interacts with control variable
- The initial relationship increases (or there was no relationship, and now a relationship appears)
  - Suppression: Relationship was suppressed, relationship is explained by its effect within control variable groups

Statistics in Action: Feelings towards Black Lives Matter

Let's take a look at one final example. This example comes from the ANES survey following the 2020 election, where eligible voters were asked how they felt about Black Lives Matter, using a feelings thermometer ranging from 0 to 100:



I collapsed the feelings thermometer into 5 categories: 0° (coldest), 1° to 49° (cold), 50° (neutral), 51° to 99° (warm), and 100° (warmest).

My independent variable is political ideology, measures as conservative, moderate, or liberal. Presumably the more liberal one is, the more likely one will be supportive of Black Lives Matter. However, I am going to add a control variable from a survey

question that asks, "How much discrimination have you personally faced because of your ethnicity or race?" To what extent will people's personal experiences of racism impact their views on Black Lives Matter, regardless of their political ideology?

Here is the SPSS output:

Feelings towards Black Lives Matter movement * Political Views * How much discrimination have you personally faced because of your ethnicity or race? Crosstabulation

How much discrimination have you personally faced because of your ethnicity or race?				Political Views: Liberal	Moderate	Conservative	Total
None at all	Feelings towards Black Lives Matter movement	Coldest	Count	3	16	175	194
			% within Political Views	0.9%	8.1%	41.3%	20.0%
		Cold	Count	18	40	147	205
			% within Political Views	5.2%	20.3%	34.7%	21.2%
		Neutral	Count	25	42	46	113
			% within Political Views	7.2%	21.3%	10.8%	11.7%
		Warm	Count	195	81	49	325
			% within Political Views	56.2%	41.1%	11.6%	33.6%
		Warmest	Count	106	18	7	131
			% within Political Views	30.5%	9.1%	1.7%	13.5%
	Total		Count	347	197	424	968
			% within Political Views	100.0%	100.0%	100.0%	100.0%
A little	Feelings towards Black Lives Matter movement	Coldest	Count	5	16	142	163
			% within Political Views	2.8%	8.6%	44.7%	23.8%
		Cold	Count	15	45	102	162
			% within Political Views	8.3%	24.3%	32.1%	23.7%
		Neutral	Count	19	10	28	57
			% within Political Views	10.5%	5.4%	8.8%	8.3%
		Warm	Count	88	99	41	228
			% within Political Views	48.6%	53.5%	12.9%	33.3%
		Warmest	Count	54	15	5	74
			% within Political Views	29.8%	8.1%	1.6%	10.8%
	Total		Count	181	185	318	684
			% within Political Views	100.0%	100.0%	100.0%	100.0%

A moderate amount	Feelings towards Black Lives Matter movement	Coldest	Count	5	18	38	61
			% within Political Views	5.6%	15.9%	38.0%	20.2%
		Cold	Count	4	12	27	43
			% within Political Views	4.5%	10.6%	27.0%	14.2%
		Neutral	Count	3	18	14	35
			% within Political Views	3.4%	15.9%	14.0%	11.6%
		Warm	Count	47	48	15	110
			% within Political Views	52.8%	42.5%	15.0%	36.4%
		Warmest	Count	30	17	6	53
			% within Political Views	33.7%	15.0%	6.0%	17.5%
	Total		Count	89	113	100	302
			% within Political Views	100.0%	100.0%	100.0%	100.0%
A lot or a great deal	Feelings towards Black Lives Matter movement	Coldest	Count	1	5	20	26
			% within Political Views	1.4%	7.9%	40.0%	14.2%
		Cold	Count	1	5	9	15
			% within Political Views	1.4%	7.9%	18.0%	8.2%
		Neutral	Count	1	9	4	14
			% within Political Views	1.4%	14.3%	8.0%	7.7%
		Warm	Count	28	17	12	57
			% within Political Views	40.0%	27.0%	24.0%	31.1%
		Warmest	Count	39	27	5	71
			% within Political Views	55.7%	42.9%	10.0%	38.8%
	Total		Count	70	63	50	183
			% within Political Views	100.0%	100.0%	100.0%	100.0%
Total	Feelings towards Black Lives Matter movement	Coldest	Count	14	55	375	444
			% within Political Views	2.0%	9.9%	42.0%	20.8%
		Cold	Count	38	102	285	425
			% within Political Views	5.5%	18.3%	32.0%	19.9%
		Neutral	Count	48	79	92	219
			% within Political Views	7.0%	14.2%	10.3%	10.2%
		Warm	Count	358	245	117	720
			% within Political Views	52.1%	43.9%	13.1%	33.7%
		Warmest	Count	229	77	23	329
			% within Political Views	33.3%	13.8%	2.6%	15.4%
	Total		Count	687	558	892	2137
			% within Political Views	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

How much discrimination have you personally faced because of your ethnicity or race?		Value	df	Asymptotic Significance (2-sided)
None at all	Pearson Chi-Square	512.036[b]	8	<.001
	N of Valid Cases	968		
A little	Pearson Chi-Square	302.107[c]	8	<.001
	N of Valid Cases	684		
A moderate amount	Pearson Chi-Square	93.557[d]	8	<.001
	N of Valid Cases	302		
A lot or a great deal	Pearson Chi-Square	69.207[e]	8	<.001
	N of Valid Cases	183		
Total	Pearson Chi-Square	933.121[a]	8	<.001
	N of Valid Cases	2137		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 57.18.
b. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 23.00.
c. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 15.08.
d. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.31.
e. 3 cells (20.0%) have expected count less than 5. The minimum expected count is 3.83.

From the chi-square test, we can see that we are over 99.9% confident there is a relationship between political ideology and feelings towards BLM among U.S. eligible voters. We can see that this relationship also holds among subgroups of U.S.

eligible voters who have experienced different levels of racism. For example, even among those who have experienced "a lot or a great deal" of racism, we are confident there is a relationship between political ideology and feelings towards BLM.

Turning to the crosstab, let's examine the actual relationship and what story the table tells us. It's always a good idea to start with the "Total" section of the table at the bottom to get a sense of the bivariate relationship.

As expected, generally the more liberal one is the warmer their feelings towards BLM. We can see that while 2.6% of conservatives have the warmest (100 degree) feelings towards BLM, this increases to 13.8% for moderates and 33.3% for liberals. Conversely, while 2.0% of liberals have the coldest (0 degree) feelings towards BLM, this increases to 9.9% for moderates and 33.3% for liberals.

Looking at all the partial crosstabs, these directions hold out among all subgroups of eligible voters with different levels of experiences of racism. However, there are some differences.

Let's zoom in on the column percentages for those who have the coldest and the warmest feelings:

Feelings towards Black Lives Matter movement * Political Views * How much discrimination have you personally faced because of your ethnicity or race? Crosstabulation

How much discrimination have you personally faced because of your ethnicity or race?			Political Views			
			Liberal	Moderate	Conservative	Total
None at all	Feelings towards Black Lives Matter movement	Coldest	0.9%	8.1%	41.3%	20.0%
		Warmest	30.5%	9.1%	1.7%	13.5%
A little	Feelings towards Black Lives Matter movement	Coldest	2.8%	8.6%	44.7%	23.8%
		Warmest	29.8%	8.1%	1.6%	10.8%
A moderate amount	Feelings towards Black Lives Matter movement	Coldest	5.6%	15.9%	38.0%	20.2%
		Warmest	33.7%	15.0%	6.0%	17.5%
A lot or a great deal	Feelings towards Black Lives Matter movement	Coldest	1.4%	7.9%	40.0%	14.2%
		Warmest	55.7%	42.9%	10.0%	38.8%
Total	Feelings towards Black Lives Matter movement	Coldest	2.0%	9.9%	42.0%	20.8%
		Warmest	33.3%	13.8%	2.6%	15.4%

Those who have not experienced any racism are pretty similar to those who have experienced a little racism. Liberals don't have much change once we get to a moderate amount of racism, but there's a jump among moderates both in having the coldest and the warmest feelings towards BLM, and while only at 6%, the percentage of conservatives with the warmest feelings towards BLM more than triples. Finally, for those who have experienced a lot or a great deal of racism, we see a substantial increase among liberals and among moderates in having the warmest feelings (for liberals from close to 1/3 to a majority, for moderates from 8.1% to 15.0% to 42.9%). Conservatives see their highest proportion having the warmest feelings towards conservatives (1/10), though across levels of experiences of racism, about 2/5 of conservatives still have the coldest feelings towards BLM.

This page titled 9: Elaborated Crosstabs is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Ezra Temko (Consortium of Academic and Research Libraries in Illinois (CARLI)) .

10: Chapter 10. Scatterplots and Correlation

The last two chapters dealt with relationships between categorical variables. But what about when you have ratio-level variables? The next six chapters all deal with ratio-level or ratio-like dependent variables. Chapter 10 focuses on bivariate (two-variable) relationships and covers scatterplots and correlation, and also introduces regression lines.

Scatterplots

Statistics in Action: U.S. State Poverty Rates & Life Expectancy

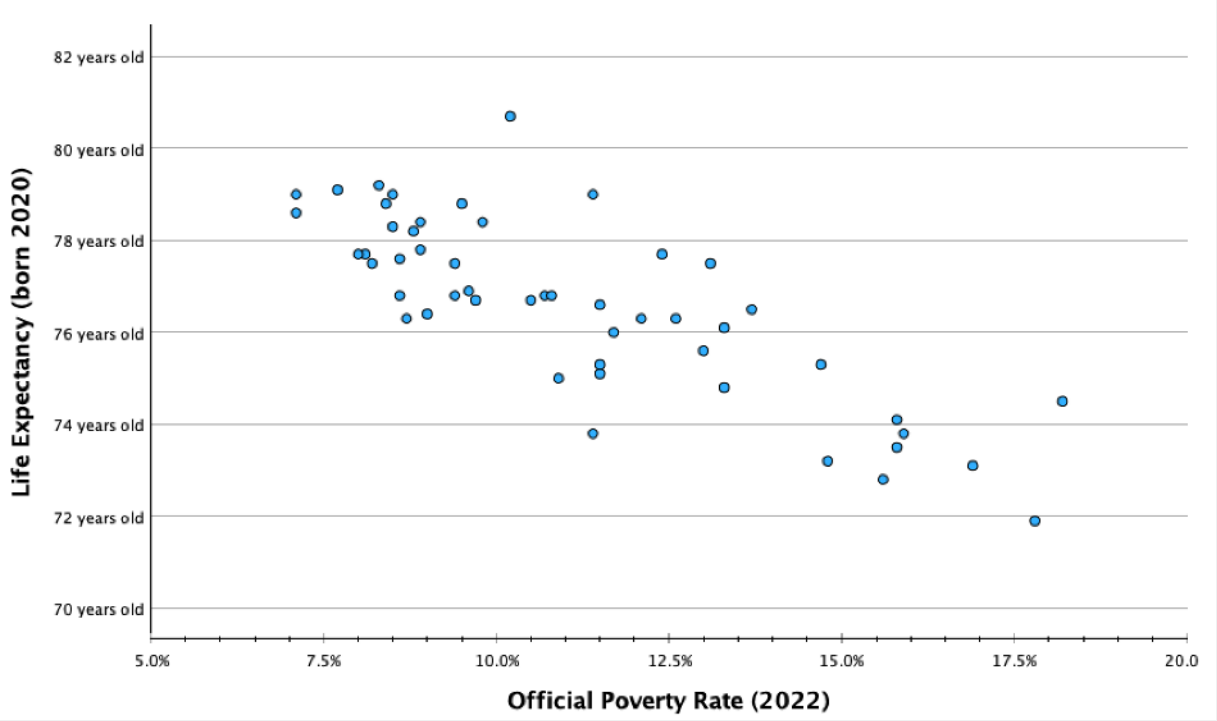
Have you heard the notion of "health = wealth"?

Below we will compare estimated U.S. states' poverty rates and life expectancies to evaluate if they have a relationship, and if so, what the relationship is.

As of 2022, 37.9 million people in the United States were officially in poverty, over one in 10 (11.5% of) U.S.-Americans. This is based on the government's official poverty measure, which in 2022 ranged from under $14,036 for a single-person household adult over the age of 65 to a family income under $20,0172 for a two-person household with both people under the age of 65 to under $35,801 for a five-person household with two related children under 17 years old. Many people struggle who have family incomes above the poverty thresholds constructed from the government's official poverty measure.

Someone born in 2020 in the United States is predicted to live 77.0 years, assuming continued patterns of mortality at that time.

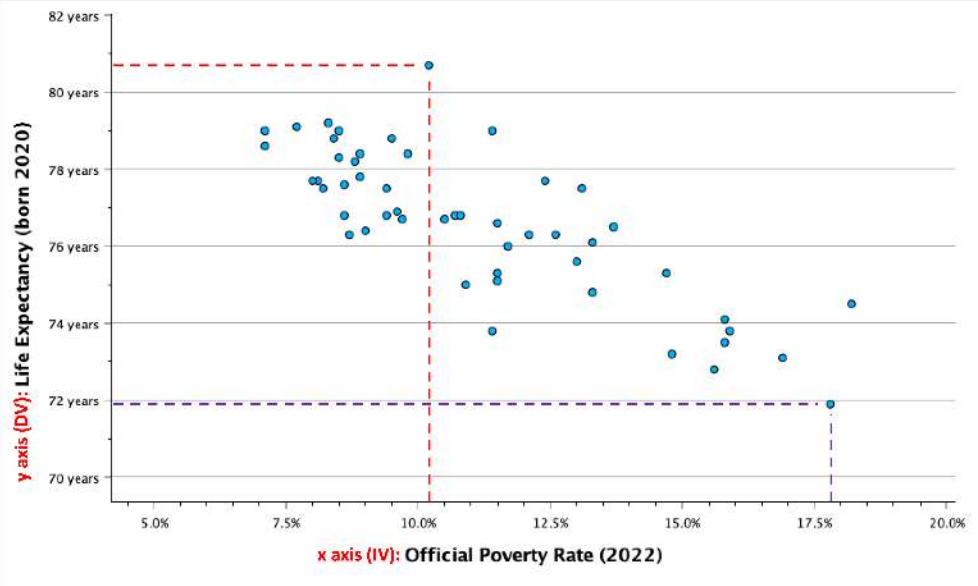
Here is a **scatterplot** of life expectancy and official poverty rate by U.S. state (as well as Washington, D.C.):



Definition: Scatterplot

A **scatterplot** is a graph that shows pairs of quantitative data using a rectangular coordinate system.

Check out this annotated version of the scatterplot:



The independent variable, x, goes on the horizontal axis, also called the x-axis.
(An axis is a fixed reference line, in this case the number line for the independent variable. The plural of axis is axes, with a long e like in cheese.)

- In this example, the independent variable is official poverty rate. You can see a number line for poverty rate on the horizontal axis.

The dependent variable, y, goes on the vertical axis, also called the y-axis.

- In this example, the dependent variable is life expectancy. You can see a number line for life expectancy on the vertical axis.

Each case has a point (marked with a dot) at a particular coordinate (x, y). X is the value of the independent variable for that case and Y is the value of the dependent variable for that case.

- In this example, there are 51 cases: each U.S. state and Washington, D.C. Each case has both a particular poverty rate and a particular life expectancy. The scatterplot shows their intersection.
- Take a look at the purple dotted lines. This is the state with the lowest life expectancy: Mississippi. Mississippi has a poverty rate of 17.8% and a life expectancy of 71.9 years. Notice how the point is located at those (x, y) coordinates (17.8%, 71.9 years).
- Take a look at the red dotted lines. This is the state with the highest life expectancy: Hawaii. Hawaii has a poverty rate of 10.2% and a life expectancy of 80.8 years.

Once these points are all mapped out, we can begin to analyze the graph for patterns/relationships.

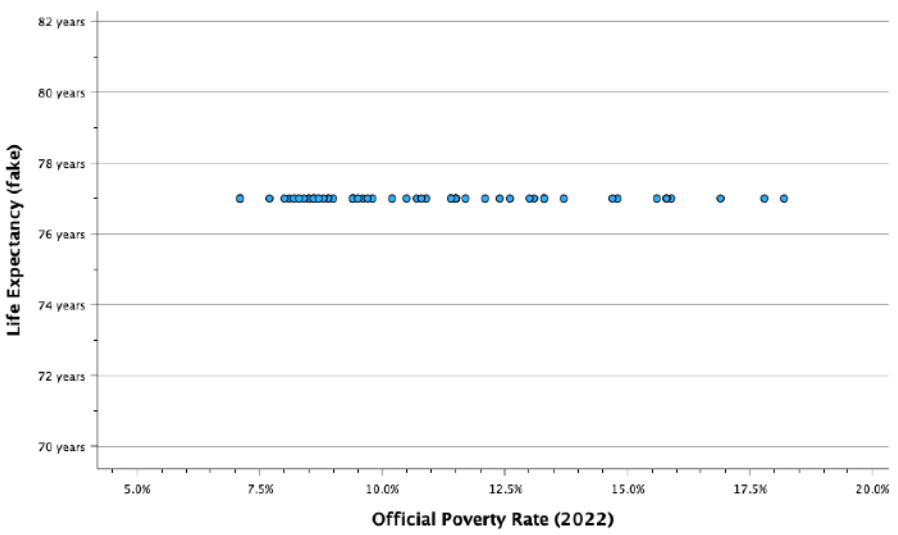
Analyzing and Interpreting Bivariate Relationships in Scatterplots

The first step is to determine whether a relationship exists.

If there is a relationship, the two variables should co-vary. Changes to the independent variable would correlate with patterned changes in the dependent variable. In this case, the independent variable is poverty rates. As we observe different poverty rates for different states, does our dependent variable, life expectancy, change in any patterned way?

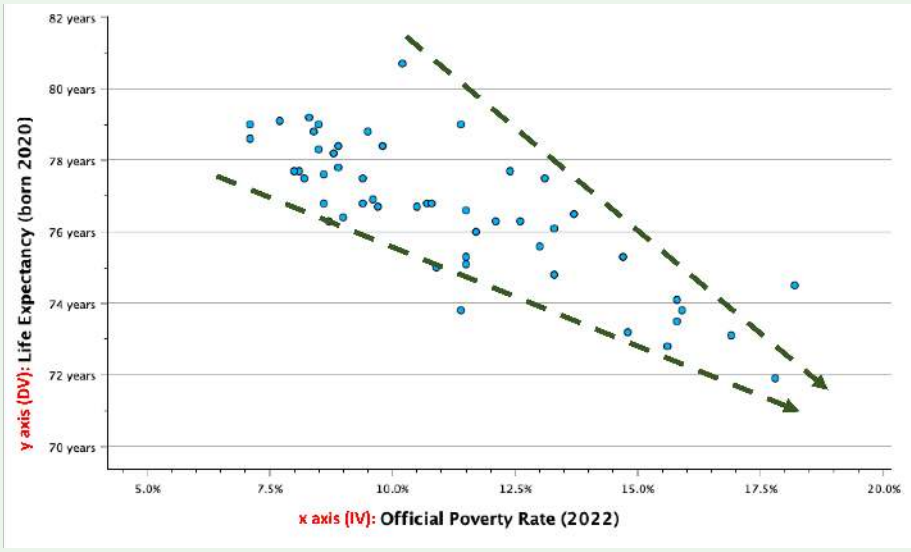
If there was no relationship, as poverty level increased, there would be no change to life expectancy.

Here is a scatterplot where every state is put at the U.S. mean life expectancy of 77.0 years old.



As you go from left to right, as poverty rates increase, there is no change to life expectancy. It's a flat horizontal line. This indicates no relationship, as changes in the independent variable are not associated with changes in the dependent variable.

As you go from left to right, if there is a clear general patterned trend to the data in terms of changes to the dependent variable values, there is a relationship.



For the actual scatterplot of poverty rate and life expectancy by state, you can see that as we go from left to right (as the independent variable values increase, as the poverty rate increases, as we go from states with lower to states with higher poverty rates), there is a general trend where the the points go down (the dependent variable values decrease, life expectancy age decreases, states tend to have lower life expectancies).

Now that we have determined there is a relationship, we have three other questions to ask:

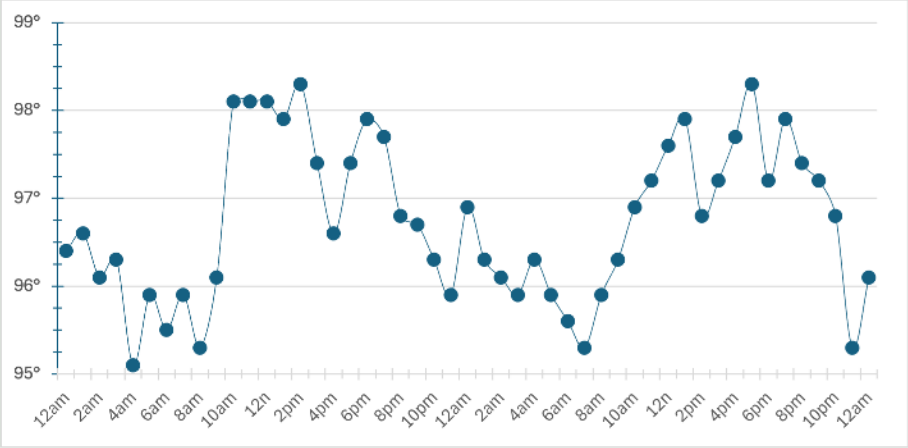
1. It is a linear relationship?

Definition: Linear relationship

A **linear relationship** is a relationship between two variables that takes the form of a **line** when graphed.
(Notice the word line in **line**ar.)

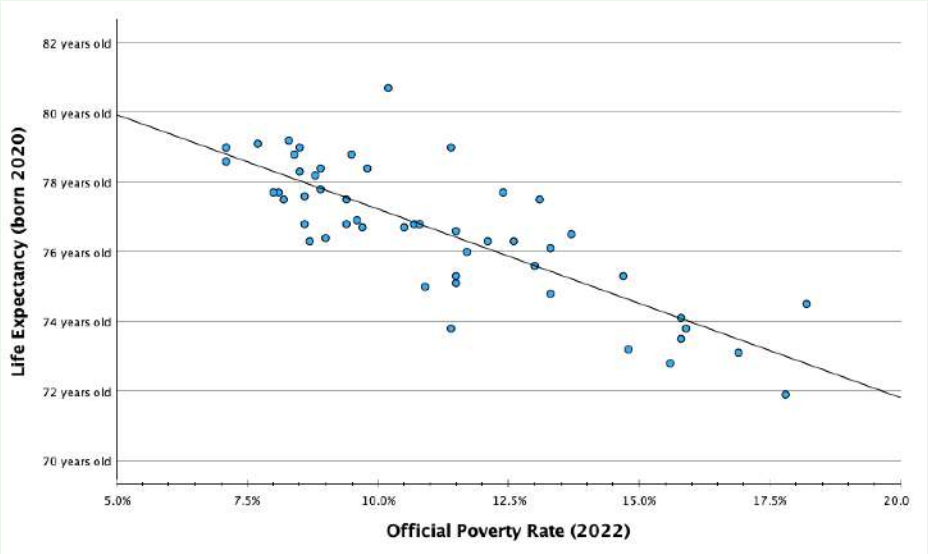
Not all relationships are linear.

For example, check out the graph below. I recorded my body temperature using an underarm thermometer every hour over a two day period.



My temperature follows a pattern. In the middle of the night, my body temperature was lower. In the afternoon, it was higher. This repeated the second day. This is a pattern, but it's not a linear pattern.

As the x-axis values increase (as you go from left to right), if the y-axis values consistently go in one general direction (either up or down), and do so at a generally consistent rate, the relationship is linear.



Our data has a linear relationship. You can see how the data generally converges onto the line drawn in the graph above.

It's not a perfect relationship --- meaning not all points are on the line. But the general pattern is linear.

Line of Best Fit

The line drawn on the scatterplot above is a "line of best fit," also called a "regression line" and "least squares line."

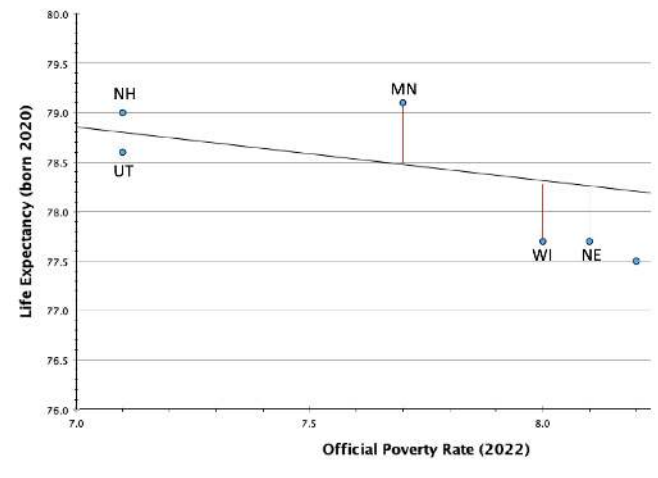
The process of fitting a line to a set of data is called linear regression, "Ordinary Least Squares regression," or "OLS regression". The next 5 chapters all focus on linear regression. The line that is the "best fit" is the one that minimizes how far each case's dependent value is from the line, calculated based on the sum of the squared difference between the data points and the line.

Definition: Regression line

The **regression line** is the line that, when drawn, is closest to the observed data. This is measured by minimizing the sum of the squared distance of each point's y-coordinate from a potential line's corresponding y-coordinate.

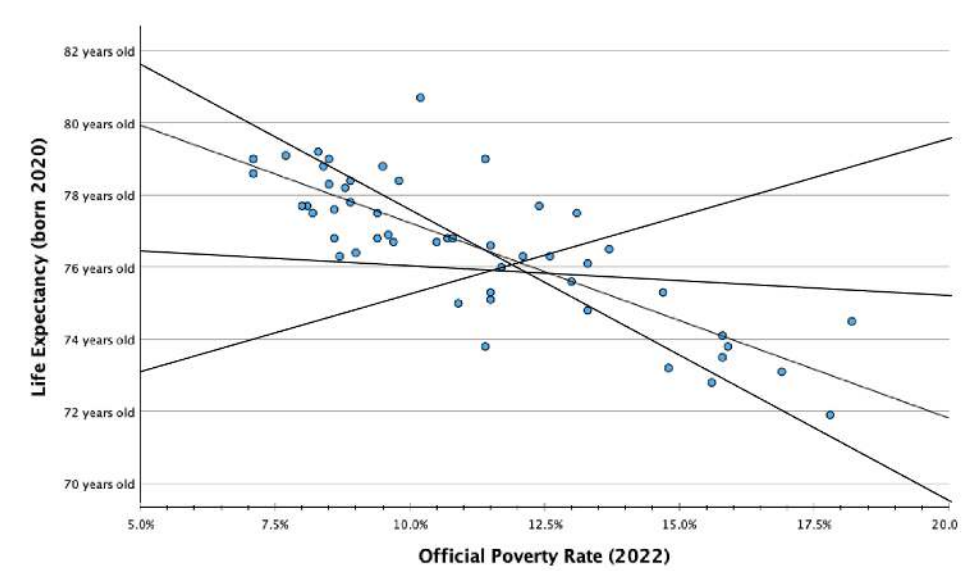
In the scatterplot above, not all points are on the line. Unless there is a perfect relationship, no single straight line can be drawn that would have all points on it. Some points are closer to the line and some points are further away.

Moving left to right, let's take the first five points.

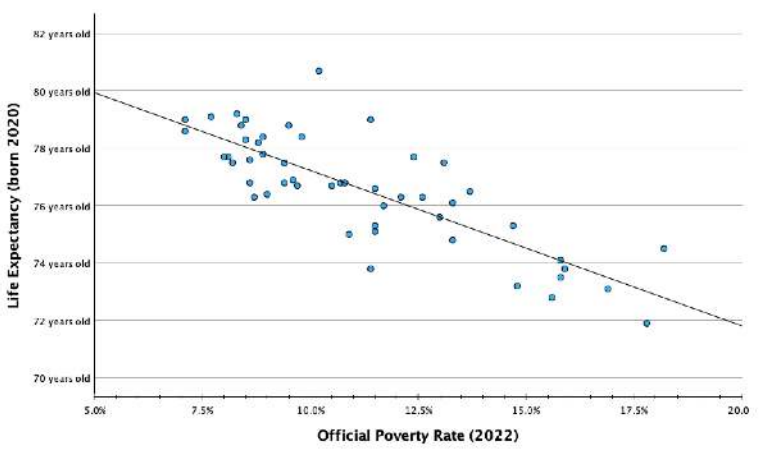


1. Utah has an official poverty rate of 7.1% and a life expectancy of 78.6 years. The regression line predicts that a state with a poverty rate of 7.1% would have a life expectancy of 78.8 years, 0.2 years off.
2. New Hampshire has an official poverty rate of 7.1% and a life expectancy of 79.0 years. The regression line predicts that a state with a poverty rate of 7.1% would have a life expectancy of 78.8 years, 0.2 years off.
3. Minnesota has an official poverty rate of 7.7% and a life expectancy of 79.1 years. The regression line predicts that a state with a poverty rate of 7.7% would have a life expectancy of 78.5 years, 0.6 years off.
4. Wisconsin has an official poverty rate of 8.0% and a life expectancy of 77.7 years. The regression line predicts that a state with a poverty rate of 8.0% would have a life expectancy of 78.3, 0.6 years off.
5. Nebraska has an official poverty rate of 8.1% and a life expectancy of 77.7 years. The regression line predicts that a state with a poverty rate of 8.1% would have a life expectancy of 78.3, 0.6 years off.

Many lines could be drawn on top of a scatterplot:



However, if you look at these lines, some are a better fit for the data then others. If you consider how far apart each point is from a line, square those differences, and add them up, the lines that are worse fits will have greater sums, and the lines that are better fits will have smaller sums. SPSS finds the line that minimizes the sum of the squared deviances as much as possible:



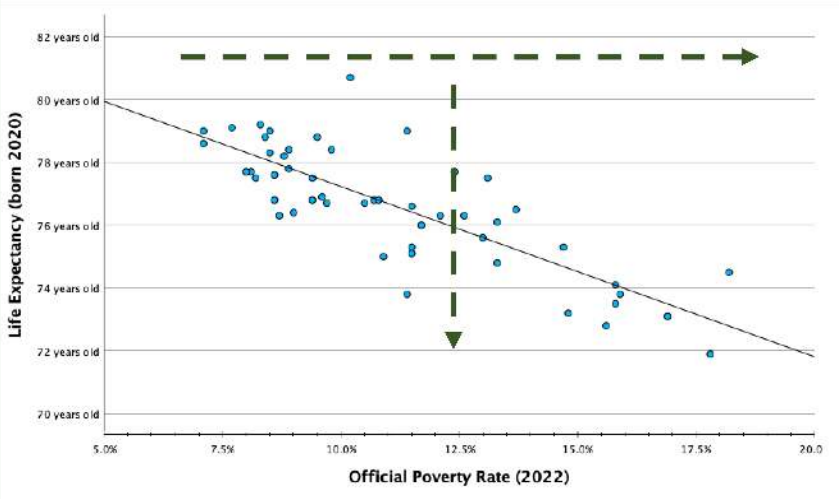
2. If it is linear, what is its direction? Is it negative or positive?

As you move from left to right, there will be a generally upward or downward movement. Upward movement reflects a positive relationship and downward movement reflects a negative relationship.

Definition: Direction: Positive and Negative relationships

In a **positive relationship**: as x (independent variable) increases, y (dependent variable) increases.

In a **negative relationship:** as x (independent variable) increases, y (dependent variable) decreases.



Our variables have a **negative** relationship. As you move from left to right, the data generally goes down. As x increases, y decreases.

Interpreting direction in real-life context:

What does it mean to say: as x increases, y increases? Here x is the official poverty rate and y is life expectancy. So we're saying, among U.S. states (and Washington D.C.), as poverty rate increases, life expectancy decreases. Higher

poverty rates are associated with lower life expectancy. States with higher poverty rates tend to have lower life expectancies.

Template:
Among [unit of analysis/target population], as [IV] increases, [DV] increases/decreases

Example:
Among U.S. states, as poverty rates increase, life expectancy decreases.

Caution

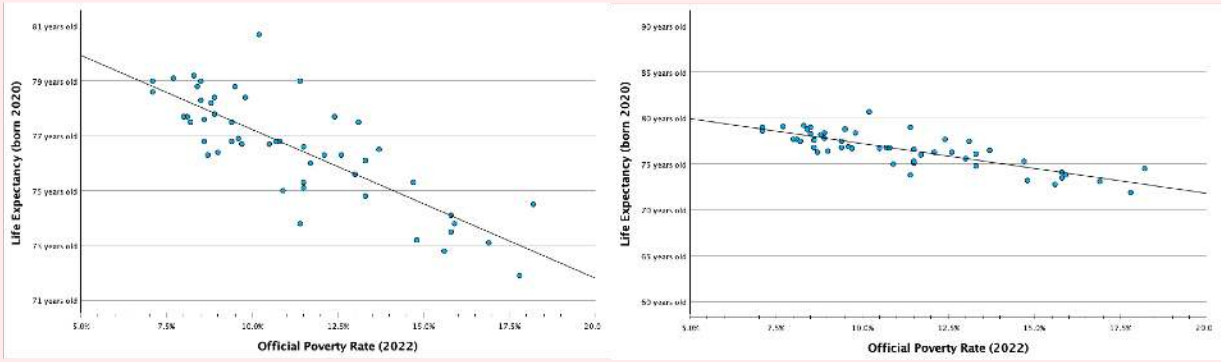
Be careful not to use causal language in describing relationships. We're just describing the pattern or co-variance of the two variables. The pattern just tells us that there is a relationship, not that the IV causes the DV. More poverty may decrease how long people live (and conversely less poverty may increase how long people live), but the scatterplot is just showing us if and how the two variables co-vary, not telling us that the IV causes change in the DV. (Remember that causality requires a significant and substantive relationship that makes theoretical sense, has the correct causal order, and is nonspurious).

3. If it is linear, how substantive is the relationship?

This is asking about how steep the line is. What is the pattern? Does a 5% increase in a state's poverty rate seem to be associated with a lowering of 1 year of the average life expectancy? 10 years? Or do you need a 10% or 20% increase in the poverty rate for any change to mean life expectancy? To evaluate how substantive a linear relationship is, we'll look at its slope (rise over run, change in y associated with a one-unit change in x). In this case, the slope of the line drawn above is -0.54, which means that the general pattern is that for every 1% increase in a state's poverty rate, there is about a 1/2 year decrease in the state's life expectancy. So for example a state with a 5% higher poverty rate (1 in 20 more people in poverty) than another would be predicted to have a life expectancy that is 2.5 years lower. Of the 13 states with a life expectancy over 78 years old (78.2 to 80.7), 9 have poverty rates below 9% and all have poverty rates below 12%. Of the 10 states with life expectancies below 75 years old (71.9 to 74.8), 8 have poverty rates above 15% and all have poverty rates above 11%.

Caution

You need to be careful analyzing how substantive a relationship is when you look at scatterplots, because like other graphs placed on number lines, the scale may make the relationship seem more or less steep. Check out the two scatterplots below. Notice how one makes it seem like the relationship is a lot more substantive than the other.



These scatterplots are the same data. The only difference are the y-axes: the scatterplot on the left shows life expectancy ranging from 71 to 81 years old and the scatterplot on the right shows life expectancy ranges from 60 to 90 years old.

We'll look more closely at slope and substantiveness in Chapter 11.

4. If it is linear, how strong is the relationship?

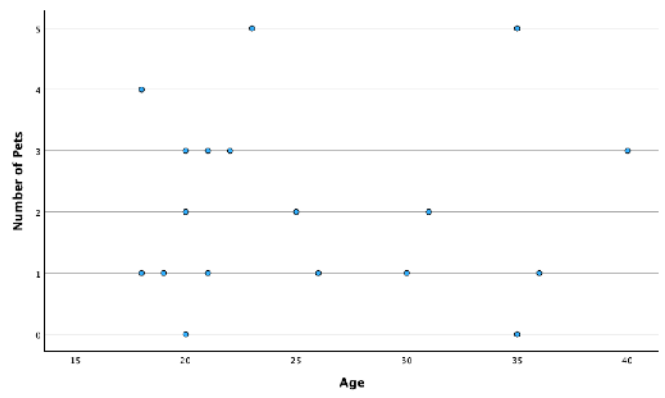
For linear relationships, the term **strength** has a particular statistical meaning.

Definition: Strength (linear relationships)

The **strength** of a linear relationship is the extent to which the points converge into a common pattern, into a particular linear relationship.

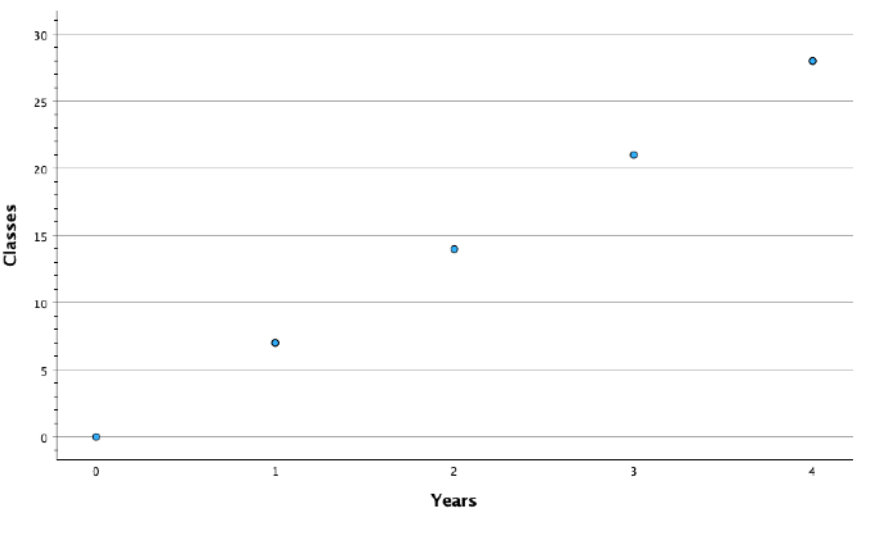
If they were scattered all over with no general trend, we would have **no relationship** and therefore **no strength.**.

Here is an example of a scatterplot with no relationship:

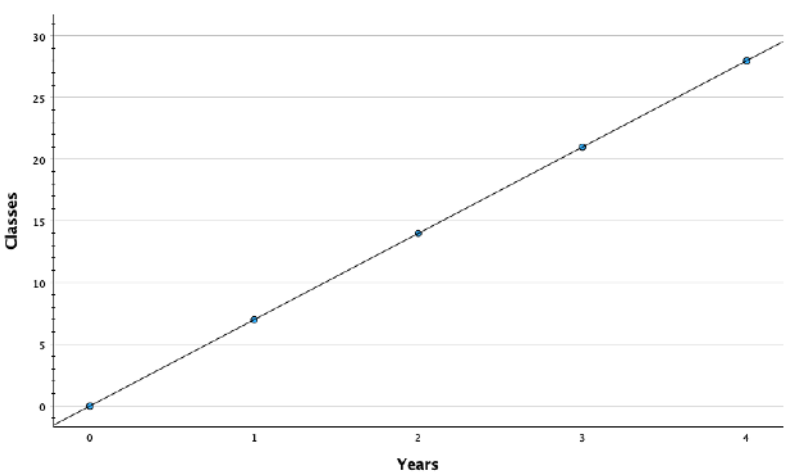


If a relationship was **perfect**, then all points would fall on the line.

For example, what if a high school required all students to take seven classes each year? Here is a graph of how many years a student has been at the high school and how many classes they have taken:



You can imagine drawing a line through the five points. Every point would go through it. It is a perfect relationship because the points converge into a common linear pattern perfectly.



Strength can be characterized as no relationship, as a perfect relationship, and in many ways in between the two.

The figure below shows six scatterplots. The ones on the left have **strong** relationships, the ones in the middle have **moderate** relationships, and the ones to the right have **weak** relationships. Take note of how the data converges closer to the line of best fit or further away from it. Consider how you could make more accurate predictions with the ones on the left, and as you move to the scatterplots to the right, as your strength decreases, your ability to make accurate predictions also decreases.

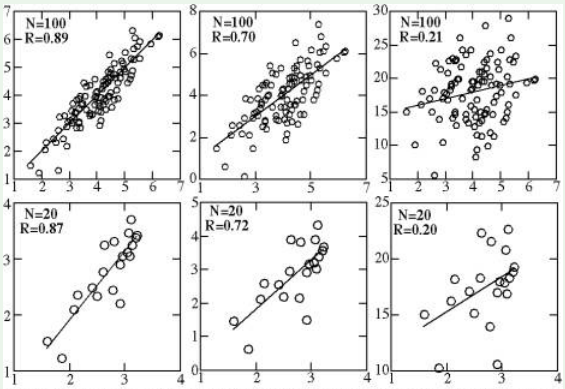
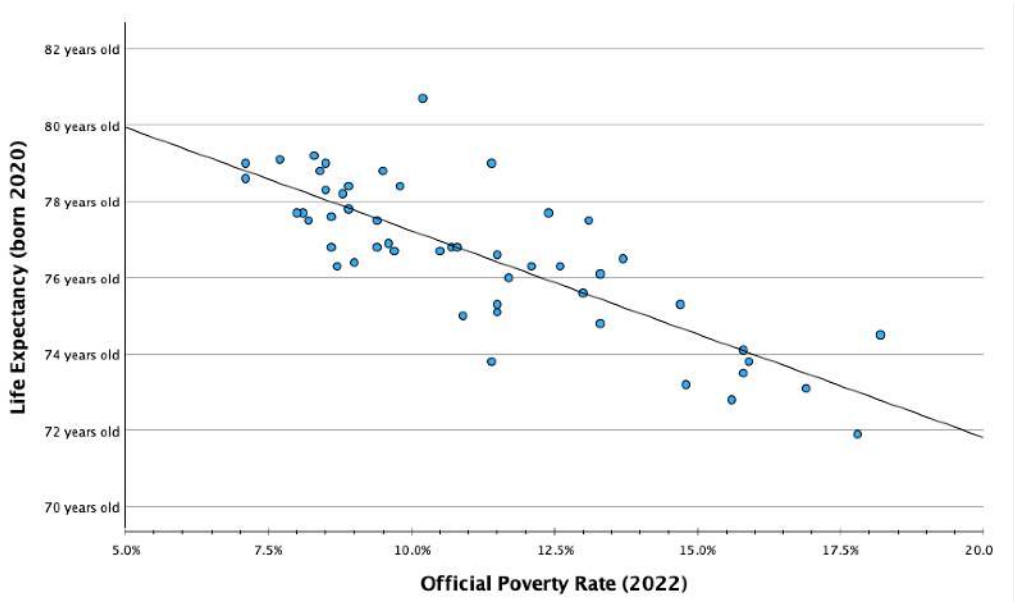


Figure 16. Examples of strong (R≈0.88), moderate (R≈0.71), and weak (R≈0.21) correlations, for N=100 points (top plots) and N=20 (bottom). Note that the linear regression line regresses toward horizontal as the correlation coefficient is reduced; only for the strongest correlation is this line as steep as would be drawn subjectively. All correlations except for that at lower right are significant at the 95% confidence level.

Let's return to our example. Take a look at the line and the points. Does this seem like no relationship, a very weak relationship, a moderate to strong relationship, or a very strong relationship?



Here, there is a general pattern, but there is also variation away from the line. There are a few states on the line, but most are not. We cannot simply use poverty rate to make accurate predictions of life expectancy. But how far away are states from the line? Are most pretty close? Pretty far away? How well could we use a state's official poverty rate to predict the state's life expectancy? We would definitely make better predictions than if we only knew that the overall U.S. life expectancy was 77 years old, but most of the time our predictions would not be accurate. However, while in some cases (like with Hawaii) we would have worse predictions, it looks like most of our predictions of life expectancy would be less than one year off.

Quantifying Strength & Direction: Correlation

Rather than just form impressions of the strength of a bivariate linear relationship, we can quantify it. This is done using a scale from -1 for a perfect negative relationship to 0 for no relationship to 1 for a perfect positive relationship. The measure is called **correlation**.

Definition: Correlation

Correlation measures the strength and direction of linear relationships.

The measure can be referred to as *Pearson's* correlation coefficient, *Pearson's* r, r, correlation, or by similar terms.

The correlation between two variables will be the same regardless of the direction of variables. It doesn't matter which is the independent variable and which is the dependent variable.

Direction:

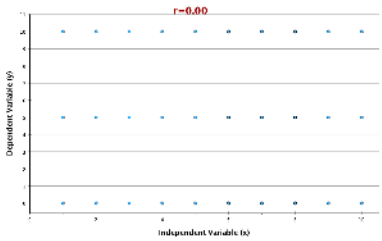
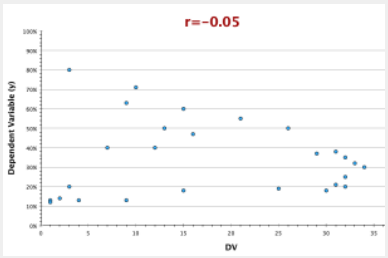
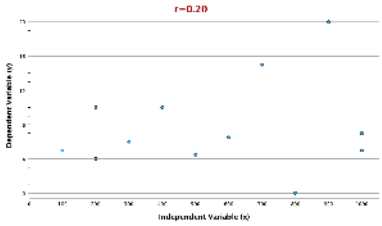
If r is negative (less than 0), this indicates a negative relationship.

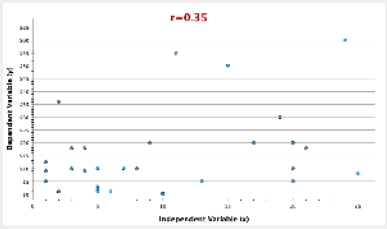
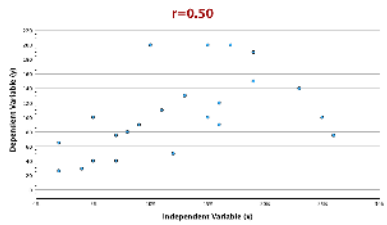
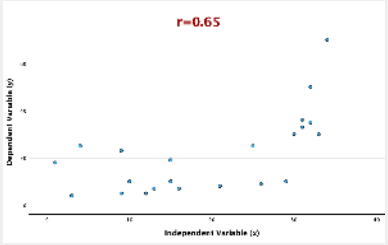
If r is positive (more than 0), this indicates a positive relationship.

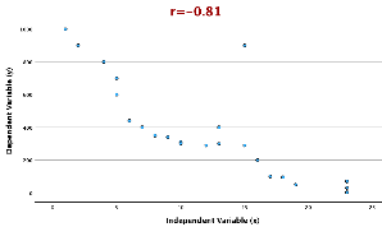
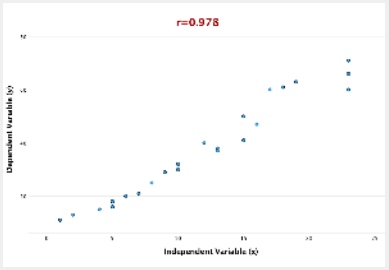
Strength:

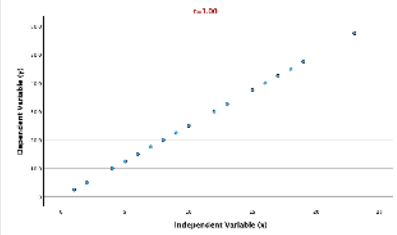
The measure of strength, of how well the data converges into a common linear pattern, ranges from 0 (no relationship) to an absolute value of 1 (perfect relationship). For example, a relationship with a correlation coefficient of -0.5 and +0.5 have the same strength, just with the former relationship being negative and the latter relationship positive.

Here is a table with a guide to interpreting relationship strength given a correlation coefficient:
(For direction, the interpretation is: "Among [unit of analysis/target population], as [IV] increases, [DV] increases/decreases.")

Correlation Coefficient	Interpretation (of strength)	Story (Real-life context)
0.00 	**No relationship** There is no (linear) relationship between [IV] and [DV].	There is a **no relationship** between [IV] and [DV]. There is **no tendency** among [target population] for [DV] to either increase or decrease based on [IV]. (Knowing [IV] does not improve our predictions of [DV] over just knowing the overall average [DV] among all [unit of analysis].)
±0.01 to ±0.09 	**Negligible, very weak** There is close to no (linear) relationship between [IV] and [DV]. A line of best fit would be close to a horizontal line, but would not be completely flat, as some co-variance or pattern can be detected (within the sample, if sample data).	There is a **barely a** relationship between [IV] and [DV]. There is a **very weak, though still faintly discernible, trend** among [target population] that [describe direction of relationship]. If we try to make predictions of [dependent variable] for [unit of analysis] in the sample [delete "in the sample" if it's population data] based on knowing [independent variable], our predictions are likely to be **very inaccurate.**
±0.10 to ±0.29 	**Weak** There is a weak (linear) relationship between [IV] and [DV], which means the data does not converge well onto the line of best fit, though there is still a definite observable pattern.	There is a **weak** relationship between [IV] and [DV]. While there is an overall observable trend among [target population] that [describe direction of relationship], there is **substantial variability**. We can only make **weak predictions** of [dependent variable] for [unit of analysis] in the sample [delete "in the sample" if it's population data] based on knowing [independent variable], though our predictions will overall be more accurate than if we ignored [IV].

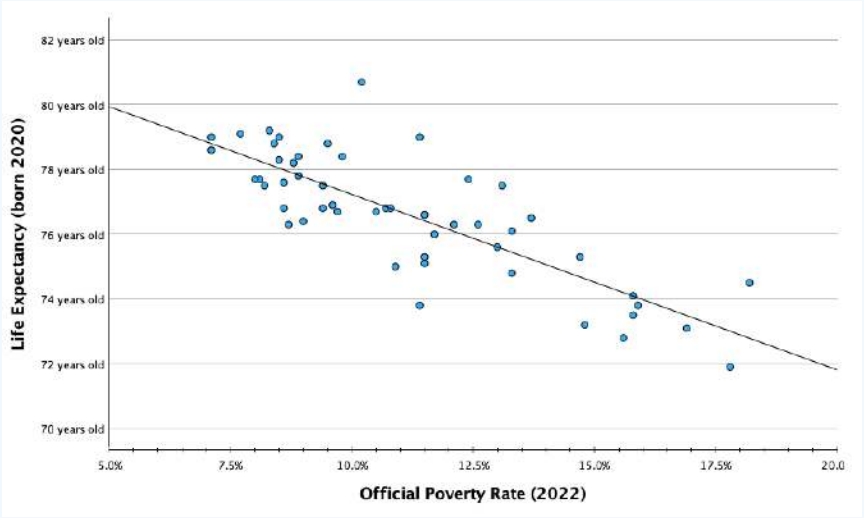
Correlation Coefficient	Interpretation (of strength)	Story (Real-life context)
±0.30 to ±0.39 	**Weak to moderate** There is a weak to moderate (linear) relationship between [IV] and [DV], which means the data converges into an observable pattern, though with a moderate to substantial amount of variability from a line of best fit.	There is a **weak to moderate** relationship between [IV] and [DV]. There is an overall observable trend among [target population] that [describe direction of relationship], there is **some variability**. We can make **weak to moderate predictions** of [dependent variable] for [unit of analysis] in the sample [delete "in the sample" if it's population data] based on knowing [independent variable], with our predictions overall noticeably more accurate than if we ignored [IV].
±0.40 to ±0.59 	**Moderate** There is a moderate relationship between [IV] and [DV], which means the data converges moderately well onto a line of best fit, with a clear observable pattern that has a moderate amount of variability.	There is a **moderate** relationship between [IV] and [DV]. There is an overall trend among [target population] where [describe direction of relationship], though there is a **fair amount of variability**. We can do a **moderately good job making predictions** of [dependent variable] for [unit of analysis] in the sample [delete "in the sample" if it's population data] based on knowing [independent variable].
±0.60 to ±0.69 	**Moderate to strong** There is a moderate to strong relationship between [IV] and [DV], which means the data converges pretty well onto a line of best fit, with a clear observable pattern that has a small to moderate amount of variability.	There is a **moderate to strong relationship** between [IV] and [DV]. There is an often consistent trend among [target population] where [describe direction of relationship], though there is a **small to fair amount of variability**. We can do a **moderately to pretty**

Correlation Coefficient	Interpretation (of strength)	Story (Real-life context)
		good job making predictions of [dependent variable] for [unit of analysis] in the sample [delete "in the sample" if it's population data] based on knowing [independent variable].
±0.70 to ±0.89 	**Strong** There is a strong relationship between [IV] and [DV], which means the data converges well onto a line of best fit, with a clear observable pattern that has a small amount of variability.	There is a **strong relationship** between [IV] and [DV]. There is a consistent trend among [target population] where [describe direction of relationship], though there is a **small amount of variability**. We can do a **pretty good job making predictions** of [dependent variable] for [unit of analysis] in the sample [delete "in the sample" if it's population data] based on knowing [independent variable].
±0.90 to ±0.99 	**Very strong** There is a very strong relationship between [IV] and [DV], which means the data converges extremely well onto the line of best fit, with a clear and relatively uniform observable pattern.	There is a **very strong relationship** between [IV] and [DV]. There is a consistent trend among [target population] where [describe direction of relationship], though there is still **a bit of variability**. We can do a **quite good job making predictions** of [dependent variable] for [unit of analysis] in the sample [delete "in the sample" if it's population data] based on knowing [independent variable].
±1.00	**Perfect**	[Describe direction of relationship.]

Correlation Coefficient	Interpretation (of strength)	Story (Real-life context)
	There is a perfect relationship between [IV] and [DV], which means all data points are on the same line. The data converges into a perfect linear pattern.	If you know [unit of analysis's] [IV], you can **predict** [DV] with **perfect accuracy.**

Begin to get a feel for correlation coefficients and how they map onto scatterplots by playing a few rounds of the game Guess the Correlation. Go to http://guessthecorrelation.com/ to play. Note that there are only positive correlations in the game.

After playing a few rounds, take a guess at the correlation coefficient for the scatterplot from our example.



Scroll down to see how accurate your guess is.

The correlation coefficient for our example is **-0.808**.

-0.808 is a negative strong relationship. There is a strong relationship between poverty rate and life expectancy, which means the data converges well onto a line of best fit, with a clear observable pattern that has a small amount of variability.

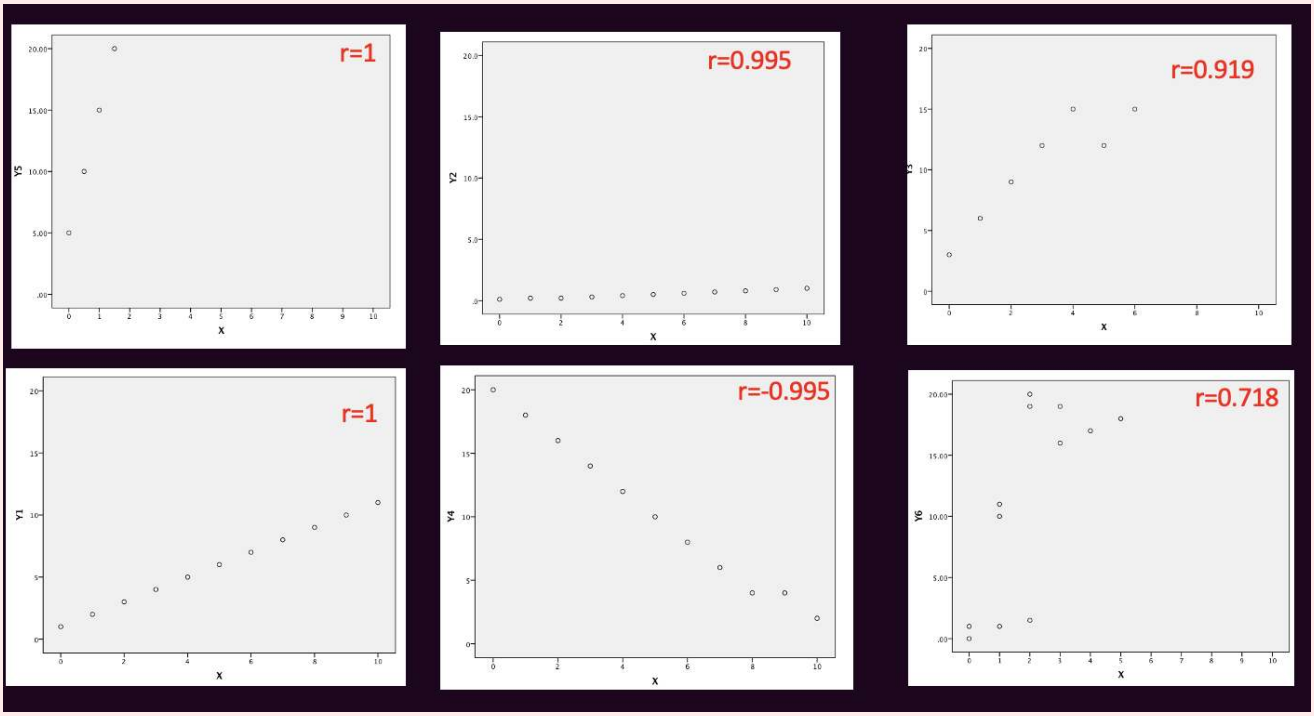
Here was the template language to interpret a correlation of -0.808:
There is a strong relationship between [IV] and [DV]. There is a consistent trend among [target population] where [describe direction of relationship], though there is a small amount of variability. We can do a pretty good job making predictions of [dependent variable] for [unit of analysis] in the sample [delete "in the sample" if it's population data] based on knowing [independent variable].

Using the template, here is what I wrote to explain the correlation coefficient in context, telling its story:
There is a strong negative relationship between poverty rates and life expectancy. There is a consistent trend among U.S. states (and DC) where, as poverty rates increase, life expectancy decreases, though there is a small amount of variability. We can do a pretty good job making predictions of states' average life expectancy based on knowing states' poverty rates.

Caution: Strength is about Pattern Consistency

The word strength has many meanings and interpretations. But when discussing correlation and linear relationships, it has a very specific and particular definition: how well the data fits the line. How well the data converges into a common linear pattern. How consistent the pattern is across the data points.

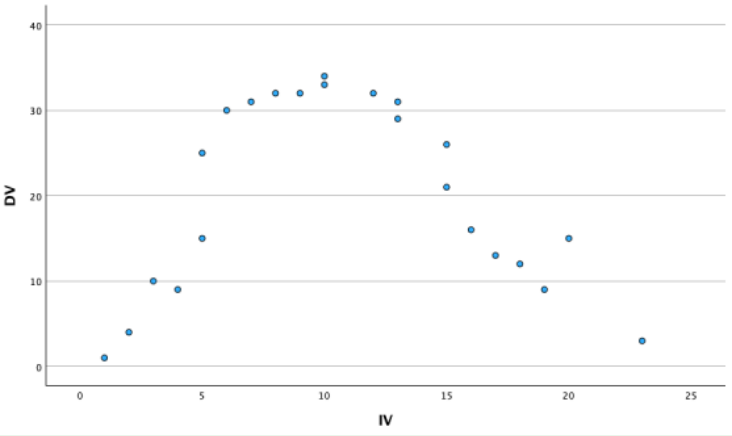
A common mistake for people first encountering this term is to confuse strength with substantiveness, to think about how steep a line's slope is. It's important to be able to differentiate these two constructs. Below are a series of scatterplots with their correlation coefficient values marked. Notice how the slopes/steepness of the patterns of these lines have no impact on the correlation coefficient values. Substantiveness and slope will be discussed more in the next chapter. For now, with the focus on strength, remember to stick to evaluating how well the data conforms into a common linear pattern.



Correlation Coefficients Are About Linear Relationships

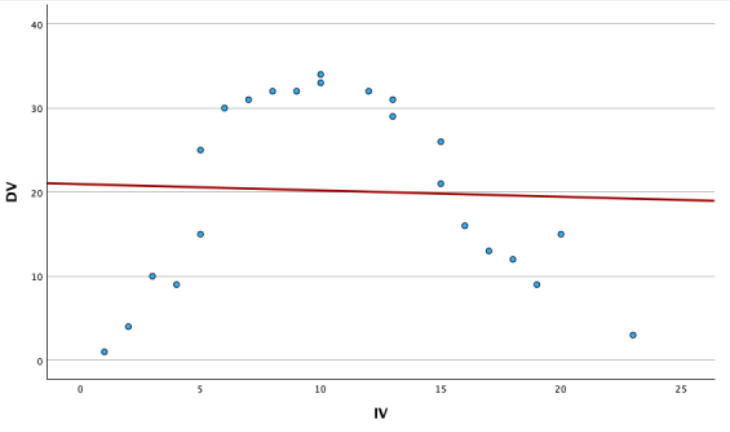
Pearson's correlation coefficient is only for linear relationships. You could have a correlation coefficient close to 0 and still have a strong relation relationship between your two variables if you have curvilinear relationship rather than a linear relationship (e.g., if as the independent variable increases, the dependent variable first increases and then decreases).

Example:



Check out this scatterplot. There is clearly a pattern between the independent and dependent variable. However, r=(-0.042).

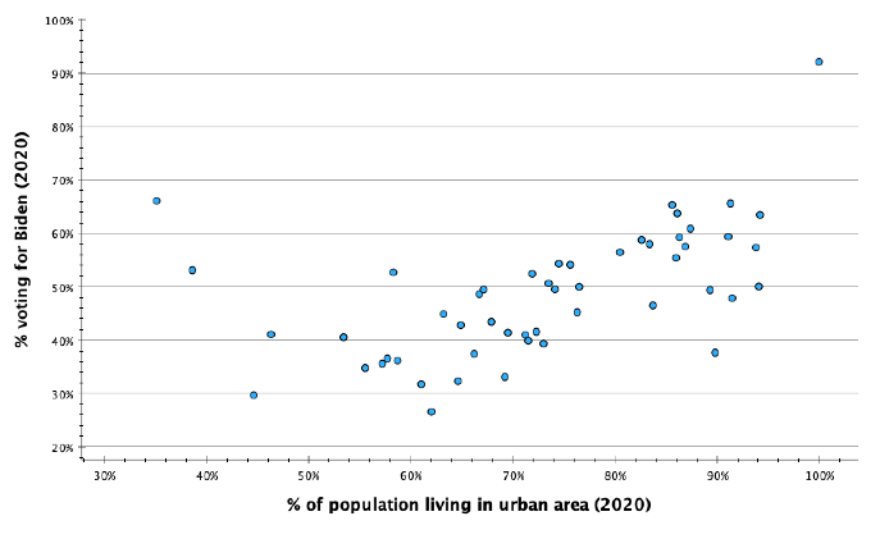
Check out the line of best fit below.



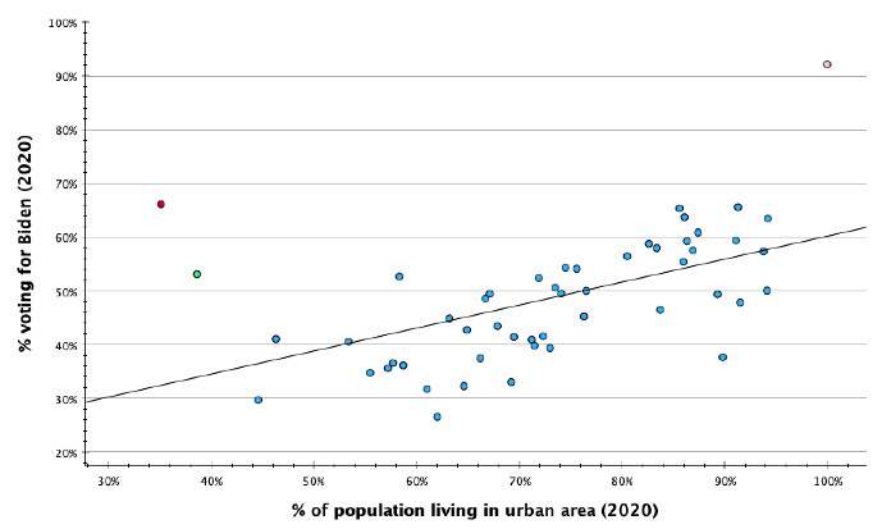
The data does not converge well into a linear pattern. When you find that there is a low correlation coefficient value, it does not mean there's necessarily a non-existent or weak relationship. It means there is a weak or non-existent **linear** relationship. There are more advanced statistical techniques that can be used to explore non-linear relationships.

Outliers

The scatterplot below is also of all 50 states as well as Washington, D.C. The independent variable is the percent of the state's population that lives in an urban area as of 2020, a measure defined by the U.S. Census based on housing unit density. The dependent variable is the percent of the state's voters who voted for Joe Biden for president in 2020.



Here is the scatterplot with a line of best fit added:



Pearson's r is 0.54, a moderate, positive relationship. There is an overall trend where States with higher percentages of their population living in urban areas have higher percentages of voters voting for Biden, though there is a fair amount of variability. We can do a moderately good job making predictions of what percent of a state voted for Biden based on knowing the percent of the state that is living in an urban area.

However, notice that while r=0.54, generally jurisdictions seem to be pretty well-clustered around the line. With the exception of three points (shaded red, green, and pink above), which are pretty far away from the line. They do not fit the general pattern.

Outliers have a substantive impact on correlation coefficients, as well as on lines of best fit. The squared difference of these jurisdictions from the lines is considerable, requiring the line to adjust to account for them, and their distance from the pattern lowers the correlation coefficient value.

If you have one or more outliers, the first thing to do is to look them up and investigate what's going on. You want to make sure there was not a data entry error, and if the data is accurate, investigate what makes these cases unique.

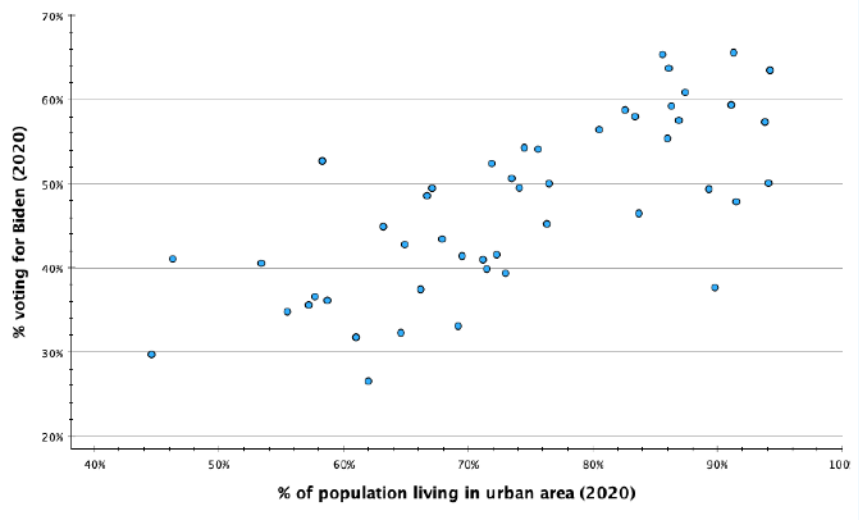
In this case, the outliers from the overall pattern include:

- Vermont (red point): 35.1% of Vermont's population is urban, but 66.1% of Vermont voters voted for Biden.
- Maine (green point): 38.6% of Maine's population is urban, but a majority of Maine voters voted for Biden (53.1%).

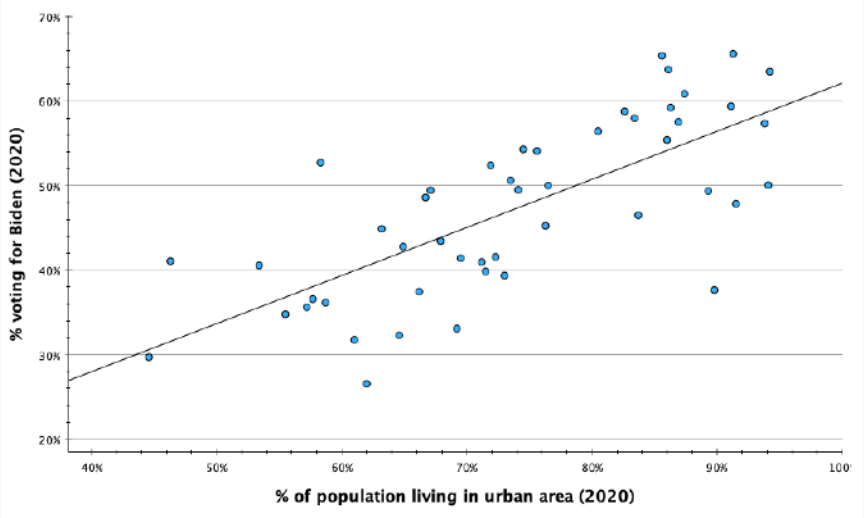
- Washington, D.C. (pink point): 100% of DC's population is urban, and 92.2% of voters voted for Biden (the next highest was Vermont, over 25% lower).

Additionally, you may want to consider whether or not it's appropriate to run a second model that excludes the outliers, to use in concert with the first model. We can't just remove data we don't like --- without these three outliers, the edited scatterplot only makes predictions for the remaining 48 states, and the predictions don't hold for Vermont, Maine, and Washington, D.C. However, we may be able to do a much better job understanding the patterns and making predictions if we leave those three jurisdictions out. For the remaining 48 states, r increases from 0.54 to 0.73.

Here is a scatterplot without those three jurisdictions, of 48 states:



Notice how the line of best fit converges better into a common linear pattern that fits more points here than it did when the outliers remained included.



Generating scatterplots, lines of best fit, using SPSS

Using SPSS

Here is how to make a scatterplot using SPSS.

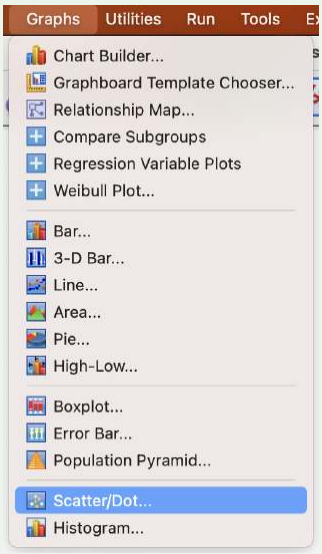
Syntax:

*Replace the IVName and DVName with the names of your independent and dependent variables.
GRAPH

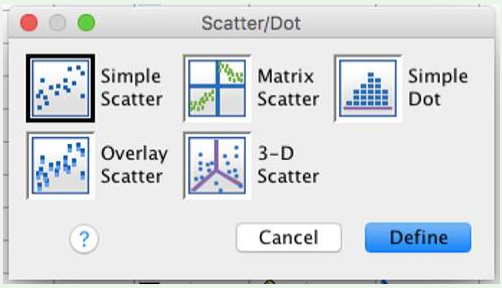
/SCATTERPLOT(BIVAR)=IVName WITH DVName
/MISSING=LISTWISE.

Menu:

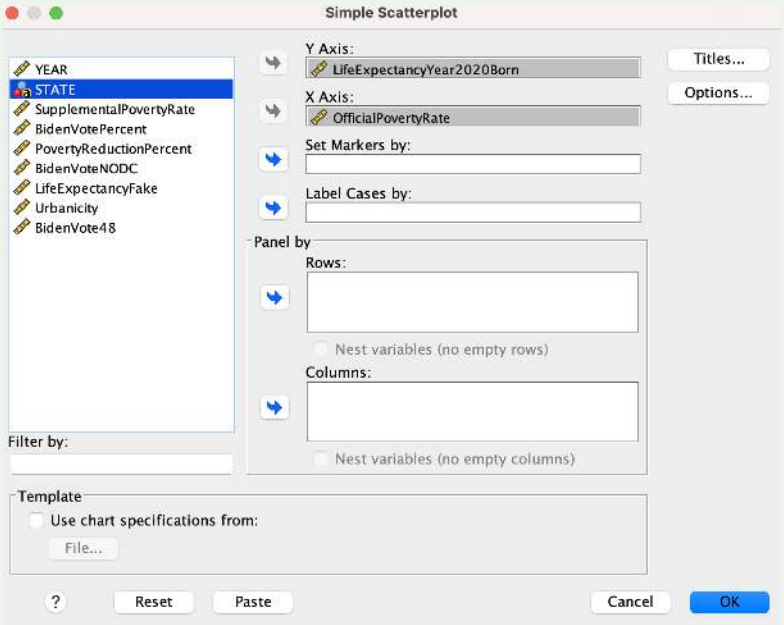
Go to: **Graphs → Scatter/Dot**



Click on the scatterplot picture to the left of "Simple Scatter" and then click Define.



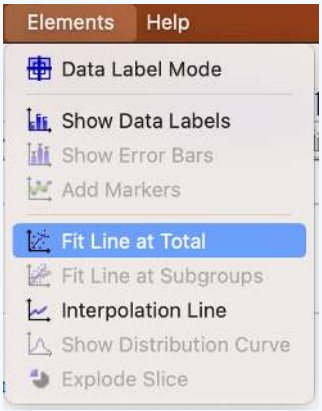
Put your independent variable in the box labeled "X Axis:" and your dependent variable in the box labeled "Y Axis:", then click OK.



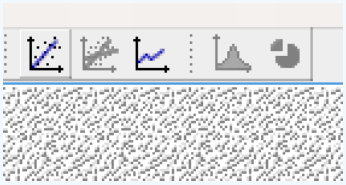
Adding a line of best fit to a scatterplot

Once you have generated a scatterplot and it is in your output, double-click on the scatterplot, or right-click and select "Edit."

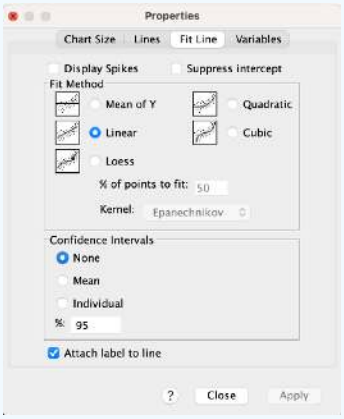
Go to: Elements → Fit line at total



Alternatively, you can click on the icon on the menu bar right above the graph that looks like a scatterplot with a fit line:



Select the circle for “Linear” for "Fit Method". Click “Apply” then “Close”.



Click on the r-squared statistic in the top right corner and then click delete.

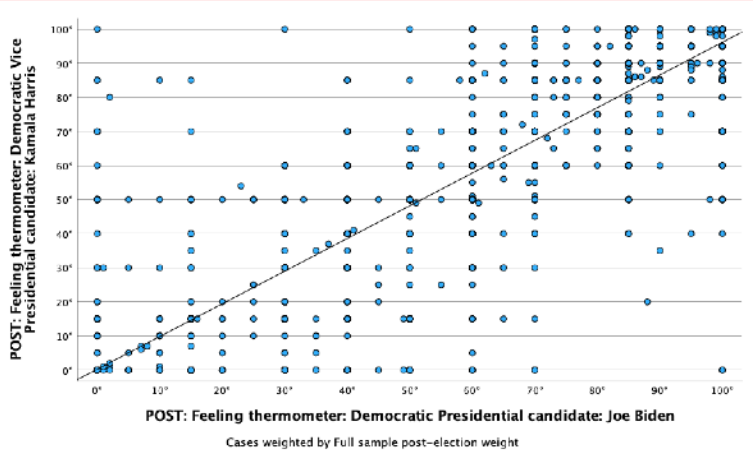
Either click on then delete the equation (y=...), or click and drag it over so it is not on top of data.

Caution: When Scatterplots Don't Work

What would you do if you were making a scatterplot of age and income and two respondents had the same age and income? When you have datasets with a smaller number of cases and more specificity and variation in values, scatterplots can work really well. However, if you have a dataset that has too many cases with the same coordinates, you are prone to misinterpret your dataset, and if you have a dataset that also has a lot of cases, you are prone to not even be able to begin an interpretation. This is because having 1 case or 100 cases at a particular coordinate may look the same.

The scatterplot below is from an American National Election Studies dataset of U.S. eligible voters following the 2020 election. The variables are feelings towards Joe Biden and feelings towards Kamala Harris, measured using a feelings

thermometer where 0 is very cold, 50 is neutral, and 100 is very warm.



The variables have a correlation coefficient of 0.91, a very strong positive relationship. Can you tell? I notice that there is an increasing sparseness of data when you get to warm feelings towards one candidate being paired with cold feelings towards the other, and I can see there are more pairings in terms of warm-warm and cold-cold groupings, so I get a sense of the positive direction, but not a correlation coefficient as high as 0.91.

This is especially because, out of the 2,578 cases, many are invisible on the scatterplot. Can you tell that there are 294 cases at (100, 100) where individuals rated both Biden and Harris at 100? How about 436 cases at (0,0)? There are over 100 cases at: (15,0), (50,50), (70,70), and (85,85). Meanwhile, there are many coordinates that only have 1 case, such as (0,85), (1,0), (35,0), (50,85), and (100,0). Do the coordinates with 1 case look any different than the coordinates with over 100 cases?

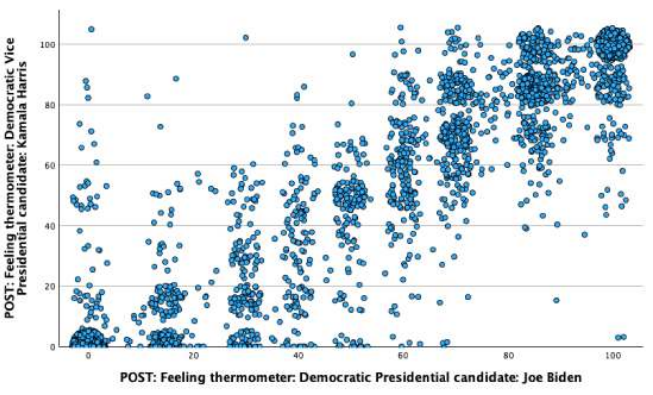
Here's another scatterplot, this one of age and feelings towards police.

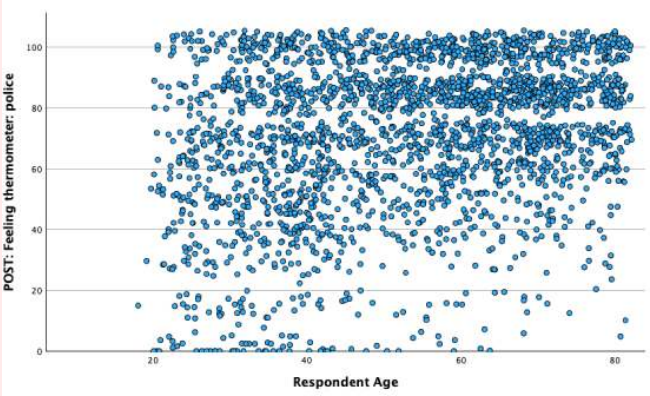


Here r=0.289, a weak positive relationship. Do you see an observable pattern? If you were to try to draw a line of best fit for that pattern, is the line you would have come up with the one pictured above? Can you see 2,550 cases?

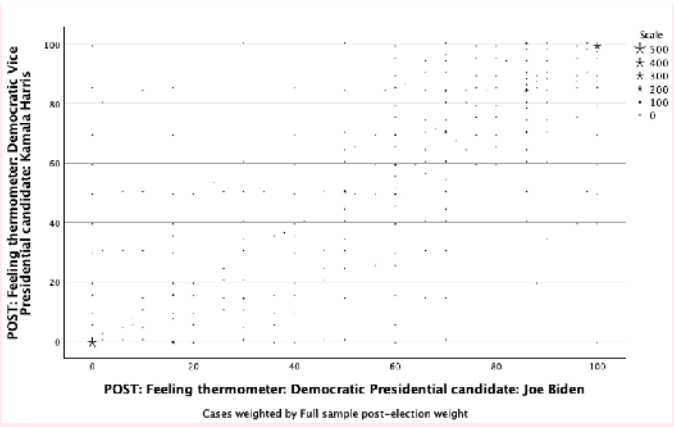
Scatterplots often do not work well for displaying data when there are repeated coordinates. You can still calculate correlation.

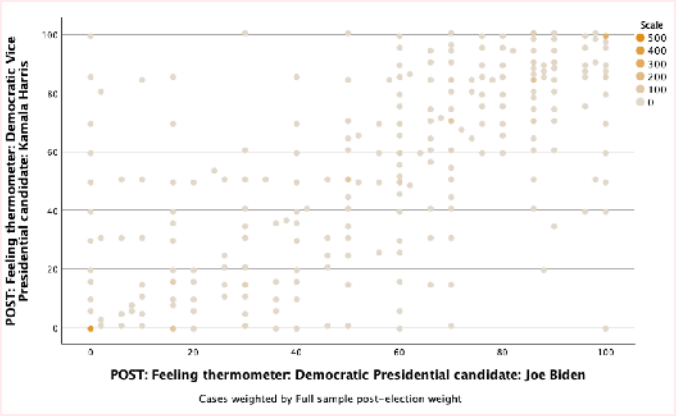
One way some people try to deal with this is through messier scatterplots that try to show repeated coordinates. One way is to use "jittering." Think about jittering as basically taking the scatterplot and giving it a little shake so that coordinate pairs are not exactly on their coordinates, reducing overlapping and allowing us to see more of the data. The easiest way I have found to add jittering in SPSS is to use "Chart Builder" under "Graphs" to generate the scatterplot, copy the syntax, and then run the syntax, changing "ELEMENT: point" to "ELEMENT: point.jitter". Here are the scatterplots from above, this time with jittering:

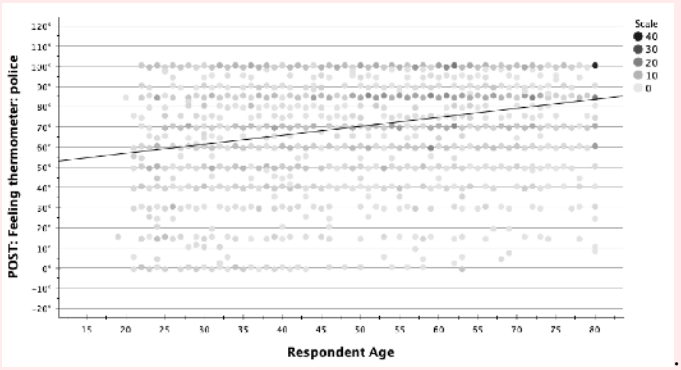


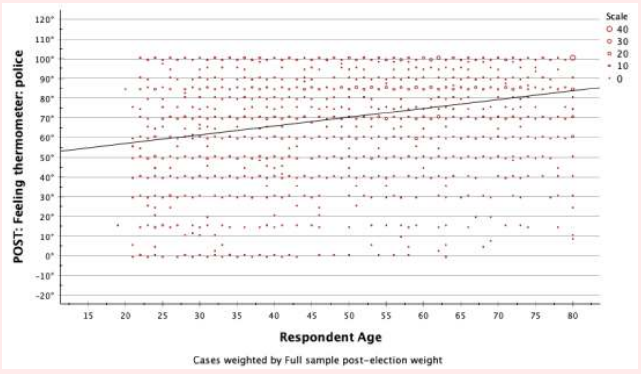


Another way people try to deal with the problem of multiple points sharing coordinate locations is by showing density of each location. For example, a sunflower plot (the first one below) has little "petals" on each point to indicate how many cases are at each coordinate. The scatterplots below were all produced in SPSS by double-clicking on the initial scatterplot and then going to Options → Bin Element.


POST: Feeling thermometer: Democratic Vice Presidential candidate: Kamala Harris
POST: Feeling thermometer: Democratic Presidential candidate: Joe Biden
Cases weighted by Full sample post-election weight
Scale


POST: Feeling thermometer: Democratic Vice Presidential candidate: Kamala Harris
POST: Feeling thermometer: Democratic Presidential candidate: Joe Biden
Cases weighted by Full sample post-election weight
Scale


POST: Feeling thermometer: police
Respondent Age
Scale


POST: Feeling thermometer: police
Respondent Age
Cases weighted by Full sample post-election weight
Scale

Correlation Matrices

For any pair of ratio-like variables, you can figure out the correlation coefficient value, and for representative samples you can evaluate statistical significance, by creating a correlation matrix.

Using SPSS

Syntax

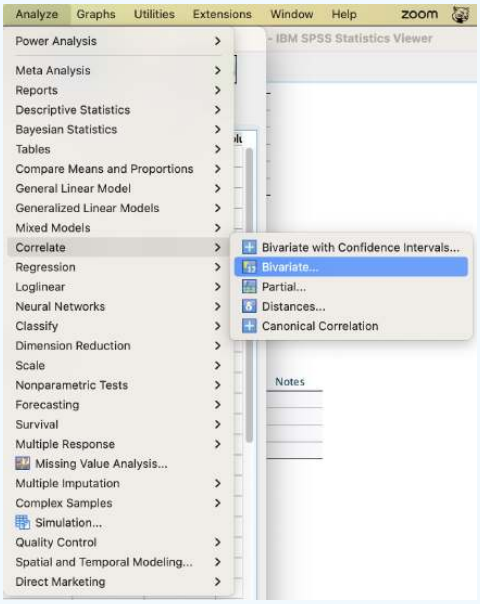
*You only need to list two variables, replacing VariableOneName etc with your variable names, but you can list as many variables as you want.

*Remember that even if you list multiple variables, the matrix will just be comparing two variables at a time, all possible pairings.

```
CORRELATIONS
/VARIABLES=VariableOneName VariableTwoName VariableThreeNameEtc
/PRINT=TWOTAIL NOSIG FULL
/STATISTICS DESCRIPTIVES /CI CILEVEL(95)
/MISSING=PAIRWISE.
```

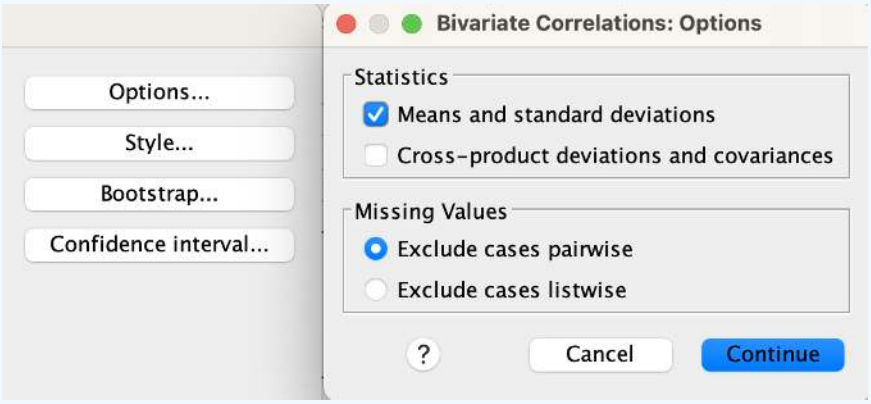
Menu

Go to Analyze → Correlate → Bivariate.

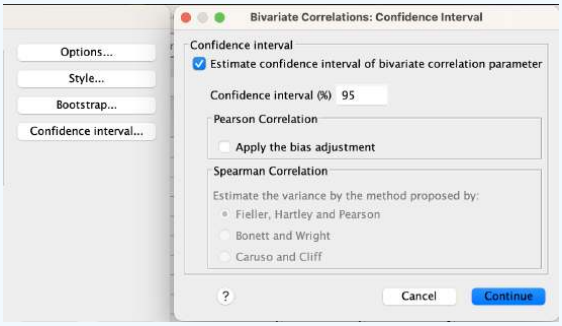


Add at least 2 variables to the "Variables:" box on the right. (You can add as many as you want, but remember that the matrix will be giving you bivariate correlations, all possible pairings of the variables.)

Click on "Options...", in the pop-up window check the box for "Means and standard deviations" and click "Continue."



Click on "Confidence Interval...", in the pop-up window check the box for "Estimate confidence interval of bivariate correlation parameter" and click "Continue."



Click OK.

Here is an SPSS output, based on one of the examples above:

Descriptive Statistics

	Mean	Std. Deviation	N
POST: Feeling thermometer: Democratic Presidential candidate: Joe Biden	51.69	35.634	2617
POST: Feeling thermometer: Democratic Vice Presidential candidate: Kamala Harris	49.94	37.621	2598

Correlations

		POST: Feeling thermometer: Democratic Presidential candidate: Joe Biden	POST: Feeling thermometer: Democratic Vice Presidential candidate: Kamala Harris
POST: Feeling thermometer: Democratic Presidential candidate: Joe Biden	Pearson Correlation	1	.909**
	Sig. (2-tailed)		<.001
	N	2617	2578
POST: Feeling thermometer: Democratic Vice Presidential candidate: Kamala Harris	Pearson Correlation	.909**	1
	Sig. (2-tailed)	<.001	
	N	2578	2598

**. Correlation is significant at the 0.01 level (2-tailed).

Confidence Intervals

	Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed)[a] Lower	Upper
POST: Feeling thermometer: Democratic Presidential candidate: Joe Biden - POST: Feeling thermometer: Democratic Vice Presidential candidate: Kamala Harris	.909	<.001	.902	.916

a. Estimation is based on Fisher's r-to-z transformation.

The descriptive statistics at the top give you summary statistics for each of the variables you selected.

The confidence intervals at the bottom tells you which 2-variable pair was analyzed, gives you the correlation coefficient (r=0.909), the p-value (p<0.001), and a 95% confidence interval for the correlation coefficient (we are 95% confident that among all U.S. eligible voters following the 2020 election, the correlation coefficient would be somewhere between 0.902 and 0.916, a positive and very strong relationship). Correlation coefficients are not inferential statistics; they are a summary of the relationship in the sample data (or population data if used).

Let's take a look at the correlation matrix in the middle. We'll go through each part. I've added letters to cells for reference.

Correlations

		POST: Feeling thermometer: Democratic Presidential candidate: Joe Biden		POST: Feeling thermometer: Democratic Vice Presidential candidate: Kamala Harris	
POST: Feeling thermometer: Democratic Presidential candidate: Joe Biden	Pearson Correlation	A	1	B	.909**
	Sig. (2-tailed)			D	<.001
	N	C	2617	C	2578
POST: Feeling thermometer: Democratic Vice Presidential candidate: Kamala Harris	Pearson Correlation	B	.909**	A	1
	Sig. (2-tailed)	D	<.001		
	N	C	2578	C	2598

E **. Correlation is significant at the 0.01 level (2-tailed).

A. This is the correlation coefficient value for the variable with itself. Because (x,y) are always identical and always match, this will always be a 1, a perfect relationship. This does not give you any useful information and can be ignored.

B. This is the correlation coefficient value for the relationship between the two variables. It is identical in both places because direction does not matter for correlation (this is about correlation, not causation). Here r=0.909, a positive strong relationship following the 2020 election between feelings towards Biden and feelings towards Harris.

C. N is the total number of cases (the sample size). Here we can observe there were 2,617 valid cases evaluating Joe Biden, 2,598 cases evaluating Kamala Harris, and 2,578 cases that had evaluations of both Biden and Harris. (The correlation matrix above is weighted, and the n here reflects weighted sample sizes. If I ran with weighting off, these numbers respectively would be 2,643, 2,635, and 2,613).

D. This is the p-value for the relationship. If we have a representative sample, we can use the p-value to evaluate significance. Here $p<0.001$, so we are over 99.9% confident that, among all U.S. eligible voters following the 2020 election, there was a relationship between how people felt about Biden and how people felt about Harris. This is our level of confidence that the correlation coefficient value is not zero, which would indicate no relationship.

E. The ** relates to the p-value from D. SPSS uses * for $p<0.05$ and ** for $p<0.01$. Oftentimes journal articles will use similar notation for significance levels, most commonly * for $p<0.05$, ** for $p<0.01$, and *** for $p<0.001$.

Correlation Matrices with three or more variables

The correlation matrix above only had two variables. However, you can also run a correlation matrix with more variables.

Here I put in 5 variables, using the data about U.S. states I was using earlier in the chapter.

Here are descriptive statistics for each variable:

Descriptive Statistics

	Mean	Std. Deviation	N
Official Poverty Rate (2022)	11.175	2.9167	51
SupplementalPovertyRate	8.708	2.2687	51
Life Expectancy (born 2020)	76.594	1.9551	51
% of population living in urban area (2020)	72.9745	15.16721	51
% voting for Biden (2020)	48.6384	12.04531	51

Here is the correlation matrix:

Correlations

		Official Poverty Rate (2022)	Supplemental PovertyRate	Life Expectancy (born 2020)	% of population living in urban area (2020)	% voting for Biden (2020)
Official Poverty Rate (2022)	Pearson Correlation	1	.724**	-.808**	-.169	-.202
	Sig. (2-tailed)		<.001	<.001	.235	.156
	N	51	51	51	51	51
SupplementalPovertyRate	Pearson Correlation	.724**	1	-.412**	.266	.262
	Sig. (2-tailed)	<.001		.003	.060	.063
	N	51	51	51	51	51
Life Expectancy (born 2020)	Pearson Correlation	-.808**	-.412**	1	.377**	.472**
	Sig. (2-tailed)	<.001	.003		.006	<.001
	N	51	51	51	51	51
% of population living in urban area (2020)	Pearson Correlation	-.169	.266	.377**	1	.538**
	Sig. (2-tailed)	.235	.060	.006		<.001
	N	51	51	51	51	51
% voting for Biden (2020)	Pearson Correlation	-.202	.262	.472**	.538**	1
	Sig. (2-tailed)	.156	.063	<.001	<.001	
	N	51	51	51	51	51

**. Correlation is significant at the 0.01 level (2-tailed).

This can be a lot to look at, but it's just more pairings. For example, the 0.724 is the r (correlation coefficient value) for the relationship between the official poverty rate and supplemental poverty rate, while -0.412 is the r for the relationship between supplemental poverty rate and life expectancy.

When you are in SPSS, you can check "Show only the lower triangle" at the bottom, or in syntax, replace the word "FULL" with "LOWER" in the "/PRINT" line, and will get a table that instead looks like this:

Correlations

		Official Poverty Rate (2022)	Supplemental PovertyRate	Life Expectancy (born 2020)	% of population living in urban area (2020)	% voting for Biden (2020)
Official Poverty Rate (2022)	Pearson Correlation	--				
	N	51				
SupplementalPovertyRate	Pearson Correlation	.724**	--			
	Sig. (2-tailed)	<.001				
	N	51	51			
Life Expectancy (born 2020)	Pearson Correlation	-.808**	-.412**	--		
	Sig. (2-tailed)	<.001	.003			
	N	51	51	51		
% of population living in urban area (2020)	Pearson Correlation	-.169	.266	.377**	--	
	Sig. (2-tailed)	.235	.060	.006		
	N	51	51	51	51	
% voting for Biden (2020)	Pearson Correlation	-.202	.262	.472**	.538**	--
	Sig. (2-tailed)	.156	.063	<.001	<.001	
	N	51	51	51	51	51

**. Correlation is significant at the 0.01 level (2-tailed).

This is the same information as the table above, just without duplications, and without showing a correlation coefficient for a variable with itself. I prefer the full table, but this is personal preference. Sometimes you will see correlation matrices like this one in journal articles.

Here is the confidence intervals output:

Confidence Intervals

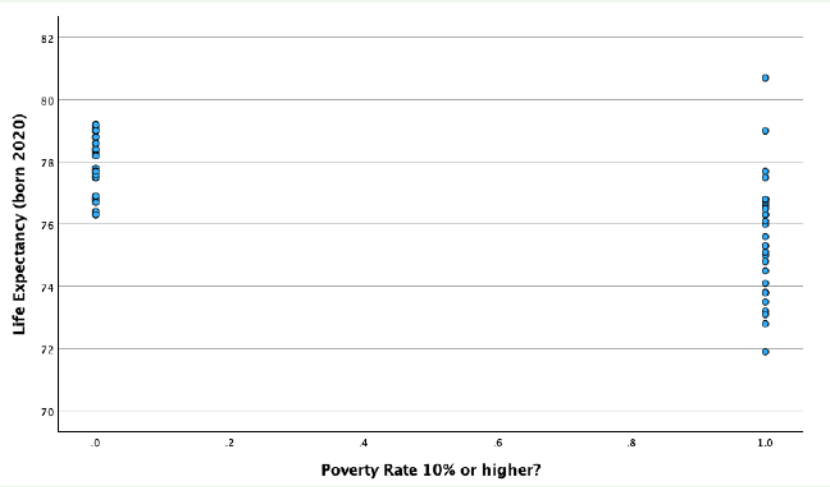
	Pearson Correlation	Sig. (2-tailed)	95% Confidence Intervals (2-tailed)[a] Lower	Upper
Official Poverty Rate (2022) - SupplementalPovertyRate	.724	<.001	.560	.833
Official Poverty Rate (2022) - Life Expectancy (born 2020)	-.808	<.001	-.886	-.685
Official Poverty Rate (2022) - % of population living in urban area (2020)	-.169	.235	-.425	.112
Official Poverty Rate (2022) - % voting for Biden (2020)	-.202	.156	-.452	.078
SupplementalPovertyRate - Life Expectancy (born 2020)	-.412	.003	-.618	-.154
SupplementalPovertyRate - % of population living in urban area (2020)	.266	.060	-.011	.504
SupplementalPovertyRate - % voting for Biden (2020)	.262	.063	-.015	.501
Life Expectancy (born 2020) - % of population living in urban area (2020)	.377	.006	.113	.591
Life Expectancy (born 2020) - % voting for Biden (2020)	.472	<.001	.225	.661
% of population living in urban area (2020) - % voting for Biden (2020)	.538	<.001	.308	.709

a. Estimation is based on Fisher's r-to-z transformation.

You can see that for each row it tells you what the two variable pair consists of (e.g., the first one is official poverty rate and supplemental poverty rate). I included the confidence intervals output so you could get a sense of it, but I would not use these confidence intervals nor the p-values in the correlation matrix above for this particular data, because the cases (50 states and DC) are not a representative sample generalizable to any broader target population.

Review: When to Use What

Scatterplots are good for ratio-like variables only. For example, check out this "scatterplot" of poverty rate and life expectancy by U.S. state, where poverty rate is an indicator variable (0=<10% poverty rate, 1=10% or higher poverty rate):



You're 10 chapters in, and we've now covered a whole host of different descriptive statistics, inferential statistics, charts, and graphs. Let's review when each should be used:

Descriptive statistics:
Mean: Needs a ratio-like variable, or an indicator variable (interpreted as proportion with value of 1)
Standard deviation: Needs a ratio-like variable
Index of Qualitative Variation: Needs a nominal variable
5 Point Summary (Median, Minimum, Maximum, Lower Quartile, Upper Quartile): Needs an ordinal, ratio-like, or interval variable
Mode: Any level of measurement (but not useful)
Range and InterQuartile Range: Needs a ratio-like variable, if ordinal can be described but usually not calculated
Gamma: Needs two ordinal-level variables (can use indicator variables)
Correlation: Needs two ratio-like variables (can use indicator variables, but they will not reflect a linear pattern)

Tables:
Frequency Distribution: Needs a categorical variable
Crosstabs: Needs 2 categorical variables (can use ratio-like if a small discrete number of answers)
Elaborated Crosstabs: Needs 3+ categorical variables (can use ratio-like if a small discrete number of answers)

Graphs:
Pie chart: Needs a categorical variable (with not too many categories)
Bar Graph: Needs a categorical variable
Histogram: Needs a ratio-like variable
Box and Whisker Plot: Needs a ratio-like variable
Scatterplot: Needs 2 ratio-like variables (with unique coordinates unless adjusting scatterplot)

Inferential statistics:
Only proceed if data is from a representative sample of your target population.
Confidence Interval: Needs a ratio-like variable, or an indicator variable (interpreted as proportion with value of 1)
Two-sample T-test: Needs 1 ratio-like variable, or 1 indicator variable (interpreted as proportion with value of 1); and 1 categorical variable with 2 selected categories
One-way ANOVA: Needs 1 ratio-like variable, or 1 indicator variable (interpreted as proportion with value of 1); and 1 categorical variable with multiple categories

Chi-square: Needs 2 categorical variables (can use ratio-like if a small discrete number of answers)
Elaborated chi-square: Needs 3+ categorical variables (can use ratio-like if a small discrete number of answers)

This page titled 10: Chapter 10. Scatterplots and Correlation is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Ezra Temko (Consortium of Academic and Research Libraries in Illinois (CARLI)) .

11: Regression Equations

Chapter 11 will continue our exploration of linear regression. In Chapter 10, we added lines of best fit (regression lines) to scatterplots. In Chapter 11 we will run linear regression using SPSS, explore lines of best fit based on the information in our outputs, and graph our regression lines using Excel.

Statistics in Action: Misbeliefs---vaccines and autism

One of the questions asked as part of the American National Election Studies to eligible U.S. voters following the 2020 election was,

"Which of these two statements do you think is most likely to be true?

- Most scientific evidence shows childhood vaccines cause autism
- Most scientific evidence shows childhood vaccines do not cause autism"

After answering, respondents were asked, "How confident are you about that?"

Here is a weighted frequency distribution table of the data:

Most scientific evidence shows childhood vaccines cause autism

		Frequency	Valid Percent	Cumulative Percent
Valid	No, Extremely confident	673	26.3	26.3
	No, very confident	634	24.8	51.2
	No, Moderately confident	630	24.7	75.9
	No, A little confident	236	9.2	85.1
	No, not at all confident	112	4.4	89.5
	Yes, not at all confident	21	.8	90.4
	Yes, a little confident	70	2.7	93.1
	Yes, moderately confident	89	3.5	96.6
	Yes, very confident	50	2.0	98.6
	Yes, extremely confident	37	1.4	100.0
	Total	2553	100.0	
Missing	System	90		
Total		2643		

I coded the variable -5 through -1 and 1 through 5 to make it ratio-like, with -5 representing respondents who were extremely confident that most scientific evidence shows childhood vaccines do not cause autism and +5 representing respondents who were extremely confident that most scientific evidence shows childhood vaccines cause autism.

We can run a correlation matrix to reveal the strength and direction of this variable in relation with others, along with evaluating significance.

The variables selected, in order of appearance below, include age, education (0 less than high school through 4 graduate degree), political views (-3 extremely liberal to 3 extremely conservative), 7 feelings thermometer variables (0 cold to 100 warm), the extent to which two statements describe the respondent's views (1 not at all to 5 extremely well), and three statements, with the respondent identifying which they think is most likely to be true, and how confident they are about their answer (-5 Disagree, extremely confident to 5 agree, extremely confident).

I only am showing the first row of the matrix, excluding the variable's correlation with itself.

Correlations

		Respondent Age	What is the highest level of school you have completed or the highest degree you have received?	Where would you place yourself on this scale?	POST: Feeling thermometer: Dr. Anthony Fauci	POST: Feeling thermometer: World Health Organization (WHO)	POST: Feeling thermometer: scientists	POST: Feeling thermometer: Jews	POST: Feeling thermometer: Muslims	POST: Feeling thermometer: Christians	POST: Feeling thermometer: Christian fundamentalists
Most scientific evidence shows childhood vaccines cause autism	Pearson Correlation	-.036	-.231	.197	-.248	-.188	-.234	-.087	-.134	.119	.261
	Sig. (2-tailed)	.069	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001	<.001
	N	2494	2528	2117	2515	2294	2537	2517	2524	2534	2492

Correlations

		Much of what people hear in schools and the media are lies designed to keep people from learning the real truth about those in power	Most business and politics in this country are secretly controlled by the same few powerful people	Russia did not try to intefere in 2016 presidential election	World temperatures have not risen on average over the last 100 years	The novel coronavirus (COVID-19) was developed intentionally in a lab
Most scientific evidence shows childhood vaccines cause autism	Pearson Correlation	.267	.137	.192	.234	.300
	Sig. (2-tailed)	<.001	<.001	<.001	<.001	<.001
	N	2545	2547	2541	2549	2545

The strongest correlation (r=0.300) was with the variable that asked whether the respondent agreed more that, "One, the novel coronavirus (COVID-19) was developed intentionally in a lab; or Two, the novel coronavirus (COVID-19) was not developed intentionally in a lab." Let's look at this one further, in relation to the question about vaccines and autism.

Let's return to the line of best fit, or regression line. To find out the equation for our line, we first will run a bivariate linear regression. Here's how to do that in SPSS:

Running Linear Regression in SPSS

Which variable is my IV and which is my DV? Timewise I'm assuming that for most people, evaluations of COVID would come after evaluations of vaccines and autism, since COVID-19 was a year or less old at the time of this survey. That would make the COVID variable the DV and the vaccines and autism variable the IV. However, I'm interested in predicting how people feel about the vaccine/autism variable based on how they feel about the COVID one, so I'm going to use the vaccine/autism variable as my DV. I don't think this is the right causal order, so I'm not arguing for a causal relationship, but I'm interested in the relationship between the two variables and whether based on someone's feelings about COVID's origins I can predict their feelings on the science being clear that vaccines causing autism, so I'm setting up my variables in this way.

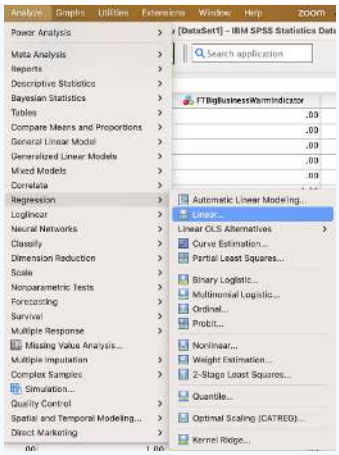
Syntax:

Here is the syntax for running a bivariate OLS linear regression in SPSS. Just replace "DVName" with the name of your dependent variable and IVName with the name of your independent variable.

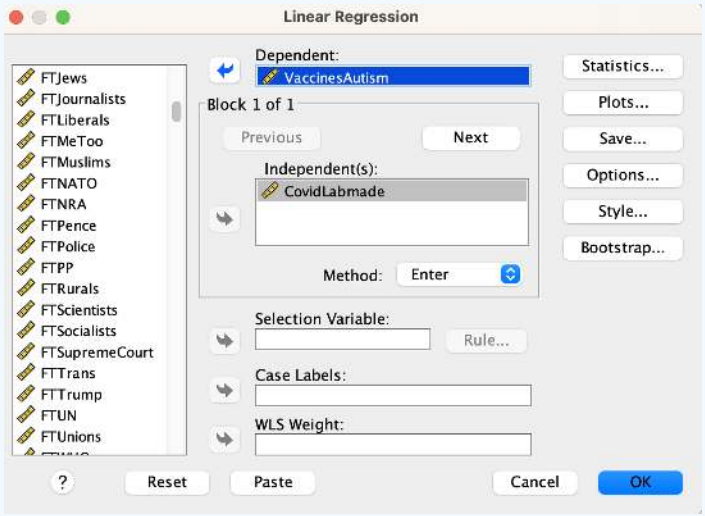
```
REGRESSION
/DESCRIPTIVES MEAN STDDEV CORR SIG N
/MISSING LISTWISE
/STATISTICS COEFF OUTS CI(95) R ANOVA
/CRITERIA=PIN(.05) POUT(.10) TOLERANCE(.0001)
/NOORIGIN
/DEPENDENT DVName
/METHOD=ENTER IVName.
```

Menu:

Go to Analyze → Regression → Linear

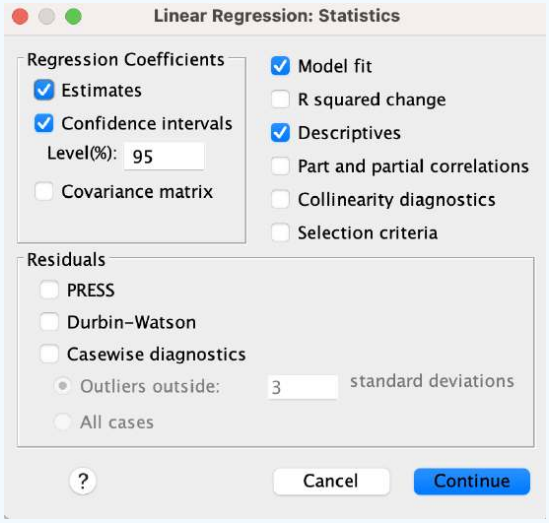


Put your independent variable in the “Independent(s)” box and your dependent variable in the “Dependent” box.



Then click on the "Statistics..." button to the right.

In the statistics window, turn on descriptives and confidence intervals.



Click the "Continue" button and then click the "OK" button.

Here's my output. Chapter 11 will only cover some of the output, but we'll come back to more of it in Chapter 12.

1. Descriptive statistics and Correlations Matrix:

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Most scientific evidence shows childhood vaccines cause autism	-2.9635	2.35387	2545
The novel coronavirus (COVID-19) was developed intentionally in a lab	.1742	3.66253	2545

Correlations

		Most scientific evidence shows childhood vaccines cause autism	The novel coronavirus (COVID-19) was developed intentionally in a lab
Pearson Correlation	Most scientific evidence shows childhood vaccines cause autism	1.000	.300
	The novel coronavirus (COVID-19) was developed intentionally in a lab	.300	1.000
Sig. (1-tailed)	Most scientific evidence shows childhood vaccines cause autism	.	<.001
	The novel coronavirus (COVID-19) was developed intentionally in a lab	.000	.
N	Most scientific evidence shows childhood vaccines cause autism	2545	2545
	The novel coronavirus (COVID-19) was developed intentionally in a lab	2545	2545

Above in the "Correlations" table you can find the correlation coefficient for the relationship. However, note that unlike the correlation matrices above or in Chapter 10, the corresponding p-values here are "1-tailed," so we won't use them.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	The novel coronavirus (COVID-19) was developed intentionally in a lab[b]	.	Enter

a. Dependent Variable: Most scientific evidence shows childhood vaccines cause autism

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.300[a]	.090	.090	2.24572

a. Predictors: (Constant), The novel coronavirus (COVID-19) was developed intentionally in a lab

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1270.612	1	1270.612	251.943	<.001[b]
	Residual	12825.462	2543	5.043		
	Total	14096.074	2544			

a. Dependent Variable: Most scientific evidence shows childhood vaccines cause autism

b. Predictors: (Constant), The novel coronavirus (COVID-19) was developed intentionally in a lab

Above in the "Model Summary" output you can also find "R" which is the correlation coefficient. However, a word of caution: get your r value from the correlation matrix near the top of the output, not from this model summary. This R is actually its absolute value. The strength is accurate, but it will be positive whether or not the correlation coefficient is positive. This is because as a "model summary" it takes into account the full model, which could include multiple independent variables, some of which may have positive and some of which may have negative relationships with the dependent variable. The model summary refers to the overall strength of the whole model. This will be useful in Chapter 13, but for now you can just ignore this output.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-2.997	.045		-67.253	<.001	-3.085	-2.910
	The novel coronavirus (COVID-19) was developed intentionally in a lab	.193	.012	.300	15.873	<.001	.169	.217

a. Dependent Variable: Most scientific evidence shows childhood vaccines cause autism

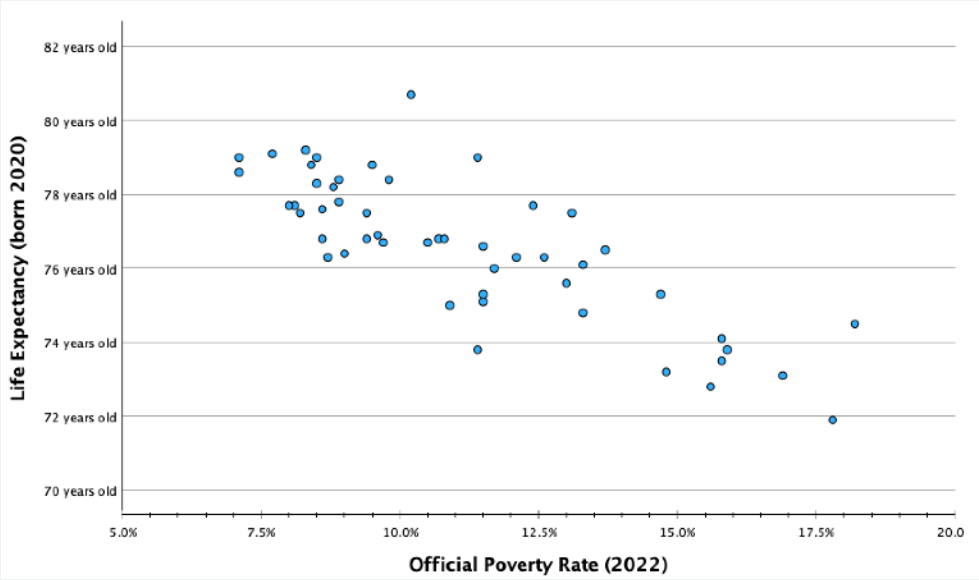
This "Coefficients" output is our focus for Chapter 11. It has the information we need to put together our regression equation, as well as additional data. After introducing you to linear equations, we'll return to this table.

Statistics in Action: U.S. State Poverty Rates & Life Expectancy

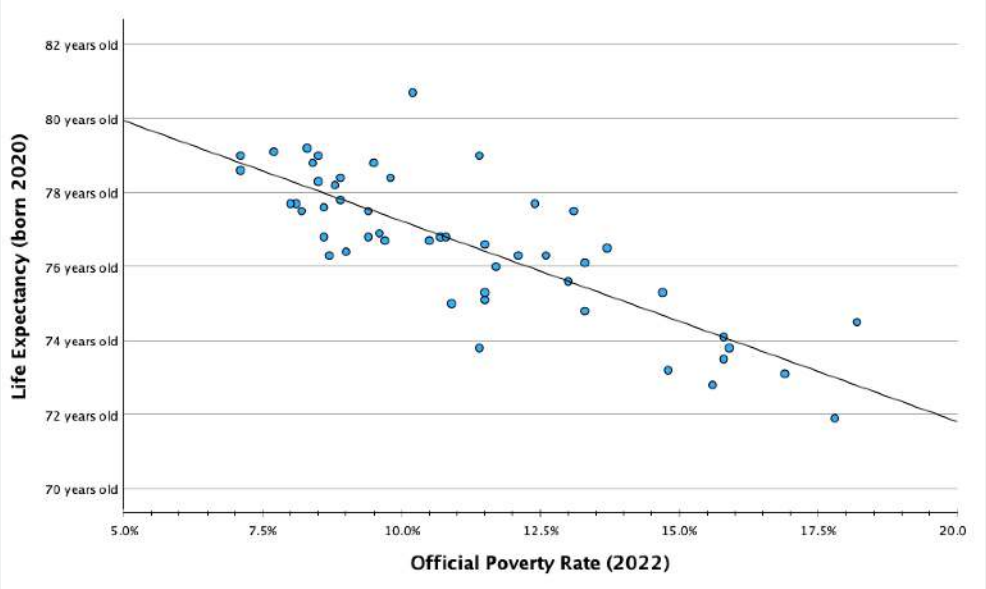
Regression Equations

Let's return to our example from Chapter 10, of the official poverty rate and life expectancy for each U.S. state (and DC).

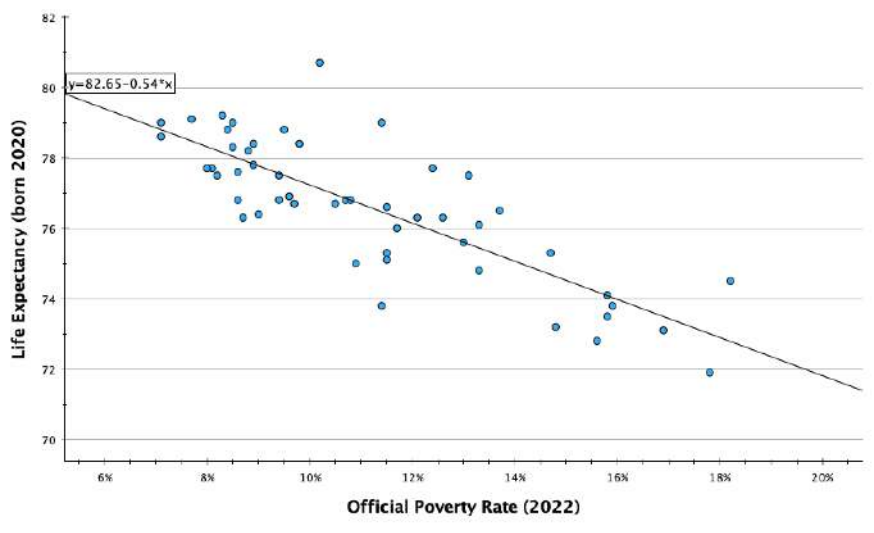
Here was our scatterplot:



Here was the scatterplot with the line of best fit:



Recall that linear regression is the process of fitting a line to the data such that it minimizes the sum of the squared difference between the data points and the line. When you add your regression line, or line of best fit, to your scatterplot in SPSS, it also gives you the equation for the line (deleted above):



y=82.65 + -0.54(x)

Linear Equation

The equation is in the form y=a+bx. (This is similar to y=mx+b that you may have learned in secondary school, but the order and letters are different.)

y: dependent variable value

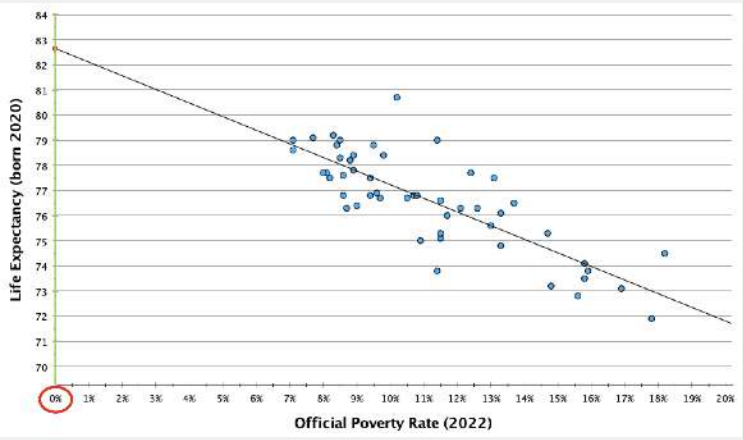
a: y-intercept (the y-value for the point on the line with an x-value of 0)

b: slope (how much the dependent variable value changes for every one-unit increase in the independent variable value)

x: independent variable value

Y-Intercept:

The y-value in the coordinate (0,y). The y-intercept is the value y would be if x=0. It is the y-value where the line crosses the y-axis.



y=a+bx
y=82.65 + -0.54(x)
The y-intercept, a, is 82.65.

For this equation, if x=0, y=82.65.

Sometimes the y-intercept can be interpreted sensically in real-life context, and has a story. In this case it doesn't have an empirical story since there are not any states that would have zero poverty (a 0% poverty rate, x=0), and we cannot assume that the best fit line would continue to follow the same pattern when it gets closer to and to zero. There could, for example, be diminishing returns or exponential returns on life expectancy as poverty rates approach zero. However, theoretically, the y-intercept here means that if a state had a 0% poverty rate, we would predict that the life expectancy for that state would average 82.65 years. The y-intercept can be interpreted whenever x=0 has real-life meaning and there is data at or nearby x=0.

Interpreting the y-intercept in context

***Template**:* I would predict that [respondents / name unit of analysis] who/that [describe what it means for the independent variable to equal 0] would [describe what it means to be at the y-intercept value for the dependent variable].

Examples:

I would predict that someone who feels absolutely cold towards labor unions (x=0) would feel fairly cold/unfavorable towards Obama (y=23.11 on the 0-100 feelings thermometer scale).

A state with a 0% poverty rate would theoretically be predicted to have a life expectancy of 82.7 years. However, all states have poverty rates of 7% or higher, so we don't have enough information to make such a prediction.

Mathematically the y-intercept means that we would predict a 0-year-old would have slightly warm feelings towards journalists (54.9 degrees). However, this survey was only of adults, so such a prediction would not make any sense.

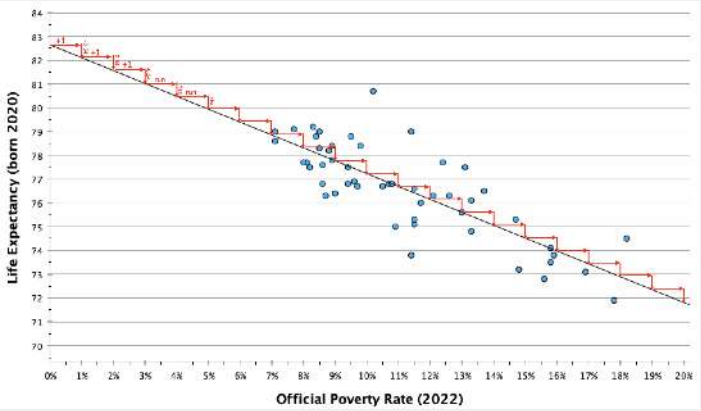
Slope:

y=a+bx
y=82.65 + -0.54(x)
The slope, b, is -0.54.

For every one-unit increase in x, how much does y increase?

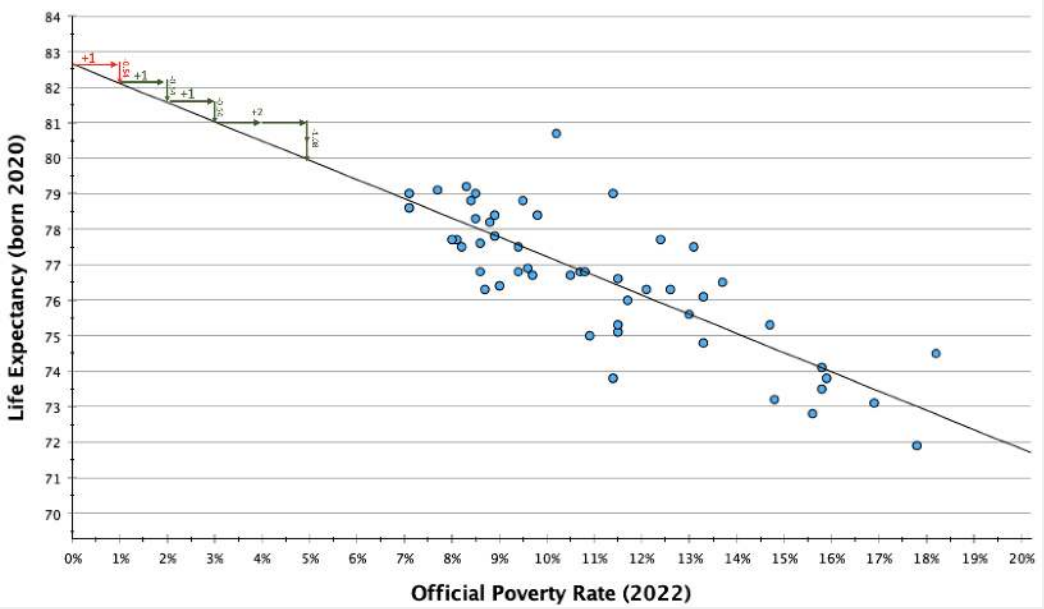
For every one-unit increase in x, meaning for every 1% increase in the official poverty rate, there is a -0.54 year increase in y, meaning a 0.54 year (about half a year) decrease in life expectancy.

You may have heard slope referred to as rise over run. It is calculated by taking the difference in y between two coordinates and dividing it by the difference in x (to make the difference in y standardized into the difference for every one unit change in x). In the graph below, you can see how for every one unit increase in x (in this case, a 1% increase in the official poverty rate), the line decreases by about half a year.



In context, -0.54 can be kind of confusing. I prefer working with whole numbers and estimates when I explain slopes. One way to do this is to calculate multiples until I get something a bit more interpretable. This can also be useful for providing a second example that scales up the change to give your audience a broader perspective of what kind of change they might expect in your DV as the IV changes.

Change in IV	DV
1	-0.54
2	-1.08
3	-1.62
4	-2.16
5	-2.7
6	-3.24
7	-3.78
8	-4.32
9	-4.86
10	-5.40



In the table and graph above, you can see that for every two-unit change in x, y decreases by about 1 unit. So every 2% increase in the official poverty rate is associated with a 1 year decrease in life expectancy.

Interpreting slope in context

Template: On average, every one [IV unit label] increase in [a respondent's/name unit of analysis] [variable description] is associated with an increase/decrease of [slope value] [DV unit label] in [DV variable description].

If not easily digestible numbers, or to scale up:
Similarly, every [multiple of one] [IV unit label] in [variable description] is associated with a [slope value times same multiple] [DV unit label] in [DV variable description].

Examples:

On average, a 1% increase in a state's poverty rate is associated with a 0.54 year decrease in the state's average life expectancy. A 2% increase in the poverty rate is associated with about a 1 year decrease in life expectancy. Similarly, a 5% increase in the poverty rate (1 in 20 more residents) is associated with a decrease in average life expectancy of about 2 1/2 (2.7) years

On average, every one degree warmer a respondent feels towards labor unions on the 0-100 feelings thermometer scale is associated with an increase of 0.68 degrees warmer in how they feel towards Obama on the 0-100 feelings thermometer scale. Every 3 degrees warmer toward labor unions is associated with 2 degrees warmer towards Obama. Similarly, every 15 degrees warmer respondents feel towards labor unions is associated with 10 degrees warmer in feelings towards Obama.

On average, every one year older a respondent's age is is associated with a marginal change of 0.086 degrees colder feelings towards journalists (using the 0-100 feelings thermometer scale). Someone who is a decade older would still be less than a degree (0.86) colder. Being 60 years older is associated with over 5 (5.16) degrees colder feelings toward journalists.

Predictions

Using the line of best fit, for any independent variable value, we can make predictions of our dependent value. However, while the linear equation is the line of "best fit," depending on the strength of the relationship, it may be a better or worse fit for the data. The more the data converges into a common linear pattern, the better our predictions will be.

To make a prediction, I select a value for the independent variable, plug it into the equation for x, and then calculate the y-value.

Let's say I wanted to predict the life expectancy for a state with a 10% poverty rate.

y=a+bx
y=82.65 + -0.54(x)
y=82.65 + -0.54(10)
y=82.65 + -5.4
y=77.25

I would predict that a state with a 10% poverty rate has a life expectancy of 77.25 years.

Let's say I wanted to predict the life expectancy for a state with a 15% poverty rate.

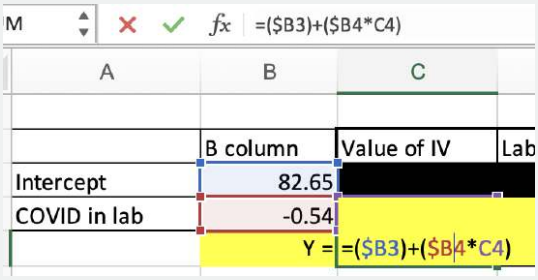
y=a+bx
y=82.65 + -0.54(x)
y=82.65 + -0.54(15)
y=82.65 + -8.1
y=74.55

I would predict that a state with a 15% poverty rate has a life expectancy of 74.55 years.

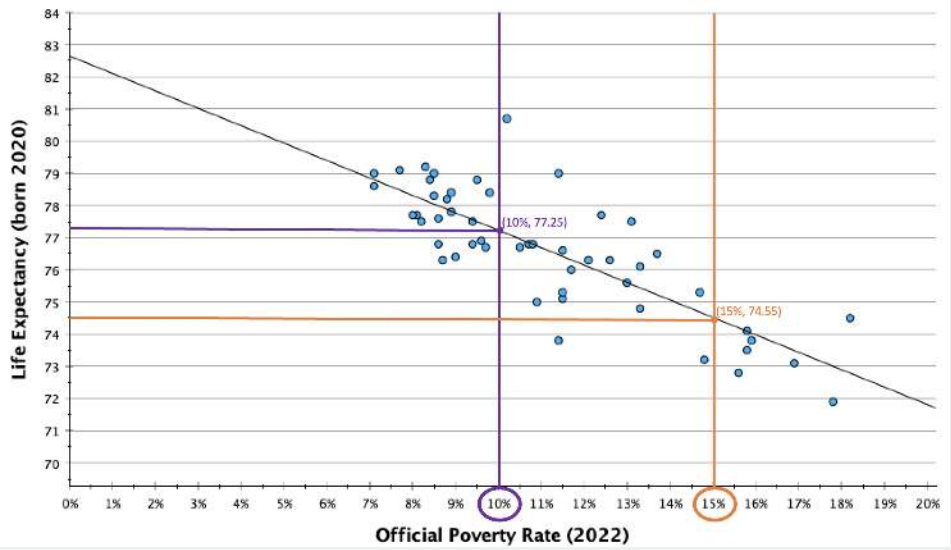
If you don't want to do this math, I've provided a spreadsheet in Excel: BivariateLinearRegressionGraphTemplate.xlsx
The fifth worksheet tab, "PredictedValueCalculator," lets you put in your intercept value, IV slope value, and IV value you want to calculate the predicted DV value for, and it will calculate it for you.

	A	B	C	D	E	F	G	H
1								
2		B column	Value of IV	Label for value	Value of IV	Label for value	Value of IV	Label for value
3	Intercept	82.65						
4	COVID in lab	-0.54	0	poverty rate %	10	poverty rate %	15	poverty rate %
5		Y =	82.650	years LE	77.250	years LE	74.550	years LE

This is doing the same thing as if you were to calculate it on your own. You can see the formula below. This is just doing the math for you.

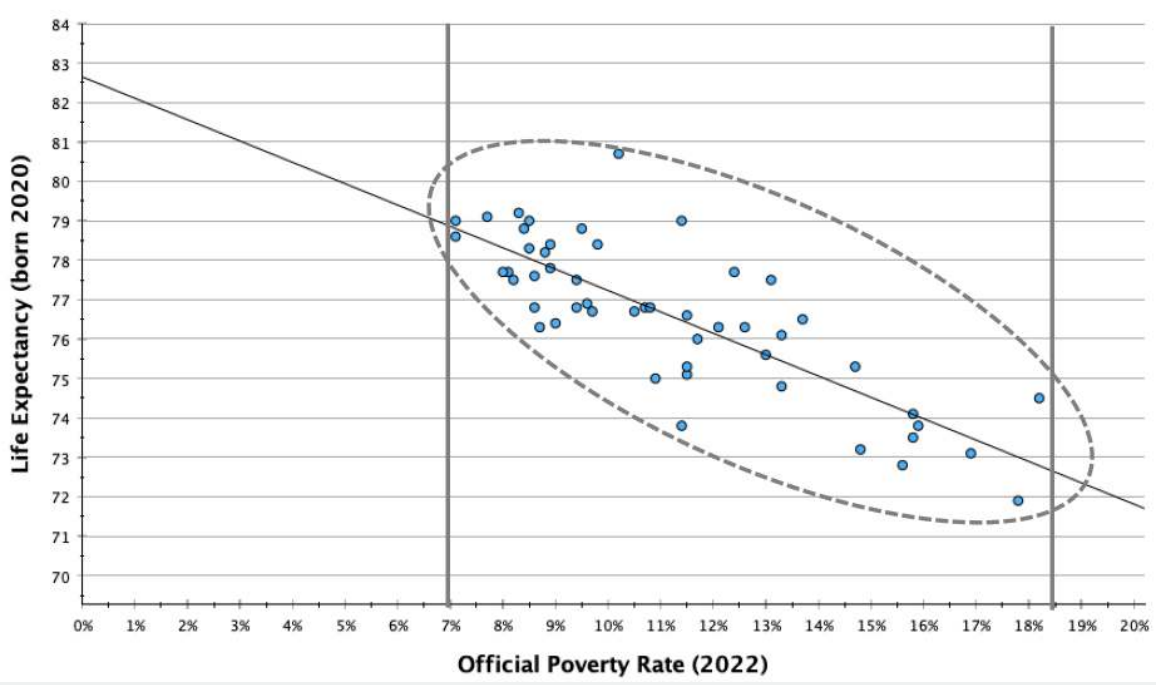


Here you can see how the life expectancy predictions for a 10% poverty rate and a 15% poverty rate match up with the line. The coordinates (10,77.25) and (15,74.55) both fall on the line.



You can also see that no states are actually at these coordinates, and most states are not directly on the line. These are our predictions, and absent a perfect relationship, we expect variability such that our predicted values will not be an exact match for what the values would actually be in real life. They are an estimate, and the stronger the relationship, the more likely our estimate is to be accurate.

I mentioned before that we wouldn't want to make a prediction for a poverty rate of 0% (our y-intercept), even though it would be great to have a state eliminate poverty. However, our line of best fit is based on the data we have available. There were no states with a 7% or lower poverty rate and no states with a 19% or higher poverty rate. We want to contain our predictions to the range of the data available, since we don't know what the trend line would look like outside of the data we have. The pattern could look the same, or it could be quite different.



In general, only make predictions between your minimum and maximum IV values (and within a narrower range if one or both of these are outliers).

Statistics in Action: Misbeliefs---vaccines and autism

Let's return to the SPSS output from our linear regression of views on whether COVID-19 was intentionally developed in a lab and views on whether most scientific evidence suggests childhood vaccines cause autism.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	a -2.997	.045		-67.253	<.001	-3.085	-2.910
	x The novel coronavirus (COVID-19) was developed intentionally in a lab	b .193	.012	.300	15.873	<.001	.169	.217

y a. Dependent Variable: Most scientific evidence shows childhood vaccines cause autism

We can write out our regression equation from values in the coefficients box.

Our two coefficients will be in the "B" column under "Unstandardized Coefficients."

The y-intercept, a, is in the row marked "(Constant)".

The slope, b, is in the row marked with the label/description of the IV.

y=a+bx
Here, y = -2.997 + 0.193x
y, our dependent variable, is agreement with the statement "Most scientific evidence shows childhood vaccines cause autism."
x, our independent variable, is agreement with the statement, "The novel coronavirus (COVID-19) was developed intentionally in a lab."

Confidence Intervals:
We can also see 95% confidence intervals for these two coefficients. The regression equation is a descriptive statistic. If you have a representative sample, you can look at the confidence interval to evaluate where, at the 95% confidence level, the y-intercept and slope would be. For example, here, while the slope was 0.193 in the sample, we are 95% confident that, among all U.S. eligible voters following the 2020 election, the slope would be somewhere between 0.169 and 0.217.

P-values:
The "Sig." column contains p-values. These are the probability we would pull this sample if B=0.

- For the y-intercept, this has meaning but is not that important. While we are predicting someone who is at a 0 for the IV (neutral on the statement) would be at about a -3 (moderately confident that the scientific evidence does *not* show childhood vaccines cause autism), this p-value is evaluating our confidence level that this is not actually a 0 (neutral on whether the scientific evidence shows childhood vaccines cause autism). If I look at the cell in the row labeled "(Constant)" and the column labeled "Sig.", I find a p-value of <0.001. I am over 99.9% confident that, among all U.S. eligible voters, when x=0, y is not also 0.
- For the slope, the p-value is evaluating whether there is a significant relationship between the IV and DV. Recall from Chapter 10 that a flat horizontal line indicates no relationship. A flat line would have a slope of 0. For every one unit increase in x, there is a 0 unit increase in y. What's the probability we would pull this sample if the slope was actually 0 (if there was actually no relationship)? If I look at the cell in the row labeled with the IV label/description and the column labeled "Sig.", I find a p-value of <0.001. I am over 99.9% confident that, among all U.S. eligible voters, there is a relationship between people's beliefs about whether COVID-19 was intentionally developed in a lab and whether most scientific evidence shows childhood vaccines cause autism. Because p<0.05 and I am over 95% confident there is a relationship among my target population, the relationship is significant.

Making predictions

Just like above, we can use the regression equation to make predictions.

For example, what if x=-5 (extremely confident COVID-19 was *not* intentionally made in a lab).
y = -2.997 + 0.193x
y = -2.997 + 0.193*-5
y = -3.962

Below I again used the Excel spreadsheet to make predictions for three different IV values:

	A	B	C	D	E	F	G	H
1								
2		B column	Value of IV	Label for value	Value of IV	Label for value	Value of IV	Label for value
3	Intercept	-2.997						
4	COVID in lab	0.193	-5	No, extremely confident	0	Neutral	5	Yes, extremely confident
5		Y =	-3.962	No, very confident	-2.997	no, moderately confident	-2.032	No, a little confident

Graphing bivariate linear equations

We're also going to use this Excel spreadsheet to graph the linear equation. This will look like the lines of best fit above, but without a scatterplot.

BivariateLinearRegressionGraphTemplate.xlsx

The first tab/worksheet in the file is labeled "BivariateGraph." Here is what it looks like:

y=a+bx

Replace cells B3 and B4 with the intercept and slope from your regression line equation.

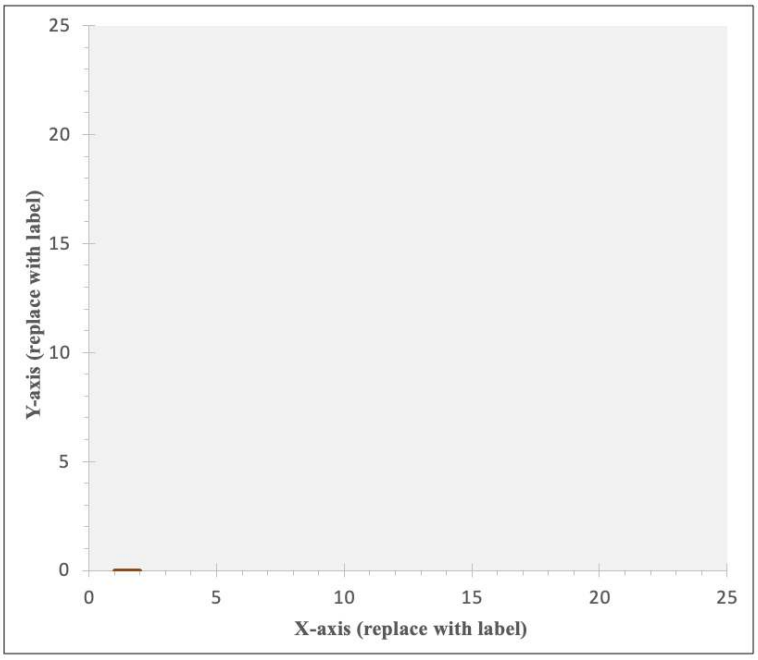
intercept a	#
slope b	#

Replace cells B9 and B10 with the minimum and maximum value for your independent variable.

	x (IV)	y (DV)
minimum x:	#	#VALUE!
maximum x:	#	#VALUE!

Make the following changes to the graph:

- ☐ Replace the x-axis label with the label for your IV
- ☐ Replace the y-axis label with the label for your DV
- ☐ Edit the minimum and maximum bounds for the x-axis to go at least from your minimum IV through your maximum IV
- ☐ Edit the major and minor units for the x-axis to reflect sensical intervals.
- ☐ Edit the minimum and maximum bounds for the y-axis to go at least from your minimum DV through your maximum DV
- ☐ Edit the major and minor units for the y-axis to reflect sensical intervals.



The first thing to do is to input the intercept, slope, minimum IV value, and maximum IV value (the cells shaded yellow). If you're not sure what the minimum and maximum IV values are, you can run descriptive statistics in SPSS to find out.

https://socialsci.libretexts.org/@go/page/190127

y=a+bx

Replace cells B3 and B4 with the intercept and slope from your regression line equation.

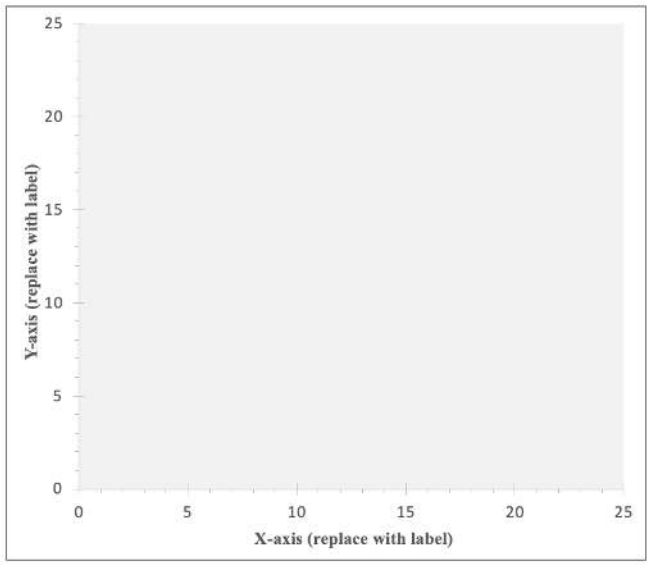
intercept a	-2.997
slope b	0.193

Replace cells B9 and B10 with the minimum and maximum value for your independent variable.

	x (IV)	y (DV)
minimum x:	-5	-3.962
maximum x:	5	-2.032

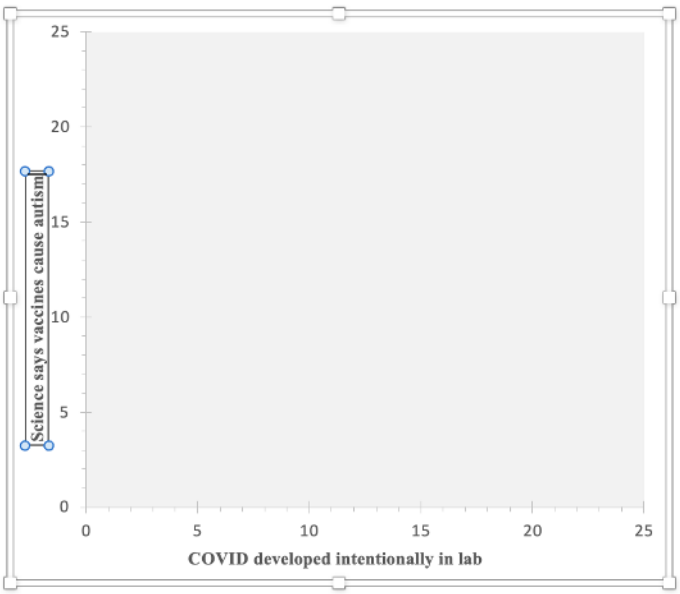
Make the following changes to the graph:

- ☐ Replace the x-axis label with the label for your IV
- ☐ Replace the y-axis label with the label for your DV
- ☐ Edit the minimum and maximum bounds for the x-axis to go at least from your minimum IV through your maximum IV
- ☐ Edit the major and minor units for the x-axis to reflect sensical intervals.
- ☐ Edit the minimum and maximum bounds for the y-axis to go at least from your minimum DV through your maximum DV
- ☐ Edit the major and minor units for the y-axis to reflect sensical intervals.

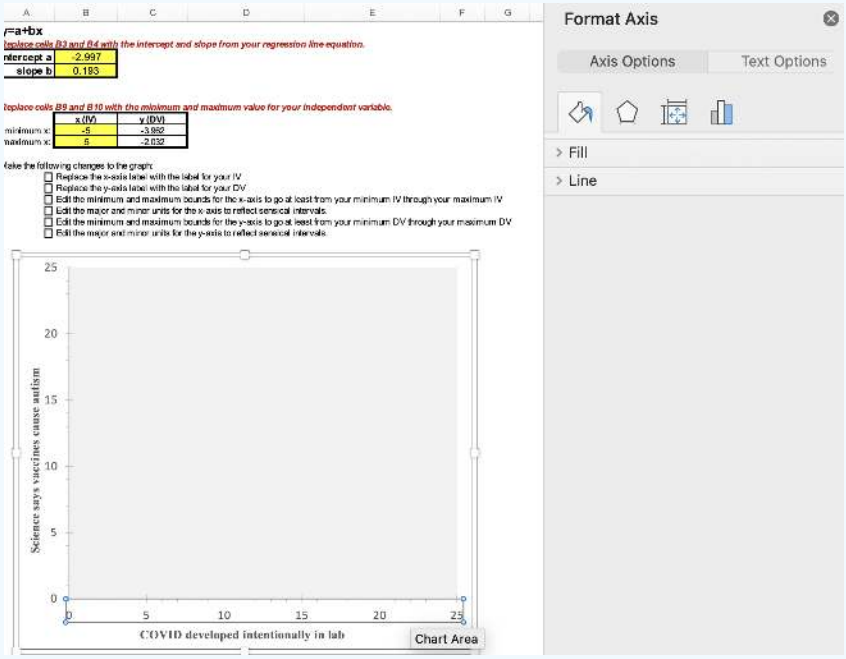


I added in the four values (the intercept of -2.997, the slope of 0.193, the minimum IV value of -5, and the maximum IV value of 5), but the graph is still blank. The line is actually there, but it's outside of the current scope of the graph. The line is connecting two points, one at (-5,-3.962) and one at (5,-2.032). Since the y-axis on the graph template goes from 0 to 25, none of those negative values are appearing within the visible area.

Before I address that, I'm going to update where it says "X-axis (replace with label)" and "Y-axis (replace with label)", by clicking on these and typing in a description of the variable.



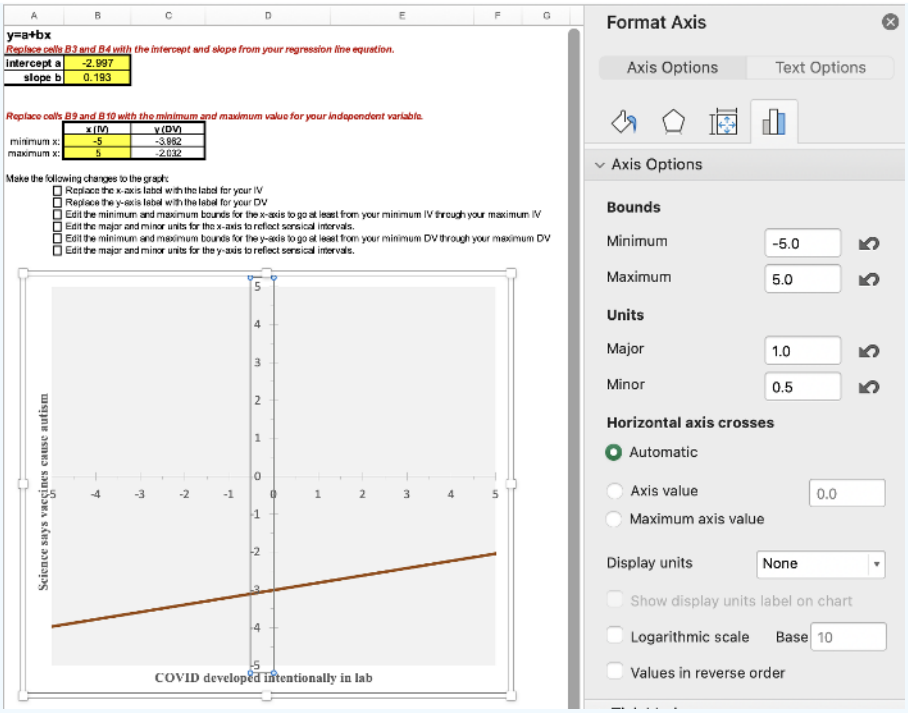
Now I am going to change the graph area so that my line shows up. Double-click one at a time on the x-axis and on the y-axis. When you do there will be a pop-up. Here's what happened when I clicked on one of the numbers on the x-axis:



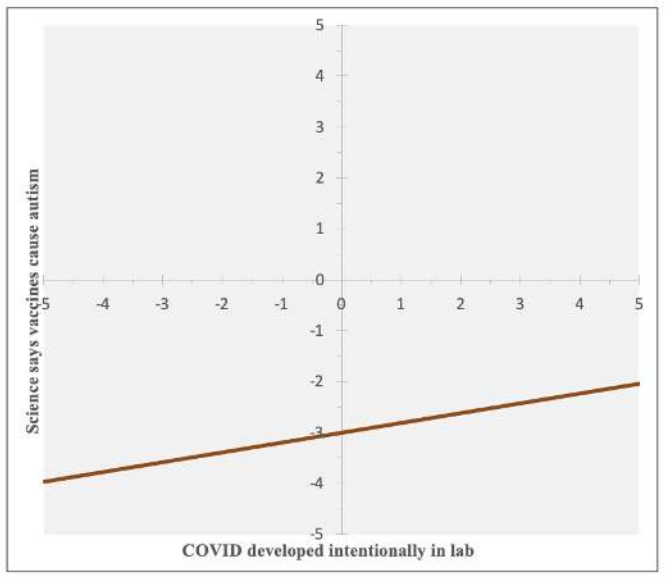
Under Axis Options, click on the picture icon that looks like a bar graph or histogram. Then you will change the bounds and units. The bounds for the x-axis need to go from at least your minimum IV value through your maximum IV value. The major units are the interval for what numbers will be displayed. For example, if your x-axis goes from 0 to 100, you could decide to show 0, 10, 20, 30, etc. with a major unit of 10. The minor units are for marks on the x-axis in between the major labeled units. You can decide how big or small you want these for ease of viewing the graph. Think about what your major unit numbers will easily break into and what will add information without overwhelming. For example, if you are doing 0, 10, 20, 30, etc., with a major unit of 10, maybe you use a minor unit of 2, so that there are 4 little dash marks in between each major unit number label. You can always experiment with the units to see what looks good on the graph.

Next, double-click on the y-axis, and again edit the bounds and units. For the y-axis, the bounds need to start at or below the y (DV) value to the right of your minimum x value (cell C9) and through the value in cell C10. Depending on the variable and its coding, you may want to display the full range of possible values or you may want to zoom in. If you do zoom in, it's important to make sure your audience is aware that the line looks more substantive/steeper than it actually is.

For this example, I set the bounds for both the IV and DV as going from -5 to +5, which are my minimum and maximum IV and DV values. With a range of 10, I decided on a major unit of 1 and then breaking each of those into a mark halfway between them.



Here is my final graph:



In this graph, you can find the coordinates for my earlier calculated predictions: (-5, -3.962), (0, -2.997), and (5,-2.032).

You can see that there is a positive relationship, that as likelihood increases in believing COVID-19 was intentionally developed in a lab, respondents are more likely to believe that most scientific evidence shows childhood vaccines cause autism. However, you can also see that the range is from about -4 to about -2, so across the entire spectrum of responses to the question about COVID, on average respondents were likely to not believe that most scientific evidence shows childhood vaccines cause autism. We would not predict that anyone would lean towards believing this based solely on their views on COVID, but our predictions would change from someone who is extremely confident that COVID-19 was not intentionally developed in a lab being very confident that most scientific evidence does not show childhood vaccines cause autism to, for someone who is extremely confident that COVID-19 was intentionally developed in a lab, predicting they would be a little confident that most scientific evidence does not show childhood vaccines cause autism.

Statistics in Action: Misbeliefs---vaccines and autism

Let's look at one more example, and also discuss here 1) **substantiveness**, 2) **what to do if your line of best fit goes above or below the bounds of where your dependent variable values can actually be**, and 3) **evaluating "ratio-like" variables as candidates for linear regression**.

Example:

We'll keep the same dependent variable, but this time use education as our independent variable. Here is the codebook information on our IV:

What is the highest level of school you have completed or the highest degree you have received?
0. Less than high school credential
1. High school credential
2. Some post-high school, no bachelor's degree
3. Bachelor's degree
4. Graduate degree

Here is the coefficients output from running a weighted linear regression:

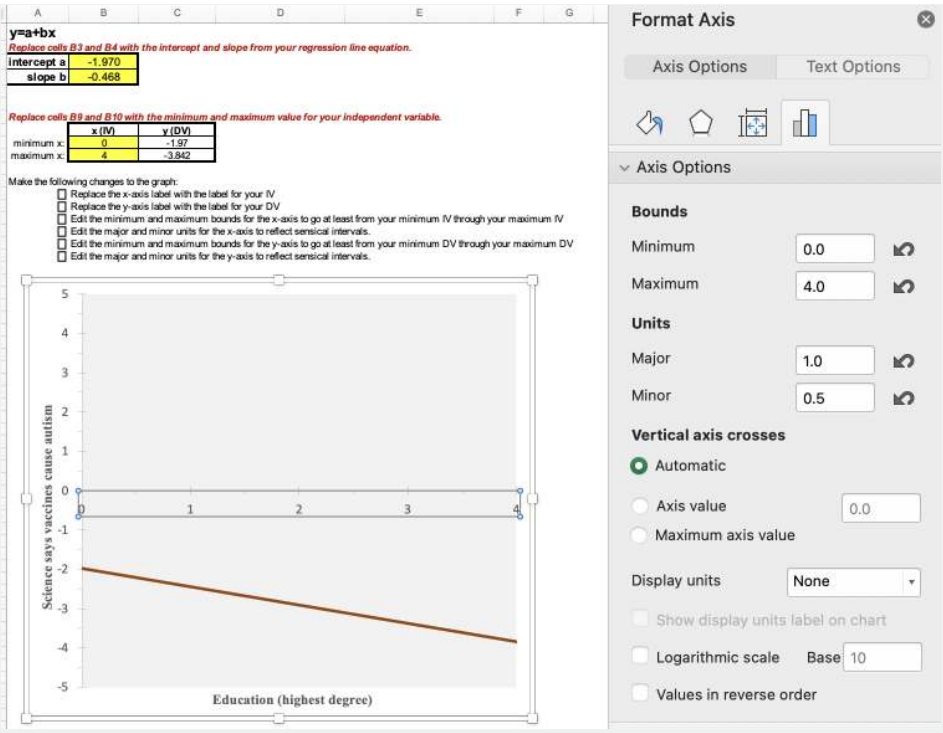
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-1.970	.095		-20.770	<.001	-2.156	-1.784
	What is the highest level of school you have completed or the highest degree you have received?	-.468	.039	-.231	-11.917	<.001	-.545	-.391

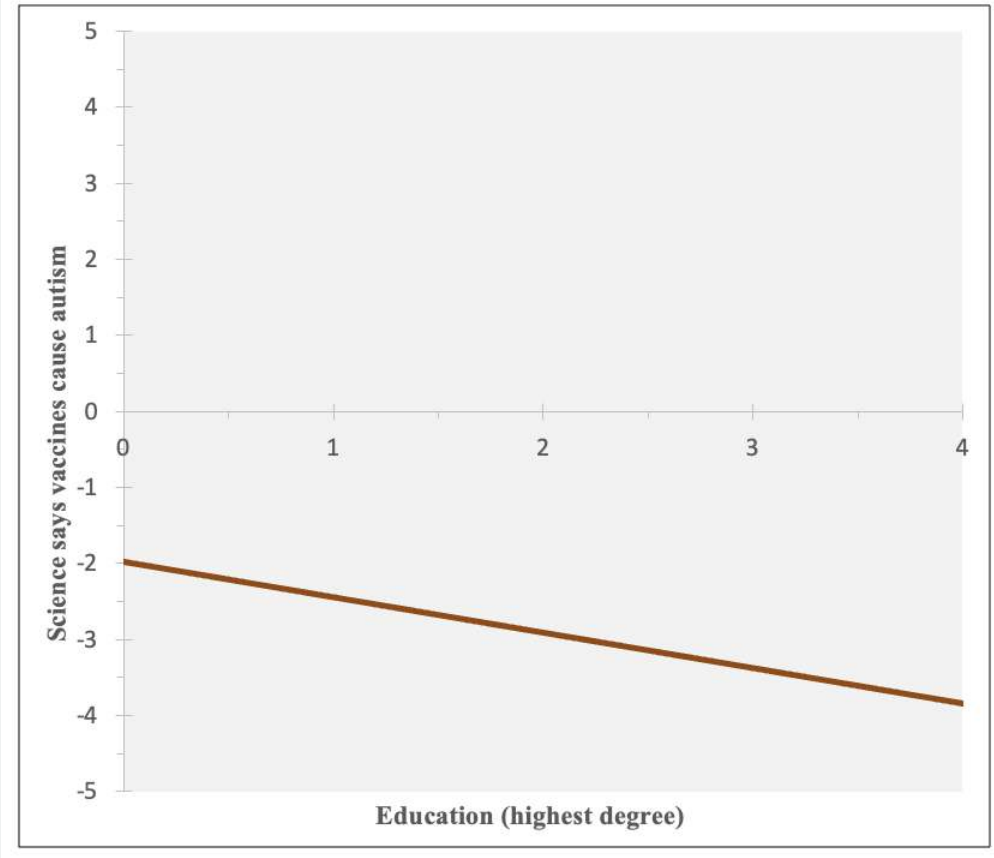
a. Dependent Variable: Most scientific evidence shows childhood vaccines cause autism

With the p-value on the bottom row (the one that starts, "What is the highest level of school...") being <0.001, we know there is a statistically significant relationship between education and whether people believe that most scientific evidence shows childhood vaccines cause autism.

Here you can see what numbers I put in for the Excel template, along with how I adjusted the x-axis bounds and units. I was able to leave the y-axis bounds and units alone since it was the same dependent variable.



Here is a graph of the equation, y = -1.970 + (-0.468)x



You can see that there is a negative relationship --- on average, as education increases, respondents are less likely to believe that most scientific evidence shows childhood vaccines cause autism.

Those with less than a high school credential are predicted at around -2 (a little confident most scientific evidence does not show childhood vaccines cause autism), while those with a graduate degree are predicted almost at -4 (very confident most scientific evidence does not show childhood vaccines cause autism).

1) **Substantiveness**

For a relationship to be substantive, it first needs to exist. If your data is of a representative sample, is the relationship significant? If you are not confident that the relationship exists among your target population, you won't claim it is substantive. (In the example above, $p<0.001$ and the relationship is significant.)

The next criterion to evaluate is the steepness of the slope. How much does the DV change as the IV changes? Is this a sizable and meaningful amount? When you go from low IV values to high IV values, what is the change to the predicted DV values? Is this a sizable and meaningful amount? (In the example above, the DV values range from very confident to a little confident that most scientific evidence does not show childhood vaccines cause autism. Do you think this is substantive?)

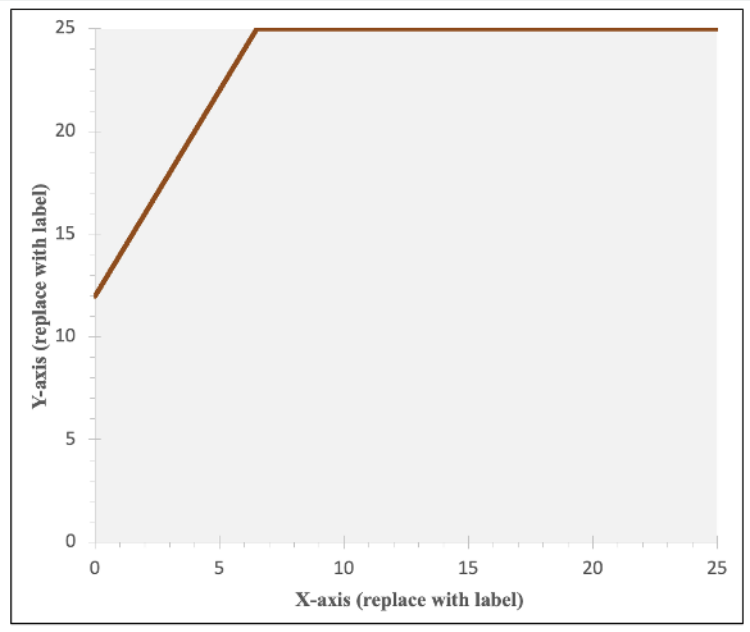
If the slope seems substantive, and you are using representative sample data, you should also look at the slope's 95% confidence interval. If you have a significant relationship, the confidence interval should not overlap zero. Use the bound closer to zero. You are 95% confident that, among your target population, the slope is at least that amount. (In the example above, while the sample slope is -0.468, we are 95% confident that the slope is at least -0.391. Going from 0 to 4 for education is over a 1.5 unit difference in the dependent variable. In context, is that substantive?)

In Chapter 12, we'll also consider r-squared, the "coefficient of determination", to make a determination of substantiveness.

And finally, remember that even if a relationship is significant and substantive, that does not mean it's necessarily causal. It still needs to have the correct causal order, make sense theoretically, and not be spurious.

2) **What to do if your line of best fit goes above or below the bounds of where your dependent variable values can actually be**

Sometimes the line of best fit goes above or below where it actually can be in real life. For example, in the graph above, y can't be below -5 or above +5. If you were using a feelings thermometer variable as your DV, in real life it can't go below 0 (very cold) or above 100 (very warm). If this happens, you can graph your regression line within the bounds of your DV range. In the template, there are three additional worksheets/tabs to use depending on your situation. These will help change the graph so that you either start, end, or start and end with a flat horizontal line at your y-value bound, keeping the predicted values within what is actually empirically possible.



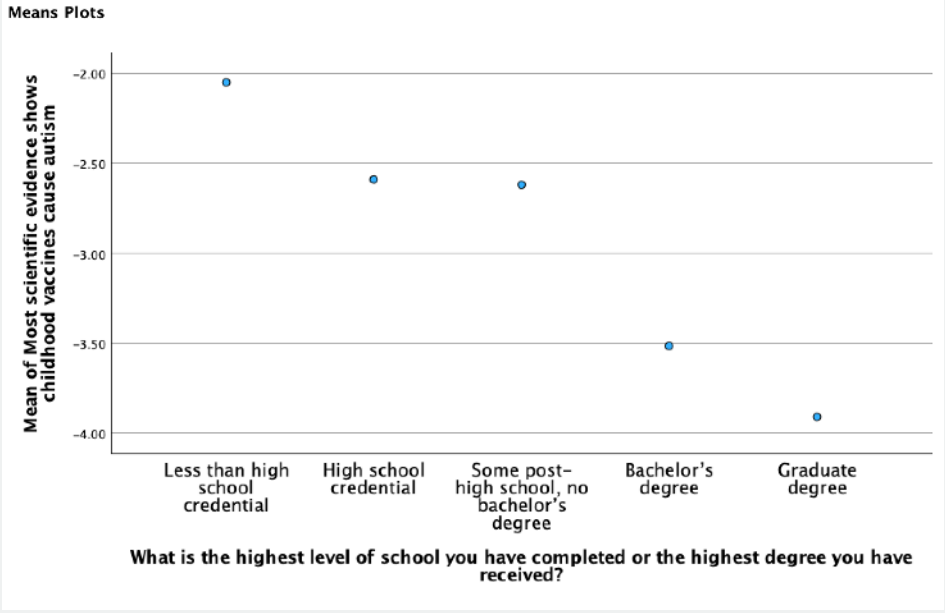
3) **Evaluating "ratio-like" variables as candidates for linear regression.**

Is the relationship between education and evaluation of what most scientific evidence on vaccines and autism concludes a linear relationship?

Sometimes with degree, relationships are not linear. For example, it may be that those with less than a high school degree are very supportive of raising the minimum wage, those with a high school degree a little less supportive, those with a college degree even a little less supportive, but then those with a graduate degree are as supportive of as those without a high school degree. This would not be ordered correctly for a linear pattern. Or what if those with some college and college are basically the same in regards to your dependent variable while otherwise the progression is linear? Should you check for this and combine those categories? Age, even as a ratio-level variable measured in years, might also have a non-linear relationship rather than the dependent variable either continuously decreasing or increasing with age. In other situations, you may have an ordinal categorical variable that you want to know if you can use a a ratio-like variable as an IV for linear regression, perhaps one that stretches the definition of ratio-like (e.g., one with only four categories).)

For the above example, I checked the mean of my dependent variable, broken down by the independent variable.

This graph comes from what we did in Chapter 7 on hypothesis tests, where we generated these together with a one-way ANOVA test:



The output below comes from what we did in Chapter 3, where we used Examine/Explore to break means for a primary variable down by another grouping variable:

Descriptives

	What is the highest level of school you have completed or the highest degree you have received?		Statistic
Most scientific evidence shows childhood vaccines cause autism	Less than high school credential	Mean	-2.0506
	High school credential	Mean	-2.5909
	Some post-high school, no bachelor's degree	Mean	-2.6188
	Bachelor's degree	Mean	-3.5151
	Graduate degree	Mean	-3.9081

What I'm looking for in these is if there is a somewhat continuous pattern of similarly sized increases or decreases, unidirectional, as the IV increases. If the pattern doesn't look linear (e.g., it changes direction, flattens out, or it appears like a curve with a changing speed), this is a sign the IV may not be appropriate in its current form for linear regression. Depending on the situation, there are different techniques you can use to still use the IV. You may transform it, collapse some categories together, exponentiate it, adjust the coding, or otherwise change it to better reflect a linear pattern. You may end up using it in

indicator form. In Chapter 12, we'll look at situations where you can use a dichotomous independent variable. Depending on the situation, sometimes you can collapse a ratio-like variable into a dichotomous variable (e.g., perhaps what matters for age is whether someone is <65 or 65+). Or, sometimes you will want to leave more categories, (e.g., maybe you want 3 or more degree categories, or age groups). In that case you can use something called reference group sets, which will be discussed in Chapter 13 when we look at nominal-level independent variables.

This page titled 11: Regression Equations is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Ezra Temko (Consortium of Academic and Research Libraries in Illinois (CARLI)) .

12: Linear Regression, continued

Building on Chapter 11, this chapter will explain the remaining components of bivariate OLS linear regression, such as r-squared, the coefficient of determination. Next, this chapter will discuss how to conduct and interpret bivariate OLS linear regression using an indicator-level independent variable.

Statistics in Action: Misbeliefs---vaccines and autism

Let's pick up with our Chapter 11 example. In Chapter 11, we looked at whether there was a relationship between people's beliefs about whether COVID-19 was intentionally developed in a lab and whether they thought most scientific evidence shows childhood vaccines cause autism. We determined from the correlation coefficient, Pearson's r, that there was a moderately weak positive relationship (r=0.300). We graphed an equation y=-2.997+0.193x, made predictions based on the equation, and interpreted the slope of the variable as well as whether the relation was significant. We found that, following the 2020 election, on average, the more likely a U.S. eligible voter believed COVID-19 was intentionally developed in a lab, the more likely they were to believe that most scientific evidence shows childhood vaccines cause autism. However, overall the average respondent, regardless of their belief about whether COVID-19 was intentionally developed in a lab, did not believe that most scientific evidence shows childhood vaccines cause autism.

There are three remaining areas we need to go over: 1) r-squared, the coefficient of determination; 2) overall model significance; and 3) sample size.

1. R-Squared, the Coefficient of Determination

Here is the first SPSS output from the regression, Model Summary, which we did not interpret in Chapter 11:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.300[a]	.090	.090	2.24572

a. Predictors: (Constant), The novel coronavirus (COVID-19) was developed intentionally in a lab

As was noted in Chapter 11, the cell R is the absolute value of Pearson's R, so it gives us strength but not direction. Instead look at the correlation matrix for the correlation value of the bivariate relationship, which gives us both strength and direction. (When we get to multivariate regression, with multiple independent variables, this r value will be the strength for the overall model, so the numbers will be different than the individual bivariate correlations in the correlation matrix.)

The next cell is "R Square."

Definition: Coefficient of Determination

The **coefficient of determination, r^2** (r-squared), is the proportion of the dependent variable's variance explained by the independent variable.

R-squared is a "goodness of fit" measure that tells us the independent variable's explanatory power. It is a descriptive sample statistic.

Let's say I was guessing your age. There's a lot of variation in ages. If all I knew was that you were a person in the United States and that the average age in the U.S. is 39.1, my best guess would be that you are 39 years old. However, if I knew something else about you, I may be able to make a more accurate guess. For example, if I knew you were a college student, I'd guess a younger age. Being or not being a college student explains some of the variation in age. It doesn't explain all of it, just some of it. In some cases my prediction would be less accurate. But overall, I can make better predictions of age. If there's a

significant relationship between college student status and age, I can make better predictions of your age if I know you are a college student.

Interpreting r^2

Template

[x] explains [r^2 x 100] of the variance in [y].

Example

Above, r^2=0.090. 0.090x100=9. I am just converting this proportion into a percent.

x, my IV, is belief that COVID-19 was intentionally developed in a lab
y, my DV, is belief that most scientific evidence shows childhood vaccines cause autism

Putting them together:
[x] explains [r^2 x 100] of the variance in [y].

Belief that COVID-19 was intentionally developed in a lab explains 9% of the variance in belief that most scientific evidence shows childhood vaccines cause autism.

Other examples:

Physical health explains about one quarter (25.4%) of the variance in mental health.

The percent of a country saying most people can be trusted explains 23.3% of the variance in the percent of a country saying being born in the country is a very important requirement for citizenship.

Behind the scenes

Where does the r^2 value come from?

First, it's literally r-squared, so for example above r=0.300, so r^2=0.300*0.300=0.090, so 9%.

Here is the actual formula for r^2:

Sum Squared Regression Error

$$R^2 = 1 - \frac{SS_{Regression}}{SS_{Total}}$$

Sum Squared Total Error

The coefficient of determination is based on explained variation divided by total variation.

Here, the sum of the squared regression error is based on how much each of the points in our scatterplot differ from the line of best fit (squared and summed). Recall that our line of best fit is based on minimizing this statistic. So these residuals are variation from our regression line.

Next, the sum of squared total error ignores the independent variable. It is how much each dependent variable value differs from the mean of the dependent variable (squared and summed) This is variance (standard deviation squared).

The unexplained variance in our regression line divided by the unexplained variance for the dependent variable as a whole gives us the proportion of unexplained variance, the proportion of the dependent variable's variance the independent variable does not explain. We subtract that from 1 to find the proportion of the dependent variable's variance that the independent variable does explain.

To evaluate whether a relationship is substantive, consider r-squared in addition to significance and the 95% confidence interval of the slope. (Or if you are using population data, consider r-squared and the slope.)

2. Overall significance of the model

Here is the second output we did not interpret in Chapter 11:

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1270.612	1	1270.612	251.943	<.001[b]
	Residual	12825.462	2543	5.043		
	Total	14096.074	2544			

a. Dependent Variable: Most scientific evidence shows childhood vaccines cause autism

b. Predictors: (Constant), The novel coronavirus (COVID-19) was developed intentionally in a lab

We use the p-value in the cell below "Sig." to evaluate significance for the overall model.

While this is an ANOVA (analysis of variance) test, the p-value should basically match the p-value of our slope in the Coefficients output, since we have a biviarate regression (so the overall model only includes the one independent variable). When we get to multivariate regression with multiple independent variables, these will no longer always match.

The ANOVA F-test here is basically determining the probability that we would draw the sample we did if the r-squared value was actually 0 among our target population. If $p<0.05$, we are over 95% confident that, among our target population, the independent variable explains more than 0% of the variance in the dependent variable. Recall that when we did a one-way ANOVA test comparing group means we had a between group sum of squares and a within group sum of squares. Here you can see the "regression sum of squares" and "residual sum of squares," concepts discussed above in the r^2 behind the scenes.

Hypothesis testing for bivariate linear regression

Null hypothesis: There is no relationship between x and y.
(Statistically, this is that knowing *x* does not improve prediction of *y* over just knowing the mean of *y*.)

Alternative hypothesis: There is a relationship between *x* and *y*.
(Statistically, this is that taking *x* into account improves prediction of *y* over just knowing the mean of *y*.)

3. Sample size

Recall that our regression output included descriptives and a correlation matrix:

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Most scientific evidence shows childhood vaccines cause autism	-2.9635	2.35387	2545
The novel coronavirus (COVID-19) was developed intentionally in a lab	.1742	3.66253	2545

Correlations

		Most scientific evidence shows childhood vaccines cause autism	The novel coronavirus (COVID-19) was developed intentionally in a lab
Pearson Correlation	Most scientific evidence shows childhood vaccines cause autism	1.000	.300
	The novel coronavirus (COVID-19) was developed intentionally in a lab	.300	1.000
Sig. (1-tailed)	Most scientific evidence shows childhood vaccines cause autism	.	<.001
	The novel coronavirus (COVID-19) was developed intentionally in a lab	.000	.
N	Most scientific evidence shows childhood vaccines cause autism	2545	2545
	The novel coronavirus (COVID-19) was developed intentionally in a lab	2545	2545

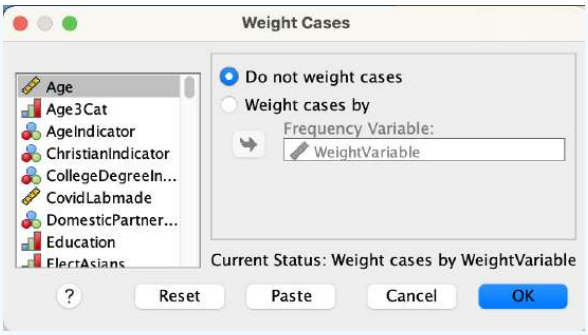
You can see a few places where N=2545 is listed. This is the weighted sample size, 2,545. Remember that weighting adjusts each case's weight to make the sample better match the target population. However, because of this, the weighted sample size is not the actual number of respondents.

To find the actual sample size, the number of respondents in the sample that we have useable data on for both our independent and dependent variables, **we need to find the unweighted sample size. Turn weighting off and then re-run your regression.**

Turning Weighting Off

Syntax: WEIGHT OFF.

Menu:
Go to Data → Weight Cases. Click "Do not weight cases" and click OK.



Here's the unweighted output. The unweighted sample size is 2,589. Use the weighted regression for everything else.

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Most scientific evidence shows childhood vaccines cause autism	-3.1089	2.24515	2589
The novel coronavirus (COVID-19) was developed intentionally in a lab	-.1421	3.69090	2589

Correlations

		Most scientific evidence shows childhood vaccines cause autism	The novel coronavirus (COVID-19) was developed intentionally in a lab
Pearson Correlation	Most scientific evidence shows childhood vaccines cause autism	1.000	.352
	The novel coronavirus (COVID-19) was developed intentionally in a lab	.352	1.000
Sig. (1-tailed)	Most scientific evidence shows childhood vaccines cause autism	.	<.001
	The novel coronavirus (COVID-19) was developed intentionally in a lab	.000	.
N	Most scientific evidence shows childhood vaccines cause autism	2589	2589
	The novel coronavirus (COVID-19) was developed intentionally in a lab	2589	2589

Display Table

The SPSS outputs from regression are not presentation-ready. They have a lot of extra information and the information included is not set up in an easily digestible format. What we'll do here is take the most relevant information and put it into a display table that can be used for your own analysis as well as for communicating to public audiences.

 Attached here are templates for bivariate regression display tables. We'll focus on the first template in the document here, the second template below, and the final templates in Chapter 13.

Table #. Estimated unstandardized coefficients from an OLS regression predicting DV, with 95% confident intervals

Independent Variable	IV Name?
Intercept	Coefficient Value? (LB, UB)
Slope	Coefficient Value? (LB, UB)
r^2	r-squared*100%
Unweighted sample size	n

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Source: Dataset Year

Your job is to fill out the red parts.

The question mark superscripts are based on your p-values. Next to IV Name you'll use the p-value from the ANOVA output, and then for the intercept and slope you'll use the p-values on those two rows of the coefficients output. Use the key below the table to figure out what symbol to use. If p<0.001, use three asterisks. If p<0.01 but not <0.001, use two asterisks. If p<0.05 but not <0.01, use one asterisk. If p<0.1 but not <0.05, use a tilda. If p≥0.1, do not use any symbol.

LB and UB are for the lower and upper bound of the 95% confidence interval.

In general, I suggest rounding to the nearest hundredths place.

Here's the display table for the regression we've been looking at:

Table One. Estimated unstandardized coefficients from an OLS regression predicting belief that most scientific evidence shows childhood vaccines cause autism, with 95% confident intervals

Independent Variable	CovidLabmade***
Intercept	-3.00*** (-3.09, -2.91)
Slope	0.19*** (0.17, 0.22)
r^2	9%
Unweighted sample size	2589

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Source: ANES 2020 post-election

The red parts are what I filled out. Here is the display table without color differentiation:

Table One. Estimated unstandardized coefficients from an OLS regression predicting belief that most scientific evidence shows childhood vaccines cause autism, with 95% confident intervals

Independent Variable	CovidLabmade***
Intercept	-3.00*** (-3.09, -2.91)
Slope	0.19*** (0.17, 0.22)
r^2	9%
Unweighted sample size	2589

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Source: ANES 2020 post-election

Caution: Assumptions for OLS

In order to trust the validity of your OLS regression results, there are some underlying assumptions that must be met. You can read more about these at https://statisticsbyjim.com/regression/ols-linear-regression-assumptions/ and https://statisticsbyjim.com/regression/gauss-markov-theorem-ols-blue/. You will likely need to test that these assumptions are met if you doing a thesis involving regression or trying to publish your regression in a scholarly publication.

Bivariate Linear Regression with an Indicator IV

So far we've only looked at bivariate linear regression with a ratio-level or ratio-like independent variable. However, what if you want to include a nominal-level variable (e.g., gender, race, geographic region, religion, etc.) as a predictor? **We can also use indicator variables as independent variables for linear regression.** Recall that indicator variables are dichotomous variables in which one response is coded as 0 and the other is coded as 1, giving them special properties. *You still cannot use them as the dependent variable (for that, you would turn to something called logistic regression).* The category coded 0 is the "reference group," the category you are comparing the other to. Sometimes people will name indicator variables with the value for the code of 1 (e.g., Voted, VoteYes, GenderMan, Worried, Midwest, POC). This can help with interpretation.

Statistics in Action: Gender pay gaps

Let's return to our example from earlier in this textbook about the gender pay gap (see Chapter 3 and Chapter 7). Again, we're using 2017 to 2021 American Community Survey data. Our target population below is full-time, year-round, non-institutionalized U.S. workers ages 16+.

SPSS:

There are no differences in running bivariate linear regression when your independent variable is an indicator variable. Just use the indicator variable for your independent variable. However, your independent variable must be coded 0 and 1. In Chapter 7, sex was coded as 1 and 2. In order to use it here for linear regression, I had to re-code it to 0 and 1.

Output:

Here is the output from the bivariate linear regression, with gender as the independent variable and income as the dependent variable.

Weighted:

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Total personal income	75269.21	81139.601	112499634
Gender (1=female)	.4291	.49495	112499634

Correlations

		Total personal income	Gender (1=female)
Pearson Correlation	Total personal income	1.000	-.134
	Gender (1=female)	-.134	1.000
Sig. (1-tailed)	Total personal income	.	<.001
	Gender (1=female)	.000	.
N	Total personal income	112499634	112499634
	Gender (1=female)	112499634	112499634

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Gender (1=female)[b]	.	Enter

a. Dependent Variable: Total personal income

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.134[a]	.018	.018	80410.685

a. Predictors: (Constant), Gender (1=female)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.325E+16	1	1.325E+16	2048844.879	<.001[b]
	Residual	7.274E+17	112499632	6465878249		
	Total	7.407E+17	112499633			

a. Dependent Variable: Total personal income

b. Predictors: (Constant), Gender (1=female)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	84676.981	10.034		8439.359	<.001	84657.315	84696.646
	Gender (1=female)	-21924.729	15.317	-.134	-1431.379	<.001	-21954.750	-21894.707

a. Dependent Variable: Total personal income

Unweighted:

Descriptive Statistics

	Mean	Std. Deviation	N
Total personal income	79121.74	85918.126	5052460
Gender (1=female)	.4344	.49568	5052460

Correlations

		Total personal income	Gender (1=female)
Pearson Correlation	Total personal income	1.000	-.141
	Gender (1=female)	-.141	1.000
Sig. (1-tailed)	Total personal income	.	<.001
	Gender (1=female)	.000	.
N	Total personal income	5052460	5052460
	Gender (1=female)	5052460	5052460

Display Table

The display table is almost identical to the one we used above. However, the y-intercept is always interpretable, as it is the predicted value when x=0, and x has meaning at 0 and at 1. Therefore, we add descriptives of what the IV categories are for the intercept and slope.

Table #. Estimated unstandardized coefficients from an OLS regression predicting DV, with 95% confident intervals

Independent Variable	IV Name?
Intercept (label=0)	Coefficient Value? (LB, UB)
Slope (label=1)	Coefficient Value? (LB, UB)
r^2	r-squared*100%
Unweighted sample size	n

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Source: Dataset Year

Here's the display table for the regression:

Table Two. Estimated unstandardized coefficients from an OLS regression predicting income, with 95% confident intervals

Independent Variable	Gender***
Intercept (male=0)	$84,676.98*** ($84,657.32, $84,696.65)
Slope (female=1)	$-21,924.73*** ($-21.954.75, $-21,894.71)
r^2	1.8%
Unweighted sample size	5,052,460

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Source: ACS 5-Year (2017-2021)

The red parts are what I filled out. Here is the display table without color differentiation:

Table Two. Estimated unstandardized coefficients from an OLS regression predicting income, with 95% confident intervals

Independent Variable	Gender***
Intercept (male=0)	$84,676.98*** ($84,657.32, $84,696.65)
Slope (female=1)	$-21,924.73*** ($-21.954.75, $-21,894.71)
r^2	1.8%
Unweighted sample size	5,052,460

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Source: ACS 5-Year (2017-2021)

These interpretations are all the same as before:

We are over 99.9% confident that there is a relationship between gender and income (for 2017 to 2021, among full-time, year-round, non-institutionalized U.S. workers ages 16+).

Here, the intercept, $84,676.98, is the predicted salary estimate for men. We are 95% confident that for all men in the target population, average salary is somewhere between $84,656.32 and $84,696.65.

Gender explains 1.8% of variance in salary.

There were over 5 million people in the sample.

What's Different?

Our interpretation of the slope and our graph will be different.

Interpreting Slope (indicator IV)

Here the slope is $-21,924.73. On the one hand, it works the same way. On average, a one unit increase in the IV is associated with a slope-value increase in the DV. Here that means that, on average a one-unit increase in gender is associated with a $21,924.73 decrease in income. However, what is a "one-unit increase in gender"? Based on the coding, this is going from 0 to 1, or from male to female. With the variable being dichotomous, there are no additional unit increases. We also would not typically describe this as an "increase." Instead, we word our interpretation as a comparison between the category coded as 1 and the reference group.

Template: [IV=1 category name]'s average [DV construct name] is predicted to be [slope value] higher than [IV=0 category name]'s average [DV construct name].
If the slope is negative, share the absolute value for the slope value and change "higher" to "lower."

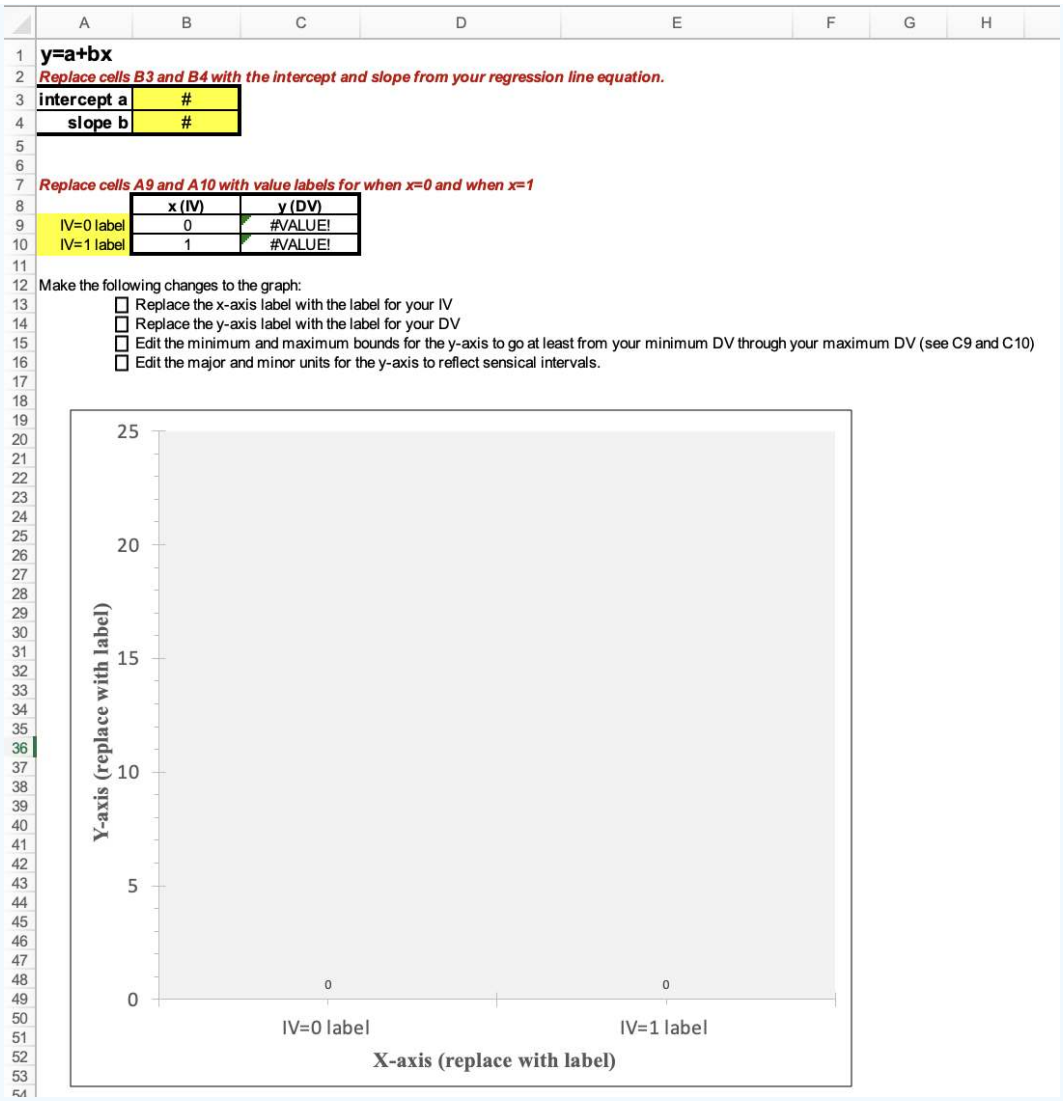
Example: Women's average income is predicted to be $21,924.73 lower than men's average income.

Graph (indicator IV)

With a ratio-level or ratio-like independent variable, we make a line graph and the x-axis is a number line. There are many predicted values all along the line.

With an indicator independent variable, x is either 0 or 1. There are only two predicted values. A line graph would inappropriately show movement where there is no x=0.25 or x=0.5 or x=0.75. Therefore, we will use a bar graph.

 This Excel template has the graph templates and predicted values worksheets from Chapter 11, but also with a new worksheet, "BivariateGraph.IndicatorIV."



Here you can see I filled in the yellow-shaded cells:

y=a+bx

Replace cells B3 and B4 with the intercept and slope from your regression line equation.

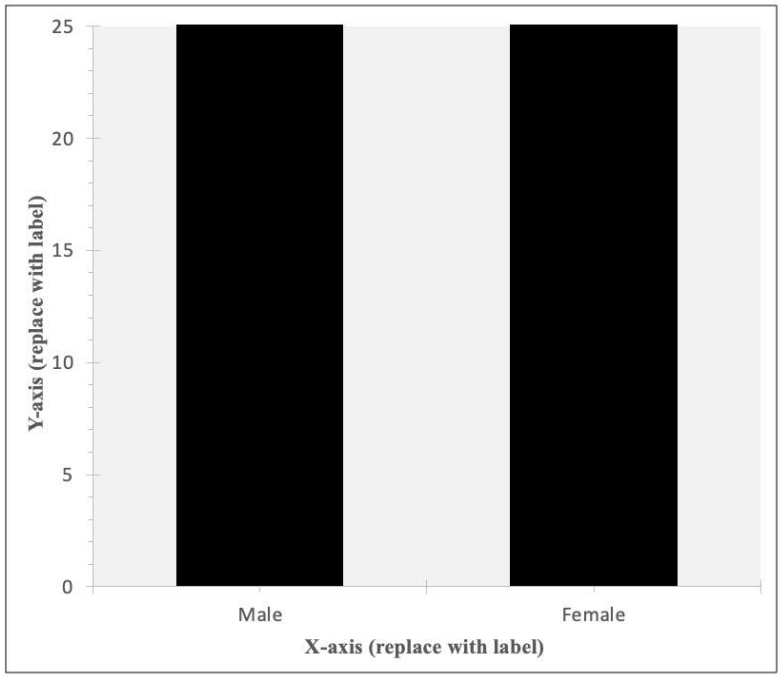
intercept a	84676.980
slope b	-21924.730

Replace cells A9 and A10 with value labels for when x=0 and when x=1

	x (IV)	y (DV)
Male	0	84676.98
Female	1	62752.25

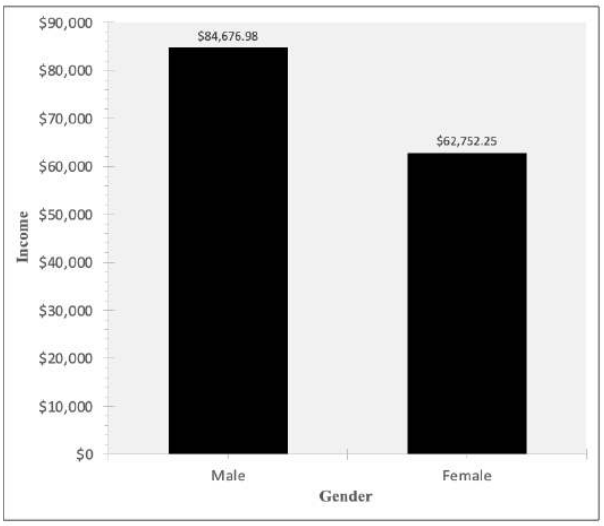
Make the following changes to the graph:

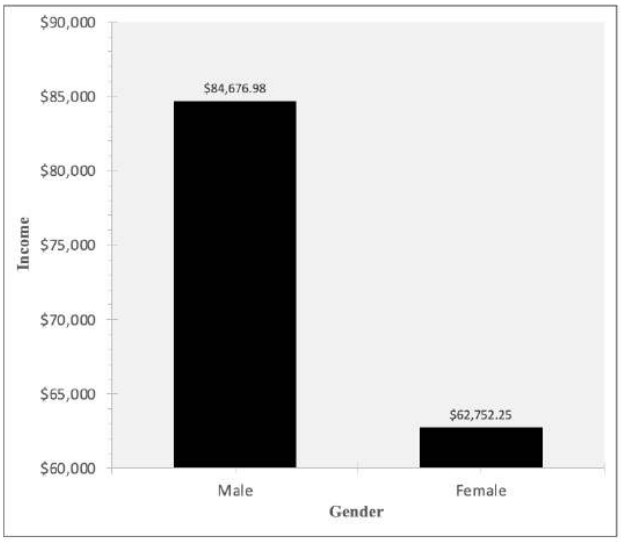
- ☐ Replace the x-axis label with the label for your IV
- ☐ Replace the y-axis label with the label for your DV
- ☐ Edit the minimum and maximum bounds for the y-axis to go at least from your minimum DV through your maximum DV (see C9 and C10)
- ☐ Edit the major and minor units for the y-axis to reflect sensical intervals.



Next I need to edit the x-axis and y-axis labels on the graph, and change the y-axis scale (bounds and units). I also changed the "Number" options to show this as "currency." I don't need to change the x-axis scale, since it is just the two categories, x=0 and x=1, each of which have a bar above them.

Again, it's up to you whether you make the y-axis scale zoom in on the differences or pan out to show the full scale. Here is what both those options look like:





Two-sample t-tests vs. Linear regression with an indicator IV

The two-sample t-test we did in Chapter 7 was simpler and gave us the same results:

T-Test

Group Statistics

	Sex	N	Mean	Std. Deviation	Std. Error Mean
Total personal income	Male	64226720	84676.98	92459.765	11.537
	Female	48272914	62752.25	60782.868	8.748

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
						Significance				95% Confidence Interval of the Difference
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	Lower	Upper
Total personal income	Equal variances assumed	2194787.322	<.001	1431.379	112499632	<.001	<.001	21924.729	15.317	21894.707	21954.750
	Equal variances not assumed			1514.254	110648749	<.001	<.001	21924.729	14.479	21896.350	21953.107

Notice how the means for each group are the same as our predicted values. The p-values match. The mean difference is the same as our slope. If this was the only analysis we were doing, a two-sample t-test would work just as well.

Even if you are doing bivariate regression, you may choose to use regression rather than a two-sample t-test if you are conducting a series of regressions and want to put them all together into one table. If some of your comparisons involve

ratio-level independent variables, you cannot use a two-sample t-test for those, but you can use regression for all of them and therefore simplify your display table to make it easier to make comparisons.

However, a two-sample t-test is limited to comparing means for one variable (in two groups). Regression is useful because we can do multivariate regression (see Chapter 14), where we have more than one independent variable. If you want to run a regression that includes multiple predictors simultaneously, you need to use regression, not a two-sample t-test.

This page titled 12: Linear Regression, continued is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Ezra Temko (Consortium of Academic and Research Libraries in Illinois (CARLI)) .

13: Linear Regression with Reference Group Set Independent Variables

So far we've covered bivariate linear regression with a ratio-level (or ratio-like) independent variable and with an indicator independent variable. But what if you want to use a nominal variable that has more than two categories? This chapter will show you how to do just that.

Reference Group Sets: Transforming Nominal-level Independent Variables

Nominal-level variables such as race, geographic region, religion, and marital status cannot be used as is for linear regression. Think back to our scatterplot. The x-axis needs to be ordered to fit the data to a regression line.

However, we already learned that we can use indicator variables (dichotomous variables coded as 0 and 1) as independent variables for linear regression. That's our solution for nominal-level variables as well.

Let's take a variable **region** with four regions: Northeast, South, Midwest, and West.

One of these regions needs to be the "reference group," which will have the value of 0.

Then we use indicator variables for each of the other categories.

Let's say I choose Northeast as the reference group.

I will then have three indicator variables:

SouthIndicator: 0=Not South (Northeast, Midwest, or West); 1=South
MidwestIndicator: 0=Not Midwest (Northeast, South, or West); 1=Midwest
WestIndicator: 0=Not West (Northeast, South, or Midwest); 1=West

If all three of these variables are zero, then the region is not South, not Midwest, and not West. Since the original region variable only had four options, if all three of these variables are zero, the region is our reference group of Northeast.

This is called a **reference group set.**

Definition: Reference Group Set

A group of indicator variables for one construct that as a collection represent its various categories. Each category, other than the comparison reference group, has an indicator variable where it is coded as 1.

Remember when we had a variable like gender (as a binary) in Chapter 12 that required **one** variable? It was coded 0=male, 1=female. Male was the reference group, the category for when the variable was zero. There were **two** categories, and we needed **one** variable, one for each category other than the reference group.

When you have nominal-level variables for multiple categories, you have to have indicator variables for each category minus one (each category other than your reference group).

- For example, if your variable was season (Fall, Winter, Spring, Summer), you would designate one season as a reference group (let's do Fall) and then make 3 indicator variables, one for each of the seasons other than Fall.
- For example, if you had a religion variable with: 1) Protestant, 2) Catholic, 3) Muslim, 4) Jewish, 5) Other or multiple religion(s), 6) No religion, you have 6 categories, so would need 5 indicator variables (one for each other than the religion you designate as the reference group).

Statistics in Action: Race Pay Gaps

Let's return to our example from earlier in this textbook about race pay gaps (see Chapter 3 and Chapter 7). Again, we're using 2017 to 2021 American Community Survey data. Our target population below is full-time, year-round, non-institutionalized U.S. workers ages 16+.

We are comparing five racial categories: non-Hispanic White only, non-Hispanic Black only, non-Hispanic Asian-American and Pacific Islander (AAPI); non-Hispanic other or multiracial; Hispanic. My variable Race5Cat has these five

mutually exclusive categories.

Let's designate one of these groups as the reference group. I am choosing non-Hispanic (NH) White only. A majority of respondents are NH White only, and this will allow me to compare the pay of other racial groups to Whites.

I don't need a variable for White, because if someone does not fit into any of the other racial groups, that means they are White.

I need a variable for each of the other categories. I started with five racial categories and designated one as my reference group. Therefore I need four indicator variables, one for each of the other categories.

- NH Black only
- NH AAPI only
- NH other/multiracial
- Hispanic

For each of these indicator variables, 0 will be a respondent who is not that category, and 1 will be a repondent who is that category.

- NH Black only (0=Not NH Black only; 1=NH Black only)
- NH AAPI (0=Not NH AAPI only; 1=NH AAPI only)
- NH other/multiracial (0=Not NH other/multiracial; 1=NH other/multiracial)
- Hispanic (0=Not Hispanic; 1=Hispanic)

Each respondent will either have a 0 for all four of these variables if they are NH White only, or a 1 for one of the variables. The coding above is derived from the following:

- NH Black only (0=NH White only, NH AAPI only, NH other/multiracial, or Hispanic; 1=NH Black only)
- NH AAPI (0=NH White only, NH Black only, NH other/multiracial, or Hispanic; 1=NH AAPI only)
- NH other/multiracial (0=NH White only, NH Black only, NH AAPI only, or Hispanic; 1=NH other/multiracial)
- Hispanic (0=NH White only, NH Black only, NH AAPI only, or NH other/multiracial; 1=Hispanic)

Choosing your reference group

Choosing your reference group is based on what you want to compare the other categories to. Usually people choose either the category with the largest sample size or the one that makes sense to compare to others based on their theory/interest/etc.

This is also a technique you can use if you have a ratio-level or ratio-like variable like age or education if they have a relationship with your dependent variable but it's not linear (e.g., it changes direction, or one group is more extremely different than the rest). You can use age groups or degree categories and then make them into a reference group set.

Running a bivariate linear regression with a reference group set using SPSS

Running the linear regression is almost identical to what we learned in Chapter 11 and did again in Chapter 12, with one difference. Instead of putting in one independent variable, we will put in multiple independent variables --- all the indicator variables that make up the reference group set.

If you mistakenly put in your reference group variable, SPSS will automatically remove one of your IVs from the regression to eliminate redundancy, but it might not be the one you wanted to use as your reference group.

Caution

Make sure to only run **one** bivariate regression, putting all the independent variables that make up the reference group set in together at the same time. You are not running multiple regressions. You want the whole reference group set in at the same time.

Statistics in Action: Race Pay Gap

Our independent variable, Race5Cat, has 5 racial categories.

There are also 5 indicator variables, one for each category:

- **RaceNHWhiteIndicator**: NH White only (0=Not NH White only; 1=NH White only)
- **RaceNHBlackIndicator:** NH Black only (0=Not NH Black only; 1=NH Black only)
- **RaceNHAAPIIndicator:** NH AAPI (0=Not NH AAPI only; 1=NH AAPI only)
- **RaceNHOtherIndicator:** NH other/multiracial (0=Not NH other/multiracial; 1=NH other/multiracial)
- **RaceHispanicIndicator:** Hispanic (0=Not Hispanic; 1=Hispanic)

I can't use Race5Cat as an IV for linear regression, because it's nominal.

I designate one of the five categories as the reference group. In this case, I chose NH White only. Therefore, RaceNHWhiteIndicator will **not** go into my regression. If all the other four variables have a value of 0, that indicates the respondent is NH White only.

The other four variable constitute my reference group set and will all go into my regression. Therefore, I have four IVs to put in: RaceNHBlackIndicator, RaceNHAAPIIndicator, RaceNHOtherIndicator, and RaceHispanicIndicator.

Syntax:

The difference in syntax is with the line:

/METHOD=ENTER IVName.

Instead of entering the name of one independent variable, you will enter however many variables make up your reference group set (total number of categories minus the reference group), with a space in between each.

/METHOD=ENTER IV1Name IV2Name IV3Name IV4Name.

Statistics in Action: Race Pay Gap

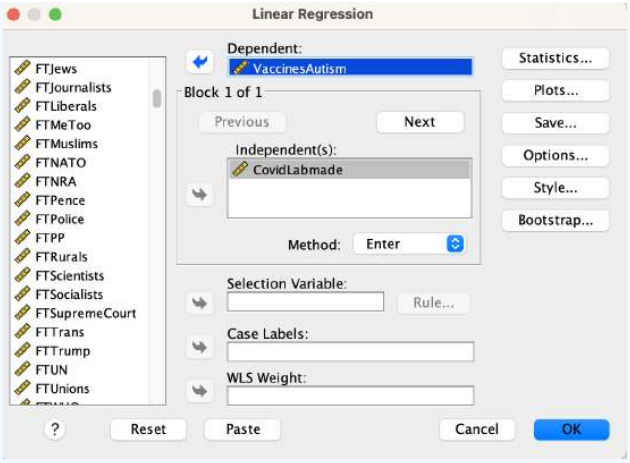
/METHOD=ENTER RaceNHBlackIndicator RaceNHAAPIIndicator RaceNHOtherIndicator RaceHispanicIndicator.

The full syntax therefore looks like this:

```
REGRESSION
/DESCRIPTIVES MEAN STDDEV CORR SIG N
/MISSING LISTWISE
/STATISTICS COEFF OUTS CI(95) R ANOVA
/CRITERIA=PIN(.05) POUT(.10) TOLERANCE(.0001)
/NOORIGIN
/DEPENDENT DVName
/METHOD=ENTER IV1Name IV2Name IV3Name IV4Name.
```

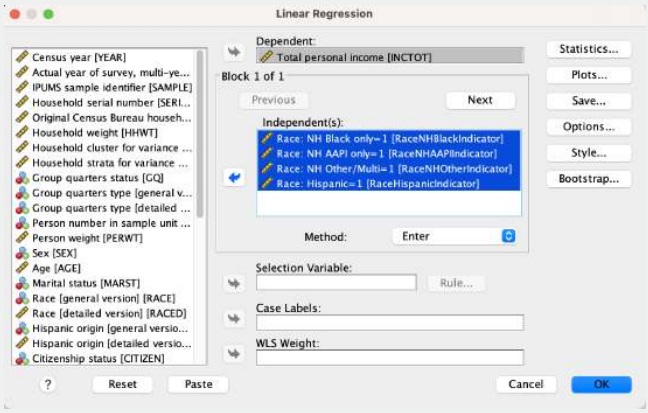
Menu:

Previously you put your independent variable in the “Independent(s)” box and your dependent variable in the “Dependent” box.



The only change is that you will put all of the independent variables that constitute your reference group set (one indicator variable for each category except for the reference group) into the "Independent(s):" box.

Statistics in Action: Race pay gap



SPSS Output

The output looks similar to what we had before, except that the coefficients output has more rows. Beyond the y-intercept row, there is one for each independent variable's slope. There are also more variables and correlation pairs in the descriptive statistics and correlations outputs. The correlation coefficient (Pearson's r) in the model summary output may not match the (absolute value of the) individual correlation values in the Correlations output, as the model summary is taking into account all the reference group set variables collectively to determine the strength of the relationship, rather than just one indicator variable on its own. Finally, the p-value in the ANOVA output may not be the same as p-values in the coefficients output, as the ANOVA output p-value is for the model as a whole, the whole construct of the reference group set collectively, whereas the p-values in the coefficients output are about whether each individual category's predicted dependent variable value is significantly different from the reference group (with the y-intercept p-value still evaluating confidence in whether the predicted dependent variable value for the reference group category is different from zero).

Statistics in Action: Race Pay Gap

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Total personal income	72891.85	80837.524	8141801
Race: NH Black only=1	.0814	.27340	8141801
Race: NH AAPI only=1	.0691	.25364	8141801
Race: NH Other/Multi=1	.0342	.18183	8141801
Race: Hispanic=1	.1623	.36875	8141801

Correlations

		Total personal income	Race: NH Black only=1	Race: NH AAPI only=1	Race: NH Other/Multi=1	Race: Hispanic=1
Pearson Correlation	Total personal income	1.000	-.070	.046	-.016	-.123
	Race: NH Black only=1	-.070	1.000	-.081	-.056	-.131
	Race: NH AAPI only=1	.046	-.081	1.000	-.051	-.120
	Race: NH Other/Multi=1	-.016	-.056	-.051	1.000	-.083
	Race: Hispanic=1	-.123	-.131	-.120	-.083	1.000
Sig. (1-tailed)	Total personal income	.	<.001	<.001	<.001	<.001
	Race: NH Black only=1	.000	.	.000	.000	.000
	Race: NH AAPI only=1	.000	.000	.	.000	.000
	Race: NH Other/Multi=1	.000	.000	.000	.	.000
	Race: Hispanic=1	.000	.000	.000	.000	.
N	Total personal income	8141801	8141801	8141801	8141801	8141801
	Race: NH Black only=1	8141801	8141801	8141801	8141801	8141801
	Race: NH AAPI only=1	8141801	8141801	8141801	8141801	8141801
	Race: NH Other/Multi=1	8141801	8141801	8141801	8141801	8141801
	Race: Hispanic=1	8141801	8141801	8141801	8141801	8141801

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Race: Hispanic=1, Race: NH Other/Multi=1, Race: NH AAPI only=1, Race: NH Black only=1[b]	.	Enter

a. Dependent Variable: Total personal income

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.155[a]	.024	.024	79857.300

a. Predictors: (Constant), Race: Hispanic=1, Race: NH Other/Multi=1, Race: NH AAPI only=1, Race: NH Black only=1

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.282E+15	4	3.206E+14	50276.770	<.001[b]
	Residual	5.192E+16	8141796	6377188289		
	Total	5.320E+16	8141800			

a. Dependent Variable: Total personal income

b. Predictors: (Constant), Race: Hispanic=1, Race: NH Other/Multi=1, Race: NH AAPI only=1, Race: NH Black only=1

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Race: Hispanic=1, Race: NH Other/Multi=1, Race: NH AAPI only=1, Race: NH Black only=1[b]	.	Enter

a. Dependent Variable: Total personal income

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.155[a]	.024	.024	79857.300

a. Predictors: (Constant), Race: Hispanic=1, Race: NH Other/Multi=1, Race: NH AAPI only=1, Race: NH Black only=1

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.282E+15	4	3.206E+14	50276.770	<.001[b]
	Residual	5.192E+16	8141796	6377188289		
	Total	5.320E+16	8141800			

a. Dependent Variable: Total personal income

b. Predictors: (Constant), Race: Hispanic=1, Race: NH Other/Multi=1, Race: NH AAPI only=1, Race: NH Black only=1

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	79789.567	34.634		2303.762	<.001	79721.684	79857.449
	Race: NH Black only=1	-25953.471	104.047	-.088	-249.441	<.001	-26157.399	-25749.544
	Race: NH AAPI only=1	6758.113	111.954	.021	60.365	<.001	6538.687	6977.539
	Race: NH Other/Multi=1	-13964.726	155.171	-.031	-89.996	<.001	-14268.855	-13660.597
	Race: Hispanic=1	-29416.066	77.620	-.134	-378.973	<.001	-29568.199	-29263.933

a. Dependent Variable: Total personal income

Unweighted:

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Total personal income	79121.74	85918.126	5052460
Race: NH Black only=1	.0811	.27300	5052460
Race: NH AAPI only=1	.0637	.24431	5052460
Race: NH Other/Multi=1	.0333	.17950	5052460
Race: Hispanic=1	.1401	.34711	5052460

Correlations

		Total personal income	Race: NH Black only=1	Race: NH AAPI only=1	Race: NH Other/Multi=1	Race: Hispanic=1
Pearson Correlation	Total personal income	1.000	-.076	.056	-.016	-.111
	Race: NH Black only=1	-.076	1.000	-.078	-.055	-.120
	Race: NH AAPI only=1	.056	-.078	1.000	-.048	-.105
	Race: NH Other/Multi=1	-.016	-.055	-.048	1.000	-.075
	Race: Hispanic=1	-.111	-.120	-.105	-.075	1.000
Sig. (1-tailed)	Total personal income	.	<.001	<.001	<.001	<.001
	Race: NH Black only=1	.000	.	.000	.000	.000
	Race: NH AAPI only=1	.000	.000	.	.000	.000
	Race: NH Other/Multi=1	.000	.000	.000	.	.000
	Race: Hispanic=1	.000	.000	.000	.000	.
N	Total personal income	5052460	5052460	5052460	5052460	5052460
	Race: NH Black only=1	5052460	5052460	5052460	5052460	5052460
	Race: NH AAPI only=1	5052460	5052460	5052460	5052460	5052460
	Race: NH Other/Multi=1	5052460	5052460	5052460	5052460	5052460
	Race: Hispanic=1	5052460	5052460	5052460	5052460	5052460

Display Tables

The display tables template from Chapter 12 BivariateOLSLinearRegression.DisplayTableTemplates.docx has three templates for bivariate linear regression with a reference group set for the independent variable(s).

The first template is for if your independent variable construct has three groups and you therefore put two indicator variables into your regression. For example, if you were using gender as your independent variable with man, woman, and genderqueer as the three categories, you would use this template:

3 groups (2 indicator variables)

Table #. Estimated unstandardized coefficients from an OLS regression predicting DV, with 95% confident intervals

Independent Variable	IV Name?
Intercept (ref. group label)	Coefficient Value? (LB, UB)
Indicator Variable 1 Name=1	Coefficient Value? (LB, UB)
Indicator Variable 2 Name=1	Coefficient Value? (LB, UB)
r^2	r-squared*100%
Unweighted sample size	n

~p<0.10, *p<0.05, **p<0.01, ***p<0.001

Source: Dataset Year

The second template is for if your variable has four categories, and you therefore put in three indicator variables (with the fourth category serving as your reference group).

4 groups (3 indicator variables)

Table #. Estimated unstandardized coefficients from an OLS regression predicting DV, with 95% confident intervals

Independent Variable	IV Name[?]
Intercept (ref. group label)	Coefficient Value[?] (LB, UB)
Indicator Variable 1 Name=1	Coefficient Value[?] (LB, UB)
Indicator Variable 2 Name=1	Coefficient Value[?] (LB, UB)
Indicator Variable 3 Name=1	Coefficient Value[?] (LB, UB)
r^2	r-squared*100%
Unweighted sample size	n

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Source: Dataset Year

The third template is for if your variable has five categories, and you therefore put in four indicator variables.

5 groups (4 indicator variables)

Table #. Estimated unstandardized coefficients from an OLS regression predicting DV, with 95% confident intervals

Independent Variable	IV Name[?]
Intercept (ref. group label)	Coefficient Value[?] (LB, UB)
Indicator Variable 1 Name=1	Coefficient Value[?] (LB, UB)
Indicator Variable 2 Name=1	Coefficient Value[?] (LB, UB)
Indicator Variable 3 Name=1	Coefficient Value[?] (LB, UB)
Indicator Variable 4 Name=1	Coefficient Value[?] (LB, UB)
r^2	r-squared*100%
Unweighted sample size	n

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Source: Dataset Year

If you have more than five categories, you may want to consider collapsing categories together so that you have fewer categories. If you have good reason to have more than five discrete categories, you will need to add another row to the template yourself.

Statistics in Action: Race Pay Gap

With the race pay gap example, we had five racial groups and put in four independent variables (with NH White only serving as our reference group), so I will use the template for "5 groups (4 indicator variables)."

I edited the summary display table below for everything that is in red:

Table One. Estimated unstandardized coefficients from an OLS regression predicting income, with 95% confident intervals

Independent Variable	Race***
Intercept (NH White only)	$79,789.57*** ($79,721.68, $79,857.45)
NH Black only=1	-$25,953.47*** ($-26,157.40, -$25,749.54)
NH AAPI only=1	$6,758.11*** ($6,538.69, $6,977.54)
NH Other/Multi=1	-$13,964.73*** (-$14,268.86, -$13,660.60)
Hispanic=1	-$29,416.07*** (-$29,568.20, -$29,263.93)
r^2	1.8%
Unweighted sample size	5,052,460

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Source: ACS 5-Year (2017-2021)

Here is the display table with all of it in black:

Table One. Estimated unstandardized coefficients from an OLS regression predicting income, with 95% confident intervals

Independent Variable	Race***
Intercept (NH White only)	$79,789.57*** ($79,721.68, $79,857.45)
NH Black only=1	-$25,953.47*** ($-26,157.40, -$25,749.54)
NH AAPI only=1	$6,758.11*** ($6,538.69, $6,977.54)
NH Other/Multi=1	-$13,964.73*** (-$14,268.86, -$13,660.60)
Hispanic=1	-$29,416.07*** (-$29,568.20, -$29,263.93)
r^2	1.8%
Unweighted sample size	5,052,460

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Source: ACS 5-Year (2017-2021)

The r-squared value is for the model as a whole, so not for the explained variance from any one category, but from all of them together, collectively. Here, this means that race explains 1.8% of variance in income.

For this example, all of the p-values were <0.001, both weighted and unweighted. This may be because there are over 5 million cases in the sample. The *** after Race is the p-value from the ANOVA output and signifies whether race as a whole is has a significant relationship with income among our target population. The other *** come from the Coefficients output. I'll briefly show you an example at the end of the chapter that has p-values that are not all identical.

Interpreting Slopes

The interpretation of the y-intercept and slopes will mirror our interpretation from Chapter 12 where we used one indicator variable as the independent variable. Each slope is in comparison to the reference group. The p-values are the probability that we would pull this sample if the slope was actually zero, meaning if there was no difference between this category's predicted dependent variable value and the reference group's predicted dependent variable value.

Statistics in Action: Race Pay Gap

Here is the coefficients output from our race pay gap linear regression.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	79789.567	34.634		2303.762	<.001	79721.684	79857.449
	Race: NH Black only=1	-25953.471	104.047	-.088	-249.441	<.001	-26157.399	-25749.544
	Race: NH AAPI only=1	6758.113	111.954	.021	60.365	<.001	6538.687	6977.539
	Race: NH Other/Multi=1	-13964.726	155.171	-.031	-89.996	<.001	-14268.855	-13660.597
	Race: Hispanic=1	-29416.066	77.620	-.134	-378.973	<.001	-29568.199	-29263.933

a. Dependent Variable: Total personal income

NH White only

- The first row, constant, is for the y-intercept. When all the other variables (NH Black Only, NH AAPI only, NH Other/Multi, and Hispanic) are zero, then we are left with our reference group, in this case NH White only.
- The value in the unstandardized coefficients B cell for the Constant row, $79,789.57, is the predicted income (years 2017 to 2021) for NH White only full-time, year-round, non-institutionalized U.S. workers ages 16+. This predicted value is based on the sample.
- For the target population, we can look at the confidence interval. We are 95% confident that the average income for this group is somewhere between $79,721.68 and $79,857.45.
- The p-value, $p<0.001$, means that we are over 99.9% confident that the sample coefficient, $79,789.57, is not actually $0 for our target population. This makes sense --- the average salary for NH white only full-time U.S. adult workers should not be $0. The p-value for the y-intercept does not tell us anything about whether or not there is a relationship between race and income.

NH Black only

- The second row, RaceNHBlackIndicator, is for our indicator variable for the NH Black only category.
- The value in the unstandardized coefficients B cell for this row, -$25,953.47, is a sample statistic comparing the average income for NH Black only respondents to the average income for the reference group, in this case NH White only respondents. NH Black only respondents are predicted to have an income $25,953.47 less than NH White only respondents.
- Since NH White only respondents are predicted to have an average income of $79,789.57, we can also calculate the predicted income for NH Black only respondents: $79,789.57-$25,953.47=$53,836.10.
- The slope and the predicted value here are based on the sample. For the target population, we can look at the confidence interval. We are 95% confident that among all NH Black only workers (full-time, year-round, non-institutionalized U.S. workers ages 16+ from 2017 to 2021), they have, on average, an income somewhere between $25,749.54 and $26,157.40 less than their NH White only worker counterparts.
- The p-value indicates the probability that we would pull this sample if this slope were actually zero. Again, the slope is comparing this category to the reference group. So the p-value is indicating whether there is a significant difference between this category's predicted income and the reference group's predicted income. With $p<0.001$, we are over 99.9% confident that there is a difference between NH Black only workers' average income and NH White only workers' average income among our target population (full-time, year-round, non-institutionalized U.S. workers ages 16+ from 2017 to 2021).

NH AAPI only

- The third row is for our NH AAPI only category.
- The slope, $6,758.11, means that among our respondents, the average income for NH AAPI only workers is $6,758.11 higher than the average income for NH White only workers.
- Our predicted income for NH AAPI only respondents, $79,789.57+$6,758.11, is $86,547.68.
- The confidence interval indicates that we are 95% confident that, among all workers in this target population, the average income for NH AAPI workers is somewhere between $6,538.69 and $6,977.54 higher than the average income for NH White only workers.

- The p-value indicates that we are over 99.9% confident, that among our target population (full-time, year-round, non-institutionalized U.S. workers ages 16+ from 2017 to 2021), NH AAPI only and NH White only workers have a different average income.

NH Other/Multiracial

- The fourth row is for our NH other race or multiracial category.
- The slope, -$13,964.73, means that among our respondents, the average income for NH other/multiracial workers is $13,964.73 lower than the average income for NH White only workers.
- Our predicted income for NH other/multiracial respondents, $79,789.57+(-$13,964.73), is $65,824.84.
- The confidence interval indicates that we are 95% confident that, among all workers in this target population, the average income for NH other/multiracial workers is somewhere between $13,660.60 and $14,268.86 lower than the average income for NH White only workers.
- The p-value indicates that we are over 99.9% confident, that among our target population (full-time, year-round, non-institutionalized U.S. workers ages 16+ from 2017 to 2021), NH other/multiracial and NH White only workers have a different average income.

Hispanic

- The fifth row is for our Hispanic category.
- The slope, -$29,416.07, means that among our respondents, the average income for Hispanic workers is $29,416.07 lower than the average income for NH White only workers.
- Our predicted income for Hispanic respondents, $79,789.57+(-$29,416.07), is $50,373.50.
- The confidence interval indicates that we are 95% confident that, among all workers in this target population, the average income for Hispanic workers is somewhere between $29,263.93 and $29,568.20 lower than the average income for NH White only workers.
- The p-value indicates that we are over 99.9% confident, that among our target population (full-time, year-round, non-institutionalized U.S. workers ages 16+ from 2017 to 2021), Hispanic and NH White only workers have a different average income.

Caution

The slopes, p-values, and confidence intervals for each of the variables is only for that category in relation to the reference group. If you want to compare two categories to each other to see if they are significantly different, you would need to figure this out using the confidence intervals of the slopes for each of those categories' predicted values and determine whether the confidence intervals of their predicted values would overlap.

Regression Equation

Regression Equation

Previously our regression equation was y=a+bx, where y is the dependent variable, a is the y-intercept, b is the slope, and x is the independent variable.

Now we have more than one independent variable, so we need to extend this equation.

If there are two independent variables, our equation will be: $y=a+b_1x_1+b_2x_2$, where x_1 is the first independent variable, b_1 the slope for that variable, x_2 the second independent variable, and b_2 the slope for that second variable.

If there are three independent variables, our equation will be: $y=a+b_1x_1+b_2x_2+b_3x_3$.

If there are four independent variables, our equation will be: $y=a+b_1x_1+b_2x_2+b_3x_3+b_4x_4$.

And so forth.

You don't have to use x_1, x_2, x_3, etc. for your notation. I prefer to use a letter that lets me know what the category is. For example, if I were using region with Northeast as my reference group set and three indicator variables: SouthIndicator,

MidwestIndicator, and WestIndicator, I might use S for South, M for Midwest, and W for West. My equation would be $y=a+b_S S+b_M M+b_W W$, or $y=a+b_1 S+b_2 M+b_3 W$ is fine as well. I could also write it without the subscripts for the slopes as long as I know that they are not all the same, and with words for the variables, e.g., y=a+b(South)+b(Midwest)+b(West). Just stay away from abbreviating variables into two letters; for example, I would not want to use MW for midwest ($y=a+b_S S+b_{MW} MW+b_W W$) because this would be confusing as to whether MW is for the midwest or for a variable M multipled by a variable W.

Statistics in Action: Race pay gap

My equation will look like: $y=a+b_1x_1+b_2x_2+b_3x_3+b_4x_4$.

I am going to use letters for each variable: B for NH Black only, A for NH AAPI only, O for NH other/multi, and H for Hispanic.

Therefore my equation will look like: $y=a+b_1B+b_2A+b_3O+b_4H$.

Filling in the coefficients, my equation is: y = $79,789.57 + -$25,953.47(B) + $6,758.11(A) + -$13,964.73(O) + -$29,416.07(H).

Again, I could also write this as: y = $79,789.57 + -$25,953.47(Black) + $6,758.11(Aapi) + -$13,964.73(Other) + -$29,416.07(Hispanic), or as

y = $79,789.57 + -$25,953.47(x_1) + $6,758.11($x_2$) + -$13,964.73(x_3) + -$29,416.07($x_4$).

Caution

Only write one equation for the entire regression, and put all the coefficients and variables into the one equation. Do **not** write multiple equations for different variables.

Making Predictions

Just like before, I can make predictions by plugging variable values into the equation and solving for y.

Statistics in Action: Race pay gap

y = $79,789.57 + -$25,953.47(B) + $6,758.11(A) + -$13,964.73(O) + -$29,416.07(H).

Prediction for NH White only
Black=0, AAPI=0, Other/Multi=0, Hispanic=0
y = $79,789.57 + -$25,953.47(0) + $6,758.11(0) + -$13,964.73(0) + -$29,416.07(0)
y = $79,789.57 + 0 + 0 + 0 + 0
y = $79,789.57

Prediction for NH Black only
Black=1, AAPI=0, Other/Multi=0, Hispanic=0
y = $79,789.57 + -$25,953.47(1) + $6,758.11(0) + -$13,964.73(0) + -$29,416.07(0)
y = $79,789.57 + (-$25,953.47) + 0 + 0 + 0
y = $53,836.10

Prediction for NH AAPI only
Black=0, AAPI=1, Other/Multi=0, Hispanic=0
y = $79,789.57 + -$25,953.47(0) + $6,758.11(1) + -$13,964.73(0) + -$29,416.07(0)
y = $79,789.57 + 0 + $6,758.11 + 0 + 0
y = $86,547.68

Prediction for NH Other/Multi
Black=0, AAPI=0, Other/Multi=1, Hispanic=0

y = $79,789.57 + -$25,953.47(0) + $6,758.11(0) + -$13,964.73(1) + -$29,416.07(0)
y = $79,789.57 + 0 + 0 + (-$13,964.73) + 0
y = $65,824.84

Prediction for Hispanic
Black=0, AAPI=0, Other/Multi=0, Hispanic=1
y = $79,789.57 + -$25,953.47(0) + $6,758.11(0) + -$13,964.73(0) + -$29,416.07(1)
y = $79,789.57 + 0 + 0 + 0 + (-$29,416.07)
y = $50,373.50

Because, with a reference group set, all the variables are 0 or one of the variables is 1 and the rest are 0, I can also just use the y-intercept and slope for the category I'm interested in, and ignore the rest of the equation (which will be zero). The reference group category's predicted value is the y-intercept, and any other category's predicted value is the y-intercept plus the slope for that category.

Statistics in Action: Race pay gap

y-intercept: $79,789.57
NH Black only slope: -$25,953.47
NH AAPI only slope: $6,758.11
NH Other/Multiracial slope: -$13,964.73
NH Hispanic slope: -$29,416.07

Prediction for NH White only
y-intercept = $79,789.57

Prediction for NH Black only
y-intercept + Black slope
y = $79,789.57 + (-$25,953.47) = $53,836.10

Prediction for NH AAPI only
y-intercept + AAPI slope
y = $79,789.57 + $6,758.11 = $86,547.68

Prediction for NH Other/Multi
y-intercept + Other/Multi slope
y = $79,789.57 + (-$13,964.73) = $65,824.84

Prediction for Hispanic
y-intercept + Hispanic slope
y = $79,789.57 + (-$29,416.07) = $50,373.50

I have added an additional worksheet to the BivariateLinearRegressionGraphTemplate.xlsx that you can use for calculating predicted values when you have multiple independent variables (see worksheet labeled PredictedValueCalc.MultipleIVs).

Because we are using a reference group set that came from an original variable with mutually exclusive categories, just make sure that either all the variables are zero if you are predicting the reference group category, or that only one IV has a value of 1 and the rest have a value of 0, with the value of 1 for the category you are making the prediction for.

Statistics in Action: Race pay gap

Here is the worksheet filled out for the first three categories.

	A	B	C	D	E	F	G	H
1								
2		B column	Value of IV	Label for value	Value of IV	Label for value	Value of IV	Label for value
3	Intercept (NH White only)	79789.57						
4	NH Black only	-25953.47	0	NH White only	1	NH Black only	0	
5	NH AAPI only	6758.11	0		0		1	NH AAPI only
6	NH Other/Multi	-13964.73	0		0		0	
7	Hispanic	-29416.07	0		0		0	
8		Y =	79789.570		53836.100		86547.680	

Graphing

Just like when we had only one indicator variable, we cannot make a line graph to graph out the equation. Each independent variable value is either 0 or 1. The independent variable construct is categorical. So instead, we make a bar graph. While last chapter our bar graph had two bars, one for each category, here we will have multiple bars, with bars for however many categories your construct has.

The graphing template in Excel: BivariateLinearRegressionGraphTemplate.xlsx, has three worksheets for this. BivariateGraph.3CatIV is for if you have three categories, BivariateGraph.4CatIV is if you have four categories, and BivariateGraph.5CatIV is for if you have five categories. If you have more than five categories you may want to consider collapsing categories together, but otherwise you will need to edit the template yourself.

The template works the same as before, except you have to add in additional slopes and category descriptions, for your y-intercept and then one for each of your independent variables.

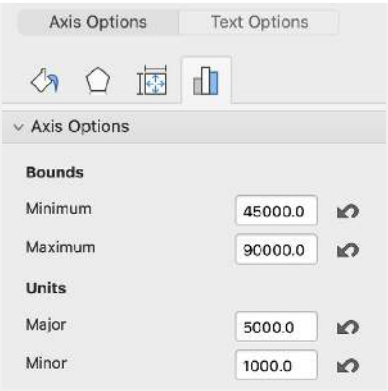
Statistics in Action: Race pay gap

I am using the graphing template for 5 categories.

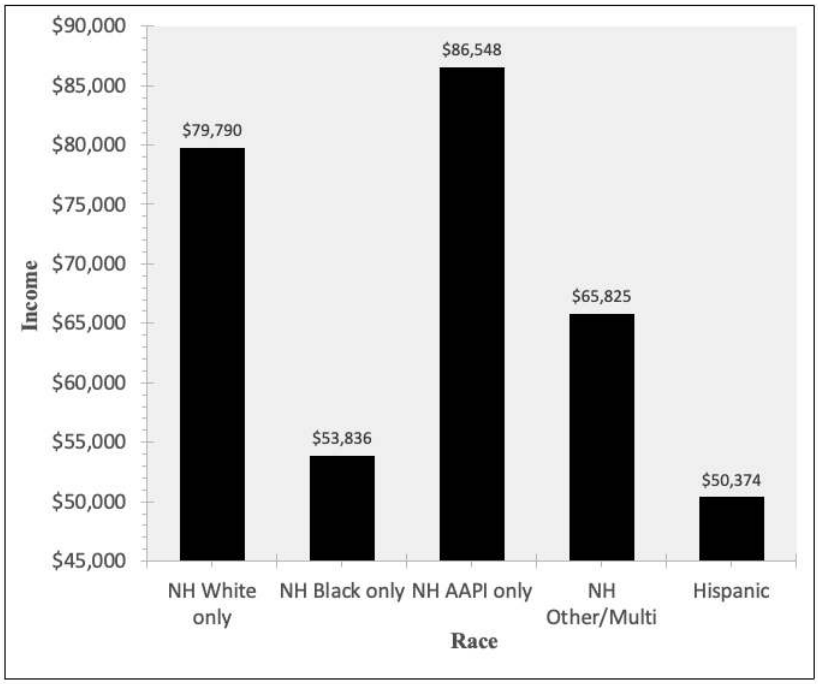
Here is the template with the yellow squares filed out:

	A	B	C	D	E
1	**y=a+bx**				
2	***Replace cells B3, B4, and B5 with the intercept and slopes from your regression line equation.***				
3	**intercept a**	79789.567			
4	**IV#1 slope b**	-25953.471			
5	**IV#2 slope b**	6758.113			
6	**IV#3 slope b**	-13964.726			
7	**IV#4 slope b**	-29416.066			
8					
9					
10	***Replace cells A9 with your reference group label and A10 through A16 with those indicator variable labels***				
11		**x (IV)**	**y (DV)**		
12	NH White only	0	79789.567		
13	NH Black only	1	53836.096		
14	NH AAPI only	1	86547.680		
15	NH Other/Multi	1	65824.841		
16	Hispanic	1	50373.501		

Next I had to replace the x-axis and y-axis labels on my graph and edit the bounds and units for the y-axis. I can tell from cells C12 through C16 above that the dependent variable values range from just over $50,000 to just under $87,000. Here is what I put in for bounds and units:



Here is the graph I ended up with:



ANOVA vs. Linear regression with a reference groups set for the IV

The one-way ANOVA (with confidence intervals) we did in Chapter 7 was simpler and gave us the same overall p-value and more accurate sample group means:

ANOVA

Total personal income

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2.227E+16	4	5.567E+15	871863.591	<.001
Within Groups	7.184E+17	112499629	6385680919		
Total	7.407E+17	112499633			

Oneway

Descriptives

Total personal income

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
NH White	69261799	83482.70	88647.450	10.652	83461.82	83503.58	-12084	1709394
NH Black	12832882	55123.06	50151.712	14.000	55095.62	55150.50	-9174	1428000
NH API	7141780	94846.17	97152.626	36.354	94774.92	94917.43	-10466	1726351
NH Other/Multi	3589218	70806.07	74604.958	39.379	70728.89	70883.25	-9279	1391860
Hispanic	19673955	53202.32	52602.763	11.859	53179.08	53225.57	-10466	1519697
Total	112499634	75269.21	81139.601	7.650	75254.22	75284.21	-12084	1726351

If the only analysis we were doing was race by income, a one-way ANOVA test would work just as well.

Even if you are doing bivariate regression, you may choose to use regression rather than ANOVA if you are conducting a series of regressions and want to put them all together into one table. If some of your comparisons involve ratio-level independent variables, you cannot use a one-way ANOVA for those, but can use regression for all of them and therefore simplify your display table to make it easier to make comparisons.

However, a one-way ANOVA is limited to comparing means for one variable (in multiple categorical groups). Regression is useful because we can do multivariate regression (see Chapter 14), where we have more than one independent variable construct. If you want to run a regression that includes multiple predictors simultaneously, you need to use regression, not a one-way ANOVA.

Statistics in Action: Political Party & Feelings Towards Congress

Let's look at one more example, briefly. This example uses the American National Election Studies post-2020 election dataset. The target population is U.S. eligible voters.

Here my dependent variable is feelings towards Congress, on a 0 to 100 scale, with 0 degrees being cold, 50 degrees neutral, and 100 degrees warm.

For my independent variable, I am using political party, in three categories: Democratic, Independent/other party, and Republican. I chose Independent/other party as my reference group, so I am left with two indicator variables to put in: one for Democratic=1 and one for Republican=1.

Here is some of the output from the linear regression:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.197[a]	.039	.038	21.665

a. Predictors: (Constant), Do you think of yourself as a Republican?, Do you think of yourself as a Democrat?

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	49425.541	2	24712.771	52.649	<.001[b]
	Residual	1221725.705	2603	469.384		
	Total	1271151.247	2605			

a. Dependent Variable: POST: Feeling thermometer: congress

b. Predictors: (Constant), Do you think of yourself as a Republican?, Do you think of yourself as a Democrat?

Coefficients[a]

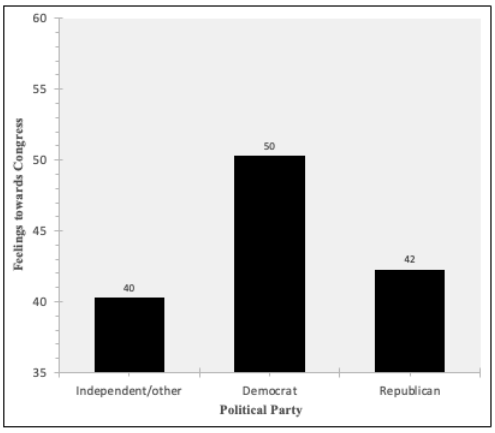
Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	40.313	.728		55.379	<.001	38.886	41.741
	Do you think of yourself as a Democrat?	9.962	1.027	.214	9.704	<.001	7.949	11.974
	Do you think of yourself as a Republican?	1.965	1.049	.041	1.874	.061	-.091	4.021

a. Dependent Variable: POST: Feeling thermometer: congress

Overall, based on the p-value in the ANOVA output, I am over 99.9% confident that political party has a significant relationship with feelings towards Congress (over 99.9% confident that the relationship exists among eligible voters following the 2020 election).

Based on my r-squared value, political party explains 3.9% of variance in feelings towards Congress.

When I get to my coefficients output, I can see that my p-values do not all match the p-value from the ANOVA output. The ANOVA output p-value is telling me whether collectively political party has a relationship with feelings towards Congress (the probability I would pull this sample if r-squared, the explained variance, is actually zero). In the coefficients output, the p-values for the Democrat and Republican rows are for whether these are significantly different from the reference group. We are over 99.9% confident that Democrats (who are eligible voters, following the 2020 election) have different feelings towards Congress than Independents/Other. We are 93.9% confident that Republicans (who are eligible voters, following the 2020 election) have different feelings towards Congress than Independents/Other. There is a significant difference between Democrats and Independents/other. There is not a significant different (or there is a marginally significant difference, $p<0.1$ but not <0.05) between Republicans and Independents/other. If I made a summary display table for this regression, the independent variable construct at the top would have ***, and then below Democrat would have ***, while GOP would be marked with ˜.



This page titled 13: Linear Regression with Reference Group Set Independent Variables is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Ezra Temko (Consortium of Academic and Research Libraries in Illinois (CARLI)) .

14: Multivariate Linear Regression

This chapter introduces multivariate linear regression.

Multivariate Regression

So far we've only looked at one predictor variable and how it can improve our prediction of and relates to a dependent variable.

However, in real life there are multiple predictors that contribute towards outcomes.

Let's say I wanted to predict your political party affiliation. Knowing your political views would help. Knowing your parents/caregivers' political affiliations would help. But if I already know your political views, does knowing your parents/caregivers' political affiliations still help? If so, what predictions would I make of your political party affiliation if I got to take into account *both* of these variables?

Definition: Multivariate Regression

Multivariate regression is when we use multiple independent variables in our regression.

You can use many independent variables with multivariate regression. You will see some journal articles with over a dozen. As we are getting started, we are going to focus on regressions that just use two independent variables. Just know that multivariate regression is not limited to two IVs.

Statistics in Action: Evaluating Trump's handling of immigration

Let's take a look at an example.

Dataset:

This data comes from the American National Election Studies (ANES) 2020 Time Series Study pre-election data, administered between August 2020 and the November election, with a target population of U.S. eligible voters.

Research Question:

How were eligible voters evaluating President Trump's handling of immigration? Was this linked to their views on immigration? To their views on Trump?

Dependent Variable:

Respondents were asked, "Do you approve or disapprove of the way Donald Trump is handling immigration?" with the options "approve" or "disapprove." Respondents were then asked, "Do you [approve/disapprove] strongly or not strongly?"

I combined these into a new variable, coded from -3 to 3, where -3 is someone who strongly disapproves, -2 someone who disapproves but did not know or refused to answer the follow-up, -1 is someone who disapproves, but not strongly, 0 is someone who volunteered "don't know" to the first question, 1 is someone who approves, but not strongly, 2 is someone who approves but did not know or refused to answer the follow-up, and 3 is someone who strongly approves.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Evaluation of the way Donald Trump is handling immigration	8243	99.5%	37	0.5%	8280	100.0%

Descriptives

			Statistic	Std. Error
Evaluation of the way Donald Trump is handling immigration	Mean		-.4385	.03019
	95% Confidence Interval for Mean	Lower Bound	-.4976	
		Upper Bound	-.3793	
	5% Trimmed Mean		-.4872	
	Median		-1.0000	
	Variance		7.512	
	Std. Deviation		2.74075	
	Minimum		-3.00	
	Maximum		3.00	
	Range		6.00	
	Interquartile Range		6.00	
	Skewness		.292	.027
	Kurtosis		-1.762	.054

Independent Variables

Policy Position on Immigration

For the independent variables, views on immigration was captured by the question, "Which comes closest to your view about what government policy should be toward unauthorized immigrants now living in the United States?"

Potential responses (with re-coding) were:

0. Make all unauthorized immigrants felons and send them back to their home country
1. Have a guest worker program that allows unauthorized immigrants to remain in US to work but only for limited time
2. Allow unauthorized immigrants to remain in US & eventually qualify for citizenship but only if they meet requirements
3. Allow unauthorized immigrants to remain in US & eventually qualify for citizenship without penalties

This is an ordinal variable with only four categories, but I tested and it had a pretty linear relationship with the dependent variable, so I felt comfortable using it as ratio-like rather than making it into a reference group set.

Favored U.S. govt. policy towards unauthorized immigrants

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Make felons and deport	1118	13.5	13.6	13.6
	Temporary guest worker program	1173	14.2	14.3	27.9
	eventual citizenship with requirements	4553	55.0	55.5	83.3
	eventual citizenship without penalties	1367	16.5	16.7	100.0
	Total	8210	99.2	100.0	
Missing	System	70	.8		
Total		8280	100.0		

Feelings towards Trump

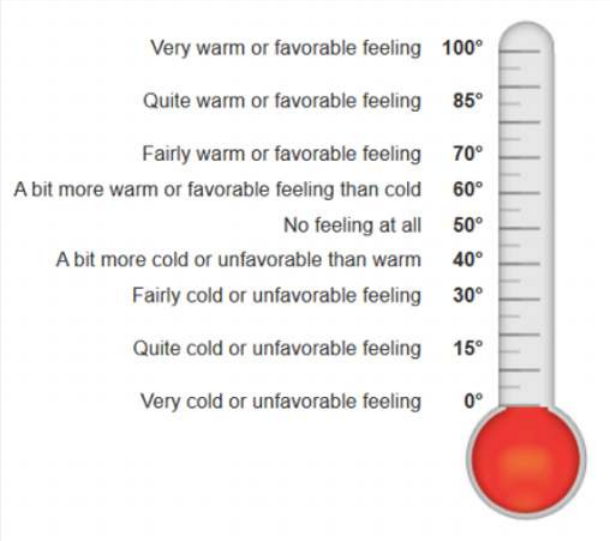
For feelings towards Trump, I used ANES' feelings thermometer question:

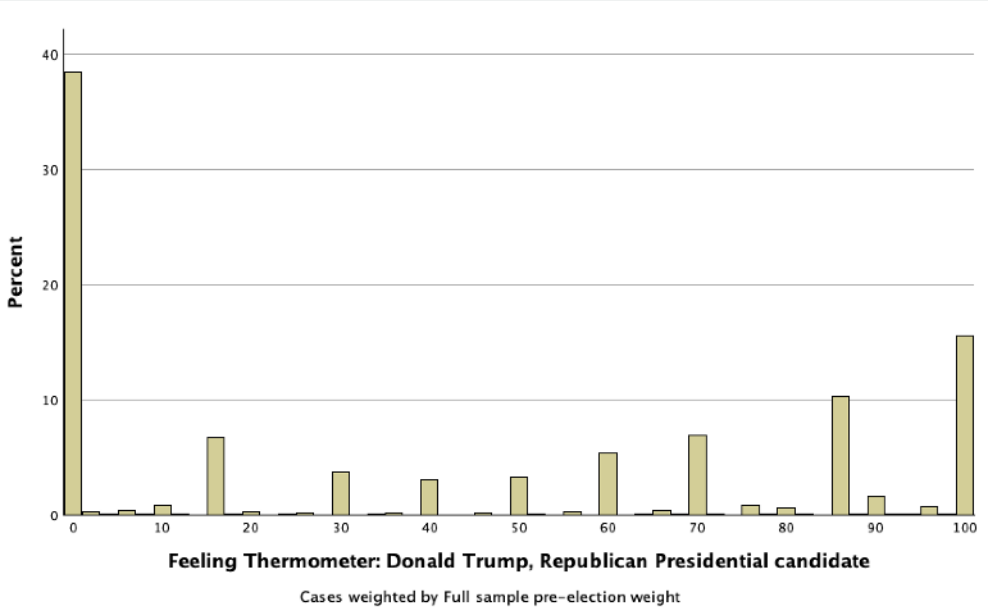
"I'd like to get your feelings toward some of our political leaders and other people who are in the news these days. I'll read the name of a person and I'd like you to rate that person using something we call the feeling thermometer.

Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the person. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the person and that you don't care too much for that person. You would rate the person at the 50 degree mark if you don't feel particularly warm or cold toward the person.

If we come to a person whose name you don't recognize, you don't need to rate that person. Just tell me and we'll move on to the next one.

How would you rate: Donald Trump."


Very warm or favorable feeling 100°
Quite warm or favorable feeling 85°
Fairly warm or favorable feeling 70°
A bit more warm or favorable feeling than cold 60°
No feeling at all 50°
A bit more cold or unfavorable than warm 40°
Fairly cold or unfavorable feeling 30°
Quite cold or unfavorable feeling 15°
Very cold or unfavorable feeling 0°


Percent
40
30
20
10
0
0
10
20
30
40
50
60
70
80
90
100
Feeling Thermometer: Donald Trump, Republican Presidential candidate
Cases weighted by Full sample pre-election weight

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Feeling Thermometer: Donald Trump, Republican Presidential candidate	8066	97.4%	214	2.6%	8280	100.0%

Descriptives

			Statistic	Std. Error
Feeling Thermometer: Donald Trump, Republican Presidential candidate	Mean		41.31	.448
	95% Confidence Interval for Mean	Lower Bound	40.43	
		Upper Bound	42.19	
	5% Trimmed Mean		40.34	
	Median		30.00	
	Variance		1618.632	
	Std. Deviation		40.232	
	Minimum		0	
	Maximum		100	
	Range		100	
	Interquartile Range		85	
	Skewness		.257	.027
	Kurtosis		-1.615	.055

Bivariate Regressions

Let's start with two bivariate regressions. We've already covered these, so I'm just going to skip the SPSS parts and give you the summary display tables and graphs.

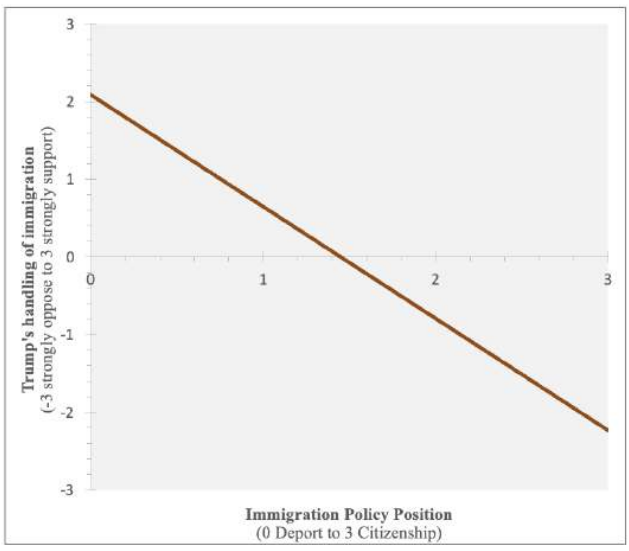
Policy Position on Immigration

Here's the first one:

Table One. Estimated unstandardized coefficients from an OLS regression predicting level of support for how Trump is handling immigration, with 95% confident intervals

Independent Variable	Immigration Policy Views***
Intercept	2.09*** (1.97, 2.20)
Slope	-1.44*** (-1.50, -1.38)
r^2	21.9%
Unweighted sample size	8,175

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Source: ANES 2020 pre-election



With a correlation of -0.47, there is a moderate negative relationship, with more liberal immigration views associated with less support for Trump's handling of immigration. Those who support unauthorized immigrants being felons and deported were predicted to support Trump's handling of immigration, while those who supported a path to citizenship without penalties were predicted to oppose Trump's handling of immigration. Views on immigration explain over 1/5 of variance in views on how Trump was handling immigration.

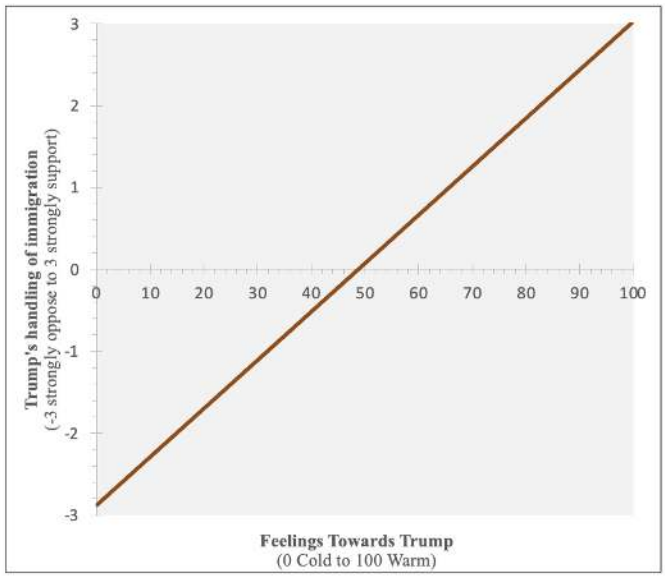
Feelings Towards Trump

Here's the second one:

Table Two. Estimated unstandardized coefficients from an OLS regression predicting level of support for how Trump is handling immigration, with 95% confident intervals

Independent Variable	Feelings Towards Trump***
Intercept	-2.87*** (-2.91, -2.83)
Slope	0.06*** (0.06, 0.06)
r^2	75.0%
Unweighted sample size	8,021

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Source: ANES 2020 pre-election



With a correlation of 0.87, there is a strong positive relationship, with warmer feelings towards Trump associated with more support for Trump's handling of immigration. Those with very cold feelings towards Trump are predicted to strongly oppose his handling of immigration and those with very warm feelings towards Trump are predicted to strongly support his handling of immigration. Respondents with a smidgeon colder than neutral feelings towards Trump (48.7 degrees) would be predicted to have neutral feelings on Trump's handling of immigration. Feelings towards Trump explain a whopping 3/4 of variance in views on how Trump was handling immigration.

Multivariate Regression

Now let's combine them and do a multivariate regression.

Running Multivariate Regression with SPSS

Running multivariate linear regression in SPSS is an identical process to what we did in Chapter 13 when we had a reference group set. All the independent variables are put in simultaneously, either into the “Independent(s)” box if you are using drop-down menus or on the /METHOD=ENTER row if you are using syntax.

SPSS Outputs and Interpretations

Here is the output for my multivariate regression, using the two IVs from above, with some explanations and commentary:

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Evaluation of the way Donald Trump is handling immigration	-.4347	2.74468	7991
Favored U.S. govt. policy towards unauthorized immigrants	1.7530	.88930	7991
Feeling Thermometer: Donald Trump, Republican Presidential candidate	41.30	40.294	7991

There is a smaller sample size than in the bivariate regressions, because we are only including cases that have valid data for all three variables.

Correlations

		Evaluation of the way Donald Trump is handling immigration	Favored U.S. govt. policy towards unauthorized immigrants	Feeling Thermometer: Donald Trump, Republican Presidential candidate
Pearson Correlation	Evaluation of the way Donald Trump is handling immigration	1.000	-.472	.866
	Favored U.S. govt. policy towards unauthorized immigrants	-.472	1.000	-.451
	Feeling Thermometer: Donald Trump, Republican Presidential candidate	.866	-.451	1.000
Sig. (1-tailed)	Evaluation of the way Donald Trump is handling immigration	.	<.001	<.001
	Favored U.S. govt. policy towards unauthorized immigrants	.000	.	.000
	Feeling Thermometer: Donald Trump, Republican Presidential candidate	.000	.000	.
N	Evaluation of the way Donald Trump is handling immigration	7991	7991	7991
	Favored U.S. govt. policy towards unauthorized immigrants	7991	7991	7991
	Feeling Thermometer: Donald Trump, Republican Presidential candidate	7991	7991	7991

This output is a correlation matrix. Note that we can see individual Pearson's r correlation coefficient values for bivariate relationships. These also tell us whether these bivariate relationships are negative or positive. You can see that one of the IVs has a negative relationship with our DV and one has a positive relationship. This is why the model summary below only reports the overall strength of the model and not its direction.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Feeling Thermometer: Donald Trump, Republican Presidential candidate, Favored U.S. govt. policy towards unauthorized immigrants[b]	.	Enter

a. Dependent Variable: Evaluation of the way Donald Trump is handling immigration

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.871[a]	.758	.758	1.35098

a. Predictors: (Constant), Feeling Thermometer: Donald Trump, Republican Presidential candidate, Favored U.S. govt. policy towards unauthorized immigrants

R-squared, explained variation, for the overall model

The model summary is based on including all of our independent variables. It does not tell us anything about either of them in isolation. This is the strength and explained variation for the overall model, accounting for both views on immigration policy and feelings towards Trump. Notice as well that **you can't just add the (absolute value of the) r-values or r-squared values from the bivariate regressions together to get the overall model summary r-value and r-squared value.** Our r-values from the bivariate regressions were -0.468 and 0.0866, which add up to a strength more than 1. Our r-squared values were 21.9% and 75.0%, which add up to 96.9%. Consider this: we can see in the bivariate correlation matrix above that our two independent variables have a moderate negative relationship. So there is overlapping explained variation. Once we already know feelings towards Trump, knowing immigration policy views does not improve our predictions by 21.9%, because some of that is already explained by feelings towards Trump. Similarly, if we already know immigration policy views, some of the 3/4 of explained variance from knowing feelings towards Trump was already accounted for, so we won't see a 75.0% boost on top of the 21.9% r-squared.

Interpretation of r: There is a strong relationship between our set of independent variables, views on immigration policy and feelings towards Trump, and support for Trump's handling of immigration.

Interpretation of r-squared: Views on immigration policy and feelings towards Trump **together/collectively** explain 75.8% of variance in support for Trump's handling of immigration (among respondents).

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	45610.446	2	22805.223	12495.096	<.001[b]
	Residual	14578.726	7988	1.825		
	Total	60189.171	7990			

a. Dependent Variable: Evaluation of the way Donald Trump is handling immigration

b. Predictors: (Constant), Feeling Thermometer: Donald Trump, Republican Presidential candidate, Favored U.S. govt. policy towards unauthorized immigrants

Overall Significance

For linear regression with one ratio-like or one indicator independent variable, our p-value from the ANOVA output was identical or near identical to the p-value for the slope of our independent variable. Now these may be different, because **the p-value here is based on the full model:** whether knowing **all of our independent variables collectively** improves our ability to predict the dependent variable over just knowing the mean of the dependent variable. This can be translated into a confidence level of whether r-squared is greater than zero. However, you could have one independent variable that helps and has a relationship and others that do not, and you could still have an ANOVA output with a p-value less than 0.05. **This does not tell you about individual variables; rather, it is about the multivariate model as a whole.**

The p-value is based on an F-test which is then converted into a p-value based on the F-distribution (which varies based on number of variables and sample size). Here is the formula that SPSS is using to calculate the F-statistic, where k is for the

$$\frac{r^2 \div k}{(1 - r^2) \div (n - k - 1)}$$

number of variables and n is for the sample size:

Interpretation: I am over 99.9% confident that there was a relationship between my independent variables, immigration policy views and feelings towards Trump, and my dependent variable, support for how Trump was handling immigration, among U.S. eligible voters leading up to the 2020 election.

This does not mean that each individual independent variable was actually helpful / was statistically significant in helping explain variation in support for how Trump was handling immigration. The ANOVA output box just speaks to the overall model. Technically $p<0.001$ means there is less than a 0.1% probability that we would have this data from a representative sample of our target population if, among our target population, knowing the collection of independent variables we put into the model does not improve our prediction of our DV over just knowing the mean of our DV.

If I did not know immigration policy views or feelings towards Trump, I would predict that a respondent would be at a -0.43 in their support for the way Trump is handling immigration, slightly unsupportive (this was the sample mean for the dependent variable). I am over 99.9% confident that, leading up to the 2020 election, for U.S. eligible voters, knowing immigration policy views and feelings towards Trump improves my prediction of support for the way Trump is handling immigration over just knowing the mean of support for the way Trump is handling immigration.

Screenshot 2023-11-27 at 1.24.35 AM.png

Slopes and Significance of Individual Independent Variables (and y-intercept), Part One

Remember with a reference group set how the constant, or y-intercept, was the value when all the indicator variables were zero? With a reference group set that was the reference group. Here, however, the y-intercept is the value when **all** the independent variables are zero. Therefore, now the y-intercept represents the point on the best fit line for a respondent who is both a 0 on immigration policy (favors deportation) and a 0 for feelings towards Trump (very cold).

We'll go over interpreting slopes below, but again note that the slopes here are different than in the bivariate regressions. It would be inaccurate to say that, for every one unit increase in support for more liberal immigration policies, support for the

way Trump is handling immigration decreases by 0.3 units. As we know from the bivariate regression, for every one unit increase in support for more liberal immigration policies, support for the way Trump is handling immigration decreases by 1.4 units.

> Therefore, whenever discussing individual variables from a multivariate model, we always add in language that notes this is while accounting for the other variable.
>
> **At the end of your explanation of the slope, or of the p-value, include language such as:**
>
> - **"while controlling for [other independent variable name]" or**
> - **"while holding [other independent variable name] constant."**

These slopes are telling us the average impact in the sample for these independent variables while taking into consideration the other independent variable simultaneously. For example, the slope for immigration policy views indicates that, if you were to consider someone who feels cold towards Trump, or feels neutral, or feels warm, that among those people with the same feelings, we would predict an average of a 0.315 unit decrease in support for Trump's handling of immigration for every 1 unit increase in support for immigration policy views. **The p-value is telling us whether we can be confident that, among our target population, the independent variable still has a relationship with the dependent variable after controlling for the other independent variable.** Once we know (control for) feelings towards Trump, do immigration policy views still matter? If we take people who have the same feelings towards Trump (e.g., both somewhat cold, or both very warm, so holding this variable constant), is there a relationship, among those with the same feelings towards Trump, between immigration policy views and evaluation of the way Trump is handling immigration?

Unweighted:

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Evaluation of the way Donald Trump is handling immigration	-.4450	2.74356	7966
Feeling Thermometer: Donald Trump, Republican Presidential candidate	40.41	40.374	7966
Favored U.S. govt. policy towards unauthorized immigrants	1.7705	.88657	7966

Just like before, I re-ran the regression without weighting to get the actual number of respondents.

Summary Display Table

I'm going to take the most useful information from above and put it into a summary display table.

Here is a Microsoft Word template for doing this. Fill out the red parts using the information from your SPSS output. MultivariateRegressionDisplayTableTemplate.docx

Here is what I filled out for the multivariate regression above. You should be able to find all the values and figure out the markers I used based on the outputs above.

Table Three. Estimated unstandardized coefficients from an OLS regression predicting level of support for how Trump is handling immigration, with 95% confident intervals

Independent Variable	Model***
Intercept	-2.19*** (-2.28, -2.10)
Immigration Policy Views	-0.32*** (-0.35, -0.28)
Feelings towards Trump	0.06*** (0.06, 0.06)
r^2	75.8%
Unweighted sample size	7,966

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Source: ANES 2020 pre-election

Here is a clean version without any parts in red:

Table Three. Estimated unstandardized coefficients from an OLS regression predicting level of support for how Trump is handling immigration, with 95% confident intervals

Independent Variable	Model***
Intercept	-2.19*** (-2.28, -2.10)
Immigration Policy Views	-0.32*** (-0.35, -0.28)
Feelings towards Trump	0.06*** (0.06, 0.06)
r^2	75.8%
Unweighted sample size	7,966

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Source: ANES 2020 pre-election

Multivariate Linear Regression: Equation, Predictions, & Graphing

Before returning to our slopes, let's graph the regression, which will also help you visualize the relationships.

Linear Regression Equation

We will write out our one regression line equation the same way we did for reference group sets in Chapter 13: $y=a+b_1x_1+b_2x_2$ (and so forth depending on how many independent variables we have). Again, only make one regression equation, and include all your variables in it. You cannot make regression equations using only some of your independent variables without running a new regression that only includes those variables.

Here, instead of x_1 I'm going to use P for policy views and instead of x_2 I'm going to use F for feelings towards Trump.

$y=a+b_PP+b_FF$
$y=(-2.19)+(-0.32)(P)+0.06(F)$

To find any predicted dependent values for a given policy position and feelings thermometer degrees, plug in the values for P and F and solve for y.

For example, someone who supports eventual citizenship for unauthorized immigrants who meet certain requirements (P=2) and has fairly cold feelings towards Trump (F=30) would be predicted to disapprove, but not strongly, of Trump's handling of immigration (y=-1):

$y=(-2.19)+(-0.32)(P)+0.06(F)$
$y=(-2.19)+(-0.32)(2)+0.06(30)$
$y=(-2.19)+(-0.64)+1.8$
$y=(-1.03)$

Here is a template you can use if you would prefer to just put in your coefficients (y-intercept and slopes) and let Excel do the calculations for you: PredictedValueCalculator.MultivariateLinearRegression.xlsx

Who would be predicted to have the lowest dependent variable value, to oppose Trump's handling of immigration the most?

To get to the lowest y-value, our value for Policy views would need to be as high as possible, since its slope is negative, and our value for Feelings towards Trump would need to be as low as possible, since its slope is positive. Policy views goes from 0 to 3, so P=3, and Feelings towars Trump goes from 0 to 100, so F=0.

y=(-2.19)+(-0.32)(P)+0.06(F)
y=(-2.19)+(-0.32)(3)+0.06(0)
y=(-3.15)

Someone who supports unauthorized immigrants eventually qualifying for citizenship without penalties (P=3) and has very cold feelings towards Trump (F=0) would be predicted to strongly oppose Trump's handling of immigration (beyond the lower bound of -3).

Who would be predicted to have the highest dependent variable value, to support Trump's handling the most?
To get to the highest y-value, our Policy views value would need to be as low as possible, since its slope is negative, and our Feelings thermometer value would need to be as high as possible, since its slope is positive. P goes from 0 to 3, so P=0, and F goes from 0 to 100, so F=100.

y=(-2.19)+(-0.32)(P)+0.06(F)
y=(-2.19)+(-0.32)(0)+0.06(100)
y=(3.81)

Someone who supports making unauthorized immigrants felons and deporting them (P=0) and has very warm feelings towards Trump (F=100) would be predicted to strongly support Trump's handling of immigration (beyond the upper bound of +3).

Graphing

Here is a template for making multivariate linear regression graphs: MutlvariateLinearRegressionGraphTemplate.xlsx

Our graphs will only be of two independent variables (if using a reference group set, that can count as one variable). If you have three or more variables, you need to hold the other variables constant (e.g., you would use two independent variables, such as age and income, and then hold other variables at a value you select, such as race=white, marital status=married, gender=woman, etc.) You can then replace the variables in your equation with those constant values to simplify your equation and graph it. The graph will show the relationship accounting for the two independent variables, while holding the others constant (e.g. a graph of the predicted relationship between age and income with the dependent variable, among white married women).

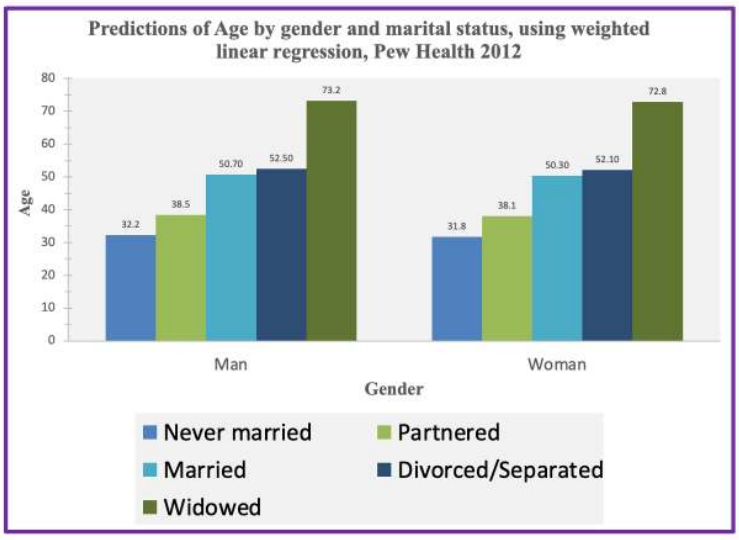
Graphing two categorical variables

If you have two categorical variables (e.g., two indicator variables, two reference group sets, one indicator variable and one reference group set), you will need to make a bar graph. (If you have more than two IVs, you only need to make a bar graph if the two IVs you want to show in terms of relationships are both categorical.) Here is an example of a bar graph where the IVs were gender (indicator) and marital status (reference group set):

Table One. Results from linear regression predicting age

Independent Variables	Model***
Intercept (married man)	50.7*** (50.3, 51.1)
Gender (Woman=1)	-0.449* (-0.9, -0.02)
Divorced/Separated=1	1.8*** (1.1, 2.4)
Never Married=1	-18.5*** (-19.0, -18.0)
Partnered=1	-12.2*** (-13.0, -11.4)
Widowed=1	22.5*** (21.6, 23.4)
r^2	34.7%
n	2945

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Sample size is unweighted, all other analyses probability-weighted
Source: Pew Health 2012



However, if at least one of your variables is ratio-like, you can make a line graph.

If only one of your independent variables is ratio-like, that is the variable that will go on the x-axis. If they are both ratio-like, you can choose which variable goes on the x-axis. In general, if you have a variable that has more of a spread (e.g., age from 18 to 89, feelings thermometer temperatures from 0 to 100, income, etc.), those are better variables to go on the x-axis than ones that are ratio-like but have fewer categories or ones that are ratio-level but still have only a few categories (e.g., number of siblings, age if everyone is 10, 11, or 12 years old, a Likert scale variable that goes from -2 to +2, etc.).

The independent variable that is not going on the x-axis will be presented by different lines. You need to choose a few different values to show for this variable. You will be holding this variable constant at these various values. For example, for a feelings thermometer variable, I might choose 0, 50, and 100. For a strongly agree to strongly disagree variable I might choose strongly agree, strongly disagree, and neutral. If there are not many respondents at these extremes, I might choose values closer to the center, e.g., for the feelings thermometer: 30 (fairly cold), 50 (neutral), and 70 (fairly warm); for a Likert Scale: agree and disagree rather than strongly agree and strongly disagree.

For this multivariate regression, I will put the feelings thermometer on the x-axis, since it goes from 0 to 100, whereas policy views only has 4 options. Because there are only 4 options for the policy views variable, I am going to graph all 4 of them rather than just choosing 2 or 3. For my graph, each line will represent a different policy view.

Here is the first part of the template for a four-line graph:

Replace cells B3, B4, and B5 with the intercept and slopes from your regression line equation.

intercept (constant)	#
IV 1 slope (x-axis)	#
IV 2 slope (lines)	#

Replace cells B11 and B12 with the minimum and maximum value for your independent variable that is going on the x-axis.
Replace cells C10, D10, E10, and F10 with the labels of the values you chose for your other independent variable.
Replace cells C11, D11, E11, and F11 with the values (numbers/codes) that correspond with the labels of the values you chose for your other independent variable.

	IV for lines:	value description	value description	value description	value description
	x (IV) for x-axis	#	#	#	#
x (IV for x-axis) minimum	#	#VALUE!	#VALUE!	#VALUE!	#VALUE!
x (IV for x-axis) maximum	#	#VALUE!	#VALUE!	#VALUE!	#VALUE!

Here are the coefficients I need:

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardi Coefficie Beta
1	(Constant)	-2.189	.047	
	Favored U.S. govt. policy towards unauthorized immigrants	-.315	.019	-
	Feeling Thermometer: Donald Trump, Republican Presidential candidate	.056	.000	

a. Dependent Variable: Evaluation of the way Donald Trump is handling im

Table Three. Estimated unstandardized coefficients from an OLS regression predicting level of support for how Trump is handling immigration, with 95% confident intervals

Independent Variable	Model***
Intercept	-2.19*** (-2.28, -2.10)
Immigration Policy Views	-0.32*** (-0.35, -0.28)
Feelings towards Trump	0.06*** (0.06, 0.06)

Here is what I put into the top part of the template:

Replace cells B3, B4, and B5 with the intercept and slopes from your regression line equation.

intercept (constant)	-2.189
IV 1 slope (x-axis)	0.056
IV 2 slope (lines)	-0.315

My independent variable for the x-axis, the feelings thermometer, goes from 0 to 100, with cases at its minimum and maximum.

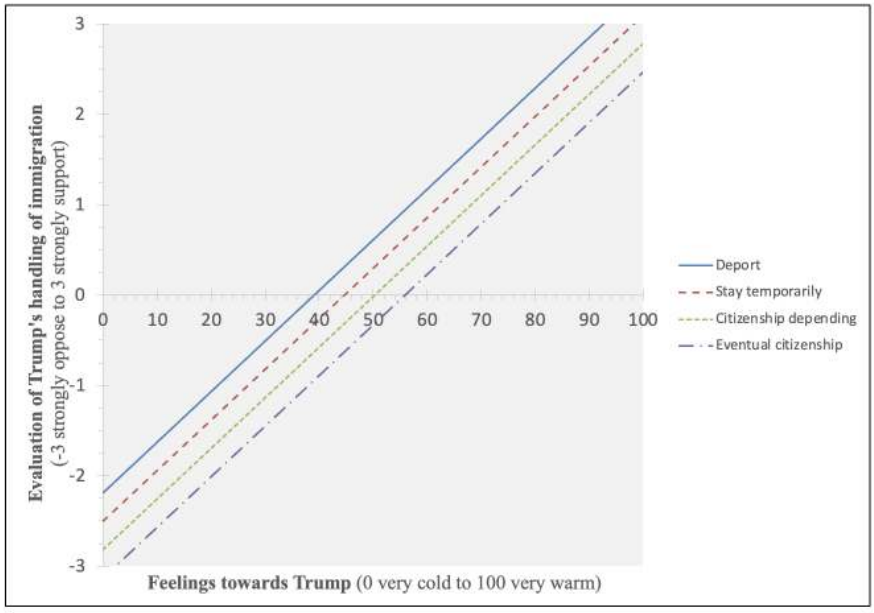
My independent variable I am holding constant for the four lines, the policy views variable, goes from 0 to 3, with cases at its minimum and maximum.

Here is what I put in for the next part of the template:

Replace cells B11 and B12 with the minimum and maximum value for your independent variable that is going on the x-axis.
Replace cells C10, D10, E10, and F10 with the labels of the values you chose for your other independent variable.
Replace cells C11, D11, E11, and F11 with the values (numbers/codes) that correspond with the labels of the values you chose for your other independent variable.

	IV for lines:	Deport	Stay temporarily	Citizenship depending	Eventual citizenship
	x (IV) for x-axis	**0**	**1**	**2**	**3**
x (IV for x-axis) minimum	0	-2.189	-2.504	-2.819	-3.134
x (IV for x-axis) maximum	100	3.411	3.096	2.781	2.466

Finally, I edited my axes labels and my x-axis and y-axis bounds and units. For the x-axis, I'm using the feelings thermometer variable, so I made the bounds 0 to 100, major units 10, and minor units 2. Here is what my final graph looks like:



https://socialsci.libretexts.org/@go/page/194404

You can see that, across policy positions, as feelings towards Trump get warmer, support for Trump's handling of immigration increases. Also, you can see that at any given feeling towards Trump, people with more liberal policy positions are less supportive of Trump's handling of immigration and people with more conservative policy positions are more supportive of Trump's handling of immigration.

Slopes and Significance of Individual Independent Variables (and y-intercept), Continued (Part Two)

Let's return to the y-intercept, slopes, and their p-values, and interpret them in context.

Y-intercept: -2.189, p-<0.001

The y-intercept is the predicted value when all IVs are zero. y=(-2.19)+(-0.32)(0)+0.06(0)=(-2.19)

In this case, this is someone at a 0 (very cold) in their feelings towards Trump and a 0 (deport) in their policy position on immigration.

In the line graph above, find the solid blue line and where it intersects with the y-axis (where the x-axis value is 0). You can see that it is at -2.19.

We would predict that someone who feels very cold towards Trump and supports deporting unauthorized immigrants would be at a -2.2, supportive of Trump's handling on immigration.

The p-value means we are over 99.9% confident that, among our target population, the y-intercept is not zero. In this case, that means that we are over 99.9% confident that, among U.S. eligible voters leading up to the 2020 election, someone who feels very cold towards Trump and supports deporting unauthorized immigrants would not be predicted to be neutral (at a 0) in their evaluation of Trump's handling of immigration.

Template for interpreting slopes and p-values

Slopes (template):

Interpreting each slope is the same as what we did in bivariate linear regression, except this time you need to add "while holding [the other IV(s) constant]" or "while controlling for [the other IVs]".

P-values (template):

Interpreting each p-value is the same as what we did in bivariate linear regression, except this time you need to add "while holding [the other IV(s) constant]" or "while controlling for [the other IVs]".

Contextualize these by naming the other IV(s).

(If you had a lot of independent variables, you would not name all the other variables.)

Immigration policy views: -0.32, p<0.001

- The slope for immigration policy views, while holding feelings towards Trump constant, is -0.32.
- That means that, while controlling for feelings towards Trump, for every one-unit increase in immigration policy views, support for Trump's immigration handling goes down by 0.32.
- Look up at the graph. Hold feelings towards Trump constant by choosing any feelings thermometer temperature and drawing an imaginary vertical line on the graph. For example, take a look at when feelings towards Trump are at 70, fairly warm. Because this variable is coded in one-unit increments and I graphed lines for every value, you can literally see the slope as you move from line to line. Each unit increase for the independent variable, which goes from 0 to 1, and then 1 to 2, and then 2 to 3, represents the change: from deport to stay temporarily, from stay temporarily to eventual citizenship with conditions, from eventual citizenship with conditions to eventual citizenship without penalties. As you go straight down at any given feelings thermometer temperature, you will see that as policy views goes from 0 to 3, it goes down by 0.32 points each time. On average, every one unit increase in immigration policy views (becoming more liberal, from 0 at deport to 3 at eventual citizenship without penalties) is associated with a respondent's support for Trump's handling of immigration decreasing by an average of 0.32 (from -3 strongly oppose to +3 strongly support), while controlling for feelings towards Trump.

- The p-value means we are over 99.9% confident that, among our target population, the slope is not zero, when controlling for the other independent variable(s). In this case, that means we are over 99.9% confident that, among U.S. eligible voters leading up to the 2020 election, there is a relationship between policy positions on immigration and support for how Trump is handling immigration, while holding feelings towards Trump constant.

Feelings towards Trump: 0.06, p<0.001

- The slope for feelings towards Trump, while holding policy positions on immigration constant, is 0.06.
- That means that, while controlling for policy positions on immigration, for every one degree warmer a respondent's feelings towards Trump are, support for Trump's immigration. handling goes up by 0.06.
- Look up at the graph. Hold policy positions constant by choosing any one line to follow. They all have the same slope of 0.06. As you go across from left to right, every one degree warmer (in feelings towards Trump) goes up by 0.06 (1% of the distance of the -3 to +3 scale). Every 10 degrees warmer goes up by 0.6. For an increase of 50 degrees warmth (e.g., from 0 to 50, or 50 to 100, or 20 to 70, etc.), the dependent variable goes up by 3 units, half the scale of -3 to +3. You can see how much support for Trump's immigration handling goes up as feelings get warmer towards Trump, while controlling for policy positions on immigration.
- The p-value means we are over 99.9% confident that, among our target population, the slope is not zero, when controlling for the other independent variable(s). In this case, that means we are over 99.9% confident that, among U.S. eligible voters leading up to the 2020 election, there is a relationship between feelings towards Trump and support for how Trump is handling immigration, while holding policy positions on immigration constant.

Beta (standardized) Slopes: Comparing IV Contributions

In the multivariate model, which has a more substantive impact? Policy positions on immigration or feelings towards Trump?

At first glance, you might assume it's policy positions on immigration, as the slope is larger (-0.32 vs. 0.06). However, policy positions on immigration only has 6 one-unit increments on its 7-category scale, whereas feelings towards Trump can vary from 0 to 100. Furthermore, respondents are not equally distributed across these scales, and so if people have more uniform responses for one of the variables, it might not explain as much about the dependent variable. So, from what we've learned so far, it's difficult to tell.

The way we can make comparisons between slopes is by using a different type of slope, called a Beta or standardized slope.

Definition: Beta (standardized) slope

Beta slopes, or standardized slopes, are slopes with standard deviations as their unit.

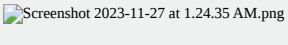
Beta or standardized slopes are slopes based on standard deviations. They make the unit for all independent variables' slopes into standard deviations.

$$Standardized\ Slope = Original\ Slope \times \frac{SD_{IV}}{SD_{DV}}$$

Whereas for our unstandardized slopes, the units are whatever they are (e.g., degrees on the feelings thermometer or dollars for income), for standardized slopes, the units are all standard deviations.

- An unstandardized slope's interpretation is: Every one unit increase in x is associated with an increase in y of (unstandardized slope value) units.
- A standardized slope's interpretation is: Every one standard deviation increase in x is associated with an increase in y of (beta slope value) standard deviations.

You can find standardized slopes in your SPSS output.



They are in the "Standardized Coefficients, Beta" column of the coefficients output. Here the standardized slope for immigration policy position is -0.102 and the standardized slope for feelings towards Trump is 0.820.

Immigration Policy Views Slopes

Unstandardized: Every one unit increase in immigration policy views (becoming more liberal on the 0 to 3 scale) is associated with a 0.315 unit decrease in support for Trump's handling of immigration (on the -3 to 3 scale), while holding feelings towards Trump constant.

Standardized: Every one standard deviation increase in immigration policy views is associated with a decrease of 0.102 standard deviations in support for Trump's handling of immigration, while holding feelings towards Trump constant.

Feelings Towards Trump Slopes

Unstandardized: Every one degree warmer in feelings towards Trump (on the 0 to 100 scale) is associated with a 0.56 unit decrease in support for Trump's handling of immigration (on the -3 to 3 scale), while holding policy positions on immigration constant.

Standardized: Every one standard deviation increase in feelings towards Trump is associated with an increase of 0.820 standard deviations in support for Trump's handling of immigration, while holding policy positions on immigration constant.

Beta slopes are not very intuitive in terms of their interpretations.
What does it look like when policy positions on immigration increase by one standard deviation, or feelings towards Trump increases by one standard deviation, or evaluations of Trump's handling on immigration decreases by 0.1 or increases by 0.8 standard deviations?

I can use the standard deviations of each variables to translate these standard deviations back into the units that the original coding of the variable uses.

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
Evaluation of the way Donald Trump is handling immigration	-.4347	2.74468	7991
Favored U.S. govt. policy towards unauthorized immigrants	1.7530	.88930	7991
Feeling Thermometer: Donald Trump, Republican Presidential candidate	41.30	40.294	7991

- A one standard deviation increase in policy positions on immigration is a 0.89 unit increase on the 0 to 3 scale.
- A one standard deviation increase in feelings towards Trump is 40 degrees warmer on the feelings thermometer.
- A decrease of 0.1 standard deviations in evaluations of Trump's handling of immigration is a 0.28 unit decrease on the -3 to +3 scale.
- An increase of 0.8 standard deviations in evaluations of Trump's handling of immigration is a 2.25 unit increase on the -3 to +3 scale.

So for example, for the Beta slope for feelings towards Trump:
Every one standard deviation increase (40 degrees warmer) in feelings towards Trump is associated with an increase of 0.820 standard deviations (a 2.25 unit decrease) in support for Trump's handling of immigration, while holding policy positions on immigration constant.

These are still not very helpful.

However, **beta slopes are useful for comparing the contribution of each independent variable to the dependent variable. We can't tell this from unstandardized slopes,** as we are often comparing apples to oranges (the variables may have different units, different variances, etc.).

> **When using beta slopes to compare which independent variable has more of an impact, just consider their absolute value (distance from zero). We don't care if the slope is positive or negative for considering its impact.**

Here, 0.82 > 0.10, so feelings towards Trump has more of an impact than policy positions on immigration. It has about 8 times more impact. This is the opposite conclusion we would have come to if we had just evaluated the unstandardized coefficients.

Comparing Bivariate and Multivariate Regressions

Flashback

Remember when we looked at elaborated/partial crosstabs in Chapter 9? We compared bivariate to multivariate relationships.

Here was a summary of what happened when we introduced a new control variable:

- The initial relationship stays the same
  - Replication: relationship still exists
- The initial relationship goes away
  - If control variable is antecedent
    - Spurious: Relationship explained by control (confounding) variable
  - If control variable is intervening
    - Relationship's mechanism explained by control (mediating) variable
- The initial relationship diminishes
  - If control variable is antecedent
    - Relationship is partially explained by control variable, but original independent variable still has its own independent effect
  - If control variable is intervening
    - Relationship's mechanism is partially but not fully explained by control variable
- The initial relationship varies
  - Moderation: relationship differs/interacts with control variable
- The initial relationship increases (or there was no relationship, and now a relationship appears)
  - Suppression: Relationship was suppressed, relationship is explained by its effect within control variable groups

Similar to when we added a second independent variable into a crosstab, we can analyze what happens when we introduce an additional independent variable or variables into our bivariate model using linear regression.

For example, I could ask: do people evaluate Trump's handling of immigration policy based on their own views on immigration (in comparison to what Trump is doing), or do their feelings about Trump trump their views on immigration? Once I already know people's policy positions on immigration, does knowing their feelings towards Trump improve our predictions further? (Or vice versa, does feelings towards Trump improve our prediction of people's evaluations of Trump's handling of immigration? And once we know how people feel overall about Trump, does knowing their policy positions on immigration improve our predictions of Trump's handling of immigration further?)

We can compare our two bivariate regressions and our multivariate regression altogether in one summary display table. We want to pay attention to what happens to the r-squared, significance of the independent variables, and slopes when we move from the bivariate to multivariate regressions.

Here is a template for doing this: LinearRegression.MultipleModels.DisplayTableTemplate.xlsx

Each column is a model/regression. For the bivariate models, you leave blank the variable that is not part of your regression. The final column is called the "full model" because it contains all of your independent variables.

- Don't forget to update the *** which are already listed as asterisks rather than question marks in this template.

- Note that in Excel, you cannot start a cell by typing a - sign or it will think it's a formula. To correct this, type ' before the - sign. For example, to write -3.5* you would input '-3.5*. This tells Excel it is just text.

Since I have two ratio-like variables, I'm using the first template worksheet, which looks like this:

Table One. Estimated unstandardized coefficients from weighted OLS regressions predicting DV, with 95% confidence intervals

Independent Variables	IV1Name***	IV2Name***	Full Model***
Intercept	Coefficient*** (LB, UB)	Coefficient*** (LB, UB)	Coefficient*** (LB, UB)
IV1Name (coding?)	Slope*** (LB, UB)		Slope*** (LB, UB)
IV2Name (coding?)		Slope*** (LB, UB)	Slope*** (LB, UB)
r^2	#.#%	#.#%	#.#%
n	###	###	###

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Sample size is unweighted, all other analyses probability-weighted
Source: Dataset name YEAR

Note: Any necessary notes here.

Coding is optional / for when relevant (e.g., "in thouands" for income if it is in thousands).
If no coding information is needed, delete "(coding?)"
For an indicator variable, add label=1 (substituting the name of the label instead of the word label)

I'm just taking the information from the summary display tables earlier in this chapter (Tables One, Two, and Three) and putting them each into one of the columns:

Table Four. Estimated unstandardized coefficients from weighted OLS regressions predicting evaluation of how Trump is handling immigration (-3 strongly oppose to +3 strongly support), with 95% confidence intervals

Independent Variables	Imm. Policy Views***	Feelings twd Trump*	Full Model***
Intercept	2.09*** (1.97, 2.20)	-2.87*** (-2.91, -2.83)	-2.19*** (-2.28, -2.10)
ImmPolicyViews (0 deport to 3 citizenship)	-1.44*** (-1.50, -1.38)		-0.32*** (-0.35, -0.28)
FTTrump (0 very cold to 100 very warm)		0.06*** (0.06, 0.06)	0.06*** (0.06, 0.06) (LB, UB)
r^2	21.9%	75.0%	75.8%
n	8,175	8,021	7,966

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Sample size is unweighted, all other analyses probability-weighted
Source: ANES 2020 pre-election

Now I can make comparisons between the bivariate and multivariate model. Both independent variables are significant in both the bivariate and multivariate regressions. This means that both are having an effect on how U.S. eligible voters prior to the 2020

election were evaluating Trump's handling of immigration, while controlling for the other (both improve our prediction, even after we know the other).

The r-squared of the full model is less than 1% greater than the r-squared of feelings towards Trump model, but over 50% more than the immigration policy views model. This means that among our respondents, if you already know immigration policy views, knowing feelings towards Trump will substantially improve the explained variation of evaluations of Trump's handling on immigration, but if you already know feelings towards Trump, also knowing immigration views will not add much explanatory power.

Finally, let's take a look at slopes. The feelings thermometer for Trump slope stays identical (when rounded to the nearest hundredths place) after controlling for immigration policy views. It is unaffected by policy views on immigration. However, we can see that the slope for immigration policy views drops from -1.44 to -0.32. The initial slope was 4.5 times larger than the slope after controlling for feelings thermometer. This means that feelings towards Trump partially explains the relationship between immigration policy views and evaluations of Trump's handling on immigration. Depending on which independent variable comes first time-order wise, this could mean that for many but not all respondents, their feelings on Trump are impacting their views on immigration based on them taking cues from Trump (or distancing themselves from Trump) for their views on immigration, or it could mean that many respondent's positions on immigration are impacting their feelings towards Trump and their evaluation of Trump's handling on immigration in a similar direction, or it could be some combination of these two. Based on the confidence intervals for the slopes, we can be confident about these claims for U.S. eligible voters leading up to the 2020 election as a whole, not just our respondents.

What is the overall story here? Originally when we looked at immigration policy views, we saw that it had a substantive impact on evaluating Trump's handling of immigration. But this is not the whole story. It's not only immigration policy views that are impacting how people evaluate Trump's handling of immigration. Once we controlled for feelings towards Trump, we saw this relationship diminish substantially, indicating that it is partially explained by feelings towards Trump. Additionally, we saw that feelings towards Trump by itself (the bivariate model) explains 3/4 of variation in evaluation of Trump's handling on immigration --- ignoring people's policy positions on immigration, which by itself explained about 1/5 of the variation in evaluation of Trump's handling on immigration. People's evaluations of Trump's handling on immigration is influenced more by feelings towards Trump than by policy positions on immigration. While there are some people whose policy positions on immigration are influencing their evaluation of Trump's handling on immigration, people's evaluation of Trump's handling of immigration are most influenced by how people overall feel about Trump, not by whether their policy positions on immigration align with Trump's handling of immigration.

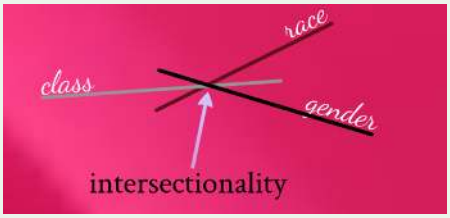
This page titled 14: Multivariate Linear Regression is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Ezra Temko (Consortium of Academic and Research Libraries in Illinois (CARLI)) .

15: Interactions

Chapter 15 will teach you about interactions with linear regression.

Intersectionality (conceptual introduction)

Are you familiar with the term **intersectionality?**



Intersectionality is about how our multiple identities interact to impact our unique experience in the social world. Our intersecting identities qualitatively change our experience, so it is necessary to take these into account when analyzing the empirical world. It's also not just an additive change where you can put the experiences of one identity and another together. They can interact in ways that produce qualitative different outcomes.

> While the idea predates the term, **intersectionality** was originally coined by Kimberlé Crenshaw. An example Crenshaw uses to explain the term is the 1970s court case of DeGraffenreid vs. General Motors. Historically women could only work at General Motors as office staff and men could only work at General Motors in manufacturing. However, only white women were allowed as office staff. When this policy changed, black women entered the workforce. However, when layoffs came around, black women were the first to be laid off, and when it came time for promotions, they were the last to be promoted. General Motors claimed this was because of their lack of seniority. Five black women sued, noting that the reason for their lack of seniority was because they were not allowed to work at General Motors previously due to their status as black women. They believed this constituted discrimination against them in both promotion and layoff policies. The courts ruled in favor of General Motors, saying they would not create a super-remedy and that the plaintiffs had failed to prove gender discrimination, since (white) women had been able to previously work in the front office, and that the plaintiffs had failed to prove racial discrimination, since black (men) had previously been able to work in manufacturing. Crenshaw's point in sharing this case is that it was not race or gender alone that impacted these employees; it was the intersection of race and gender–the status of being a black woman–that meant they could not gain earlier employment.

Statistics in Action: Gender and Race Pay Gaps

Earlier we looked at gender and race pay gaps for full-time civilian workers, using U.S. Census data (ACS data for 2017 to 2021).

Here is the coefficients output and a graph of a multivariate regression with both gender and race as IVs.

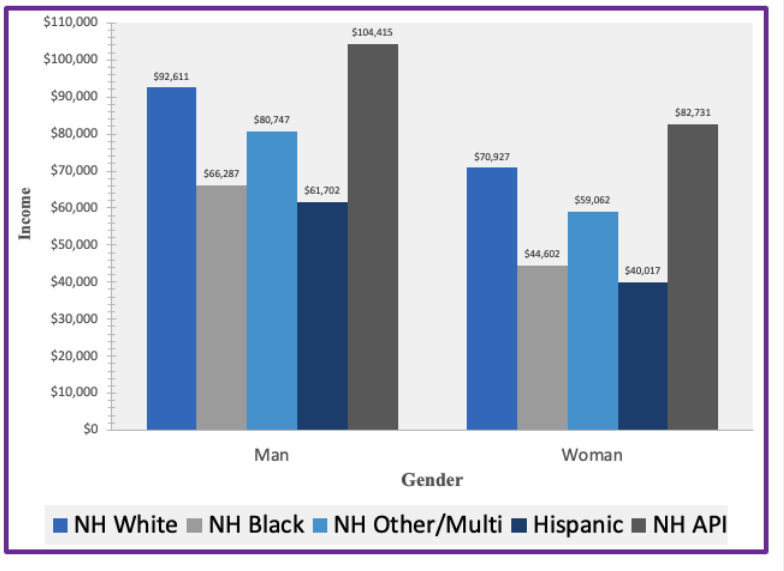
Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	92611.366	11.448		8089.706	<.001	92588.928	92633.804
	Gender (1=female)	-21684.680	15.120	-.132	-1434.133	<.001	-21714.315	-21655.044
	Race: NH Black only=1	-26324.188	24.109	-.103	-1091.900	<.001	-26371.440	-26276.936
	Race: NH Other/Multi=1	-11864.237	42.872	-.026	-276.733	<.001	-11948.266	-11780.209
	Race: Hispanic=1	-30909.187	20.236	-.145	-1527.463	<.001	-30948.849	-30869.526
	Race: NH API=1	11803.950	31.124	.035	379.254	<.001	11742.948	11864.952

a. Dependent Variable: Total personal income

(The p-values were also all significant at p<0.001 with weighting off.)

Altogether, gender and race explain 4.7% percent of variance in income among these full-time civilian worker respondents. We can see that gender has a particular effect (on average, women's income was about $21,700 less) regardless of race / while holding race constant / for any particular race, and that race has a particular effect (on average, compared to NH White, NH Black only $26K less, NH other/multi $12K less, Hispanic $30K less, and NH API $12K more) regardless of gender / holding gender constant / for men and for women.



However, does gender have the same effect across racial groups? Does race have the same effect for men and women? Exploring their intersection can help us answer that.

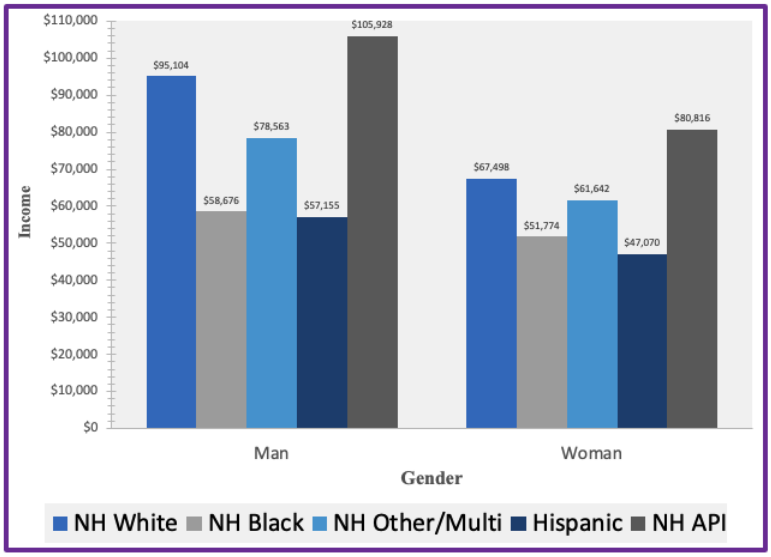
Here's some results from a new regression that can answer that, this time including "interactions" between race and gender. While we did not add any new demographics, now race and gender and their interaction explain 5.0% percent of variance in income among our respondents.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	95104.458	12.487		7616.108	<.001	95079.984	95128.933
	Gender (1=female)	-27606.894	19.246	-.168	-1434.422	<.001	-27644.616	-27569.173
	Race: NH Black only=1	-36428.002	34.064	-.143	-1069.398	<.001	-36494.767	-36361.238
	Race: NH Other/Multi=1	-16541.078	58.079	-.036	-284.804	<.001	-16654.910	-16427.245
	Race: Hispanic=1	-37949.050	26.052	-.178	-1456.665	<.001	-38000.111	-37897.989
	Race: NH API=1	10823.233	41.511	.033	260.732	<.001	10741.873	10904.593
	NHBlackGenderInteraction	20704.937	48.181	.060	429.737	<.001	20610.505	20799.369
	NHAPIGenderInteraction	2495.012	62.625	.005	39.840	<.001	2372.269	2617.755
	NHOtherGenderInteraction	10685.679	85.955	.016	124.318	<.001	10517.211	10854.147
	HispanicGenderInteraction	17521.851	41.281	.055	424.455	<.001	17440.942	17602.759

a. Dependent Variable: Total personal income

(The p-values were also all significant at p<0.001 with weighting off.)



Before, gender had a uniform gap for each racial group, and race had a uniform gap for each gender, but now we can see in the graph above that subgroups are not uniform. While on average a NH White Woman earns 71 cents for every dollar earned by a NH White Man, a NH Black Woman earns 88 cents for every dollar earned by a NH Black Man (and 54 cents for every dollar earned by a NH White Man). There is not simply a gender pay gap and a race pay gap. You can't take the two gaps and add them together, or just control for the other like we did in the multivariate model that did not include the interaction. Gender and race intersect to create different experiences and outcomes for different subgroups.

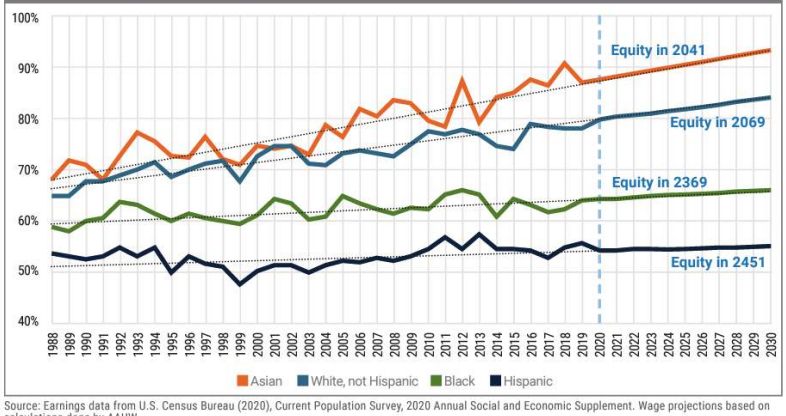
Here are a couple graphs of intersectional wage gaps from the American Association of University Women (AAUW), using median income. You can read more about its causes at the link as well.

Figure 1. Women's Earnings as a Percentage of White Men's Earnings, by Race/Ethnicity

Race/Ethnicity	Earnings Ratio
White alone, not Hispanic or Latino	79%
Black or African American	63%
American Indian and Alaska Native	60%
Asian	87%
Native Hawaiian and Other Pacific Islander	63%
Hispanic or Latino origin	55%

Source: Figures for Black, Latina and Asian women are AAUW calculations based on CPS 2020 ASEC; figures for Native Hawaiian and Pacific Islander and Native women are AAUW calculations based on ACS 2019.

Figure 6. Wage Gap Closing Projections, by Race/Ethnicity



Source: Earnings data from U.S. Census Bureau (2020), Current Population Survey, 2020 Annual Social and Economic Supplement. Wage projections based on calculations done by AAUW.

Interactions

Traditionally intersectionality is used to look at social identities to help us understand systems of oppression. In statistics, we can also broaden this concept to look at the intersection of any variables and how that intersection may influence our dependent variable. While region and age might both impact cereal preferences, while controlling for the other, it could be that the intersection of region and age gives us improved information, perhaps older people from a particular region have a particular cereal preference that younger people in that region don't have and older people in other regions don't have.

In statistics, the variables we create showing the intersection of two variables are called **interaction** variables and the outcomes are **interaction effects**.

Definition: Interaction

A statistical **interaction** occurs when the effects of one predictor variable varies based on another predictor variable.

Interaction variables are computed by multiplying the two independent variables together. The two independent variables, such as race and gender, are the parent variables, or main effect variables, and their product is the interaction variable(s). Generally you will not see codes with values listed for interactions. You can also create three-way interactions (e.g., race*gender*sexuality), but these get more complicated, and depending on your variables you have to be careful about the sample size of your subgroups. In this chapter we are sticking to two-way interactions (interactions with two parent variables).

SPSS: Run multivariate regressions with interactions the same way you ran multivariate regressions in Chapter 14. This time, just make sure to include the parent variables **and** the interaction variable as your independent variables. This isolates the effect of the interaction variable so you can see if it makes a difference. If you don't include the parent variables, the interaction will also be showing you some of the main effects of the parent variables, so you won't know if the interaction matters or what its independent effect might be.

When we test for interactions, we are testing to see if the effect of one independent variable varies among subgroups categorized by a second independent variable. While you can use your SPSS output to figure out if interactions are significant,

beyond that, interpreting interactions from your SPSS output is not intuitive. You can't just look at the interaction slopes without also considering the slopes of the main effects, and the coding of multiplied variables is complicated. If you start to calculate predicted values, you can start to observe what the interaction's story is, but **the easiest way to figure out its pattern is to graph the interaction.** Once you've graphed an interaction, you can figure out its story.

Caution

Don't try to tell a story about your parent/main effect or your interaction variable slopes without also including the others simultaneously.

Remember that the overall model significance in the ANOVA output is about whether collectively all of your IVs improve your predictions of your DV. The p-value in the ANOVA indicates whether the overall model is significant, not whether any specific variables, including any interactions, are significant. To figure out if an interaction is significant, you have to go to the coefficients output and look at the p-value on the row with the interaction variable. (And this will only mean there is a significant interaction if you also remembered to include your parent/main effect variables in the regression.)

Statistics in Action: Political Party, Income, and Perceptions of Income Inequality

Using the 2022 General Social Survey (representative of U.S. adults), let's take a look at how political party and income influence perspectives on the need for government action regarding income inequality.

Our dependent variable asks:

> *Some people think that the government in Washington ought to reduce the income differences between the rich and the poor, perhaps by raising the taxes of wealthy families or by giving income assistance to the poor. Others think that the government should not concern itself with reducing this income difference between the rich and the poor. Here is a card with a scale from 1 to 7. Think of a score of 1 as meaning that the government ought to reduce the income differences between rich and poor, and a score of 7 meaning that the government should not concern itself with reducing income differences. What score between 1 and 7 comes closest to the way you feel?*

I re-coded this so that -3 is "the government should not concern itself with reducing income differences," 0 is neutral, and +3 is "the government should reduce income differences."

Political party asks, "Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent, or what?" I re-coded this to:

- -3 Strong Republican
- -2 Not very strong Republican
- -1 Independent, close to Republican
- 0 Independent (neither, no response); Other party
- 1 Independent, close to Democrat
- 2 Not very strong Democrat
- 3 Strong Democrat

For income, I used respondent's estimated individual income, coded in thousands of dollars.

Using SPSS

Syntax:
Here is the syntax I used with SPSS to run the regression. Note how it's the same as for multivariate regression, just with one additional independent variable (the interaction variable):

```
REGRESSION
/DESCRIPTIVES MEAN STDDEV CORR SIG N
/MISSING LISTWISE
/STATISTICS COEFF OUTS CI(95) R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT IncomeInequality
/METHOD=ENTER PARTYID RIncomeK PartyIDRIncomeKInteraction.
```

Menu:
Run your regression the same way you did for multivariate regression in Chapter 14, but include both the main effect/parent and interaction variables in your "Independent(s):" box.

Here is the coefficients output:

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	.673	.049		13.802	<.001	.577	.768
	Party identification	.486	.025	.440	19.169	<.001	.437	.536
	Respondent's estimated income (in thousands)	-.004	.001	-.105	-5.580	<.001	-.005	-.002
	PartyIDRIncomeKInteraction	.001	.000	.066	2.885	.004	.000	.002

a. Dependent Variable: Should the government reduce income differences?

If you were to write a linear regression equation, it would be:
y=0.67 + 0.49P + -0.004I + 0.001(interaction)
The interaction is party multiplied by income, so if you wanted to make predictions of the dependent variable based on your predictors, you could also write it: y=0.67 + 0.49(party) + -0.004(income) + 0.001(party)(income)
Notice that party and income are in the equation twice. That's why you can't interpret one of the slopes or variables without considering the others.

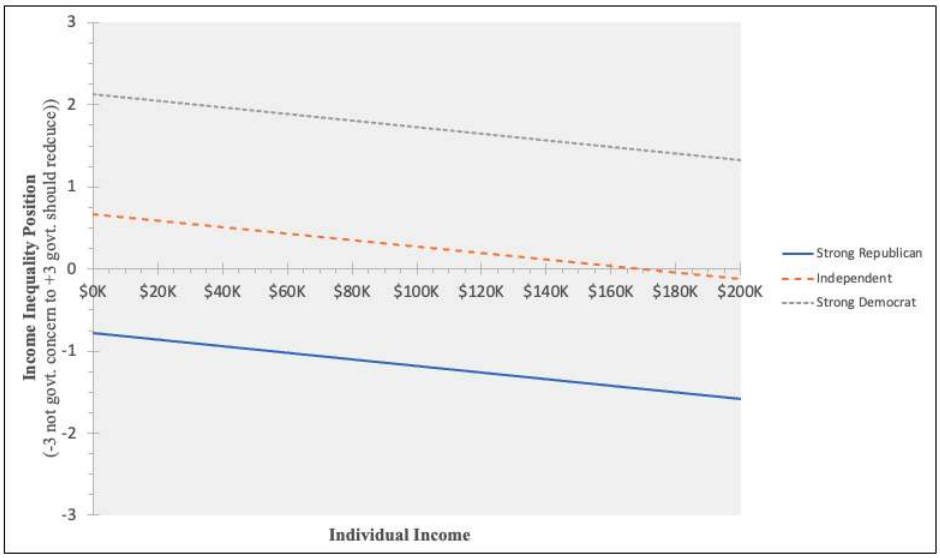
Here's a summary display table with bivariate models, a multivariate model with both IVs, and a multivariate model that also includes the interaction. The interaction gets its own row as its own independent variable.

Table: Estimated unstandardized coefficients from weighted OLS regressions predicting feelings on income inequality (-3 not govt. concern to +3 govt. should reduce), with 95% confidence intervals

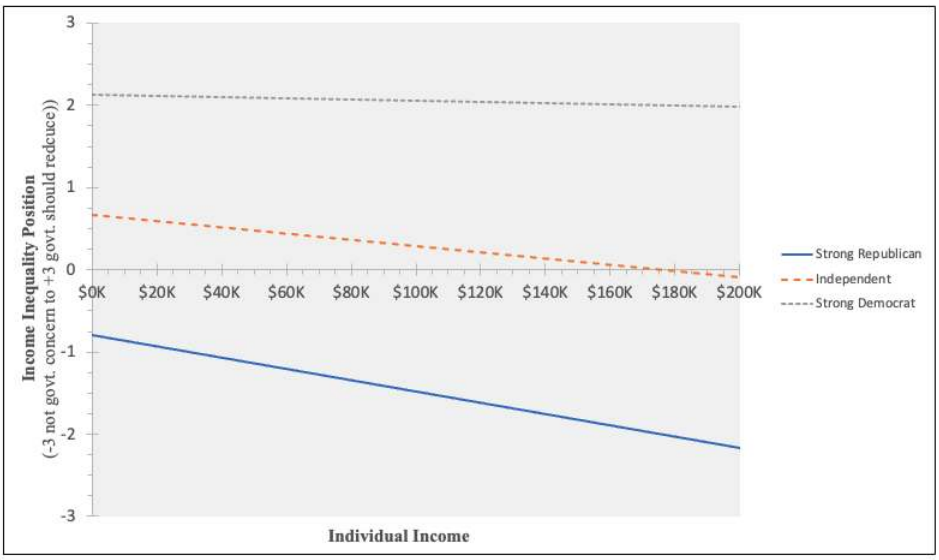
Independent Variables	PartyID***	Income***	Additive Model***	Interaction Model***
Intercept	0.53*** (0.45, 0.60)	0.72*** (0.61, 0.83)	0.67*** (0.58, 0.77)	0.67*** (0.58, 0.77)
PartyID (-3 strong GOP to +3 strong Dem)	0.53*** (0.49, 0.57))		0.53*** (0.49, 0.57)	0.49*** (0.44, 0.54)
Income (individual, in thousands)		-0.004*** (-0.006, -0.003)	-0.004*** (-0.005, -0.003)	-0.004*** (-0.005, -0.002)
Interaction (PartyID*Income)				0.0010** (0.0003, 0.0017)
r^2	22.7%	1.3%	24.1%	24.4%
n	2,330	2,193	2,177	2,177

~p<0.10, *p<0.05, **p<0.01, ***p<0.001
Sample size is unweighted, all other analyses probability-weighted
Source: GSS 2022

Here is a graph of the multivariate additive model (just the two IVs without the interaction). Notice how the lines are all parallel. The effect of income is the same across party, and the differences for party are the same across income levels:



Here's a graph of the interaction model. What do you notice? What's the story?



Democrats are supportive of the government tackling income inequality. Independents start out slightly supportive and by high income levels are neutral to a smidgen unsupportive. Republicans are not supportive, but low-income Republicans are only slightly unsupportive, whereas high-income Republicans are very unsupportive.

You can see that, as income increases, support for government reducing income inequality decreases ---- among independents and among Republicans. However, among strong Democrats, their support does not change much regardless of income.

The template for making the interaction graph looks pretty similar to the last one (though the formulas embedded are a little different). There is just one more row for your interaction slope.

Replace cells B3, B4, B5, & B6 with the intercept and slopes from your regression line equation.

intercept (constant)	0.673
IV 1 slope (x-axis)	-0.004
IV 2 slope (lines)	0.486
Interaction slope	0.001

Replace cells B13 and B14 with the minimum and maximum value for your independent variable that is going on the x-axis.
Replace cells C11, D11, and E11 with the labels of the values you chose for your other independent variable.
Replace cells C12, D13, and E12 with the values (numbers/codes) that correspond with the labels of the values you chose for your other independent variable.

	IV for lines:	Strong Republican	Independent	Strong Democrat
	x (IV) for x-axis	-3	0	3
x (IV for x-axis) minimum	0	-0.786179	0.67267	2.131519
x (IV for x-axis) maximum	200	-2.168079	-0.09333	1.981419

You can download the Excel spreadsheet with worksheets for making the summary table, using a predicted value calculator, and making your graph here. For graphing, make sure to use the worksheets that end in .Interaction or .Int.

Possible Interaction Effects

Interactions can have varying effects. If you graph two lines, you might see:

- Two parallel lines (no interaction effect)
- Two lines that both have positive (or negative) slopes, but one line has a steeper slope than another.
- One line with a negative slope and one line with a positive slope.
- One line has a positive or negative slope (there is a relationship with the DV), but the other line is horizontal (with a slope of zero, no relationship).

Statistics in Action: Evaluating Trump's handling of immigration

Last chapter we looked at U.S. eligible voters' evaluations of Trump's handling of immigration based on their policy positions on immigration and feelings towards Trump, using ANES 2020 pre-election data.

Here it is again, this time with an interaction between policy positions on immigration and their feelings towards Trump

If you look at the coefficients output, you can see that **the interaction is not significant** (or is marginally significant, p=0.075, so <0.1 but not <0.05).

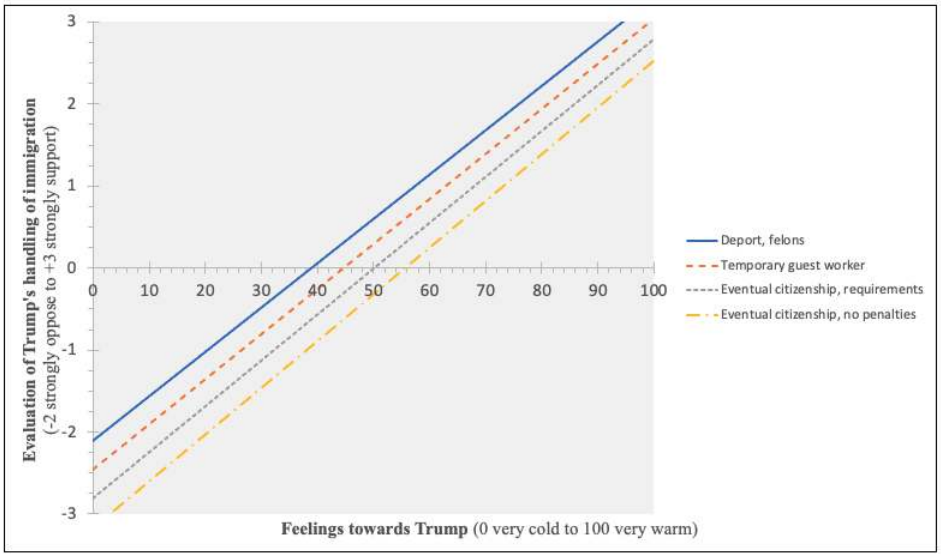
We are not confident that among all U.S. eligible voters there is a significant interaction. We know these other two things matter, even while controlling for the other, but we are not confident that this intersection matters (e.g., the main effects don't differ among subgroups, e.g., on average, among U.S. eligible voters leading up to the 2020 election, immigration policy views have the same effect while controlling for feelings towards Trump among people with different feelings towards Trump, and feelings towards Trump has the same effect while controlling for policy positions even among people with different policy positions). We don't need to consider the interaction for this model and could just use the multivariate model from Chapter 14 without losing explanatory/prediction power.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	-2.105	.066		-31.796	<.001	-2.235	-1.975
	Feeling Thermometer: Donald Trump, Republican Presidential candidate	.054	.001	.799	61.501	<.001	.053	.056
	Favored U.S. govt. policy towards unauthorized immigrants	-.356	.030	-.115	-11.934	<.001	-.414	-.297
	Interaction	.001	.000	.021	1.782	.075	.000	.002

a. Dependent Variable: Evaluation of the way Donald Trump is handling immigration

Here's the graph for respondents. You can see that the lines are almost parallel. The graph looks very similar to the graph from Chapter 14 when there was no interaction term included.



This page titled 15: Interactions is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Ezra Temko (Consortium of Academic and Research Libraries in Illinois (CARLI)) .

16: Appendix- Resources

Summary Display Table Templates

- Chapter 12: OLS Linear Regression
- Chapter 14: Display Table Template for Multivariate Regression with multiple models

Graphing and predicted value templates - currently available by chapter, check back for appendix

- Chapter 11: Bivariate Linear Regression
- Chapter 12: Bivariate Linear Regression (including an indicator IV bar graph)
- Chapter 14: Predicted value calculator for multivariate linear regression
- Chapter 14: Multivariate linear regression graph templates
- Chapter 15: Graph templates including linear regression with interactions

Interpretation Templates: **Templates for Contextual Interpretation.docx**

Glossary is available in "Back Matter" of textbook.

Reading Notes Template: This is a document with a Cornell-style template and cues for taking notes on each chapter of the textbook.

Syntax File (annotated)

Datasets

ANES 2020 post-election:
This dataset contains many variables from the ANES 2020 post-election study, including variables re-coded for use as indicator variables, reference group sets, and interactions.
Here is the dataset: ANES2020PE.Temko2024.Dataset.sav
Here is the codebook: ANES2020PE.Temko2024.Codebook.docx

GSS 2022:
This dataset contains a number of variables from the 2022 General Social Survey.
Here is the dataset: GSS.2022.SPSS.Temko2024.sav
Here is the corresponding codebook: GSS2022.Temko2024Codebook.docx

Sample Lab Assignments

Independent labs focus on one unique feelings thermometer variable. This is from a survey question where respondents were asked about how warm/cold they felt towards a particular person or organization on a scale of 0 to 100. I assigned students (or student pairs) each their own feelings thermometer variable to focus on for the semester for their independent labs. They take a groupwork interest survey that I use to make these assignments. The demo labs here mirror the independent labs but use the General Social Survey dataset with different variables. There is an assignment and a video of me completing the demonstration lab for each independent lab.

Chapter 1: Data, Datasets, SPSS, and Frequency Distribution Tables
This assignment focuses on getting SPSS installed and successfully working.
Demo Lab: Assignment, Video
Independent lab: Assignment

Chapter 2: Levels of Measurement; Five Point Summaries
This assignment focuses on running and interpreting a five-point summary and box plot.
Demo Lab: Assignment, Syntax file, Video
Independent lab: Assignment, Syntax file

Chapter 3: More Measures of Central Tendency and Spread- Mean, Standard Deviation, Z-scores
This assignment focuses on running and interpreting mean and standard deviation.

Demo Lab: Assignment, Syntax file, Video
Independent lab: Assignment, Syntax file

Chapter 4: Histograms, Normal Distributions, and the Central Limit Theorem
This assignment focuses on creating a histogram, interpreting its distribution, and understanding inferential statistics.
Demo Lab: Assignment, Syntax file, Video
Independent lab: Assignment, Syntax file

Chapter 5: Confidence Intervals
This assignment focuses on creating and interpreting confidence intervals.
Demo Lab: Assignment, Syntax file, Video
Independent lab: Assignment, Syntax file

Chapter 6: Comparing Confidence Intervals
This assignment focuses on comparing confidence intervals.
Demo Lab: Assignment, Syntax file, Video
Independent lab: Assignment, Syntax file

Chapter 7: Hypothesis Tests
This assignment focuses on two-sample t-tests and one-way ANOVA.
Demo Lab: Assignment, Syntax file, Two-sample t-test video, One-way ANOVA video
Independent lab: Assignment, Syntax file

Chapter 8: Crosstabs and Chi-square
This assignment focuses on crosstabs and chi-square.
Demo Lab: Assignment, Syntax file, Video
Independent lab: Assignment, Syntax file

Chapter 9: Elaborated Crosstabs
This assignment focuses on elaborated crosstabs.
Demo Lab: Assignment, Syntax file, Video
Independent lab: Assignment, Syntax file

Chapter 10: Scatterplots and Correlation
This assignment focuses on scatterplots and correlation.
Demo Lab: Assignment, WVS 2014 country-level dataset and codebook (from Linneman textbook student resources), Syntax file, Video
Independent lab: Assignment, IL County Data 2024 dataset and variables list, Syntax file

Chapter 11: Regression Equations
This assignment focuses on regression equations.
Demo Lab: Assignment, Syntax file, Video, Excel spreadsheet
Independent lab: Assignment, Syntax file, Excel spreadsheet

Chapter 12: Linear Regression, continued
This assignment focuses on bivariate OLS linear regression with a ratio-like independent variable and an indicator independent variable.
Demo Lab: Assignment, Syntax file, Video, Excel spreadsheet for graphing
Independent lab: Assignment, Syntax file, Excel spreadsheet for graphing

Chapter 13: Linear Regression with Reference Group Set Independent Variables
This assignment focuses on bivariate OLS linear regression with a reference group set for the independent variable(s)
Demo Lab: Assignment, Syntax file, Video
Independent lab: Assignment, Syntax file

Chapter 14: Multivariate Linear Regression
This assignment focuses on multivariate linear regression.

Demo Lab: Assignment, Syntax file, Video
Independent lab: Assignment, Syntax file

Chapter 15: Interactions

This assignment focuses on interactions in linear regression.
Demo Lab: Assignment, Syntax file, Graphing Template, Video
Independent lab: Assignment, Syntax file

Practice Quizzes and Quizzes

If you are an instructor using this text and would like these resources, please reach out to the author.
If you are a student using this text for self-study and would like access to practice quizzes, please reach out to the author.

This page titled 16: Appendix- Resources is shared under a CC BY-NC-SA license and was authored, remixed, and/or curated by Ezra Temko (Consortium of Academic and Research Libraries in Illinois (CARLI)) .

Index

A

accessible education
- InfoPage

alternative hypothesis
- 7: Hypothesis Tests, T-test, & ANOVA

ANOVA
- 13: Linear Regression with Reference Group Set Independent Variables
- Glossary

B

bivariate linear regression
- 16: Appendix- Resources

bivariate OLS linear regression
- 12: Linear Regression, continued

bivariate regression
- 11: Regression Equations

bivariate relationships
- 10: Chapter 10. Scatterplots and Correlation

C

Causation
- Glossary

Central Limit Theorem
- 4: Histograms, Normal Distributions, and the Central Limit Theorem
- Glossary

central tendency
- 2: Levels of Measurement; Five Point Summaries
- 3: More Measures of Central Tendency and Spread- Mean, Standard Deviation, Z-scores
- 4: Histograms, Normal Distributions, and the Central Limit Theorem
- 16: Appendix- Resources

collaborative learning
- InfoPage

Confidence interval
- 13: Linear Regression with Reference Group Set Independent Variables
- Glossary

confidence intervals
- 4: Histograms, Normal Distributions, and the Central Limit Theorem
- 5: Confidence Intervals
- 6: Comparing Confidence Intervals
- 7: Hypothesis Tests, T-test, & ANOVA
- 16: Appendix- Resources

confidence level
- 5: Confidence Intervals

control variable
- 9: Elaborated Crosstabs

correlation
- 10: Chapter 10. Scatterplots and Correlation

Cronbach's alpha
- Glossary

crosstabs
- 9: Elaborated Crosstabs

crosstabulation
- 8: Crosstabs and Chi-square

customizable textbooks
- InfoPage

D

data
- 1: Data, Datasets, SPSS, and Frequency Distribution Tables

data analysis
- 1: Data, Datasets, SPSS, and Frequency Distribution Tables
- 3: More Measures of Central Tendency and Spread- Mean, Standard Deviation, Z-scores
- 10: Chapter 10. Scatterplots and Correlation
- 11: Regression Equations
- 12: Linear Regression, continued
- 13: Linear Regression with Reference Group Set Independent Variables
- 14: Multivariate Linear Regression
- 15: Interactions
- 16: Appendix- Resources

data collection
- 1: Data, Datasets, SPSS, and Frequency Distribution Tables
- 8.1: Crosstabs for Research Methods

data distribution
- 3: More Measures of Central Tendency and Spread- Mean, Standard Deviation, Z-scores

data interpretation
- 5: Confidence Intervals

data visualization
- 4: Histograms, Normal Distributions, and the Central Limit Theorem

demographics
- 1: Data, Datasets, SPSS, and Frequency Distribution Tables

dependent variable
- 10: Chapter 10. Scatterplots and Correlation
- 11: Regression Equations
- 12: Linear Regression, continued
- 14: Multivariate Linear Regression

descriptive statistics
- 2: Levels of Measurement; Five Point Summaries
- 4: Histograms, Normal Distributions, and the Central Limit Theorem
- 8: Crosstabs and Chi-square

discrimination
- 15: Interactions

E

econometrics
- 13: Linear Regression with Reference Group Set Independent Variables

education
- 9: Elaborated Crosstabs

educational disparities
- 8: Crosstabs and Chi-square

educational statistics
- 1: Data, Datasets, SPSS, and Frequency Distribution Tables

empirical analysis
- 15: Interactions

Excel
- 11: Regression Equations

F

Film Industry
- 8: Crosstabs and Chi-square
- 8.1: Crosstabs for Research Methods

G

gender representation
- 8: Crosstabs and Chi-square
- 8.1: Crosstabs for Research Methods

government action
- 15: Interactions

H

health and wealth
- 10: Chapter 10. Scatterplots and Correlation

higher education
- InfoPage

histograms
- 4: Histograms, Normal Distributions, and the Central Limit Theorem

household income
- 9: Elaborated Crosstabs

Hypothesis test
- Glossary

hypothesis testing
- 7: Hypothesis Tests, T-test, & ANOVA
- 11: Regression Equations
- 12: Linear Regression, continued

hypothesis tests
- 16: Appendix- Resources

I

identities
- 15: Interactions

Income Inequality
- 15: Interactions

independent variable
- 10: Chapter 10. Scatterplots and Correlation
- 11: Regression Equations

independent variables
- 14: Multivariate Linear Regression

inferential claims
- 6: Comparing Confidence Intervals

inferential statistics
- 4: Histograms, Normal Distributions, and the Central Limit Theorem
- 5: Confidence Intervals
- 8: Crosstabs and Chi-square

interactive learning
- InfoPage

interpretation
- 12: Linear Regression, continued

intersectionality
- 15: Interactions

L

levels of measurement
- 16: Appendix- Resources

LibreTexts
- InfoPage

linear regression
- 10: Chapter 10. Scatterplots and Correlation
- 11: Regression Equations
- 14: Multivariate Linear Regression

M

margin of error
- 6: Comparing Confidence Intervals

mean
- 2: Levels of Measurement; Five Point Summaries
- 3: More Measures of Central Tendency and Spread- Mean, Standard Deviation, Z-scores

measures of spread
- 2: Levels of Measurement; Five Point Summaries

median
- 2: Levels of Measurement; Five Point Summaries
- 3: More Measures of Central Tendency and Spread- Mean, Standard Deviation, Z-scores

multivariate
- 12: Linear Regression, continued

multivariate analysis
- 9: Elaborated Crosstabs

multivariate linear regression
- 16: Appendix- Resources

Multivariate regression
- 13: Linear Regression with Reference Group Set Independent Variables
- 14: Multivariate Linear Regression
- Glossary

N

nominal variables
- 2: Levels of Measurement; Five Point Summaries

Normal distribution
- Glossary

normal distributions
- 4: Histograms, Normal Distributions, and the Central Limit Theorem

null hypothesis
- 7: Hypothesis Tests, T-test, & ANOVA

O

OER
- InfoPage

open education
- InfoPage

ordinal variables
- 2: Levels of Measurement; Five Point Summaries

outliers
- 3: More Measures of Central Tendency and Spread- Mean, Standard Deviation, Z-scores

P

political party
- 15: Interactions

political party affiliation
- 14: Multivariate Linear Regression

political polls
- 6: Comparing Confidence Intervals

polling
- 5: Confidence Intervals

polling data
- 6: Comparing Confidence Intervals

population mean
- 5: Confidence Intervals

predictive modeling
- 13: Linear Regression with Reference Group Set Independent Variables

predictor variable
- 12: Linear Regression, continued

probability
- 4: Histograms, Normal Distributions, and the Central Limit Theorem

public opinion research
- 6: Comparing Confidence Intervals

Q

quantitative analysis
- 1: Data, Datasets, SPSS, and Frequency Distribution Tables
- 2: Levels of Measurement; Five Point Summaries

R

race
- 9: Elaborated Crosstabs

Racial Representation
- 8.1: Crosstabs for Research Methods

ratio variables
- 2: Levels of Measurement; Five Point Summaries

regression
- 13: Linear Regression with Reference Group Set Independent Variables

regression analysis
- 14: Multivariate Linear Regression

regression equation
- 11: Regression Equations

regression equations
- 16: Appendix- Resources

regression line
- 10: Chapter 10. Scatterplots and Correlation

regression output
- 12: Linear Regression, continued

representation
- 8: Crosstabs and Chi-square

research methods
- 1: Data, Datasets, SPSS, and Frequency Distribution Tables
- 8.1: Crosstabs for Research Methods

S

Sample
- Glossary

sample data
- 5: Confidence Intervals

sample proportions
- 6: Comparing Confidence Intervals

sampling distribution
- 7: Hypothesis Tests, T-test, & ANOVA

sampling error
- 4: Histograms, Normal Distributions, and the Central Limit Theorem

scatterplot
- 10: Chapter 10. Scatterplots and Correlation

social identities
- 15: Interactions

social sciences
- 1: Data, Datasets, SPSS, and Frequency Distribution Tables

social world
- 15: Interactions

socioeconomic status
- 9: Elaborated Crosstabs

SPSS
- 11: Regression Equations
- 16: Appendix- Resources

standard deviation
- 2: Levels of Measurement; Five Point Summaries
- 3: More Measures of Central Tendency and Spread- Mean, Standard Deviation, Z-scores

statistical analysis
- 5: Confidence Intervals
- 8: Crosstabs and Chi-square
- 8.1: Crosstabs for Research Methods
- 11: Regression Equations
- 12: Linear Regression, continued
- 14: Multivariate Linear Regression

statistical significance
- 6: Comparing Confidence Intervals
- 7: Hypothesis Tests, T-test, & ANOVA
- 12: Linear Regression, continued
- 13: Linear Regression with Reference Group Set Independent Variables

statistical software
- 1: Data, Datasets, SPSS, and Frequency Distribution Tables

statistical test
- 7: Hypothesis Tests, T-test, & ANOVA

statistics
- 1: Data, Datasets, SPSS, and Frequency Distribution Tables
- 3: More Measures of Central Tendency and Spread- Mean, Standard Deviation, Z-scores
- 5: Confidence Intervals
- 8.1: Crosstabs for Research Methods
- 9: Elaborated Crosstabs
- 13: Linear Regression with Reference Group Set Independent Variables

Studio Analysis
- 8.1: Crosstabs for Research Methods

summary statistics
- 3: More Measures of Central Tendency and Spread- Mean, Standard Deviation, Z-scores

survey methods
- 6: Comparing Confidence Intervals

survey research
- 5: Confidence Intervals

U

U.S. eligible voters
- 14: Multivariate Linear Regression

U.S. states
- 10: Chapter 10. Scatterplots and Correlation

U.S. students
- 9: Elaborated Crosstabs

U.S. voters
- 6: Comparing Confidence Intervals

USC Annenberg Inclusion Initiative
- 8: Crosstabs and Chi-square

V

Variable
- Glossary

variance
- 7: Hypothesis Tests, T-test, & ANOVA

variation
- 3: More Measures of Central Tendency and Spread- Mean, Standard Deviation, Z-scores

voting behavior
- 14: Multivariate Linear Regression

W
Women in Film
8.1: Crosstabs for Research Methods

Glossary

ANOVA, one-way: A statistical test used to evaluate claims about whether three or more population means are different.

Beta (or standardized) slope: Slopes with standard deviations as their unit.

Causality / Causation: There is a causal relationship when one variable has an impact or effect on variation of another variable. This is evaluated through significance, substantiveness, causal order, theory, and nonspuriousness.

Central Limit Theorem: Given a random sample with enough cases, the distribution of sample means will be normally distributed.

Cronbach's alpha (α): A common measure of index reliability (from 0 unreliable to 1 perfect reliability, with 0.7 or higher being reliable enough).

Codebook: A document that tells you the codes for each variable in your dataset and what attributes they represent

Codes: Numbers that represent variable attributes

Coefficient of determination (r^2): The proportion of the dependent variable's variance explained by the independent variable.

Confidence interval: The values within which we are claiming a population mean or proportion falls, at a particular confidence level, based on a representative sample of that population.

Control variable: An additional variable added to a model and held constant in order to improve our evaluation of relationships with the dependent variable.

Correlation: A measure of the strength and direction of linear relationships.

Crosstabulation: A crosstab is a table that shows subgroup distribution of data, enabling examination of potential relationships between variables.

> **Partial (or elaborated) crosstabulation:** Crosstabs that add a control variable to the original bivariate crosstab, producing separate bivariate crosstabs for each subgroup.

Data: Information. In statistics, usually the values for variable(s).
Data is a plural word. The singular is datum.

Dataset: A collection of data organized for analysis

Histogram: A frequency distribution graph for numeric data.

Hypothesis: An inferential claim about a population.

> **Alternative hypothesis (H_a):** There is a difference/relationship (among the target population).
>
> **Null hypothesis ($H_\varnothing$)**: There is no difference/relationship (among the target population).

Hypothesis test: A statistical procedure that tests an inferential claim about a population using representative sample data. Also known as "statistical test."

Index: An index is when multiple variables are combined together into one construct.

Inferential statistics: Statistical procedures where we make claims about a population based on corresponding sample data.

Interaction: A statistical interaction occurs when the effects of one predictor variable varies based on another predictor variable.

Interquartile range (IQR): The distance spanning the middle 50% of data, the distance between the lower and upper quartile values.

Level of measurement: a variable characteristic classifying the relationship among a variable's values

> **Nominal**: Nominal variables have values that are qualitative categories. They do not have any meaningful rank or order.
>
> **Ordinal**: Ordinal variables have values that are qualitative categories. They do have a meaningful rank or order.
>
> **Ratio**: Ratio-level variables have values that are numeric. They have a true zero point, meaningful ratios, consistent intervals between them, and the numbers function arithmetically.
>
> **Interval**: Interval-level variables have equal intervals between values, but do not meet all the criteria to be classified as ratio-level variables. Some definitions of interval-level variables require interval-level variables to have numeric values.
>
> **Indicator:** Indicator variables, also called **dummy** variables, are dichotomous/binary, meaning they only have two categories, and these two values are coded as 0 and 1.

Linear relationship: A relationship between two variables that takes the form of a line when graphed.

Mean: The average value; the value each case would have if all cases' values were redistributed equally among all the cases.

Median: The middle number in a group of ordered numbers.

Mode: The repeated value that appears most frequently

Multivariate regression: Regression that uses multiple independent variables.

Negative relationship: As x (independent variable) increases, y (dependent variable) decreases.

Normal distribution: A bell-shaped symmetrical histogram with data concentrated in the middle and evenly tapering off to both sides.

One-sample t-test: A statistical test used to evaluate claims about whether a

population mean is different from a particular hypothesized value.

Outlier: An extreme value; a case with a value that substantively differs from most other values.

Percentile: The proportion of data that falls below a given value. The nth percentile is the value below which n% of observations fall.

Population: All individuals, groups, objects, or cases of interest.

Positive relationship: As x (independent variable) increases, y (dependent variable) increases.

Q_1, the lower quartile, is the median of the lower half of the data. It is the 25th percentile, greater than 1/4 or 25% of the data (and less than 3/4 or 75% of the data).

Q_3, the upper quartile, is the median of the upper half of the data. It is the 75th percentile, greater than 3/4 or 75% of the data (and less than 1/4 or 25% of the data).

Recoding: The process of changing a variable's number-attribute assignments so that it has new/different coding.

Range: The distance between the minimum and maximum values.

Reference group set: A group of indicator variables for one construct that as a collection represent its various categories. Each category, other than the comparison reference group, has an indicator variable where it is coded as 1.

Regression line (or least squares line, or line of best fit): The line that, when drawn, is closest to the observed data. This is measured by minimizing the sum of the squared distance of each point's y-coordinate from a potential line's corresponding y-coordinate.

P-value: The probability (likelihood) of getting a particular sample statistic as extreme (or more extreme than) the outcome we observed, based on what we would expect from the sampling distribution.

Sample: A subset of cases from the target population.

Sampling distribution: A distribution of sample means from the same population.

Sampling error: The difference between sample statistics and population parameters resulting from samples not being perfectly representative of the populations they are drawn from.

Scatterplot: A graph that shows pairs of quantitative data using a rectangular coordinate system.

Significant difference: If we are over 95% confident that two proportions or means are different, we label them as significantly different.

Standard deviation: An estimate of the average distance of cases' data values from their overall mean.

Standard error: The standard deviation for a sampling distribution. This is the average distance of sample means from the population mean.

Standard normal distribution: Also known as a z-distribution, this is a special normal distribution standardized by z-scores.

Statistical test: A statistical procedure that tests an inferential claim about a population using representative sample data. Also known as "hypothesis test."

Statistics: data, the practice of using data for research

Substantive difference: A difference that is big, sizable, notable.

Skewed distribution: An asymmetrical histogram.

Strength (for linear relationships): The extent to which the points converge into a common pattern, into a particular linear relationship.

T-score: How many standard errors a value (sample mean) is away from the population mean.

Two-sample t-test: A statistical test used to evaluate claims about whether two population means are different.

Unit of analysis: Who or what you are analyzing

Variable: Anything that can vary (differ), as opposed to something that must be constant (the same).

Other words you will come across that have a form of the word variable in them (variate) include:
univariate: having to do with one (uni) variable
bivariate: having to do with two (bi) variables
multivariate: having to do with multiple (multi) variables

Antecedent variables: Variables that precede (come before) both the independent and dependent variables.

Dependent Variable: The outcome; the variable potentially being impacted by the independent variable.

Independent Variable: The determinant; the variable potentially impacting the dependent variable.

Intervening variables: Variables that come in between the independent and dependent variables.

Weighting: A process of adjusting statistical calculations to compensate for sampling error.

Z-score: The number of standard deviations a value is from its mean.

Z-distribution: See "standard normal distribution."

Detailed Licensing

Overview

Title: The Stories Behind Social Statistics: Data Analysis, Interpretation, and Communication

Webpages: 27

Applicable Restrictions: Noncommercial

All licenses found:

- CC BY-NC-SA 4.0: 66.7% (18 pages)
- Undeclared: 33.3% (9 pages)

By Page

- The Stories Behind Social Statistics: Data Analysis, Interpretation, and Communication - *CC BY-NC-SA 4.0*
  - Front Matter - *Undeclared*
    - TitlePage - *Undeclared*
    - InfoPage - *Undeclared*
    - Table of Contents - *Undeclared*
    - Licensing - *Undeclared*
  - 1: Data, Datasets, SPSS, and Frequency Distribution Tables - *CC BY-NC-SA 4.0*
  - 2: Levels of Measurement; Five Point Summaries - *CC BY-NC-SA 4.0*
  - 3: More Measures of Central Tendency and Spread- Mean, Standard Deviation, Z-scores - *CC BY-NC-SA 4.0*
  - 4: Histograms, Normal Distributions, and the Central Limit Theorem - *CC BY-NC-SA 4.0*
  - 5: Confidence Intervals - *CC BY-NC-SA 4.0*
  - 6: Comparing Confidence Intervals - *CC BY-NC-SA 4.0*
  - 7: Hypothesis Tests, T-test, & ANOVA - *CC BY-NC-SA 4.0*
  - 8: Crosstabs and Chi-square - *CC BY-NC-SA 4.0*
    - 8.1: Crosstabs for Research Methods - *CC BY-NC-SA 4.0*
  - 9: Elaborated Crosstabs - *CC BY-NC-SA 4.0*
  - 10: Chapter 10. Scatterplots and Correlation - *CC BY-NC-SA 4.0*
  - 11: Regression Equations - *CC BY-NC-SA 4.0*
  - 12: Linear Regression, continued - *CC BY-NC-SA 4.0*
  - 13: Linear Regression with Reference Group Set Independent Variables - *CC BY-NC-SA 4.0*
  - 14: Multivariate Linear Regression - *CC BY-NC-SA 4.0*
  - 15: Interactions - *CC BY-NC-SA 4.0*
  - 16: Appendix- Resources - *CC BY-NC-SA 4.0*
  - Back Matter - *Undeclared*
    - Index - *Undeclared*
    - Glossary - *Undeclared*
    - Detailed Licensing - *Undeclared*