UCSD Political Science Math Camp

Table of contents

List of Figures

List of Tables

About this Booklet

This booklet is adapted from the Harvard Gov Prefresher booklet, maintained by Shiro Kuriwaki. It will serve as as the text for the UCSD Math Camp, taught by Bertrand Wilden and Keng-Chi Chang, and we have reordered and adapted it to fit the structure of our course. For information about the role of this math camp as an introduction to graduate school, you may be interested in "The Math Prefresher and The Collective Future of Political Science Graduate Training", in PS: Political Science & Politics, by Gary King, Shiro Kuriwaki, and Yon Soo Park.

Authors and Contributors

For information about the authors of the Harvard Gov Prefresher booklet, see here.
We have also updated it to include some material from previous instructors of the class at UCSD, including Rachel Schoner (2021), Luke Sanford (2019-2020), Kathryn Baragwanath (2019-2020), Brandon Merrell (2018), Inbok Rhee (2018).

Contributing

We transitioned the booklet into a bookdown github repository in 2021. As we update this version, we appreciate any bug reports or fixes appreciated.

All changes should be made in the `.Rmd` files in the project root. To contribute a change, please make a pull request and set the repository maintainer as the reviewer.

Warmup Questions

Before our first meeting, please try solving these questions. They are a sample of the very beginning of each math section. We have provided links to the parts of the book you can read if the concepts are new to you.

The goal of this “pre”-math camp assignment is not to intimidate you but to set common expectations so you can make the most out of the actual Math Camp. Even if you do not understand some or all of these questions after skimming through the linked sections, your effort will pay off, and you will be better prepared for the math camp. We are also open to adjusting these expectations based on feedback (this class is for *you*), so please do not hesitate to write to the instructors for feedback.

Operations

Summation

Simplify the following

1. $\sum\limits_{i=1}^{3} i$
2. $\sum\limits_{k=1}^{3} (3k+2)$
3. $\sum\limits_{i=1}^{4} (3k+i+2)$

Products

1. $\prod\limits_{i=1}^{3} i$
2. $\prod\limits_{k=1}^{3} (3k+2)$

Logs and exponents

Simplify the following:

1. 4^2
2. $4^2 2^3$
3. $\log_{10} 100$
4. $\log_2 4$
5. $\log e$, where log is the natural log (also written as ln) – a log with base$e$, and e is Euler’s constant
6. e^a, e^b, e^c, where a, b, c are each constants
7. $\log 0$

8. e^0
9. e^1
10. $\log e^2$

Limits

Find the limit of the following.

1. $\lim_{x\to 2}(x-1)$
2. $\lim_{x\to 2}\frac{(x-2)(x-1)}{(x-2)}$
3. $\lim_{x\to 2}\frac{x^2-3x+2}{x-2}$

Linear Algebra

Vectors

Define the vectors

$u = \begin{pmatrix}1\\2\\3\end{pmatrix}$,

$v = \begin{pmatrix}4\\5\\6\end{pmatrix}$,

and the scalar $c = 2$.

Calculate the following:

1. $u + v$
2. cv
3. $u \cdot v$

Are the following sets of vectors linearly independent?

1. $u = \begin{pmatrix}1\\2\end{pmatrix}$, $v = \begin{pmatrix}2\\4\end{pmatrix}$
2. $u = \begin{pmatrix}1\\2\\5\end{pmatrix}$, $v = \begin{pmatrix}3\\7\\9\end{pmatrix}$
3. $a = \begin{pmatrix}2\\-1\\1\end{pmatrix}$, $b = \begin{pmatrix}3\\-4\\-2\end{pmatrix}$, $c = \begin{pmatrix}5\\-10\\-8\end{pmatrix}$ (this requires some guesswork)

Matrices

Given that

$$\mathbf{A} = \begin{bmatrix} 7 & 5 & 1 \\ 11 & 9 & 3 \\ 2 & 14 & 21 \\ 4 & 1 & 5 \end{bmatrix}$$

What is the dimensionality of matrix $\mathbf{A}$?

What is the element a_{23} of $\mathbf{A}$?

Given that

$$\mathbf{B} = \begin{bmatrix} 1 & 2 & 8 \\ 3 & 9 & 11 \\ 4 & 7 & 5 \\ 5 & 1 & 9 \end{bmatrix}$$

What is $\mathbf{A} + \mathbf{B}$?

Given that

$$\mathbf{C} = \begin{bmatrix} 1 & 2 & 8 \\ 3 & 9 & 11 \\ 4 & 7 & 5 \end{bmatrix}$$

What is $\mathbf{A} + \mathbf{C}$?

Given that

$$c = 2$$

What is $c\ \mathbf{A}$?

Calculus

For each of the following functions $f(x)$, find the derivative $f'(x)$ or $\frac{d}{dx}f(x)$

1. $f(x) = c$
2. $f(x) = x$
3. $f(x) = x^2$
4. $f(x) = x^3$
5. $f(x) = 3x^2 + 2x^{1/3}$
6. $f(x) = (x^3)(2x^4)$

Optimization

For each of the followng functions $f(x)$, does a maximum and minimum exist in the domain $x \in \mathbf{R}$? If so, for what are those values and for which values of x?

1. $f(x) = x$
2. $f(x) = x^2$
3. $f(x) = -(x-2)^2$

If you are stuck, please try sketching out a picture of each of the functions.

Probability

1. If there are 12 cards, numbered 1 to 12, and 4 cards are chosen, how many distinct possible choices are there? (unordered, without replacement)
2. Let $A = \{1, 3, 5, 7, 8\}$ and $B = \{2, 4, 7, 8, 12, 13\}$. What is $A \cup B$? What is $A \cap B$? If A is a subset of the Sample Space $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, what is the complement A^C?
3. If we roll two fair dice, what is the probability that their sum would be 11?
4. If we roll two fair dice, what is the probability that their sum would be 12?

Prerequisites

The syllabus lists the assumed knowledge for Math Camp (Chapters 1-4 of Moore and Siegel). Below, you will find additional review material, examples, and sample exercises.

Operators

Addition (+), Subtraction (−), multiplication and division are basic operations of arithmetic − combining numbers. In statistics and calculus, we want to add a *sequence* of numbers that can be expressed as a pattern without needing to write down all its components. For example, how would we express the sum of all numbers from 1 to 100 without writing a hundred numbers?

For this we use the summation operator $\sum$ and the product operator $\prod$.

Summation

$$\sum_{i=1}^{100} x_i = x_1 + x_2 + x_3 + \cdots + x_{100}$$

The bottom of the $\sum$ symbol indicates an index (here, i), and its start value 1. At the top is where the index ends. The notion of "addition" is part of the $\sum$ symbol. The content to the right of the summation is the meat of what we add. While you can pick your favorite index, start, and end values, the content must also have the index.

Corollary 0.1.

- $\sum\limits_{i=1}^{n} cx_i = c\sum\limits_{i=1}^{n} x_i$
- $\sum\limits_{i=1}^{n}(x_i + y_i) = \sum\limits_{i=1}^{n} x_i + \sum\limits_{i=1}^{n} y_i$
- $\sum\limits_{i=1}^{n} c = nc$

Product

$$\prod_{i=1}^{n} x_i = x_1 x_2 x_3 \cdots x_n$$

Corollary 0.2.

- $\prod\limits_{i=1}^{n} cx_i = c^n \prod\limits_{i=1}^{n} x_i$
- $\prod\limits_{i=k}^{n} cx_i = c^{n-k+1} \prod\limits_{i=k}^{n} x_i$

- $\prod_{i=1}^{n}(x_i + y_i) = a\ total\ mess$
- $\prod_{i=1}^{n} c = c^n$

Factorials

Definition 0.1 (Factorials)**.**

$$x! = x \cdot (x-1) \cdot (x-2) \cdots (1)$$

Modulo

Tells you the remainder when you divide the first number by the second.

- $17 \mod 3 = 2$
- $100 \ \% \ 30 = 10$

Exercise 0.1 (Operators)**.** Let $x_1 = 4, x_2 = 3, x_3 = 7, x_4 = 11, x_5 = 2$. Find the following:

1. $\sum_{i=1}^{5} i$
2. $\prod_{i=1}^{5} i$
3. $14 \mod 4$
4. $4!$
5. $\sum_{i=1}^{3}(7)x_i$
6. $\sum_{i=1}^{5} 2$
7. $\prod_{i=3}^{5}(2)x_i$

Functions

A **function** is a mapping that relates members of one set to members of another set.

For instance, if you have two sets: set A and set B, a function f from A to B maps every value a in set A such that $f(a) \in B$.

- The set A is called the **domain** of function f
- The set B is called the **range** of function f

Functions can be **many-to-one**, where many values from set A produce a single output in set B, or they can be **one-to-one**, where each value in set A corresponds to a single value in set B.

A function by definition has a single function value for each element of its domain. This means, there cannot be "one-to-many" mapping.

Dimensionality

$\mathbb{R}^1$ is the set of all real numbers extending from $-\infty$ to $+\infty$ — i.e., the real number line. $\mathbb{R}^n$ is an n-dimensional space, where each of the n axes extends from $-\infty$ to $+\infty$.

- $\mathbb{R}^1$ is a one dimensional line.
- $\mathbb{R}^2$ is a two dimensional plane.
- $\mathbb{R}^3$ is a three dimensional space.

Points in $\mathbb{R}^n$ are ordered n-tuples (just means an combination of n elements where order matters), where each element of the n-tuple represents the coordinate along that dimension.

For example:

- $\mathbb{R}^1$: (3)
- $\mathbb{R}^2$: $(-15, 5)$
- $\mathbb{R}^3$: $(86, 4, 0)$

! Notation (Functions)

Function of one variable:

$$f : \mathbb{R}^1 \to \mathbb{R}^1$$

- Example: $f(x) = x + 1$
  - For each x in $\mathbb{R}^1$, $f(x)$ assigns the number $x + 1$

Function of two variables:

$$f : \mathbb{R}^2 \to \mathbb{R}^1$$

- Example: $f(x, y) = x^2 + y^2$
  - For each ordered pair (x, y) in $\mathbb{R}^2$, $f(x, y)$ assigns the number $x^2 + y^2$

We often use variable x as input and another y as output, e.g. $y = x + 1$

Exercise 0.2 (Functions)**.** For each of the following, state whether they are one-to-one or many-to-one functions.

1. For $x \in [0, \infty]$, $f : x \to x^2$ (this could also be written as $f(x) = x^2$).
2. For $x \in [-\infty, \infty]$, $f : x \to x^2$.
3. For $x \in [-3, \infty]$, $f : x \to x^2$.
4. For $x \in [0, \infty]$, $f : x \to \sqrt{x}$

Some functions are defined only on proper subsets of $\mathbb{R}^n$.

- **Domain**: the set of numbers in X at which $f(x)$ is defined.
- **Range**: elements of Y assigned by $f(x)$ to elements of X, or we can use the notation $f(X)$ to denote the range, where
$$f(X) = \{y : y = f(x), x \in X\}$$

Types of Functions

Monomials: $f(x) = ax^k$

a is the coefficient. k is the degree.

Examples: $y = x^2$, $y = -\frac{1}{2}x^3$

Polynomials: sum of monomials.

Examples: $y = -\frac{1}{2}x^3 + x^2$, $y = 3x + 5$

The degree of a polynomial is the highest degree of its monomial terms. Also, it's often a good idea to write polynomials with terms in decreasing degree.

Exponential Functions: Example: $y = 2^x$

Logs and Exponents

Relationship of logarithmic and exponential functions:

$$y = \log_a(x) \iff a^y = x$$

The log function can be thought of as an inverse for exponential functions. a is referred to as the "base" of the logarithm.

Common Bases: The two most common logarithms are base 10 and base e.

1. Base 10: $y = \log_{10}(x) \iff 10^y = x$. The base 10 logarithm is often simply written as "$\log(x)$" with no base denoted.
2. Base e: $y = \log_e(x) \iff e^y = x$. The base e logarithm is referred to as the "natural" logarithm and is written as "$\ln(x)$".

Properties of exponential functions:

- $a^x a^y = a^{x+y}$
- $a^{-x} = 1/a^x$
- $a^x/a^y = a^{x-y}$
- $(a^x)^y = a^{xy}$
- $a^0 = 1$

Properties of logarithmic functions (any base):

Generally, when statisticians or social scientists write $\log(x)$ they mean $\log_e(x)$. In other words: $\log_e(x) \equiv \ln(x) \equiv \log(x)$

$$\log_a(a^x) = x$$

and

$$a^{\log_a(x)} = x$$

- $\log(xy) = \log(x) + \log(y)$
- $\log(x^y) = y\log(x)$
- $\log(1/x) = \log(x^{-1}) = -\log(x)$
- $\log(x/y) = \log(x \cdot y^{-1}) = \log(x) + \log(y^{-1}) = \log(x) - \log(y)$
- $\log(1) = \log(e^0) = 0$

Change of Base Formula: Use the change of base formula to switch bases as necessary:

$$\log_b(x) = \frac{\log_a(x)}{\log_a(b)}$$

Example:

$$\log_{10}(x) = \frac{\ln(x)}{\ln(10)}$$

You can use logs to go between sum and product notation. This will be particularly important when you're learning maximum likelihood estimation.

$$\begin{aligned}
\log\left(\prod_{i=1}^{n} x_i\right) &= \log(x_1 \cdot x_2 \cdot x_3 \cdots \cdot x_n) \\
&= \log(x_1) + \log(x_2) + \log(x_3) + \cdots + \log(x_n) \\
&= \sum_{i=1}^{n} \log(x_i)
\end{aligned}$$

Therefore, you can see that the log of a product is equal to the sum of the logs. We can write this more generally by adding in a constant, c:

$$\begin{aligned}
\log\left(\prod_{i=1}^{n} cx_i\right) &= \log(cx_1 \cdot cx_2 \cdots cx_n) \\
&= \log(c^n \cdot x_1 \cdot x_2 \cdots x_n) \\
&= \log(c^n) + \log(x_1) + \log(x_2) + \cdots + \log(x_n) \\
&= n\log(c) + \sum_{i=1}^{n} \log(x_i)
\end{aligned}$$

Exercise 0.3 (Log)**.** Evaluate each of the following logarithms

1. $\log_4(16)$
2. $\log_2(16)$
3. $\log_{\frac{3}{2}}(\frac{27}{8})$

Simplify the following logarithm. By "simplify", we actually really mean - use as many of the logarithmic properties as you can.

4. $\log_4(x^3y^5)$
5. $\log(\frac{x^9y^5}{z^3})$
6. $\ln\sqrt{xy}$

Graphing Functions

What can a graph tell you about a function?

- Is the function increasing or decreasing? Over what part of the domain?
- How "fast" does it increase or decrease?
- Are there global or local maxima and minima? Where?
- Are there inflection points?
- Is the function continuous?
- Is the function differentiable?
- Does the function tend to some limit?
- Other questions related to the substance of the problem at hand.

Solving for Variables

Sometimes we're given a function $y = f(x)$ and we want to find how x varies as a function of y. Use algebra to move x to the left hand side (LHS) of the equation and so that the right hand side (RHS) is only a function of y.

Example 0.1 (Solving for Variables)**.** Solve for x:

1. $y = 3x + 2$
2. $y = e^x$

Solving for variables is especially important when we want to find the **roots** of an equation: those values of variables that cause an equation to equal zero. Especially important in finding equilibria and in doing maximum likelihood estimation.

Procedure: Given $y = f(x)$, set $f(x) = 0$. Solve for x.

Multiple Roots:

$$f(x) = x^2 - 9 \quad \Longrightarrow \quad 0 = x^2 - 9 \quad \Longrightarrow \quad 9 = x^2 \quad \Longrightarrow \quad \pm\sqrt{9} = \sqrt{x^2} \quad \Longrightarrow \quad \pm 3 = x$$

Quadratic Formula: For quadratic equations $ax^2 + bx + c = 0$, use the quadratic formula:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Exercise 0.4 (Solving for Variables)**.** Solve for x:

1. $f(x) = 3x + 2 = 0$
2. $f(x) = x^2 + 3x - 4 = 0$
3. $f(x) = e^{-x} - 10 = 0$

Answers to Examples and Exercises

Answer to Exercise **??**:

1. $1 + 2 + 3 + 4 + 5 = 15$
2. $1 * 2 * 3 * 4 * 5 = 120$
3. 2
4. $4 * 3 * 2 * 1 = 24$
5. $7(4 + 3 + 7) = 98$
6. $2 + 2 + 2 + 2 + 2 = 10$
7. $2^3(7)(11)(2) = 1232$

Answer to Exercise **??**:

1. one-to-one
2. many-to-one
3. many-to-one
4. one-to-one

Answer to Exercise **??**:

1. 2
2. 4
3. 3
4. $3\log_4(x) + 5\log_4(y)$
5. $9\log(x) + 5\log(y) - 3\log(z)$
6. $\frac{1}{2}(\ln x + \ln y)$

Answer to Example **??**:

1. $y = 3x + 2 \quad \Longrightarrow \quad -3x = 2 - y \quad \Longrightarrow \quad 3x = y - 2 \quad \Longrightarrow \quad x = \frac{1}{3}(y-2)$
2. $x = \ln y$

Answer to Exercise **??**:

1. $\frac{-2}{3}$
2. x = {1, -4}
3. x = - $\ln 10$

Part I

I Introduction to R

Chapter 1

Orientation and Reading in Data

Welcome to Math Camp! The first day we will get to know each other and get acquainted with R. We will be working with R throughout this course, and today is the beginning, where you will learn about this program. It can be overwhelming to learn any programming language, but we will spend a lot of time and practice with it. Please work through the following tutorials and exercises. There is a lot of information here, and it will take time to understand everything. In fact, programming is a continual learning experience. Please come to class with areas of confusion and questions; we will walk through more examples together.

You should download the latest versions of R and R Studio.

R Studio Cloud has great resources we will be using throughout this course. Please complete the following primer tutorial for the first session: The Basics: Programming Basics

Where are we starting?

Today we'll cover:

- What's what in RStudio
- What R is, at a high level
- How to read in data
- How to create cool graphs using the ggplot2 package
- Best practices for reproducible workflow using R Markdown

1.1 Orienting

1.1.1 Getting to know RStudio

RStudio is an Integrated Development Environment for the programming language R. An IDE is a piece of software which allows you to more easily interface with programming languages. You can program with other languages (such as Python or SQL) using RStudio, and you can use other IDEs to program with R, but typically R users use RStudio—and vice versa. This is because RStudio was specifically created to facilitate R programming and comes with a ton of helpful features.

There are four main sections, or "panes", in RStudio.

1. **Bottom Left**. This is the *Console*. It is the place that executes your R code. You can type directly in your Console and it will display the output immediately below. If your code generates errors or warning

messages they will be displayed here too. You should rarely type code directly into your Console, however, because it can be difficult to keep track of things line by line. Also, every time you restart RStudio your Console will refresh itself which means you will lose your previous work! This is not only a problem for your own analysis, but it also makes it impossible for anyone else to replicate your steps. I only write in the Console directly if I'm testing small code snippets or for quick data exploration (for example using the `summary()` function on a variable).

Useful tip: use the Up-Arrow on your keyboard to quickly re-run pieces of code which were previously sent to the Console.

2. **Top Left**. If you go to "File > New File" you can open up a script file which will then appear in your *Source* window. Script files are text documents containing written R code which then gets sent down to the Console when executed. Script files are great because we can save them onto our computer to rerun our work or to continue writing later. Plus we can send these files to anyone else with R installed on their computer and they should be able to rerun our analysis too. This is a core concept in ensuring your research is reproducible.

   There are two types of script files people typically use when programming in R: .R and .Rmd files. Most R users do the bulk of their programming in .R files. These are plain text files containing only commands for R to execute. If you want to write something else in .R files (such as English sentences), you can do so by starting a line with a `#` character. These are called **comments** and are a great way to explain what your R code is supposed to be doing. Comments make it helpful for other people to understand your code, as well as for yourself if you revisit a project months or years later!

   The other common type of file used to write R code is an .Rmd, or R Markdown, file. We will cover R Markdown in much greater detail in a later section, but these are the basics for now. R Markdown allows you to seamlessly combine R code with written text to create a wide variety of possible documents. All of my papers in grad school have been written in R Markdown, as well as my presentation slides and website. This book you're reading right now was written in R Markdown too. The strength of combining R code and written text comes from how easy it is to update your document when something in your analysis changes. New data? Simply plug it in to the top of your R Markdown document and every graph and table will be automatically updated once you compile a new document. No more copy and pasting figures into Word.

3. **Top Right**. This is your *Environment* tab and it keeps track of which object you've created in R. Objects in R are things like data frames, vectors, and functions. Many people refer to R as an "object-oriented language", by which they mean that most of your code either creates new objects or modifies existing objects. This aspect is probably the largest difference between R and statistical programs like Stata. Each object in R has a specific type which defines what you can and cannot do with it. For example, you cannot add objects that are character types together `"UC" + "SD"`, but you can add two objects with numeric types `3 + 5`. Getting to know the rules surrounding object types is one of the trickier aspects of learning R. But overall, this method of programming is generally pretty intuitive.

4. **Bottom Right**. Three important tabs live here. When you execute the code which makes a plot in a .R file it will, appropriately enough, show up down in your *Plot* tab.

   The *Help* tab is where you can read the documentation for a specific function in R. To find the right help file you can either enter the function name into the search bar, or you can type `?function_name` in the Console. Use the Help tab often! This should be the first place you go when you encounter a problem with your R code.

   Lastly, the *Files* tab is very handy for keeping track of the various R scripts and data you may be working with on a particular project. It acts like a replacement File Explorer (if using a PC) or Finder (if using a Mac) without you having to have multiple windows open on your computer.

1.1.2 RStudio Projects

Speaking of files and file paths…

Throughout your time in the Political Science PhD program you will likely load hundreds of data sets into R. The first step to loading data into R is locating the file holding on the data on your computer. This is done using a file path–a string of characters which points to the file. For example, `"/Users/bertrandwilden/Documents/UCSD/amazing_paper/data/cool_data.csv"` tells us the location of the file "cool_data.csv" located in the folder "data" which is a sub-directory of "amazing_paper" and so on. Getting a hang of using file paths to locate files can be one of the most frustrating parts of using a programming language like R. Modern computer systems, such as your phone, have made file paths invisible to most users. So don't worry if any of this is confusing to you in the beginning!

There are two aspects of file paths that make them particularly annoying/difficult to work with. First, a file path that correctly locates a file on one computer will not locate the same file in another computer. This is because everyone has their own unique folder structures on their computer. Computer-specific file paths make it difficult to share code with others or to collaborate on the same project. They also lead to headaches when revisiting old code on a new computer.

The second, but related, issue with file paths is that they differ between Mac and PC computers. Mac file paths use the forward slash "/" between folders whereas PCs use the back slash "\". Back slashes in R strings are not processed literally–instead they are considered "escape" characters and serve a different purpose. This means you have to manually change all your backslashes to forward slashes to locate files when using a PC. Or you can manually add an extra back slash in front of each folder (`"C:\\Users\\bertrandwilden\\Documents\\UCSD\\amazing_paper\\data\\cool_data.csv"`). What a hassle!

Luckily tools have been developed to solve all these file path annoyances. The first solution is to use RStudio Projects. When you create a new RStudio Project it adds a .Rproj file with your project name to a folder on your computer. The .Rproj file now serves as the top level directory for any R code or data files in folders below it. So instead of using the full path `"/Users/bertrandwilden/Documents/UCSD/amazing_paper/data/cool_data.csv"` to locate your data, you now only need to type `"data/cool_data.csv"`! This is sometimes called a *relative path*. Each RStudio Project is self-contained and easily portable to other machines. RStudio Projects also have the benefit of making it easy to switch between various projects–giving you a clean slate with which to work from every time.[1]

Awesome—RStudio Projects helped solve one of our file path issues by using relative paths, but what about the Mac vs PC problem with different slashes? That's where the R package {here} shines. The `{here}` package allows you to simply put each folder name in quotes and stitches the full path together behind the scenes. The file path `"data/cool_data.csv"` becomes `here("data", "cool_data.csv")`. This code will now point to the correct file on any computer.[2]

1.1.3 Installing Packages

While it is technically feasible to use only the functions that come with R when you install it (aka "base R"), thousands of open source packages have been written to provide extra functionality. The term "open source" means that the underlying code for these packages lives on online repositories, such as GitHub, and can be viewed publicly. While open source packages can be written by anyone (including you someday!), there is a special process packages must undergo in order to be hosted officially by CRAN. Packages that have passed this systematic review by CRAN can be installed on your computer using the following command:

[1] Link to more on why you should be using R Studio Projects
[2] For more on the {here} package see: Link

```
install.packages("package_name_here")
```

It is good practice to only use the `install.packages` command in your Console, rather than in a .R or .qmd script file. This is because you only need to install a particular R package once and then you can then use it forever. Putting `install.packages` in your script file will make R attempt to download the package each time your code is run.

After using the `install.packages` command, you then need to use the following command to access the package's functions:

```
library(package_name_here)
```

Unlike the `install.packages` command, the `library` command *should* be included at the top of any script files which then make use of the package's functions. Another thing to note: the `install.packages` command requires the package name to be in quotes, whereas the `library` command requires the package name to *not* be in quotes. Don't worry—mixing up when to use quotes and when not to is a common error you might encounter when starting out!

1.1.4 Installing the Tidyverse and Here Packages

In this course we will be making extensive use of the packages included in the Tidyverse. The Tidyverse is a set of packages designed to make data analysis in R easier and more streamlined.

•

This approach is usually contrasted against "base R" functions, which do not require external packages. While everything Tidyverse can do, base R can do too, I find the Tidyverse approach much more intuitive. In fact, almost all my script files begin with the command `library(tidyverse)`. There are also a multitude of packages that, although not technically part of the Tidyverse, share the same coding conventions as the core Tidyverse. So learning the Tidyverse will help when you when using more advanced packages.

•

Install the Tidyverse using:

```
# This might take a couple minutes to download all packages
install.packages("tidyverse")
```

I also highly recommend using the "here" package for file path management as explained earlier.

```
install.packages("here")
```

Once both packages are done installing, run the following lines of code to load them into R and make them available for use.

```
library(tidyverse)
library(here)
```

1.2 Reading in Data

For our data visualization exercises we will be using the data set "county_elections.csv" which you can download at the Math Bootcamp GitHub repository here. After you download "county_elections.csv", either copy or move it into your "data" folder in your Math Camp R Project directory. The source for this data comes from the MIT Election Data Science Lab and from the US Census accessed via IPUMS NHGIS.

Now read the "county_elections.csv" data set into R using the following command:

```
county_elections <- read_csv(here("data", "county_elections.csv"))
```

Let's break down this line of code.

- The function `read_csv` is from the package readr which is part of the Tidyverse. It loads a .csv data set file into R. The suffix .csv stands for "comma separated value" and is a very common format for storing tabular data. Inside `read_csv(...)` we put the file path pointing to the file we want to read into R.
- We saw the `here` function earlier. Remember that this is just a convenient way of dealing with file paths. Try only running the command `here("data", "county_elections.csv")` in your console to see how it creates an automatic file path for you.
- In R, `<-` is the operator we use when we want to assign a value to an object. So the expression `county_elections <- ...` can be read as "take the thing on the right side of the arrow and assign it to an object named `county_elections`". Check your Environment tab and verify that there is now an object called `county_elections` there. We can now take the `county_elections` object and do a bunch of stuff with it!

When you read a new data set into R it's often a good idea to do a quick visual inspection. Does the data look like what we'd expect? To do this, either click on the `county_elections` object in your Environment tab or type `View(county_elections)` in your Console. This will make the raw data pop up in a spreadsheet that you can scroll through and check out.

Chapter 2

Objects, Functions, Loops

Where are we? Where are we headed?

Up till now, you should have covered:

- R basic programming
- Data Import

Today we'll cover

- Objects
- Functions
- Loops

2.1 What is an object?

Now that we have covered how to load in data and some basic information about R, let's dive into some fundamentals of the R language.

Let's first set up

```
library(dplyr)
library(readr)
library(haven)
library(ggplot2)
```

```
cen10 <- read_csv("data/input/usc2010_001percent.csv", col_types = cols())
```

Objects are abstract symbols in which you store data. Here we will create an object from `copy`, and assign `cen10` to it.

```
copy <- cen10
```

This looks the same as the original dataset:

```r
copy
```

```
# A tibble: 30,871 x 4
   state         sex      age race
   <chr>         <chr>  <dbl> <chr>
 1 New York      Female     8 White
 2 Ohio          Male      24 White
 3 Nevada        Male      37 White
 4 Michigan      Female    12 White
 5 Maryland      Female    18 Black/Negro
 6 New Hampshire Male      50 White
 7 Iowa          Female    51 White
 8 Missouri      Female    41 White
 9 New Jersey    Male      62 White
10 California    Male      25 White
# ... with 30,861 more rows
# i Use `print(n = ...)` to see more rows
```

What happens if you do this next?

```r
copy <- ""
```

It got reassigned:

```r
copy
```

```
[1] ""
```

2.1.1 lists

Lists are one of the most generic and flexible type of object. You can make an empty list by the function `list()`

```r
my_list <- list()
my_list
```

```
list()
```

And start filling it in. Slots on the list are invoked by double square brackets `[[]]`

```r
my_list[[1]] <- "contents of the first slot -- this is a string"
my_list[["slot 2"]] <- "contents of slot named slot 2"
my_list
```

```
[[1]]
[1] "contents of the first slot -- this is a string"

$`slot 2`
[1] "contents of slot named slot 2"
```

each slot can be anything. What are we doing here? We are defining the 1st slot of the list `my_list` to be a vector `c(1, 2, 3, 4, 5)`

```
my_list[[1]] <- c(1, 2, 3, 4, 5)
my_list
```

```
[[1]]
[1] 1 2 3 4 5

$`slot 2`
[1] "contents of slot named slot 2"
```

You can even make nested lists. Let's say we want the 1st slot of the list to be another list of three elements.

```
my_list[[1]][[1]] <- "subitem 1 in slot 1 of my_list"
my_list[[1]][[2]] <- "subitem 1 in slot 2 of my_list"
my_list[[1]][[3]] <- "subitem 1 in slot 3 of my_list"

my_list
```

```
[[1]]
[1] "subitem 1 in slot 1 of my_list" "subitem 1 in slot 2 of my_list"
[3] "subitem 1 in slot 3 of my_list" "4"
[5] "5"

$`slot 2`
[1] "contents of slot named slot 2"
```

2.2 Making your own objects

We've covered one type of object, which is a list. You saw it was quite flexible. How many types of objects are there?

There are an infinite number of objects, because people make their own class of object. You can detect the type of the object (the class) by the function `class`

Object can be said to be an instance of a class.

Analogies:

class - Pokemon, **object** - Pikachu

class - Book, **object** - To Kill a Mockingbird

class - DataFrame, **object** - 2010 census data

class - Character, **object** - "Programming is Fun"

What is type (class) of object is `cen10`?

```
class(cen10)
```

```
[1] "spec_tbl_df" "tbl_df"      "tbl"         "data.frame"
```

What about this text?

```
class("some random text")
```

```
[1] "character"
```

To change or create the class of any object, you can *assign* it. To do this, assign the name of your class to character to an object's `class()`.

We can start from a simple list. For example, say we wanted to store data about pokemon. Because there is no pre-made package for this, we decide to make our own class.

```
pikachu <- list(name = "Pikachu",
                number = 25,
                type = "Electric",
                color = "Yellow")
```

and we can give it any class name we want.

```
class(pikachu) <- "Pokemon"
str(pikachu)
```

```
List of 4
 $ name  : chr "Pikachu"
 $ number: num 25
 $ type  : chr "Electric"
 $ color : chr "Yellow"
 - attr(*, "class")= chr "Pokemon"
```

```
pikachu$type
```

```
[1] "Electric"
```

2.2.1 Seeing R through objects

Most of the R objects that you will see as you advance are their own objects. For example, here's a linear regression object (which you will learn more about in 204B):

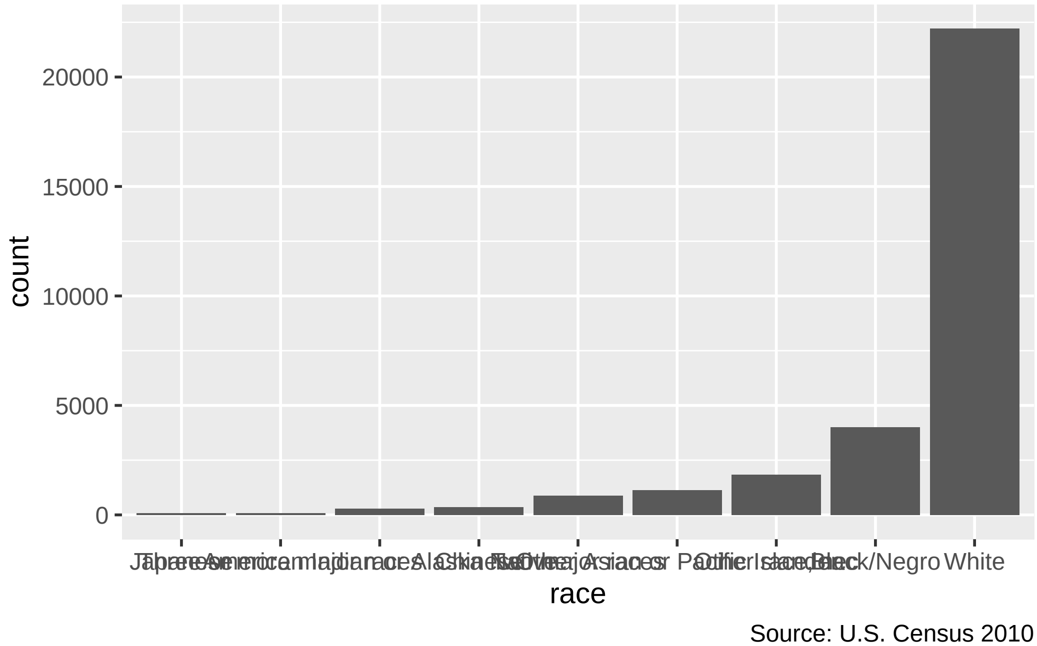
```
ols <- lm(mpg ~ wt + vs + gear + carb, mtcars)
class(ols)
```

```
[1] "lm"
```

Anything can be an object! Even graphs (in `ggplot`) can be assigned, re-assigned, and edited.

```
grp_race <- group_by(cen10, race)%>%
  summarize(count = n())

grp_race_ordered <- arrange(grp_race, count) %>%
  mutate(race = forcats::as_factor(race))
```

```
gg_tab <- ggplot(data = grp_race_ordered) +
  aes(x = race, y = count) +
  geom_col() +
  labs(caption = "Source: U.S. Census 2010")

gg_tab
```



You can change the orientation

```
gg_tab<- gg_tab + coord_flip()
```

2.2.2 Parsing an object by `str()`s

It can be hard to understand an `R` object because its contents are unknown. The function `str`, short for structure, is a quick way to look into the innards of an object

```
str(my_list)
```

```
List of 2
 $       : chr [1:5] "subitem 1 in slot 1 of my_list" "subitem 1 in slot 2 of my_list" "subitem 1 in slot
 $ slot 2: chr "contents of slot named slot 2"
```

```
class(my_list)
```

```
[1] "list"
```

Same for the object we just made

```
str(pikachu)
```

```
List of 4
 $ name  : chr "Pikachu"
 $ number: num 25
 $ type  : chr "Electric"
 $ color : chr "Yellow"
 - attr(*, "class")= chr "Pokemon"
```

What does a `ggplot` object look like? Very complicated, but at least you can see it:

```
# enter this on your console
str(gg_tab)
```

2.3 Types of variables

In the social science we often analyze variables. As you saw in the tutorial, different types of variables require different care.

A key link with what we just learned is that variables are also types of R objects.

2.3.1 scalars

One number. How many people did we count in our Census sample?

```
nrow(cen10)
```

```
[1] 30871
```

Question: What proportion of our census sample is Native American? This number is also a scalar

```
# Enter yourself
unique(cen10$race)
```

```
[1] "White"                          "Black/Negro"
[3] "Other race, nec"                "American Indian or Alaska Native"
[5] "Chinese"                        "Other Asian or Pacific Islander"
[7] "Two major races"                "Three or more major races"
[9] "Japanese"
```

```
mean(cen10$race == "American Indian or Alaska Native")
```

```
[1] 0.009555894
```

Hint: you can use the function `mean()` to calcualte the sample mean. The sample proportion is the mean of a sequence of number, where your event of interest is a 1 (or `TRUE`) and others are 0 (or `FALSE`).

2.3.2 numeric vectors

A sequence of numbers.

```
grp_race_ordered$count
```

```
[1]    77    88   295   354   869  1129  1839  4013 22207
```

```
class(grp_race_ordered$count)
```

```
[1] "integer"
```

Or even, all the ages of the millions of people in our Census. Here are just the first few numbers of the list.

```
head(cen10$age)
```

```
[1]  8 24 37 12 18 50
```

2.3.3 characters (aka strings)

This can be just one stretch of characters

```
my_name <- "Meg"
my_name
```

```
[1] "Meg"
```

```
class(my_name)
```

```
[1] "character"
```

or more characters. Notice here that there's a difference between a vector of individual characters and a length-one object of characters.

```
my_name_letters <-  c("M","e","g")
my_name_letters
```

```
[1] "M" "e" "g"
```

```
class(my_name_letters)
```

```
[1] "character"
```

Finally, remember that lower vs. upper case matters in R!

```
my_name2 <- "shiro"
my_name == my_name2
```

```
[1] FALSE
```

2.4 What is a function?

Most of what we do in R is executing a function. `read_csv()`, `nrow()`, `ggplot()` .. pretty much anything with a parentheses is a function. And even things like `<-` and `[` are functions as well.

A function is a set of instructions with specified ingredients. It takes an **input**, then **manipulates** it – changes it in some way – and then returns the manipulated product.

One way to see what a function actually does is to enter it without parentheses.

```
# enter this on your console
table
```

You'll see below that the most basic functions are quite complicated internally.

You'll notice that functions contain other functions. *wrapper* functions are functions that "wrap around" existing functions. This sounds redundant, but it's an important feature of programming. If you find yourself repeating a command more than two times, you should make your own function, rather than writing the same type of code.

2.4.1 Write your own function

It's worth remembering the basic structure of a function. You create a new function, call it `my_fun` by this:

```
my_fun <- function() {

}
```

If we wanted to generate a function that computed the number of men in your data, what would that look like?

```
count_men <- function(data) {

  nmen <- sum(data$sex == "Male")

  return(nmen)
}
```

Then all we need to do is feed this function a dataset

```
count_men(cen10)
```

```
[1] 15220
```

The point of a function is that you can use it again and again without typing up the set of constituent manipulations. So, what if we wanted to figure out the number of men in California?

```
count_men(cen10[cen10$state == "California",])
```

```
[1] 1876
```

Let's go one step further. What if we want to know the proportion of non-whites in a state, just by entering the name of the state? There's multiple ways to do it, but it could look something like this

```
nw_in_state <- function(data, state) {

  s.subset <- data[data$state == state,]
  total.s <- nrow(s.subset)
  nw.s <- sum(s.subset$race != "White")

  nw.s / total.s
}
```

The last line is what gets generated from the function. To be more explicit you can wrap the last line around `return()`. (as in `return(nw.s/total.s)`. `return()` is used when you want to break out of a function in the middle of it and not wait till the last line.

Try it on your favorite state!

```
nw_in_state(cen10, "Massachusetts")
```

```
[1] 0.2040185
```

Checkpoint

1

Try making your own function, `average_age_in_state`, that will give you the average age of people in a given state.

```
# Enter on your own
```

2

Try making your own function, `asians_in_state`, that will give you the number of `Chinese`, `Japanese`, and `Other Asian or Pacific Islander` people in a given state.

```
# Enter on your own
```

3

Try making your own function, 'top_10_oldest_cities', that will give you the names of cities whose population's average age is top 10 oldest.

```
# Enter on your own
```

2.5 What is a package?

You can think of a package as a suite of functions that other people have already built for you to make your life easier.

```
help(package = "ggplot2")
```

To use a package, you need to do two things: (1) install it, and then (2) load it.

Installing is a one-time thing

```
install.packages("ggplot2")
```

But you need to load each time you start a R instance. So always keep these commands on a script.

```
library(ggplot2)
```

In `rstudio.cloud`, we already installed a set of packages for you. But when you start your own R instance, you need to have installed the package at some point.

2.6 Conditionals

Sometimes, you want to execute a command only under certain conditions. This is done through the almost universal function, `if()`. Inside the `if` function we enter a logical statement. The line that is adjacent to, or follows, the `if()` statement only gets executed if the statement returns `TRUE`.

For example,

For example,

```
x <- 5
if (x >0) {
  print("positive number")
} else if (x == 0)  {
  print ("zero")
} else {
  print("negative number")
}
```

```
[1] "positive number"
```

You can wrap that whole things in a function

```
is_positive <- function(number) {
  if (number >0) {
    print("positive number")
  } else if (number == 0)  {
    print ("zero")
  } else {
    print("negative number")
  }
}

is_positive(5)
```

```
[1] "positive number"
```

```
is_positive(-3)
```

```
[1] "negative number"
```

2.7 For-loops

Loops repeat the same statement, although the statement can be "the same" only in an abstract sense. Use the `for(x in X)` syntax to repeat the subsequent command as many times as there are elements in the right-hand object `X`. Each of these elements will be referred to the left-hand index `x`

First, come up with a vector.

```
fruits <- c("apples", "oranges", "grapes")
```

Now we use the `fruits` vector in a `for` loop.

```
for (fruit in fruits) {
  print(paste("I love", fruit))
}
```

```
[1] "I love apples"
[1] "I love oranges"
[1] "I love grapes"
```

Here `for()` and `in` must be part of any for loop. The right hand side `fruits` must be a thing that exists. Finally the `left-hand` side object is "Pick your favor name." It is analogous to how we can index a sum with any letter. $\sum_{i=1}^{10} i$ and `sum_{j = 1}^{10}j` are in fact the same thing.

```
for (i in 1:length(fruits)) {
  print(paste("I love", fruits[i]))
}
```

```
[1] "I love apples"
[1] "I love oranges"
[1] "I love grapes"
```

```
states_of_interest <- c("California", "Massachusetts", "New Hampshire", "Washington")

for( state in states_of_interest){
  state_data <- cen10[cen10$state == state,]
  nmen <- sum(state_data$sex == "Male")

  n <- nrow(state_data)
  men_perc <- round(100*(nmen/n), digits=2)
  print(paste("Percentage of men in",state, "is", men_perc))
```

```
}
```

```
[1] "Percentage of men in California is 49.85"
[1] "Percentage of men in Massachusetts is 47.6"
[1] "Percentage of men in New Hampshire is 48.55"
[1] "Percentage of men in Washington is 48.19"
```

Instead of printing, you can store the information in a vector

```
states_of_interest <- c("California", "Massachusetts", "New Hampshire", "Washington")
male_percentages <- c()
iter <-1

for( state in states_of_interest){
  state_data <- cen10[cen10$state == state,]
  nmen <- sum(state_data$sex == "Male")
  n <- nrow(state_data)
  men_perc <- round(100*(nmen/n), digits=2)

  male_percentages <- c(male_percentages, men_perc)
  names(male_percentages)[iter] <- state
  iter <- iter + 1
}

male_percentages
```

```
   California Massachusetts New Hampshire    Washington
        49.85         47.60         48.55         48.19
```

2.8 Nested Loops

What if I want to calculate the population percentage of a race group for all race groups in states of interest? You could probably use tidyverse functions to do this, but let's try using loops!

```
states_of_interest <- c("California", "Massachusetts", "New Hampshire", "Washington")
for (state in states_of_interest) {
  for (race in unique(cen10$race)) {
    race_state_num <- nrow(cen10[cen10$race == race & cen10$state == state, ])
    state_pop <- nrow(cen10[cen10$state == state, ])
    race_perc <- round(100*(race_state_num/(state_pop)), digits=2)
    print(paste("Percentage of ", race , "in", state, "is", race_perc))
  }
}
```

```
[1] "Percentage of  White in California is 57.61"
[1] "Percentage of  Black/Negro in California is 6.72"
[1] "Percentage of  Other race, nec in California is 15.55"
[1] "Percentage of  American Indian or Alaska Native in California is 1.12"
[1] "Percentage of  Chinese in California is 3.75"
```

```
[1] "Percentage of  Other Asian or Pacific Islander in California is 9.54"
[1] "Percentage of  Two major races in California is 4.62"
[1] "Percentage of  Three or more major races in California is 0.37"
[1] "Percentage of  Japanese in California is 0.72"
[1] "Percentage of  White in Massachusetts is 79.6"
[1] "Percentage of  Black/Negro in Massachusetts is 5.87"
[1] "Percentage of  Other race, nec in Massachusetts is 4.02"
[1] "Percentage of  American Indian or Alaska Native in Massachusetts is 0.77"
[1] "Percentage of  Chinese in Massachusetts is 2.32"
[1] "Percentage of  Other Asian or Pacific Islander in Massachusetts is 4.33"
[1] "Percentage of  Two major races in Massachusetts is 2.78"
[1] "Percentage of  Three or more major races in Massachusetts is 0"
[1] "Percentage of  Japanese in Massachusetts is 0.31"
[1] "Percentage of  White in New Hampshire is 93.48"
[1] "Percentage of  Black/Negro in New Hampshire is 0.72"
[1] "Percentage of  Other race, nec in New Hampshire is 0.72"
[1] "Percentage of  American Indian or Alaska Native in New Hampshire is 0.72"
[1] "Percentage of  Chinese in New Hampshire is 0.72"
[1] "Percentage of  Other Asian or Pacific Islander in New Hampshire is 2.17"
[1] "Percentage of  Two major races in New Hampshire is 0.72"
[1] "Percentage of  Three or more major races in New Hampshire is 0"
[1] "Percentage of  Japanese in New Hampshire is 0.72"
[1] "Percentage of  White in Washington is 76.05"
[1] "Percentage of  Black/Negro in Washington is 2.9"
[1] "Percentage of  Other race, nec in Washington is 5.37"
[1] "Percentage of  American Indian or Alaska Native in Washington is 2.03"
[1] "Percentage of  Chinese in Washington is 1.31"
[1] "Percentage of  Other Asian or Pacific Islander in Washington is 6.68"
[1] "Percentage of  Two major races in Washington is 4.79"
[1] "Percentage of  Three or more major races in Washington is 0.29"
[1] "Percentage of  Japanese in Washington is 0.58"
```

Exercises

Exercise 1: Write your own function

Write your own function that makes some task of data analysis simpler. Ideally, it would be a function that helps you do either of the previous tasks in fewer lines of code. You can use the three lines of code that was provided in exercise 1 to wrap that into another function too!

```
# Enter yourself
```

Exercise 2: Using Loops

Using a loop, create a crosstab of sex and race for each state in the set "states_of_interest"

```
states_of_interest <- c("California", "Massachusetts", "New Hampshire", "Washington")
# Enter yourself
```

Exercise 3: Storing information derived within loops in a global dataframe

Recall the following nested loop

```
states_of_interest <- c("California", "Massachusetts", "New Hampshire", "Washington")
for (state in states_of_interest) {
  for (race in unique(cen10$race)) {
    race_state_num <- nrow(cen10[cen10$race == race & cen10$state == state, ])
    state_pop <- nrow(cen10[cen10$state == state, ])
    race_perc <- round(100*(race_state_num/(state_pop)), digits=2)
    print(paste("Percentage of ", race , "in", state, "is", race_perc))
  }
}
```

```
[1] "Percentage of  White in California is 57.61"
[1] "Percentage of  Black/Negro in California is 6.72"
[1] "Percentage of  Other race, nec in California is 15.55"
[1] "Percentage of  American Indian or Alaska Native in California is 1.12"
[1] "Percentage of  Chinese in California is 3.75"
[1] "Percentage of  Other Asian or Pacific Islander in California is 9.54"
[1] "Percentage of  Two major races in California is 4.62"
[1] "Percentage of  Three or more major races in California is 0.37"
[1] "Percentage of  Japanese in California is 0.72"
[1] "Percentage of  White in Massachusetts is 79.6"
[1] "Percentage of  Black/Negro in Massachusetts is 5.87"
[1] "Percentage of  Other race, nec in Massachusetts is 4.02"
[1] "Percentage of  American Indian or Alaska Native in Massachusetts is 0.77"
[1] "Percentage of  Chinese in Massachusetts is 2.32"
[1] "Percentage of  Other Asian or Pacific Islander in Massachusetts is 4.33"
[1] "Percentage of  Two major races in Massachusetts is 2.78"
[1] "Percentage of  Three or more major races in Massachusetts is 0"
[1] "Percentage of  Japanese in Massachusetts is 0.31"
[1] "Percentage of  White in New Hampshire is 93.48"
[1] "Percentage of  Black/Negro in New Hampshire is 0.72"
[1] "Percentage of  Other race, nec in New Hampshire is 0.72"
[1] "Percentage of  American Indian or Alaska Native in New Hampshire is 0.72"
[1] "Percentage of  Chinese in New Hampshire is 0.72"
[1] "Percentage of  Other Asian or Pacific Islander in New Hampshire is 2.17"
[1] "Percentage of  Two major races in New Hampshire is 0.72"
[1] "Percentage of  Three or more major races in New Hampshire is 0"
[1] "Percentage of  Japanese in New Hampshire is 0.72"
[1] "Percentage of  White in Washington is 76.05"
[1] "Percentage of  Black/Negro in Washington is 2.9"
[1] "Percentage of  Other race, nec in Washington is 5.37"
[1] "Percentage of  American Indian or Alaska Native in Washington is 2.03"
[1] "Percentage of  Chinese in Washington is 1.31"
[1] "Percentage of  Other Asian or Pacific Islander in Washington is 6.68"
[1] "Percentage of  Two major races in Washington is 4.79"
[1] "Percentage of  Three or more major races in Washington is 0.29"
[1] "Percentage of  Japanese in Washington is 0.58"
```

Instead of printing the percentage of each race in each state, create a dataframe, and store all that information in that dataframe. (Hint: look at how I stored information about male percentage in each state of interest in a vector.)

Chapter 3

Visualization

This lesson is about creating effective data visualizations using the ggplot2 package (part of the Tidyverse). Becoming good at graphing your data is a key skill you will want to develop while in the PhD program. Each graph you make should clearly communicate an insight without overloading your audience with too much information. Today we will practice the nuts and bolts of the coding necessary to accomplish this.

Let's start by loading in our external packages: the Tidyverse, and here.

```
library(tidyverse)
library(here)
```

We will load in the same "county_elections.csv" data set from the previous chapter. Note: we will also remove each row in the data set containing missing values so that we avoid being spammed with warning messages from R. In a real data analysis project, you will want to investigate the source of missing data rather than blanket-removing everything.

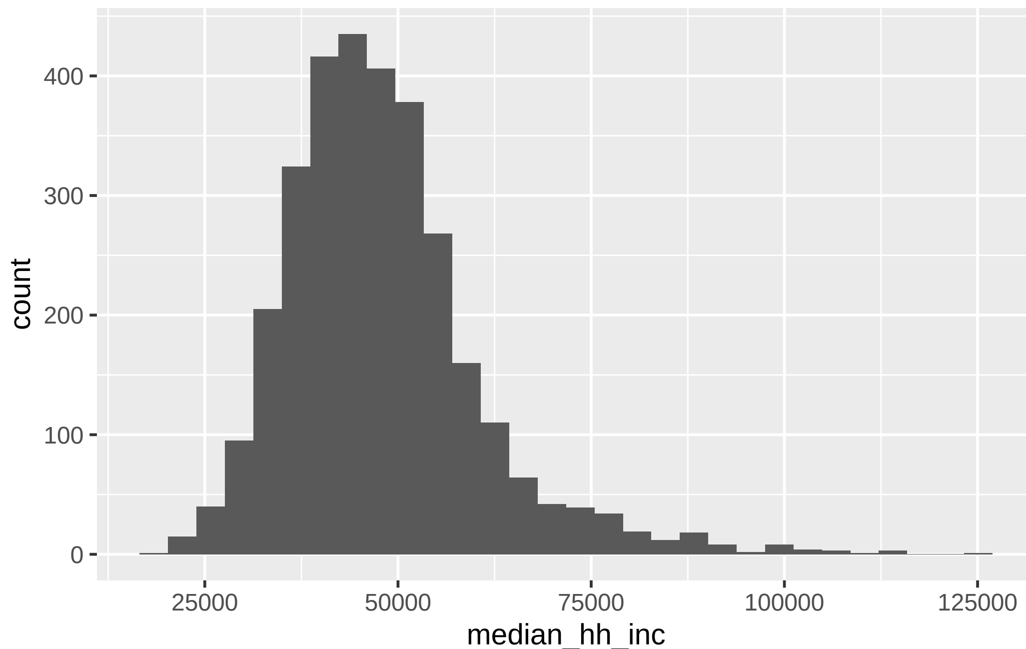
```
county_elections <- read_csv(here("data", "county_elections.csv"))

# Remove any rows with missing values to avoid warning messages
county_elections <- na.omit(county_elections)
```

3.1 Univariate Graphs

The first graph we will make is a histogram. Histograms are the most common type of graph for continuous variables and make it easy to see the spread and central tendency of the data. Let's plot the distribution of county median household income using the variable `median_hh_inc` in `county_elections`.

```
ggplot(county_elections) +
  aes(x = median_hh_inc) +
  geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



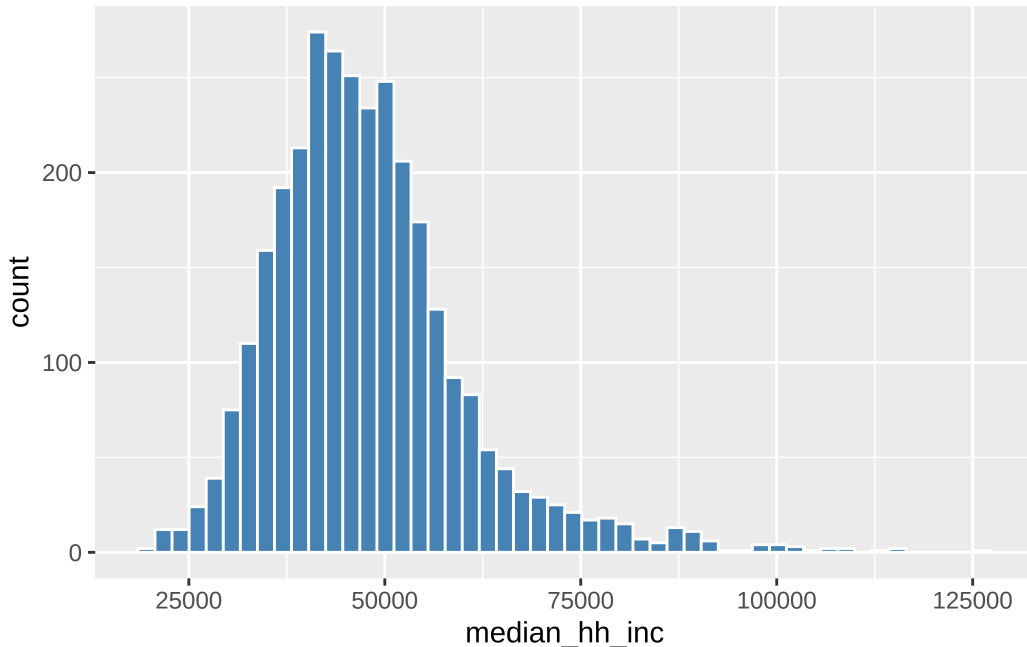
Each graph you create using ggplot will contain the following three elements

1. **Data**. You need to tell ggplot which data set the variables that you want to graph come from. This section is `ggplot(county_elections)` in the code above.
2. **Aesthetics**. Now that we know which data set we're working with, which variables do you want to use and in what way do we want them to be used? This information goes in the `aes()` section. Because histograms typically view the distribution of a single variable along the x-axis of a graph, we specify our aesthetic `aes(x = median_hh_inc)` in the code above.
3. **Geoms**. The "geom" we choose defines the type of graph we're ultimately creating (histogram, scatter plot, bar graph, etc). As you might expect, `geom_histogram()` creates a histogram for us!

In ggplot we combine these elements together using the `+` symbol. You could put the data, aesthetics, and geom sections all in the same line of code. But it is good practice to put each on its own line to make your code more readable.

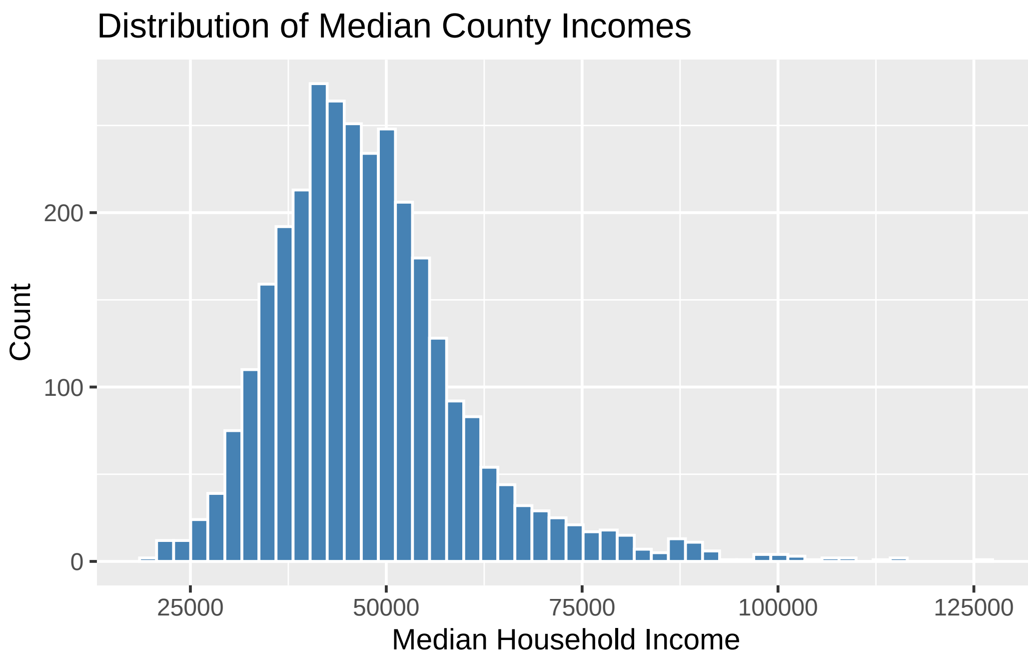
Each geom in ggplot has tons of extra options (also called arguments), which you can specify to make your graphs more pretty. Let's begin to customize our histogram!

```
ggplot(county_elections) +
  aes(x = median_hh_inc) +
  geom_histogram(bins = 50,
                 color = "white",
                 fill = "steelblue")
```



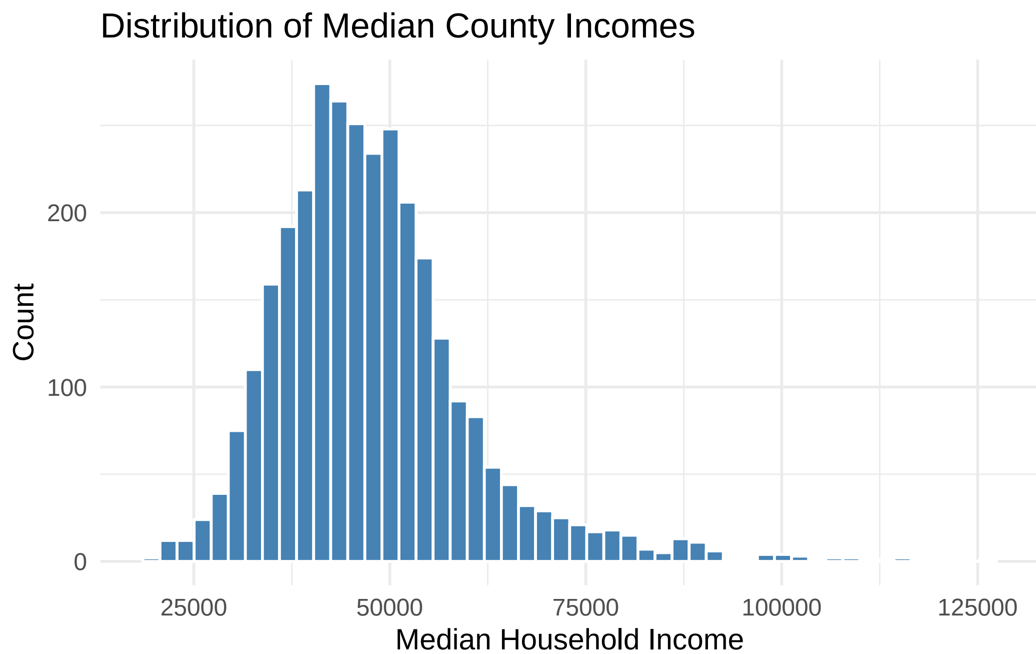
Wow look at that! Now let's fix the ugly default names on the x and y axes, and add an informative title for our graph. We add custom labels to a ggplot graph by adding another `+` followed by a `labs()` section.

```
ggplot(county_elections) +
  aes(x = median_hh_inc) +
  geom_histogram(bins = 50,
                 color = "white",
                 fill = "steelblue") +
  labs(title = "Distribution of Median County Incomes",
       x = "Median Household Income",
       y = "Count")
```



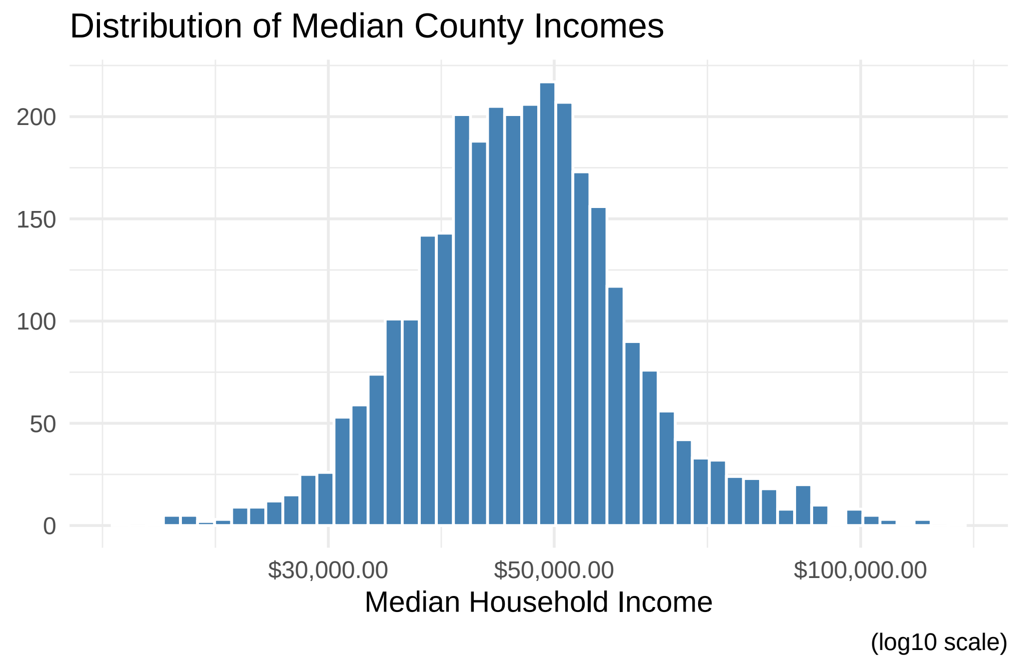
Themes in ggplot control the overall look and background style of our graphs. For a complete list of themes: Link. There is also a package with a bunch of additional cool themes you can check out here: Link. Personally I'm a big fan of `theme_minimal()`.

```
ggplot(county_elections) +
  aes(x = median_hh_inc) +
  geom_histogram(bins = 50,
                 color = "white",
                 fill = "steelblue") +
  labs(title = "Distribution of Median County Incomes",
       x = "Median Household Income",
       y = "Count") +
  theme_minimal()
```



Our county median household income variable looks like it's a bit right-skewed—with a few extremely high income counties shown on the right hand side of the graph. Depending on your research question, it might make more sense to view this distribution on the log scale. It's very easy to do this in ggplot using `scale_x_log10()`.

```
ggplot(county_elections) +
  aes(x = median_hh_inc) +
  geom_histogram(bins = 50,
                 color = "white",
                 fill = "steelblue") +
  labs(title = "Distribution of Median County Incomes",
       x = "Median Household Income",
       y = "",
       caption = "(log10 scale)") +
  theme_minimal() +
  scale_x_log10(labels = scales::dollar)
```

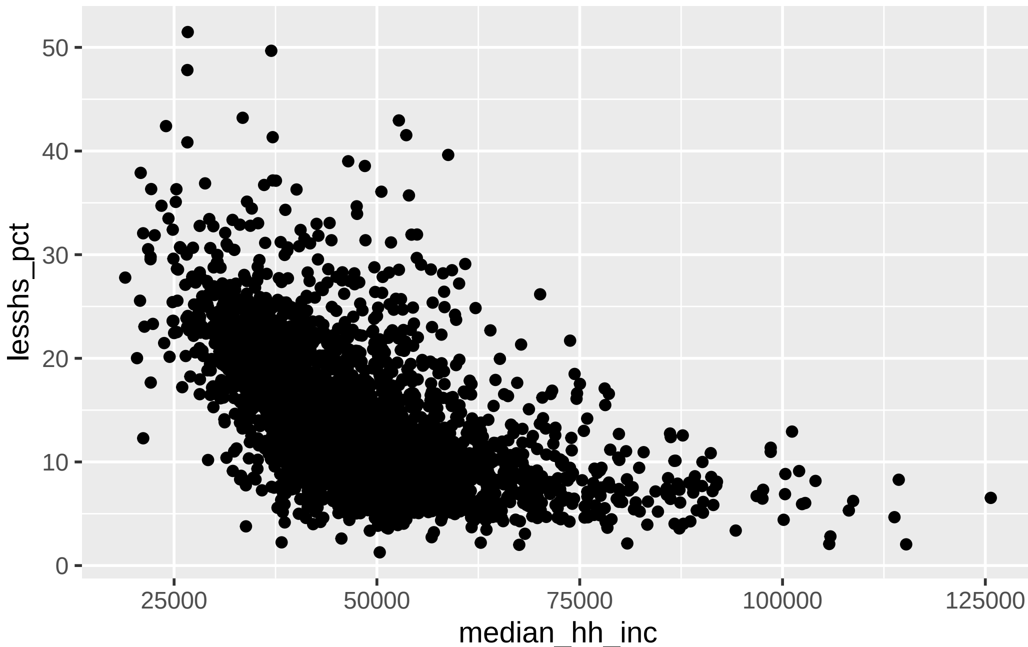


Now the data looks almost normally distributed. Also note the use of `scales::dollar` to make our x-axis a little easier to read. The scales package provides a ton of handy functions to deal with ugly default scales in ggplot. The `::` operator is a way of accessing a single function from a package without loading all its other functions into R. It's also a way of being explicit about which package's function you are using. Sometimes you will run across situations where multiple packages have functions with the same name, but which do different things! Speaking from personal experience, this can lead to some really frustrating debugging sessions.

3.2 Bivariate Graphs

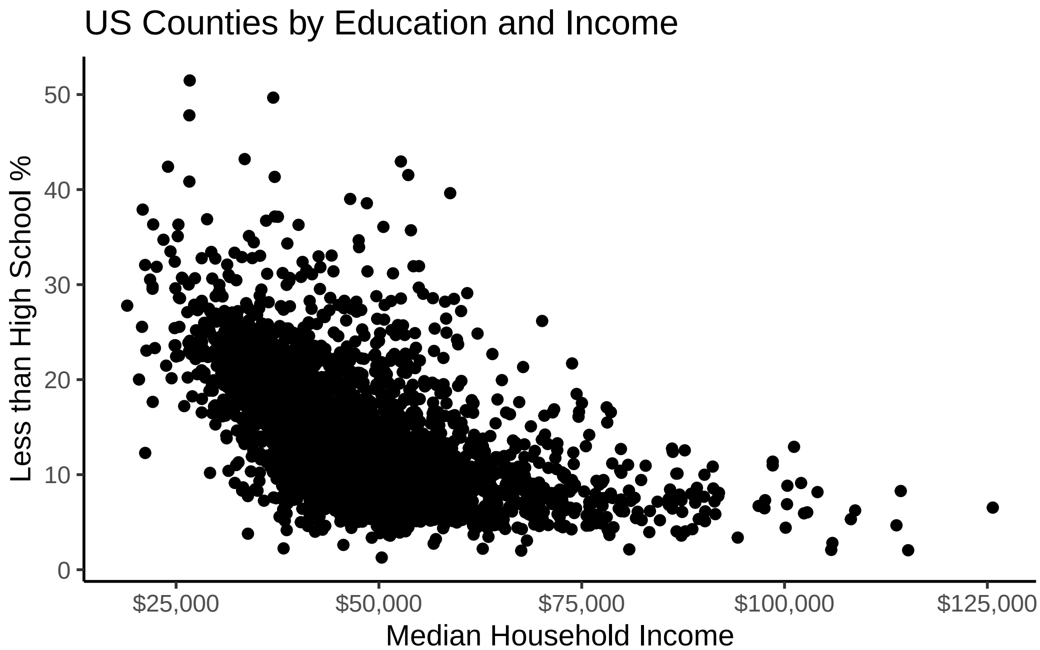
If histograms are the most common way to plot the distribution of a single continuous variable, scatter plots are the most common way to show the relationship between two continuous variables. Translating our histogram ggplot code to a scatter plot is straightforward: simply add a y-axis variable to `aes()`, and change the geom to `geom_point()`. The graph below displays the relationship between median household income and the population percentage in a county who did not complete high school.

```
ggplot(county_elections) +
  aes(x = median_hh_inc, y = lesshs_pct) +
  geom_point()
```



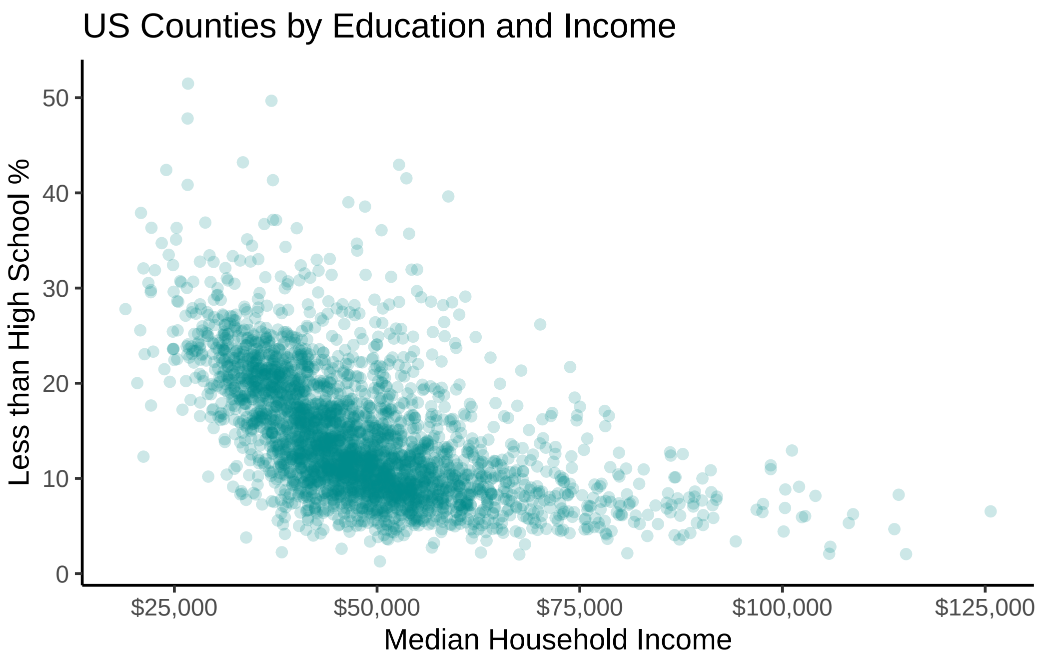
It looks like richer counties have lower rates of their population having less than a high school education. We can use the same customization options from histograms on our scatter plot to make things prettier.

```
ggplot(county_elections) +
  aes(x = median_hh_inc, y = lesshs_pct) +
  geom_point() +
  labs(title = "US Counties by Education and Income",
       x = "Median Household Income",
       y = "Less than High School %") +
  scale_x_continuous(label = scales::dollar) +
  theme_classic()
```



A lot of our data seems to be clustered up together. The solid points in `geom_point()` obscure this density so let's fix this using the `alpha` argument. A geom's `alpha` level specifies its transparency and ranges from 1 (solid) to 0 (invisible).

```
ggplot(county_elections) +
  aes(x = median_hh_inc, y = lesshs_pct) +
  geom_point(alpha = 0.2, color = "darkcyan") +
  labs(title = "US Counties by Education and Income",
       x = "Median Household Income",
       y = "Less than High School %") +
  scale_x_continuous(label = scales::dollar) +
  theme_classic()
```

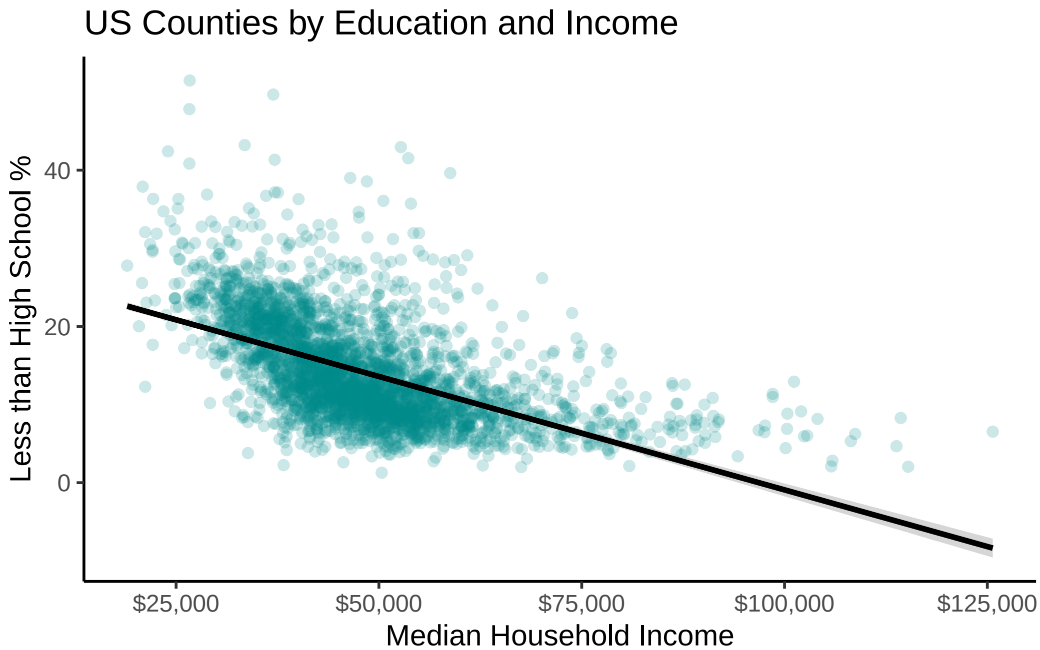


The negative relationship between our two variables is clear just by eyeballing it, but if we want to be real scientists we need to add the magic regression line. If you are unfamiliar with regression lines, don't worry—we will be covering them extensively in your introductory quantitative methods course. A linear regression line is essentially just the "best fitting" straight line to the data.

Adding a regression line to the graph gives us our first opportunity to combine multiple *geoms*. In the code chunk below, notice how we simply use `+` to add `geom_smooth()` to our ggplot object. This overlays a fitted line on top of the dots from `geom_point()`. The argument `method = "lm"` tells ggplot to use a linear regression line (`lm` = "linear model") as opposed to some other type of fitted line.

```
ggplot(county_elections) +
  aes(x = median_hh_inc, y = lesshs_pct) +
  geom_point(alpha = 0.2, color = "darkcyan") +
  geom_smooth(method = "lm", color = "black") +
  labs(title = "US Counties by Education and Income",
       x = "Median Household Income",
       y = "Less than High School %") +
  scale_x_continuous(label = scales::dollar) +
  theme_classic()
```

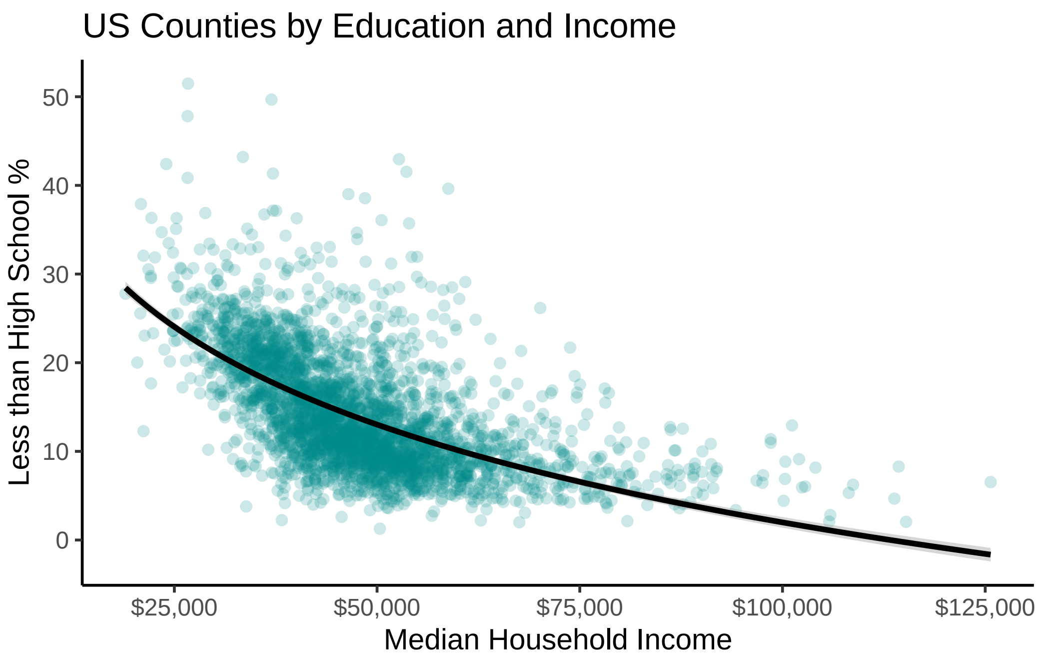
```
`geom_smooth()` using formula 'y ~ x'
```



This graph would look way more professional if it didn't mistakenly predict negative high school percentage values for high income counties. Linear regression is clearly not flexible enough to reflect the true relationship between our two variables.

What if we re-scaled median household income to the log10 scale again?

```
ggplot(county_elections) +
  aes(x = median_hh_inc, y = lesshs_pct) +
  geom_point(alpha = 0.2, color = "darkcyan") +
  geom_smooth(method = "lm", color = "black",
              formula = "y ~ log(x)") +
  labs(title = "US Counties by Education and Income",
       x = "Median Household Income",
       y = "Less than High School %") +
  scale_x_continuous(label = scales::dollar) +
  theme_classic()
```



Not perfect, but now our fitted line is looking better!

3.3 Trivariate(!) Graphs

We are now experts are graphing one variable at a time, or two variables together, but what if we want to graph three or more variables at once? There are a number of ways to do this in ggplot as we will see below. However, first a word of caution: beware of cluttering your plots with *too* much information! It can be tempting to throw everything into a graph, but doing so can obscure the main point you're trying to make. Always keep this in mind when going beyond graphing two variables at once.

3.3.1 Using Colors and Shapes

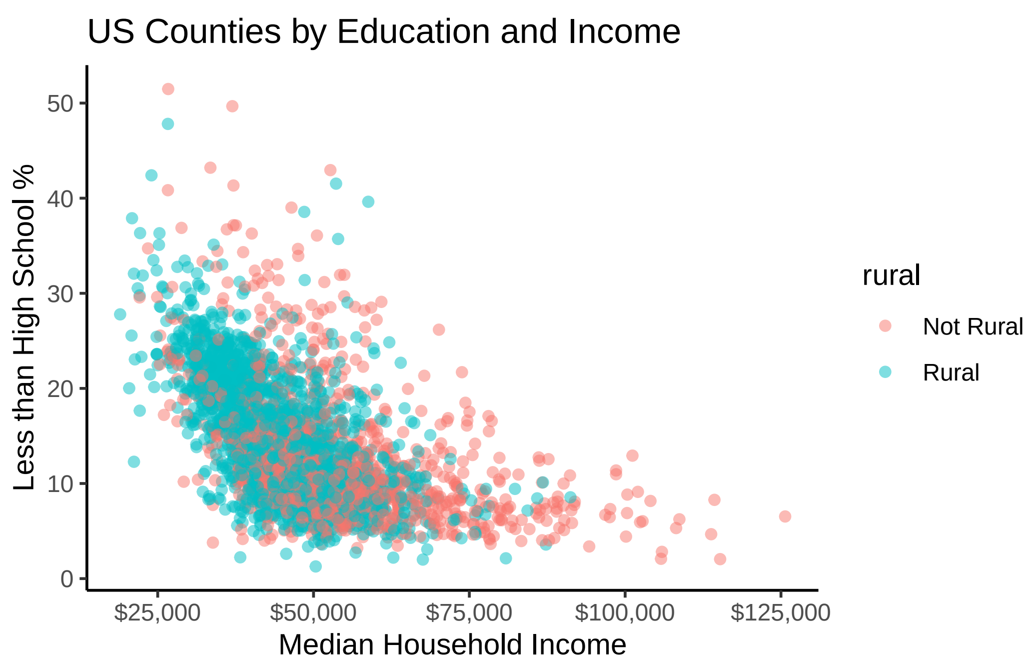
The `county_elections` data set does not have a lot of categorical variables for us to work with. So let's create one!

```
county_elections <- county_elections |>
  mutate(rural = ifelse(rural_pct > 50, "Rural", "Not Rural"))
```

This code chunk uses the `mutate` function to create a new variable in the `county_elections` data set called `rural`. The variable `rural` takes the value "Rural" if `rural_pct` is greater than 50 and takes the value "Not Rural" if `rural_pct` is less than or equal to 50. It is usually not a good idea to dichotomize a continuous variable in this way (using a binary Rural/Not Rural as opposed to the county's rural percentage). Doing so throws away valuable information that is almost always relevant to the final analysis. In this case we can justify our choice to create a categorical variable because it will make plotting multiple variables much easier.

Let's now take our scatter plot showing the relationship between county median household income and education level, and color the points based on whether the county is rural or not. Doing so is as easy as adding `color = rural` to the `aes()` section in ggplot.

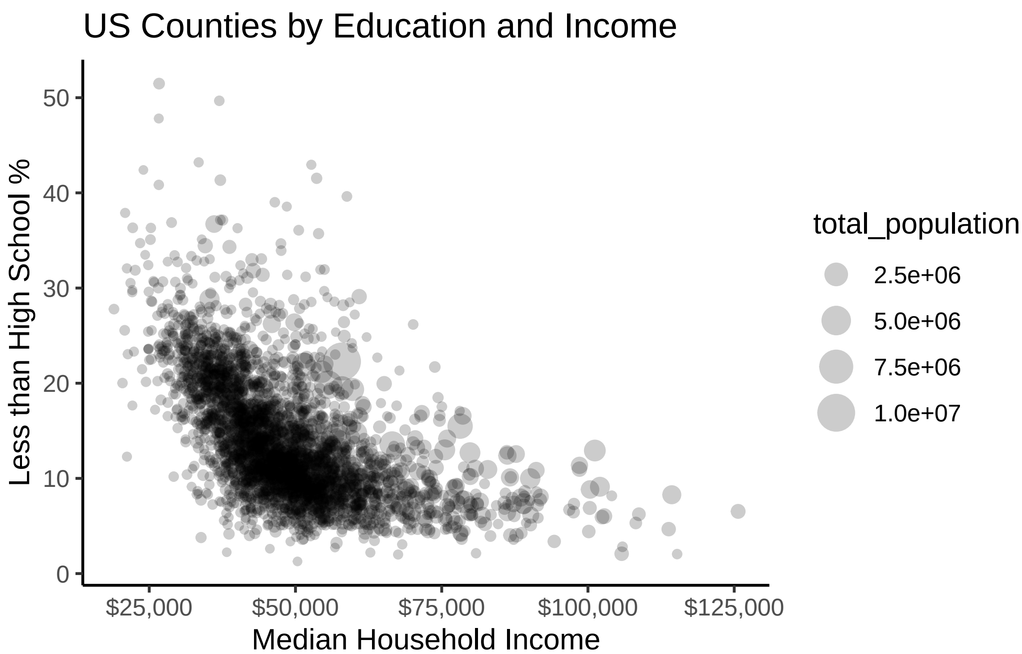
```
ggplot(county_elections) +
  aes(x = median_hh_inc, y = lesshs_pct,
      color = rural) + # Coloring points based on rural variable
  geom_point(alpha = 0.5) + # Removed color from geom
  labs(title = "US Counties by Education and Income",
       x = "Median Household Income",
       y = "Less than High School %") +
  scale_x_continuous(label = scales::dollar) +
  theme_classic()
```



By default, ggplot even gives us a handy legend to tell us which color points correspond to which value of `rural`.

Now let's try adding a third variable to our scatter plot which is continuous. One way to do this is with the `size` option in `aes()`.

```
ggplot(county_elections) +
  aes(x = median_hh_inc, y = lesshs_pct,
      size = total_population) + # Changing size of points
  geom_point(alpha = 0.2) +
  labs(title = "US Counties by Education and Income",
       x = "Median Household Income",
       y = "Less than High School %") +
  scale_x_continuous(label = scales::dollar) +
  theme_classic()
```

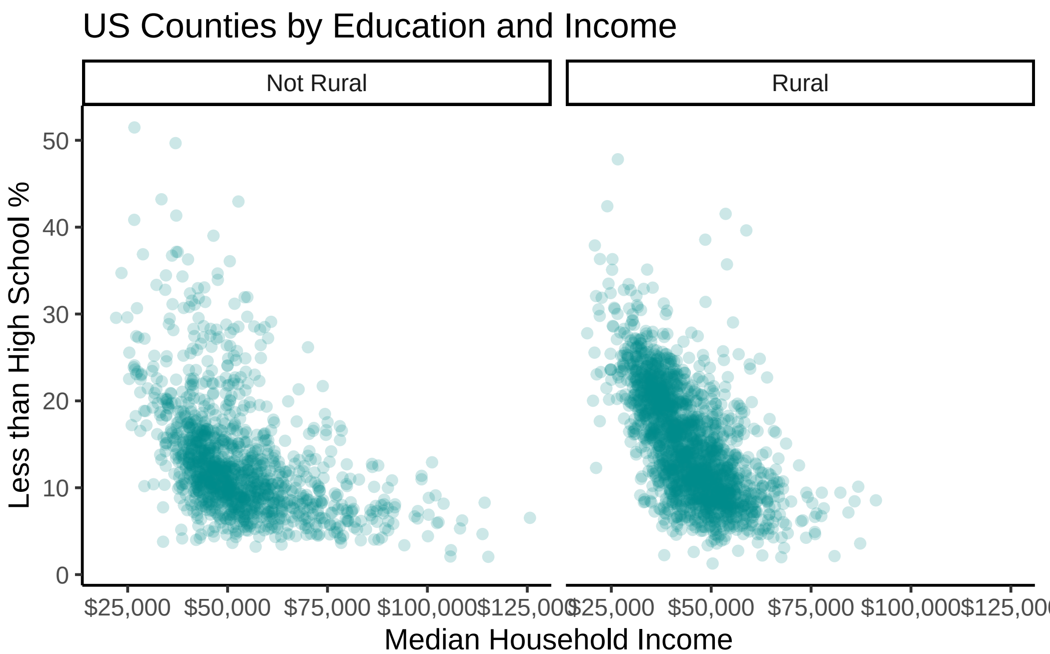


Are we overdoing things with adding too much information to our graph? Possibly!

3.3.2 Using Facets to Graph Comparisons

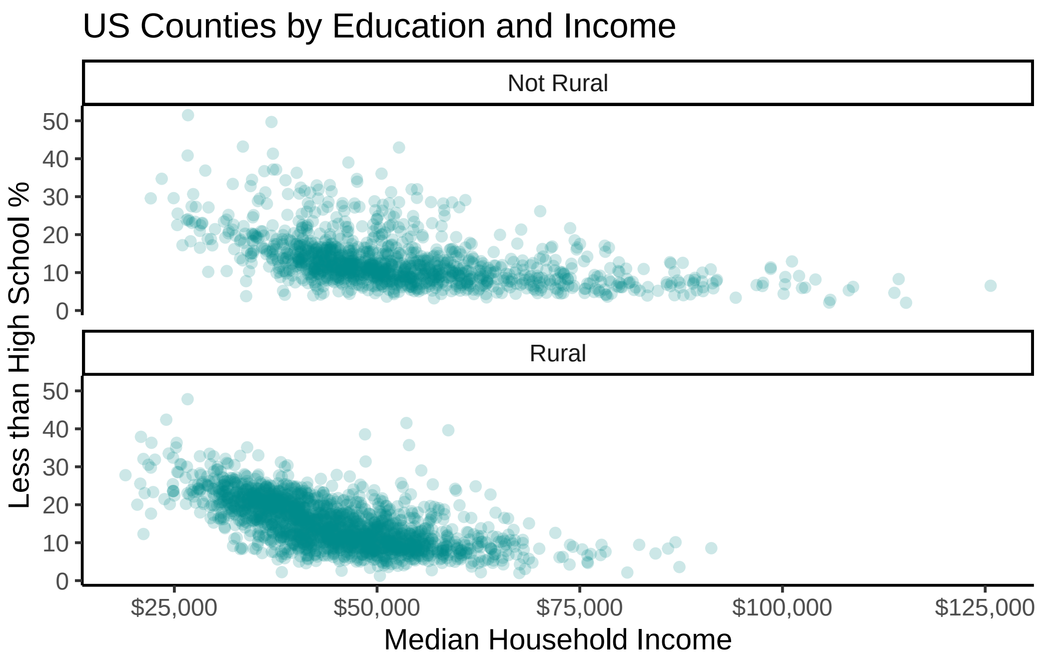
One of ggplot's most powerful features is "faceting". Facets allow you to easily graph comparisons between different levels of a categorical variable in a clear manner by creating side by side subgraphs. To apply a facet to our ggplot graph we can simply add `+ facet_wrap(~ facet_variable)`.

```
ggplot(county_elections) +
  aes(x = median_hh_inc, y = lesshs_pct) +
  geom_point(alpha = 0.2, color = "darkcyan") +
  labs(title = "US Counties by Education and Income",
       x = "Median Household Income",
       y = "Less than High School %") +
  scale_x_continuous(label = scales::dollar) +
  theme_classic() +
  facet_wrap(~ rural) # Adding faceting
```



As you can see, faceting is so powerful for showing comparisons because it preserves the scale in each subplot. This might be a better choice rather than coloring each point and overlapping everything. We can control whether we want the subgraphs side-by-side or on top of each other with the `nrow` argument.

```
ggplot(county_elections) +
  aes(x = median_hh_inc, y = lesshs_pct) +
  geom_point(alpha = 0.2, color = "darkcyan") +
  labs(title = "US Counties by Education and Income",
       x = "Median Household Income",
       y = "Less than High School %") +
  scale_x_continuous(label = scales::dollar) +
  theme_classic() +
  facet_wrap(~ rural, nrow = 2)
```



3.4 Choropleth Maps

Making maps in ggplot is relatively straightforward—and a much better idea than copying and pasting your data back and forth between R and a specialized program like ArcGIS. Choropleth maps show data broken down by geographic unit (in this case US counties). We will need to install an additional package urbnmapr to help ggplot make this type of graph. To install urbnmapr, run the following command in your Console.

```
devtools::install_github("UrbanInstitute/urbnmapr")
```

We use the command `devtools::install_github()` because the developers of urbnmapr have not submitted their package to the official CRAN repository. So rather than using `install.packages` like we're used to, we instead need to install the package directly from GitHub. A lot of excellent packages are not available on CRAN, but be aware that they might not have all the quality-control checks CRAN packages have.

Once you have the urbnmapr package installed, you can load it into R using:

```
library(urbnmapr)
```

We need to perform a couple data cleaning steps before the data is ready to map in ggplot. The first step is making our `countyCode` variable match the format of the corresponding US county code in the urbnmapr data. US counties are each given a unique 5-digit number called a FIPS code. However, at some point the "county_elections.csv" file was opened in Excel, which read the FIPS codes as numeric values thereby removing any 0's from the start of each code. Never open your data in Excel! Now a bunch of the FIPS codes in our data are only 4-digits long instead of 5, which means they will not match the FIPS codes in the urbnmapr data. Luckily we can fix this using the Tidyverse. The function `str_pad` from the stringr package can be used to "pad" out a variable with a specific character until it becomes a specific size.

```
county_elections <- county_elections |>
  mutate(county_fips = str_pad(countyCode, width = 5, pad = "0"))
```

Next we need to join our `county_elections` data with the mapping data from urbnmapr. We will do this using a `left_join` command, which, if you are not familiar with, we will cover in much greater detail in a future lesson. The big idea here is that we have one data set with county-level variables, such as median household income, that we need to merge with a data set containing the geographic coordinate information for each US county.

```
map_data <- left_join(county_elections, counties)
```

Awesome! Now we are ready to make a map! Let's check out the geographic distribution of population percentage without a high school diploma.

```
ggplot(map_data) +
  aes(x = long, y = lat,
      group = group, fill = lesshs_pct) +
  geom_polygon(color = NA) +
  # This second geom_polygon shows the state borders
  geom_polygon(data = states, mapping = aes(long, lat, group = group),
               fill = NA, size = 0.1, color = "white") +
  # Making maps requires you to choose a geographic projection
  coord_map(projection = "albers", lat0 = 39, lat1 = 45) +
  # theme_void gives us a blank canvas
  theme_void()
```

Don't worry if you are not yet able to understand every aspect of the ggplot code that produced this map. Try playing around with some of the arguments and see what happens to the map!

```
ggplot(map_data) +
  aes(x = long, y = lat,
      group = group, fill = lesshs_pct) +
  geom_polygon(color = NA) +
  geom_polygon(data = states, mapping = aes(long, lat, group = group),
               fill = NA, size = 0.1, color = "white") +
  coord_map(projection = "albers", lat0 = 39, lat1 = 45) +
  # This creates a diverging color scale
  # that is also colorblind friendly
  scale_fill_viridis_c() +
  labs(fill = "Less than High School %") +
  theme_void() +
  theme(legend.position = "bottom")
```

Sometimes a diverging color scale is better for contrasting high and low value areas.

Chapter 4

Simulation

Module originally written by Connor Jerzak and Shiro Kuriwaki

Motivation: Simulation as an Analytical Tool

An increasing amount of political science contributions now include a simulation.

- Axelrod (1977) demonstrated via simulation how atomized individuals evolve to be grouped in similar clusters or countries, a model of culture.[1]
- Chen and Rodden (2013) argued in a 2013 article that the vote-seat inequality in U.S. elections that is often attributed to intentional partisan gerrymandering can actually attributed to simply the reality of "human geography" – Democratic voters tend to be concentrated in smaller area. Put another way, no feasible form of gerrymandering could spread out Democratic voters in such a way to equalize their vote-seat translation effectiveness. After demonstrating the empirical pattern of human geography, they advance their key claim by simulating thousands of redistricting plans and record the vote-seat ratio.[2]
- Gary King, James Honaker, and multiple other authors propose a way to analyze missing data with a method of multiple imputation, which uses a lot of simulation from a researcher's observed dataset.[3] (Software: Amelia[4])

Statistical methods also incorporate simulation:

- The bootstrap: a statistical method for estimating uncertainty around some parameter by re-sampling observations.
- Bagging: a method for improving machine learning predictions by re-sampling observations, storing the estimate across many re-samples, and averaging these estimates to form the final estimate. A variance reduction technique.
- Statistical reasoning: if you are trying to understand a quantitative problem, a wonderful first-step to understand the problem better is to simulate it! The analytical solution is often very hard (or impossible), but the simulation is often much easier :-)

---

[1] Axelrod, Robert. 1997. "The Dissemination of Culture." *Journal of Conflict Resolution* 41(2): 203–26.

[2] Chen, Jowei, and Jonathan Rodden. "Unintentional Gerrymandering: Political Geography and Electoral Bias in Legislatures. *Quarterly Journal of Political Science*, 8:239-269"

[3] King, Gary, et al. "Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation". *American Political Science Review*, 95: 49-69.

[4] James Honaker, Gary King, Matthew Blackwell (2011). Amelia II: A Program for Missing Data. Journal of Statistical Software, 45(7), 1-47.

Where are we? Where are we headed?

Up till now, you should have covered:

- `R` basics
- Visualization
- Matrices and vectors
- Functions, objects, loops
- Joining real data

In this module, we will start to work with generating data within R, from thin air, as it were. Doing simulation also strengthens your understanding of Probability.

Check your Understanding

- What does the `sample()` function do?
- What does `runif()` stand for?
- What is a `seed`?
- What is a Monte Carlo?

Check if you have an idea of how you might code the following tasks:

- Simulate 100 rolls of a die
- Simulate one random ordering of 25 numbers
- Simulate 100 values of white noise (uniform random variables)
- Generate a "bootstrap" sample of an existing dataset

We're going to learn about this today!

4.1 Pick a sample, any sample

4.2 The `sample()` function

The core functions for coding up stochastic data revolves around several key functions, so we will simply review them here.

Suppose you have a vector of values `x` and from it you want to randomly sample a sample of length `size`. For this, use the `sample` function

```r
sample(x = 1:10, size = 5)
```

```
[1]  2  4 10  5  1
```

There are two subtypes of sampling – with and without replacement.

1. Sampling without replacement (`replace = FALSE`) means once an element of `x` is chosen, it will not be considered again:

```r
sample(x = 1:10, size = 10, replace = FALSE) ## no number appears more than once
```

```
 [1]  6  7  5  3  2 10  9  8  4  1
```

2. Sampling with replacement (`replace = TRUE`) means that even if an element of `x` is chosen, it is put back in the pool and may be chosen again.

```r
sample(x = 1:10, size = 10, replace = TRUE) ## any number can appear more than once
```

```
 [1]  4  8  7  6  1  5 10  3  8  3
```

It follows then that you cannot sample without replacement a sample that is larger than the pool.

```r
sample(x = 1:10, size = 100, replace = FALSE)
```

```
Error in sample.int(length(x), size, replace, prob): cannot take a sample larger than the population when
```

So far, every element in `x` has had an equal probability of being chosen. In some application, we want a sampling scheme where some elements are more likely to be chosen than others. The argument `prob` handles this.

For example, this simulates 20 fair coin tosses (each outcome is equally likely to happen)

```r
sample(c("Head", "Tail"), size = 20, prob = c(0.5, 0.5), replace = TRUE)
```

```
 [1] "Tail" "Tail" "Head" "Head" "Tail" "Head" "Head" "Head" "Tail" "Tail"
[11] "Head" "Head" "Head" "Tail" "Head" "Head" "Head" "Tail" "Head" "Head"
```

But this simulates 20 biased coin tosses, where say the probability of Tails is 4 times more likely than the number of Heads

```r
sample(c("Head", "Tail"), size = 20, prob = c(0.2, 0.8), replace = TRUE)
```

```
 [1] "Head" "Tail" "Tail" "Tail" "Tail" "Tail" "Tail" "Tail" "Tail" "Tail"
[11] "Tail" "Tail" "Tail" "Head" "Tail" "Head" "Tail" "Tail" "Tail" "Tail"
```

4.2.1 Sampling rows from a dataframe

In tidyverse, there is a convenience function to sample rows randomly: `sample_n()` and `sample_frac()`.

For example, load the dataset on cars, `mtcars`, which has 32 observations.

```r
mtcars
```

```
                     mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4           21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag       21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710          22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive      21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout   18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant             18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360          14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D           24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230            22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280            19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C           17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE          16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL          17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC         15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
```

```
Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
Fiat 128            32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
Honda Civic         30.4   4  75.7  52 4.93 1.615 18.52  1  1    4    2
Toyota Corolla      33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
Toyota Corona       21.5   4 120.1  97 3.70 2.465 20.01  1  0    3    1
Dodge Challenger    15.5   8 318.0 150 2.76 3.520 16.87  0  0    3    2
AMC Javelin         15.2   8 304.0 150 3.15 3.435 17.30  0  0    3    2
Camaro Z28          13.3   8 350.0 245 3.73 3.840 15.41  0  0    3    4
Pontiac Firebird    19.2   8 400.0 175 3.08 3.845 17.05  0  0    3    2
Fiat X1-9           27.3   4  79.0  66 4.08 1.935 18.90  1  1    4    1
Porsche 914-2       26.0   4 120.3  91 4.43 2.140 16.70  0  1    5    2
Lotus Europa        30.4   4  95.1 113 3.77 1.513 16.90  1  1    5    2
Ford Pantera L      15.8   8 351.0 264 4.22 3.170 14.50  0  1    5    4
Ferrari Dino        19.7   6 145.0 175 3.62 2.770 15.50  0  1    5    6
Maserati Bora       15.0   8 301.0 335 3.54 3.570 14.60  0  1    5    8
Volvo 142E          21.4   4 121.0 109 4.11 2.780 18.60  1  1    4    2
```

sample_n picks a user-specified number of rows from the dataset:

```r
sample_n(mtcars, 3)
```

```
                mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Fiat 128       32.4   4  78.7  66 4.08 2.200 19.47  1  1    4    1
Toyota Corolla 33.9   4  71.1  65 4.22 1.835 19.90  1  1    4    1
Mazda RX4      21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
```

Sometimes you want a X percent sample of your dataset. In this case use `sample_frac()`

```r
sample_frac(mtcars, 0.10)
```

```
               mpg cyl  disp  hp drat   wt qsec vs am gear carb
Ferrari Dino  19.7   6 145.0 175 3.62 2.77 15.5  0  1    5    6
Maserati Bora 15.0   8 301.0 335 3.54 3.57 14.6  0  1    5    8
Porsche 914-2 26.0   4 120.3  91 4.43 2.14 16.7  0  1    5    2
```

As a side-note, these functions have very practical uses for any type of data analysis:

- Inspecting your dataset: using `head()` all the same time and looking over the first few rows might lead you to ignore any issues that end up in the bottom for whatever reason.
- Testing your analysis with a small sample: If running analyses on a dataset takes more than a handful of seconds, change your dataset upstream to a fraction of the size so the rest of the code runs in less than a second. Once verifying your analysis code runs, then re-do it with your full dataset (by simply removing the `sample_n` / `sample_frac` line of code in the beginning). While three seconds may not sound like much, they accumulate and eat up time.

4.3 Random numbers from specific distributions

`rbinom()`

`rbinom` builds upon `sample` as a tool to help you answer the question – what is the *total number of successes* I would get if I sampled a binary (Bernoulli) result from a test with `size` number of trials each, with a event-wise probability of `prob`. The first argument `n` asks me how many such numbers I want.

For example, I want to know how many Heads I would get if I flipped a fair coin 100 times.

```
rbinom(n = 1, size = 100, prob = 0.5)
```

```
[1] 55
```

Now imagine this I wanted to do this experiment 10 times, which would require I flip the coin 10 x 100 = 1000 times! Helpfully, we can do this in one line

```
rbinom(n = 10, size = 100, prob = 0.5)
```

```
 [1] 50 58 47 49 54 60 54 50 53 51
```

`runif()`

`runif` also simulates a stochastic scheme where each event has equal probability of getting chosen like `sample`, but is a continuous rather than discrete system. We will cover this more in the next math module.

The intuition to emphasize here is that one can generate potentially infinite amounts (size `n`) of noise that is a essentially random

```
runif(n = 5)
```

```
[1] 0.57343306 0.60872460 0.52434672 0.09688019 0.46674553
```

`rnorm()`

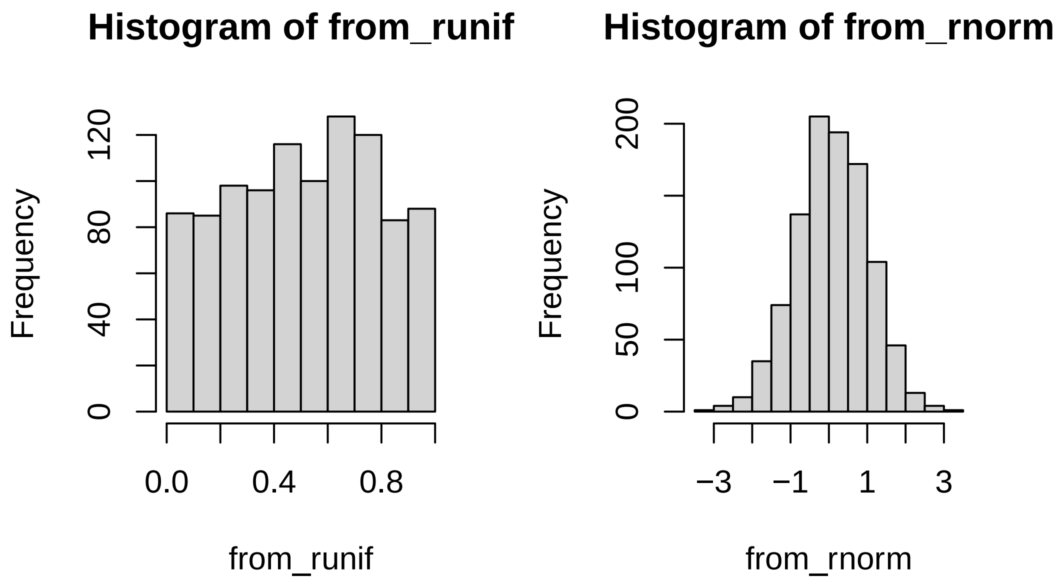
`rnorm` is also a continuous distribution, but draws from a Normal distribution – perhaps the most important distribution in statistics. It runs the same way as `runif`

```
rnorm(n = 5)
```

```
[1]  0.04014912 -1.33034698 -1.24651827  1.82108817  0.95616967
```

To better visualize the difference between the output of `runif` and `rnorm`, let's generate lots of each and plot a histogram.

```
from_runif <- runif(n = 1000)
from_rnorm <- rnorm(n = 1000)

par(mfrow = c(1, 2)) ## base-R parameter for two plots at once
hist(from_runif)
hist(from_rnorm)
```



4.4 r, p, and d

Each distribution can do more than generate random numbers (the prefix `r`). We can compute the cumulative probability by the function `pbinom()`, `punif()`, and `pnorm()`. Also the density – the value of the PDF – by `dbinom()`, `dunif()` and `dnorm()`.

4.5 `set.seed()`

`R` doesn't have the ability to generate truly random numbers! Random numbers are actually very hard to generate. (Think: flipping a coin –> can be perfectly predicted if I know wind speed, the angle the coin is flipped, etc.). Some people use random noise in the atmosphere or random behavior in quantum systems to generate "truly" (?) random numbers. Conversely, R uses deterministic algorithms which take as an input a "seed" and which then perform a series of operations to generate a sequence of random-seeming numbers (that is, numbers whose sequence is sufficiently hard to predict).

Let's think about this another way. Sampling is a stochastic process, so every time you run `sample()` or `runif()` you are bound to get a different output (because different random seeds are used). This is intentional in some cases but you might want to avoid it in others. For example, you might want to diagnose a coding discrepancy by setting the random number generator to give the same number each time. To do this, use the function `set.seed()`.

In the function goes any number. When you run a sample function in the same command as a preceding `set.seed()`, the sampling function will always give you the same sequence of numbers. In a sense, the sampler is no longer random (in the sense of unpredictable to use; remember: it never was "truly" random in the first place)

```r
set.seed(02138)
runif(n = 10)
```

```
[1] 0.51236144 0.61530551 0.37451441 0.43541258 0.21166530 0.17812129
[7] 0.04420775 0.45567854 0.88718264 0.06970056
```

The random number generator should give you the exact same sequence of numbers if you precede the function by the same seed,

```
set.seed(02138)
runif(n = 10)
```

```
 [1] 0.51236144 0.61530551 0.37451441 0.43541258 0.21166530 0.17812129
 [7] 0.04420775 0.45567854 0.88718264 0.06970056
```

Exercises

Census Sampling

What can we learn from surveys of populations, and how wrong do we get if our sampling is biased?[5] Suppose we want to estimate the proportion of U.S. residents who are non-white (`race != "White"`). In reality, we do not have any population dataset to utilize and so we *only see the sample survey*. Here, however, to understand how sampling works, let's conveniently use the Census extract in some cases and pretend we didn't in others.

(a) First, load `usc2010_001percent.csv` into your R session. After loading the `library(tidyverse)`, browse it. Although this is only a 0.01 percent extract, treat this as your population for pedagogical purposes. What is the population proportion of non-White residents?

(b) Setting a seed to `1669482`, sample 100 respondents from this sample. What is the proportion of non-White residents in this *particular* sample? By how many percentage points are you off from (what we labelled as) the true proportion?

(c) Now imagine what you did above was one survey. What would we get if we did 20 surveys?

To simulate this, write a loop that does the same exercise 20 times, each time computing a sample proportion. Use the same seed at the top, but be careful to position the `set.seed` function such that it generates the same sequence of 20 samples, rather than 20 of the same sample.

Try doing this with a `for` loop and storing your sample proportions in a new length-20 vector. (Suggestion: make an empty vector first as a container). After running the loop, show a histogram of the 20 values. Also what is the average of the 20 sample estimates?

(d) Now, to make things more real, let's introduce some response bias. The goal here is not to correct response bias but to induce it and see how it affects our estimates. Suppose that non-White residents are 10 percent less likely to respond to enter your survey than White respondents. This is plausible if you think that the Census is from 2010 but you are polling in 2018, and racial minorities are more geographically mobile than Whites. Repeat the same exercise in (c) by modeling this behavior.

You can do this by creating a variable, e.g. `propensity`, that is 0.9 for non-Whites and 1 otherwise. Then, you can refer to it in the propensity argument.

(e) Finally, we want to see if more data ("Big Data") will improve our estimates. Using the same unequal response rates framework as (d), repeat the same exercise but instead of each poll collecting 100 responses, we collect 10,000.

(f) Optional - visualize your 2 pairs of 20 estimates, with a bar showing the "correct" population average.

[5]This example is inspired from Meng, Xiao-Li (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics* 12:2, 685–726. doi:10.1214/18-AOAS1161SF.

Conditional Proportions

This example is not on simulation, but is meant to reinforce some of the probability discussion from math lecture.

Read in the Upshot Siena poll from Fall 2016, `data/input/upshot-siena-polls.csv`.

In addition to some standard demographic questions, we will focus on one called `vt_pres_2` in the csv. This is a two-way presidential vote question, asking respondents who they plan to vote for President if the election were held today – Donald Trump, the Republican, or Hilary Clinton, the Democrat, with options for Other candidates as well. For this problem, use the two-way vote question rather than the 4-way vote question.

(a) Drop the the respondents who answered the November poll (i.e. those for which `poll == "November"`). We do this in order to ignore this November population in all subsequent parts of this question because they were not asked the Presidential vote question.

(b) Using the dataset after the procedure in (a), find the proportion of *poll respondents* (those who are in the sample) who support Donald Trump.

(c) Among those who supported Donald Trump, what proportion of them has a Bachelor's degree or higher (i.e. have a Bachelor's, Graduate, or other Professional Degree)?

(d) Among those who did not support Donald Trump (i.e. including supporters of Hilary Clinton, another candidate, or those who refused to answer the question), what proportion of them has a Bachelor's degree or higher?

(e) Express the numbers in the previous parts as probabilities of specified events. Define your own symbols: For example, we can let T be the event that a randomly selected respondent in the poll supports Donald Trump, then the proportion in part (b) is the probability $P(T)$.

(f) Suppose we randomly sampled a person who participated in the survey and found that he/she had a Bachelor's degree or higher. Given this evidence, what is the probability that the same person supports Donald Trump? Use Bayes Rule and show your work – that is, do not use data or R to compute the quantity directly. Then, verify this is the case via R.

The Birthday problem

Write code that will answer the well-known birthday problem via simulation.[6]

The problem is fairly simple: Suppose k people gather together in a room. What is the probability at least two people share the same birthday?

To simplify reality a bit, assume that (1) there are no leap years, and so there are always 365 days in a year, and (2) a given individual's birthday is randomly assigned and independent from each other.

Step 1: Set `k` to a concrete number. Pick a number from 1 to 365 randomly, `k` times to simulate birthdays (would this be with replacement or without?).

```
# Your code
```

Step 2: Write a line (or two) of code that gives a `TRUE` or `FALSE` statement of whether or not at least two people share the same birth date.

```
# Your code
```

[6] This exercise draws from Imai (2017)

Step 3: The above steps will generate a `TRUE` or `FALSE` answer for your event of interest, but only for one realization of an event in the sample space. In order to estimate the *probability* of your event happening, we need a "stochastic", as opposed to "deterministic", method. To do this, write a loop that does Steps 1 and 2 repeatedly for many times, call that number of times `sims`. For each of `sims` iteration, your code should give you a `TRUE` or `FALSE` answer. Code up a way to store these estimates.

```
# Your code
```

Step 4: Finally, generalize the function further by letting `k` be a user-defined number. You have now created a *Monte Carlo simulation*!

```
# Your code
```

Step 5: Generate a table or plot that shows how the probability of sharing a birthday changes by `k` (fixing `sims` at a large number like `1000`). Also generate a similar plot that shows how the probability of sharing a birthday changes by `sims` (fixing `k` at some arbitrary number like `10`).

```
# Your code
```

Extra credit: Give an "analytical" answer to this problem, that is an answer through deriving the mathematical expressions of the probability.

```
# Your equations
```

Chapter 5

LaTeX and Markdown

Where are we? Where are we headed?

Up till now, you should have covered:

- Statistical Programming in `R`

This is only the beginning of `R` – programming is like learning a language, so learn more as we use it. And yet `R` is of likely not the only programming language you will want to use. While we cannot introduce everything, we'll pick out a few that we think are particularly helpful.

Here will cover

- Markdown
- LaTeX (and BibTeX)

as examples of a non-WYSIWYG editor

command-line are a basic set of tools that you may have to use from time to time. It also clarifies what more complicated programs are doing. Markdown is an example of compiling a plain text file. LaTeX is a typesetting program and git is a version control program – both are useful for non-quantitative work as well.

Please familiarize yourself closing with Markdown, and be sure you know how to open an .Rmd file as described below. In class, we will walk through an Rmd file together. LaTeX is included here for your future reference as this is a popular typesetting program among political scientists. This is not needed for Math Camp and is never required for any course. In fact, many prefer R Markdown's integration rather than a separate typesetting program. This depends on your background and interests but exposure to the range of popular programs and techniques will be helpful moving forward.

Check your understanding

Check if you have an idea of how you might code the following tasks:

- What does "WYSIWYG" stand for? How would a non-WYSIWYG format text?
- How do you start a header in markdown?
- What are some "plain text" editors?
- How do you start a document in `.tex`?
- How do you start a environment in `.tex`?
- How do you insert a figure in `.tex`?
- How do you reference a figure in `.tex`?

- What is a `.bib` file?
- Say you came across a interesting journal article. How would you want to maintain this reference so that you can refer to its citation in all your subsequent papers?

5.1 Motivation

Statistical programming is a fast-moving field. The beta version of `R` was released in 2000, `ggplot2` was released on 2005, and `RStudio` started around 2010. Of course, some programming technologies are quite "old": (`C` in 1969, `C++` around 1989, `TeX` in 1978, `Linux` in 1991, Mac OS in 1984). But it is easy to feel you are falling behind in the recent developments of programming. Today we will do a **brief** and rough overview of some fundamental and new tools other than `R`, with the general aim of having you break out of your comfort zone so you won't be shut out from learning these tools in the future.

5.2 Markdown

At its core markdown is just plain text. Plain text does not have any formatting embedded in it. Instead, the formatting is coded up as text. Markdown is *not* a WYSIWYG (What you see is what you get) text editor like Microsoft Word or Google Docs. This will mean that you need to explicitly code for `bold{text}` rather than hitting Command+B and making your text look **bold** on your own computer.

Markdown is known as a "light-weight" editor, which means that it is relatively easy to write code that will compile. It is quick and easy and satisfies most presentation purposes; you might want to try `LaTeX` for more involved papers.

5.2.1 markdown commands

For italic and bold, use either the asterisks or the underlines,

```
*italic*   **bold**
_italic_   __bold__
```

And for headers use the hash symbols,

```
# Main Header
## Sub-headers
```

5.2.2 your own markdown

RStudio makes it easy to compile your very first markdown file by giving you templates. Got to `New > R Markdown`, pick a document and click Ok. This will give you a skeleton of a document you can compile – or "knit".

Rmd is actually a slight modification of real markdown. It is a type of file that R reads and turns into a proper `md` file. Then, it uses a document-conversion called pandoc to compile your `md` into documents like PDF or HTML.

5.2.3 A note on plain-text editors

Multiple software exist where you can edit plain-text (roughly speaking, text that is not WYSIWYG).

- RStudio (especially for R-related links)
- TeXMaker, TeXShop (especially for TeX)
- emacs, aquamacs (general)

Figure 5.1: How Rmds become PDFs or HTMLs

- vim (general)
- Sublime Text (general)
- Atom (general)

Each has their own keyboard shortcuts and special features. You can browse a couple and see which one(s) you like.

5.3 LaTeX

LaTeX is a typesetting program. You'd engage with LaTeX much like you engage with your `R` code. You will interact with LaTeX in a text editor, and will writing code which will be interpreted by the LaTeX compiler and which will finally be parsed to form your final PDF.

5.3.1 compile online

1. Go to https://www.overleaf.com
2. Scroll down and go to "CREATE A NEW PAPER" if you don't have an account.
3. Let's discuss the default template.
4. Make a new document, and set it as your main document. Then type in the Minimal Working Example (MWE):

```
\documentclass{article}
\begin{document}
Hello World
\end{document}
```

5.3.2 compile your first LaTeX document locally

LaTeX is a very stable system, and few changes to it have been made since the 1990s. The main benefit: better control over how your papers will look; better methods for writing equations or making tables; overall pleasing aesthetic.

1. Open a plain text editor. Then type in the MWE

```
\documentclass{article}
\begin{document}
Hello World
\end{document}
```

2. Save this as `hello_world.tex`. Make sure you get the file extension right.

3. Open this in your "LaTeX" editor. This can be `TeXMaker`, `Aqumacs`, etc..
4. Go through the click/dropdown interface and click compile.

5.3.3 main LaTeX commands

LaTeX can cover most of your typesetting needs, to clean equations and intricate diagrams.

Some main commands you'll be using are below, and a very concise cheat sheet here: https://wch.github.io/latexsheet/latexsheet.pdf

Most involved features require that you begin a specific "environment" for that feature, clearly demarcating them by the notation `\begin{figure}` and then `\end{figure}`, e.g. in the case of figures.

```
\begin{figure}
\includegraphics{histogram.pdf}
\end{figure}
```

where `histogram.pdf` is a path to one of your files.

Notice that each line starts with a backslash `\` – in LaTeX this is the symbol to run a command.

The following syntax at the endpoints are shorthand for math equations.

```
\[\int x^2 dx\]
```

these compile math symbols: $\int x^2 dx$.[1]

The `align` environment is useful to align your multi-line math, for example.

```
\begin{align}
P(A \mid B) &= \frac{P(A \cap B)}{P(B)}\\
&= \frac{P(B \mid A)P(A)}{P(B)}
\end{align}
```

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} && (5.1)\\ &= \frac{P(B \mid A)P(A)}{P(B)} && (5.2) \end{aligned}$$

Regression tables should be outputted as `.tex` files with packages like `xtable` and `stargazer`, and then called into LaTeX by `\input{regression_table.tex}` where `regression_table.tex` is the path to your regression output.

Figures and equations should be labelled with the tag (e.g. `label{tab:regression}` so that you can refer to them later with their tag `Table \ref{tab:regression}`, instead of hard-coding `Table 2`).

For some LaTeX commands you might need to load a separate package that someone else has written. Do this in your preamble (i.e. before `\begin{document}`):

```
\usepackage[options]{package}
```

where `package` is the name of the package and `options` are options specific to the package.

[1] Enclosing with `$$` instead of `\[` also has the same effect, so you may see it too. But this is now discouraged due to its inflexibility.

Further Guides

For a more comprehensive listing of LaTeX commands, Mayya Komisarchik has a great tutorial set of folders: https://scholar.harvard.edu/mkomisarchik/tutorials-0

There is a version of LaTeX called Beamer, which is a popular way of making a slideshow. Slides in markdown is also a competitor. The language of Beamer is the same as LaTeX but has some special functions for slides.

5.4 BibTeX

BibTeX is a reference system for bibliographical tests. We have a `.bib` file separately on our computer. This is also a plain text file, but it encodes bibliographical resources with special syntax so that a program can rearrange parts accordingly for different citation systems.

5.4.1 what is a `.bib` file?

For example, here is the Nunn and Wantchekon article entry in `.bib` form.

```
@article{nunn2011slave,
  title={The Slave Trade and the Origins of Mistrust in Africa},
  author={Nunn, Nathan and Wantchekon, Leonard},
  journal={American Economic Review},
  volume={101},
  number={7},
  pages={3221--3252},
  year={2011}
}
```

The first entry, `nunn2011slave`, is "pick your favorite" – pick your own name for your reference system. The other slots in this `@article` entry are entries that refer to specific bibliographical text.

5.4.2 what does LaTeX do with .bib files?

Now, in LaTeX, if you type

```
  \textcite{nunn2011slave} argue that current variation in the trust among citizens of African countries h
```

as part of your text, then when the `.tex` file is compiled the PDF shows something like

> Nunn and Wantchekon (2011) argue that current variation in the trust among citizens of African countries has historical roots in the European slave trade in the 1600s.

in whatever citation style (APSA, APA, Chicago) you pre-specified!

Also at the end of your paper you will have a bibliography with entries ordered and formatted in the appropriate citation.

Bibliography

Nunn, Nathan and Wantchekon, Leonard (2011). "The Slave Trade and the Origins of Mistrust in Africa". *American Economic Review* 101 (7), pp. 3221–3252.

This is a much less frustrating way of keeping track of your references – no need to hand-edit formatting the bibliography to conform to citation rules (which biblatex already knows) and no need to update your bibliography as you add and drop references (biblatex will only show entries that are used in the main text).

5.4.3 stocking up on your .bib files

You should keep your own `.bib` file that has all your bibliographical resources. Storing entries is cheap (does not take much memory), so it is fine to keep all your references in one place (but you'll want to make a new one for collaborative projects where multiple people will compile a `.tex` file).

For example, Gary's BibTeX file is here: https://github.com/iqss-research/gkbibtex/blob/master/gk.bib

Citation management software (Mendeley or Zotero) automatically generates .bib entries from your library of PDFs for you, provided you have the bibliography attributes right.

Extension: Optional Exercise

Create a LaTeX document for a hypothetical research paper on your laptop and, once you've verified it compiles into a PDF, come show it to either one of the instructors.

You can also use overleaf if you have preference for a cloud-based system. But don't swallow the built-in templates without understanding or testing them.

Each student will have slightly different substantive interests, so we won't impose much of a standard. But at a minimum, the LaTeX document should have:

- A title, author, date, and abstract
- Sections
- Italics and boldface
- A figure with a caption and in-text reference to it.

Depending on your subfield or interests, try to implement some of the following:

- A bibliographical reference drawing from a separate `.bib` file
- A table
- A math expression
- A different font
- Different page margins
- Different line spacing

Part II

II Linear Algebra

Chapter 6

Linear Algebra

6.1 Working with Vectors

Vector: A vector in n-space is an ordered list of n numbers. These numbers can be represented as either a row vector or a column vector:

$$\mathbf{v} = \begin{bmatrix} v_1 & v_2 & \dots & v_n \end{bmatrix}$$

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

We can also think of a vector as defining a point in n-dimensional space, usually $\mathbb{R}^n$; each element of the vector defines the coordinate of the point in a particular direction.

Vector Addition and Subtraction: If two vectors, $\mathbf{u}$ and $\mathbf{v}$, have the same length (i.e. have the same number of elements), they can be added (subtracted) together:

$$\mathbf{u} + \mathbf{v} = \begin{bmatrix} u_1 + v_1 \\ u_2 + v_2 \\ \vdots \\ u_k + v_n \end{bmatrix}$$

$$\mathbf{u} - \mathbf{v} = \begin{bmatrix} u_1 - v_1 \\ u_2 - v_2 \\ \vdots \\ u_k - v_n \end{bmatrix}$$

Scalar Multiplication: The product of a scalar c (i.e. a constant) and vector $\mathbf{v}$ is:

$$c\mathbf{v} = \begin{bmatrix} cv_1 \\ cv_2 \\ \cdots \\ cv_n \end{bmatrix}$$

Vector Inner Product: The inner product (also called the dot product or scalar product) of two vectors $\mathbf{u}$ and $\mathbf{v}$ is again defined if and only if they have the same number of elements

$$\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n = \sum_{i=1}^{n} u_i v_i$$

If $\mathbf{u} \cdot \mathbf{v} = 0$, the two vectors are orthogonal (or perpendicular).

Vector Norm: The norm of a vector is a measure of its length. There are many different ways to calculate the norm, but the most common is the Euclidean norm (which corresponds to our usual conception of distance in three-dimensional space):

$$||\mathbf{v}|| = \sqrt{\mathbf{v} \cdot \mathbf{v}} = \sqrt{v_1 v_1 + v_2 v_2 + \cdots + v_n v_n}$$

Example 6.1. Let $a = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}$, $b = \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}$. Calculate the following:

1. $a - b$
2. $a \cdot b$

Exercise 6.1. Let $u = \begin{bmatrix} 7 \\ 1 \\ -5 \\ 3 \end{bmatrix}$, $v = \begin{bmatrix} 9 \\ -3 \\ 2 \\ 8 \end{bmatrix}$, $w = \begin{bmatrix} 1 \\ 13 \\ -7 \\ 2 \\ 15 \end{bmatrix}$, and $c = 2$. Calculate the following:

1. $u - v$
2. cw
3. $u \cdot v$
4. $w \cdot v$

6.2 Linear Independence

Linear combinations: The vector $\mathbf{u}$ is a linear combination of the vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k$ if

$$\mathbf{u} = c_1 \mathbf{v}_1 + c_2 \mathbf{v}_2 + \cdots + c_k \mathbf{v}_k$$

For example, $\begin{bmatrix} 9 \\ 13 \\ 17 \end{bmatrix}$ is a linear combination of the following three vectors: $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$, $\begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$, and $\begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}$. This is because $\begin{bmatrix} 9 \\ 13 \\ 17 \end{bmatrix} = (2) \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + (-1) \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix} + 3 \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}$

Linear independence: A set of vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k$ is linearly independent if the only solution to the equation

$$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_k\mathbf{v}_k = 0$$

is $c_1 = c_2 = \cdots = c_k = 0$. If another solution exists, the set of vectors is linearly dependent.

A set S of vectors is linearly dependent if and only if at least one of the vectors in S can be written as a linear combination of the other vectors in S.

Linear independence is only defined for sets of vectors with the same number of elements; any linearly independent set of vectors in n-space contains at most n vectors.

Since $\begin{bmatrix} 9 \\ 13 \\ 17 \end{bmatrix}$ is a linear combination of $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$, $\begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}$, and $\begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}$, these 4 vectors constitute a linearly dependent set.

Example 6.2. Are the following sets of vectors linearly independent?

1. $\begin{bmatrix} 2 \\ 3 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} 4 \\ 6 \\ 1 \end{bmatrix}$
2. $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix}$, and $\begin{bmatrix} 10 \\ 10 \\ 0 \end{bmatrix}$

Exercise 6.2. Are the following sets of vectors linearly independent?

1. $$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

2. $$\mathbf{v}_1 = \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}, \mathbf{v}_2 = \begin{bmatrix} -4 \\ 6 \\ 5 \end{bmatrix}, \mathbf{v}_3 = \begin{bmatrix} -2 \\ 8 \\ 6 \end{bmatrix}$$

6.3 Basics of Matrix Algebra

Matrix: A matrix is an array of real numbers arranged in m rows by n columns. The dimensionality of the matrix is defined as the number of rows by the number of columns, $m \times n$.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

Note that you can think of vectors as special cases of matrices; a column vector of length k is a $k \times 1$ matrix, while a row vector of the same length is a $1 \times k$ matrix.

It's also useful to think of matrices as being made up of a collection of row or column vectors. For example,

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_\mathrm{m} \end{bmatrix}$$

Matrix Addition: Let $\mathbf{A}$ and $\mathbf{B}$ be two $m \times n$ matrices.

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \cdots & a_{mn} + b_{mn} \end{bmatrix}$$

Note that matrices $\mathbf{A}$ and $\mathbf{B}$ must have the same dimensionality, in which case they are **conformable for addition**.

Example 6.3.

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 2 \end{bmatrix}$$

Find $\mathbf{A} + \mathbf{B}$

Scalar Multiplication: Given the scalar s, the scalar multiplication of $s\mathbf{A}$ is

$$s\mathbf{A} = s \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = \begin{bmatrix} sa_{11} & sa_{12} & \cdots & sa_{1n} \\ sa_{21} & sa_{22} & \cdots & sa_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ sa_{m1} & sa_{m2} & \cdots & sa_{mn} \end{bmatrix}$$

Example 6.4. $s = 2$,

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$$

Find $s\mathbf{A}$

Matrix Multiplication: If $\mathbf{A}$ is an $m \times k$ matrix and $\mathbf{B}$ is a $k \times n$ matrix, then their product $\mathbf{C} = \mathbf{AB}$ is the $m \times n$ matrix where

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{ik}b_{kj}$$

Example 6.5.

1. Find $\begin{bmatrix} a & b \\ c & d \\ e & f \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix}$

2. Find $\begin{bmatrix} 1 & 2 & -1 \\ 3 & 1 & 4 \end{bmatrix} \begin{bmatrix} -2 & 5 \\ 4 & -3 \\ 2 & 1 \end{bmatrix}$

Note that the number of columns of the first matrix must equal the number of rows of the second matrix, in which case they are **conformable for multiplication**. The sizes of the matrices (including the resulting product) must be

$$(m \times k)(k \times n) = (m \times n)$$

Also note that if $\mathbf{AB}$ exists, $\mathbf{BA}$ exists only if $\dim(\mathbf{A}) = m \times n$ and $\dim(\mathbf{B}) = n \times m$.

This does not mean that $\mathbf{AB} = \mathbf{BA}$. $\mathbf{AB} = \mathbf{BA}$ is true only in special circumstances, like when $\mathbf{A}$ or $\mathbf{B}$ is an identity matrix or $\mathbf{A} = \mathbf{B}^{-1}$.

6.3.1 Laws of Matrix Algebra

1. Associative: $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$
2. $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
3. Commutative: $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
4. Distributive: $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
5. $(\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC}$

Commutative law for multiplication does not hold – the order of multiplication matters:

$$\mathbf{AB} \neq \mathbf{BA}$$

For example,

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ -1 & 3 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$$

$$\mathbf{AB} = \begin{bmatrix} 2 & 3 \\ -2 & 2 \end{bmatrix}, \qquad \mathbf{BA} = \begin{bmatrix} 1 & 7 \\ -1 & 3 \end{bmatrix}$$

Transpose: The transpose of the $m \times n$ matrix $\mathbf{A}$ is the $n \times m$ matrix $\mathbf{A}^\top$ (also written $\mathbf{A}'$) obtained by interchanging the rows and columns of $\mathbf{A}$.

For example,

$$\mathbf{A} = \begin{bmatrix} 4 & -2 & 3 \\ 0 & 5 & -1 \end{bmatrix}, \qquad \mathbf{A}^\top = \begin{bmatrix} 4 & 0 \\ -2 & 5 \\ 3 & -1 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 2 \\ -1 \\ 3 \end{bmatrix}, \qquad \mathbf{B}^\top = \begin{bmatrix} 2 & -1 & 3 \end{bmatrix}$$

The following rules apply for transposed matrices:

1. $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$
2. $(\mathbf{A}^\top)^\top = \mathbf{A}$
3. $(s\mathbf{A})^\top = s\mathbf{A}^\top$
4. $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$; and by induction $(\mathbf{ABC})^\top = \mathbf{C}^\top\mathbf{B}^\top\mathbf{A}^\top$

Example of $(\mathbf{AB})^\top = \mathbf{B}^\top\mathbf{A}^\top$:

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 2 \\ 2 & -1 & 3 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 & 1 \\ 2 & 2 \\ 3 & -1 \end{bmatrix}$$

$$(\mathbf{AB})^\top = \left[\begin{bmatrix} 1 & 3 & 2 \\ 2 & -1 & 3 \end{bmatrix}\begin{bmatrix} 0 & 1 \\ 2 & 2 \\ 3 & -1 \end{bmatrix}\right]^\top = \begin{bmatrix} 12 & 7 \\ 5 & -3 \end{bmatrix}$$

$$\mathbf{B}^\top\mathbf{A}^\top = \begin{bmatrix} 0 & 2 & 3 \\ 1 & 2 & -1 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 3 & -1 \\ 2 & 3 \end{bmatrix} = \begin{bmatrix} 12 & 7 \\ 5 & -3 \end{bmatrix}$$

Exercise 6.3. Let

$$A = \begin{bmatrix} 2 & 0 & -1 & 1 \\ 1 & 2 & 0 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & 5 & -7 \\ 1 & 1 & 0 \\ 0 & -1 & 1 \\ 2 & 0 & 0 \end{bmatrix}$$

$$C = \begin{bmatrix} 3 & 2 & -1 \\ 0 & 4 & 6 \end{bmatrix}$$

Calculate the following:

1. $$AB$$

2. $$BA$$

3. $$(BC)^\top$$

4. $$BC^\top$$

6.4 Systems of Linear Equations

6.4.1 Linear Equation

Linear equations take form of $a_1x_1 + a_2x_2 + \cdots + a_nx_n = b$

a_i are parameters or coefficients. x_i are variables or unknowns.

Linear because only one variable per term and degree is at most 1.

We are often interested in solving linear systems like

$$\begin{cases} x - 3y &= -3 \\ 2x + y &= 8 \end{cases}$$

More generally, we might have a system of m equations in n unknowns

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ \vdots & & & & \vdots & & & & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m \end{array}$$

A **solution** to a linear system of m equations in n unknowns is a set of n numbers $x_1, x_2, \cdots, x_n$ that satisfy each of the m equations.

Example: $x = 3$ and $y = 2$ is the solution to the above 2×2 linear system. If you graph the two lines, you will find that they intersect at $(3, 2)$.

Does a linear system have one, no, or multiple solutions? For a system of 2 equations with 2 unknowns (i.e., two lines):

- **One solution:** The lines intersect at exactly one point.
- **No solution:** The lines are parallel.
- **Infinite solutions:** The lines coincide.

Methods to solve linear systems:

1. Substitution
2. Elimination of variables
3. Matrix methods

Exercise 6.4. Provide a system of 2 equations with 2 unknowns that has

1. one solution
2. no solution
3. infinite solutions

6.5 Systems of Equations as Matrices

Matrices provide an easy and efficient way to represent linear systems such as

$$\begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & \cdots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \cdots & + & a_{2n}x_n & = & b_2 \\ \vdots & & & & \vdots & & & & \vdots \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \cdots & + & a_{mn}x_n & = & b_m \end{array}$$

as

$$\mathbf{Ax} = \mathbf{b}$$

where

The $m \times n$ **coefficient matrix A** is an array of mn real numbers arranged in m rows by n columns:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

The unknown quantities are represented by the vector $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$.

The right hand side of the linear system is represented by the vector $\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$.

Augmented Matrix: When we append $\mathbf{b}$ to the coefficient matrix $\mathbf{A}$, we get the augmented matrix $\widehat{\mathbf{A}} = [\mathbf{A}|\mathbf{b}]$

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & | & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & | & b_2 \\ \vdots & & \ddots & \vdots & | & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & | & b_m \end{bmatrix}$$

Exercise 6.5. Create an augmented matrix that represent the following system of equations:

$$2x_1 - 7x_2 + 9x_3 - 4x_4 = 8$$

$$41x_2 + 9x_3 - 5x_6 = 11$$

$$x_1 - 15x_2 - 11x_5 = 9$$

6.6 Finding Solutions to Augmented Matrices and Systems of Equations

Row Echelon Form: Our goal is to translate our augmented matrix or system of equations into row echelon form. This will provide us with the values of the vector $\mathbf{x}$ which solve the system. We use the row operations to change coefficients in the lower triangle of the augmented matrix to 0. An augmented matrix of the form

$$\begin{bmatrix} \boxed{a'_{11}} & a'_{12} & a'_{13} & \cdots & a'_{1n} & | & b'_1 \\ 0 & \boxed{a'_{22}} & a'_{23} & \cdots & a'_{2n} & | & b'_2 \\ 0 & 0 & \boxed{a'_{33}} & \cdots & a'_{3n} & | & b'_3 \\ 0 & 0 & 0 & \ddots & \vdots & | & \vdots \\ 0 & 0 & 0 & 0 & \boxed{a'_{mn}} & | & b'_m \end{bmatrix}$$

is said to be in row echelon form — each row has more leading zeros than the row preceding it.

Reduced Row Echelon Form: We can go one step further and put the matrix into reduced row echelon form. Reduced row echelon form makes the value of $\mathbf{x}$ which solves the system very obvious. For a system of m equations in m unknowns, with no all-zero rows, the reduced row echelon form would be

$$\begin{bmatrix} \boxed{1} & 0 & 0 & 0 & 0 & | & b_1^* \\ 0 & \boxed{1} & 0 & 0 & 0 & | & b_2^* \\ 0 & 0 & \boxed{1} & 0 & 0 & | & b_3^* \\ 0 & 0 & 0 & \ddots & 0 & | & \vdots \\ 0 & 0 & 0 & 0 & \boxed{1} & | & b_m^* \end{bmatrix}$$

Gaussian and Gauss-Jordan elimination: We can conduct elementary row operations to get our augmented matrix into row echelon or reduced row echelon form. The methods of transforming a matrix or system into row echelon and reduced row echelon form are referred to as Gaussian elimination and Gauss-Jordan elimination, respectively.

Elementary Row Operations: To do Gaussian and Gauss-Jordan elimination, we use three basic operations to transform the augmented matrix into another augmented matrix that represents an equivalent linear system – equivalent in the sense that the same values of x_j solve both the original and transformed matrix/system:

Interchanging Rows: Suppose we have the augmented matrix

$$\widehat{\mathbf{A}} = \begin{bmatrix} a_{11} & a_{12} & | & b_1 \\ a_{21} & a_{22} & | & b_2 \end{bmatrix}$$

If we interchange the two rows, we get the augmented matrix

$$\begin{bmatrix} a_{21} & a_{22} & | & b_2 \\ a_{11} & a_{12} & | & b_1 \end{bmatrix}$$

which represents a linear system equivalent to that represented by matrix $\widehat{\mathbf{A}}$.

Multiplying by a Constant: If we multiply the second row of matrix $\widehat{\mathbf{A}}$ by a constant c, we get the augmented matrix

$$\begin{bmatrix} a_{11} & a_{12} & | & b_1 \\ ca_{21} & ca_{22} & | & cb_2 \end{bmatrix}$$

which represents a linear system equivalent to that represented by matrix $\widehat{\mathbf{A}}$.

Adding (subtracting) Rows: If we add (subtract) the first row of matrix $\widehat{\mathbf{A}}$ to the second, we obtain the augmented matrix

$$\begin{bmatrix} a_{11} & a_{12} & | & b_1 \\ a_{11} + a_{21} & a_{12} + a_{22} & | & b_1 + b_2 \end{bmatrix}$$

which represents a linear system equivalent to that represented by matrix $\widehat{\mathbf{A}}$.

Example 6.6. Solve the following system of equations by using elementary row operations:

$$\begin{array}{ccccc} x & - & 3y & = & -3 \\ 2x & + & y & = & 8 \end{array}$$

Exercise 6.6. Put the following system of equations into augmented matrix form. Then, using Gaussian or Gauss-Jordan elimination, solve the system of equations by putting the matrix into row echelon or reduced row echelon form.

1.
$$\begin{cases} x + y + 2z = 2 \\ 3x - 2y + z = 1 \\ y - z = 3 \end{cases}$$

2.

$$\begin{cases} 2x + 3y - z = -8 \\ x + 2y - z = 12 \\ -x - 4y + z = -6 \end{cases}$$

6.7 Rank of a Matrix

To determine how many solutions exist, we can use information about (1) the number of equations m, (2) the number of unknowns n, and (3) the **rank** of the matrix representing the linear system.

Rank: The maximum number of linearly independent row or column vectors in the matrix. This is equivalent to the number of nonzero rows of a matrix in row echelon form. For any matrix $\mathbf{A}$, the row rank always equals column rank, and we refer to this number as the rank of $\mathbf{A}$.

For example

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 6 \end{bmatrix}$$

Rank = 3

$$\begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 5 \\ 0 & 0 & 0 \end{bmatrix}$$

Rank = 2

Exercise 6.7. Find the rank of each matrix below:

(Hint: transform the matrices into row echelon form. Remember that the number of nonzero rows of a matrix in row echelon form is the rank of that matrix)

1.

$$\begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 3 \\ 1 & 2 & 3 \end{bmatrix}$$

2.

$$\begin{bmatrix} 1 & 3 & 3 & -3 & 3 \\ 1 & 3 & 1 & 1 & 3 \\ 1 & 3 & 2 & -1 & -2 \\ 1 & 3 & 0 & 3 & -2 \end{bmatrix}$$

Answer to Exercise **??**:

1. rank is 2
2. rank is 3

6.8 The Inverse of a Matrix

Identity Matrix: The $n \times n$ identity matrix $\mathbf{I}_n$ is the matrix whose diagonal elements are 1 and all off-diagonal elements are 0. Examples:

$$\mathbf{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{I}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Inverse Matrix: An $n \times n$ matrix $\mathbf{A}$ is **nonsingular** or **invertible** if there exists an $n \times n$ matrix $\mathbf{A}^{-1}$ such that

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}_n$$

where $\mathbf{A}^{-1}$ is the inverse of $\mathbf{A}$. If there is no such $\mathbf{A}^{-1}$, then $\mathbf{A}$ is singular or not invertible.

Example: Let

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 2 & 2 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} -1 & \frac{3}{2} \\ 1 & -1 \end{bmatrix}$$

Since

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$$

we conclude that $\mathbf{B}$ is the inverse, $\mathbf{A}^{-1}$, of $\mathbf{A}$ and that $\mathbf{A}$ is nonsingular.

Properties of the Inverse:

- If the inverse exists, it is unique.
- If $\mathbf{A}$ is nonsingular, then $\mathbf{A}^{-1}$ is nonsingular.
- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
- If $\mathbf{A}$ and $\mathbf{B}$ are nonsingular, then $\mathbf{AB}$ is nonsingular
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
- If $\mathbf{A}$ is nonsingular, then $(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top$

Procedure to Find $\mathbf{A}^{-1}$: We know that if $\mathbf{B}$ is the inverse of $\mathbf{A}$, then

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$$

Looking only at the first and last parts of this

$$\mathbf{AB} = \mathbf{I}_n$$

Solving for $\mathbf{B}$ is equivalent to solving for n linear systems, where each column of $\mathbf{B}$ is solved for the corresponding column in $\mathbf{I}_n$. We can solve the systems simultaneously by augmenting $\mathbf{A}$ with $\mathbf{I}_n$ and performing Gauss-Jordan elimination on $\mathbf{A}$. If Gauss-Jordan elimination on $[\mathbf{A}|\mathbf{I}_n]$ results in $[\mathbf{I}_n|\mathbf{B}]$, then $\mathbf{B}$ is the inverse of $\mathbf{A}$. Otherwise, $\mathbf{A}$ is singular.

To summarize: To calculate the inverse of $\mathbf{A}$

1. Form the augmented matrix $[\mathbf{A}|\mathbf{I}_n]$
2. Using elementary row operations, transform the augmented matrix to reduced row echelon form.
3. The result of step 2 is an augmented matrix $[\mathbf{C}|\mathbf{B}]$.
   a. If $\mathbf{C} = \mathbf{I}_n$, then $\mathbf{B} = \mathbf{A}^{-1}$.
   b. If $\mathbf{C} \neq \mathbf{I}_n$, then $\mathbf{C}$ has a row of zeros. This means $\mathbf{A}$ is singular and $\mathbf{A}^{-1}$ does not exist.

Example 6.7. Find the inverse of the following matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 3 \\ 5 & 5 & 1 \end{bmatrix}$$

Exercise 6.8. Find the inverse of the following matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 4 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

6.9 Linear Systems and Inverses

Let's return to the matrix representation of a linear system

$$\mathbf{Ax} = \mathbf{b}$$

If $\mathbf{A}$ is an $n \times n$ matrix,then $\mathbf{Ax} = \mathbf{b}$ is a system of n equations in n unknowns. Suppose $\mathbf{A}$ is nonsingular. Then $\mathbf{A^{-1}}$ exists. To solve this system, we can multiply each side by $\mathbf{A^{-1}}$ and reduce it as follows:

$$\begin{aligned} \mathbf{A^{-1}}(\mathbf{Ax}) &= \mathbf{A^{-1}b} \\ (\mathbf{A^{-1}A})\mathbf{x} &= \mathbf{A^{-1}b} \\ \mathbf{I_n x} &= \mathbf{A^{-1}b} \\ \mathbf{x} &= \mathbf{A^{-1}b} \end{aligned}$$

Hence, given $\mathbf{A}$ and $\mathbf{b}$ and given that $\mathbf{A}$ is nonsingular, then $\mathbf{x} = \mathbf{A^{-1}b}$ is a unique solution to this system.

Exercise 6.9. Use the inverse matrix to solve the following linear system:

$$\begin{aligned} -3x + 4y &= 5 \\ 2x - y &= -10 \end{aligned}$$

Hint: the linear system above can be written in the matrix form

$\mathbf{Az} = \mathbf{b}$

given

$$\mathbf{A} = \begin{bmatrix} -3 & 4 \\ 2 & -1 \end{bmatrix},$$

$$\mathbf{z} = \begin{bmatrix} x \\ y \end{bmatrix},$$

and

$$\mathbf{b} = \begin{bmatrix} 5 \\ -10 \end{bmatrix}$$

6.10 Determinants

Singularity: Determinants can be used to determine whether a square matrix is nonsingular.

A square matrix is nonsingular if and only if its determinant is not zero.

Determinant of a 1×1 matrix, equals $|\mathbf{A}| = |a_{11}| = a_{11}$

Determinant of a 2×2 matrix,

$$\mathbf{A} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

$$\begin{aligned} \det(\mathbf{A}) &= |\mathbf{A}| \\ &= a_{11}|a_{22}| - a_{12}|a_{21}| \\ &= a_{11}a_{22} - a_{12}a_{21} \end{aligned}$$

We can extend the second to last equation above to get the definition of the determinant of a 3×3 matrix:

$$\begin{aligned} \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} &= a_{11} \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12} \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13} \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix} \\ &= a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}) \end{aligned}$$

Let's extend this now to any $n \times n$ matrix. Let's define $\mathbf{A}_{ij}$ as the $(n-1) \times (n-1)$ submatrix of $\mathbf{A}$ obtained by deleting row i and column j. Let the (i, j)th **minor** of $\mathbf{A}$ be the determinant of $\mathbf{A}_{ij}$:

$$M_{ij} = |\mathbf{A}_{ij}|$$

Then for any $n \times n$ matrix $\mathbf{A}$

$$|\mathbf{A}| = a_{11}M_{11} - a_{12}M_{12} + \cdots + (-1)^{n+1}a_{1n}M_{1n}$$

For example, in figuring out whether the following matrix has an inverse?

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 3 \\ 5 & 5 & 1 \end{bmatrix}$$

1. Calculate its determinant.

$$\begin{aligned} |\mathbf{A}| &= 1(2-15) - 1(0-15) + 1(0-10) \\ &= -13 + 15 - 10 \\ &= -8 \end{aligned}$$

2. Since $|\mathbf{A}| \neq 0$, we conclude that $\mathbf{A}$ has an inverse.

Exercise 6.10. Determine whether the following matrices are nonsingular:

1.
$$\begin{bmatrix} 1 & 0 & 1 \\ 2 & 1 & 2 \\ 1 & 0 & -1 \end{bmatrix}$$

2.
$$\begin{bmatrix} 2 & 1 & 2 \\ 1 & 0 & 1 \\ 4 & 1 & 4 \end{bmatrix}$$

6.11 Getting Inverse of a Matrix using its Determinant

Thus far, we have a number of algorithms to

1. Find the solution of a linear system,
2. Find the inverse of a matrix

but these remain just that — algorithms. At this point, we have no way of telling how the solutions x_j change as the parameters a_{ij} and b_i change, except by changing the values and "rerunning" the algorithms.

With determinants, we can provide an explicit formula for the inverse and therefore provide an explicit formula for the solution of an $n \times n$ linear system.

Hence, we can examine how changes in the parameters and b_i affect the solutions x_j.

Determinant Formula for the Inverse of a 2×2:

The determinant of a 2×2 matrix $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is defined as:

$$\frac{1}{\det(\mathbf{A})} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

For example, Let's calculate the inverse of matrix A from Exercise **??** using the determinant formula.

Recall,

$$\mathbf{A} = \begin{bmatrix} -3 & 4 \\ 2 & -1 \end{bmatrix}$$

$$\det(\mathbf{A}) = (-3)(-1) - (4)(2) = 3 - 8 = -5$$

$$\frac{1}{\det(\mathbf{A})} \begin{bmatrix} -1 & -4 \\ -2 & -3 \end{bmatrix}$$

$$\frac{1}{-5} \begin{bmatrix} -1 & -4 \\ -2 & -3 \end{bmatrix}$$

$$\begin{bmatrix} \frac{1}{5} & \frac{4}{5} \\ \frac{2}{5} & \frac{3}{5} \end{bmatrix}$$

Exercise 6.11. Caculate the inverse of **A**

$$\mathbf{A} = \begin{bmatrix} 3 & 5 \\ -7 & 2 \end{bmatrix}$$

Answers to Examples and Exercises

Answer to Example **??**:

1. $\begin{bmatrix} -1 & -3 & -3 \end{bmatrix}$
2. $6 + 4 + 10 = 20$

Answer to Exercise **??**:

1. $\begin{bmatrix} -2 & 4 & -7 & -5 \end{bmatrix}$
2. $\begin{bmatrix} 2 & 26 & -14 & 4 & 30 \end{bmatrix}$
3. 63 -3 -10 + 24 = 74
4. undefined

Answer to Example **??**:

1. yes
2. no

Answer to Exercise **??**:

1. yes
2. no $(-v_1 - v_2 + v_3 = 0)$

Answer to Example **??**:

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 2 & 4 & 4 \\ 6 & 6 & 8 \end{bmatrix}$$

Answer to Example **??**:

$$s\mathbf{A} = \begin{bmatrix} 2 & 4 & 6 \\ 8 & 10 & 12 \end{bmatrix}$$

Answer to Example **??**:

1. $\begin{bmatrix} aA + bC & aB + bD \\ cA + dC & cB + dD \\ eA + fC & eB + fD \end{bmatrix}$
2. $\begin{bmatrix} 1(-2) + 2(4) - 1(2) & 1(5) + 2(-3) - 1(1) \\ 3(-2) + 1(4) + 4(2) & 3(5) + 1(-3) + 4(1) \end{bmatrix} = \begin{bmatrix} 4 & -2 \\ 6 & 16 \end{bmatrix}$

Answer to Exercise **??**:

1. $AB = \begin{bmatrix} 4 & 11 & -15 \\ 5 & 7 & -7 \end{bmatrix}$
2. $BA =$ undefined
3. $(BC)^\top =$ undefined
4. $BC^\top = \begin{bmatrix} 1 & 5 & -7 \\ 1 & 1 & 0 \\ 0 & -1 & 1 \\ 2 & 0 & 0 \end{bmatrix} \begin{bmatrix} 3 & 0 \\ 2 & 4 \\ -1 & 6 \end{bmatrix} = \begin{bmatrix} 20 & -22 \\ 5 & 4 \\ -3 & 2 \\ 6 & 0 \end{bmatrix}$

Answer to Exercise **??**:

There are many answers to this. Some possible simple ones are as follows:

1. One solution:
$$\begin{array}{rcrcl} -x & + & y & = & 0 \\ x & + & y & = & 2 \end{array}$$

2. No solution:
$$\begin{array}{rcrcl} -x & + & y & = & 0 \\ x & - & y & = & 2 \end{array}$$

3. Infinite solutions:
$$\begin{array}{rcrcl} -x & + & y & = & 0 \\ 2x & - & 2y & = & 0 \end{array}$$

Answer to Exercise **??**:

$$\begin{bmatrix} 2 & -7 & 9 & -4 & 0 & 0| & 8 \\ 0 & 41 & 9 & 0 & 0 & 5| & 11 \\ 1 & -15 & 0 & 0 & -11 & 0| & 9 \end{bmatrix}$$

Answer to Example **??**:

$$\begin{array}{rcrcr} x & - & 3y & = & -3 \\ 2x & + & y & = & 8 \end{array}$$

$$\begin{array}{rcrcr} x & - & 3y & = & -3 \\ & & 7y & = & 14 \end{array}$$

$$\begin{array}{rcrcr} x & - & 3y & = & -3 \\ & & y & = & 2 \end{array}$$

$$\begin{array}{rcl} x & = & 3 \\ y & = & 2 \end{array}$$

Answer to Exercise **??**:

1. x = 2, y = 2, z = -1
2. x = -17, y = -3, z = -35

Answer to Exercise **??**:

1. rank is 2
2. rank is 3

Answer to Example **??**:

$$\left(\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 2 & 3 & 0 & 1 & 0 \\ 5 & 5 & 1 & 0 & 0 & 1 \end{array}\right)$$

$$\left(\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 2 & 3 & 0 & 1 & 0 \\ 0 & 0 & -4 & -5 & 0 & 1 \end{array}\right)$$

$$\left(\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 2 & 3 & 0 & 1 & 0 \\ 0 & 0 & 1 & 5/4 & 0 & -1/4 \end{array}\right)$$

$$\left(\begin{array}{ccc|ccc} 1 & 1 & 0 & -1/4 & 0 & 1/4 \\ 0 & 2 & 0 & -15/4 & 1 & 3/4 \\ 0 & 0 & 1 & 5/4 & 0 & -1/4 \end{array}\right)$$

$$\left(\begin{array}{ccc|ccc} 1 & 1 & 0 & -1/4 & 0 & 1/4 \\ 0 & 1 & 0 & -15/8 & 1/2 & 3/8 \\ 0 & 0 & 1 & 5/4 & 0 & -1/4 \end{array}\right)$$

$$\left(\begin{array}{ccc|ccc} 1 & 0 & 0 & 13/8 & -1/2 & -1/8 \\ 0 & 1 & 0 & -15/8 & 1/2 & 3/8 \\ 0 & 0 & 1 & 5/4 & 0 & -1/4 \end{array}\right)$$

$$\mathbf{A}^{-1} = \begin{pmatrix} 13/8 & -1/2 & -1/8 \\ -15/8 & 1/2 & 3/8 \\ 5/4 & 0 & -1/4 \end{pmatrix}$$

Answer to Exercise **??**:

1. $\mathbf{A}^{-1} = \begin{bmatrix} 1 & 0 & -4 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 \end{bmatrix}$

Answer to Exercise **??**:

$$\mathbf{z} = \mathbf{A^{-1}b} = \begin{bmatrix} 1/5 & 4/5 \\ 2/5 & 3/5 \end{bmatrix} \begin{bmatrix} 5 \\ -10 \end{bmatrix} = \begin{bmatrix} -7 \\ -4 \end{bmatrix} = \begin{bmatrix} \mathrm{x} \\ \mathrm{y} \end{bmatrix}$$

Answer to Exercise **??**:

1. nonsingular
2. singular

Answer to Exercise **??**:

$$\begin{bmatrix} \frac{2}{41} & \frac{-5}{41} \\ \frac{7}{41} & \frac{3}{41} \end{bmatrix}$$

Chapter 7

Manipulating Vectors and Matrices

Motivation

Nunn and Wantchekon (2011) – "The Slave Trade and the Origins of Mistrust in Africa"[1] – argues that across African countries, the distrust of co-ethnics fueled by the slave trade has had long-lasting effects on modern day trust in these territories. They argued that the slave trade created distrust in these societies in part because as some African groups were employed by European traders to capture their neighbors and bring them to the slave ships.

Nunn and Wantchekon use a variety of statistical tools to make their case (adding controls, ordered logit, instrumental variables, falsification tests, causal mechanisms), many of which will be covered in future courses. In this module we will only touch on their first set of analysis that use Ordinary Least Squares (OLS). OLS is likely the most common application of linear algebra in the social sciences. We will cover some linear algebra, matrix manipulation, and vector manipulation from this data.

Where are we? Where are we headed?

Up till now, you should have covered:

- R basic programming
- Data Import
- Data Visualization
- R Markdown

Today we'll cover

- Matrices & Dataframes in R
- Manipulating variables
- And other `R` tips

[1] Nunn, Nathan, and Leonard Wantchekon. 2011. "The Slave Trade and the Origins of Mistrust in Africa." American Economic Review 101(7): 3221–52.

7.1 Read Data

```r
library(haven)
nunn_full <- read_dta("data/input/Nunn_Wantchekon_AER_2011.dta")
```

Nunn and Wantchekon's main dataset has more than 20,000 observations. Each observation is a respondent from the Afrobarometer survey.

```r
head(nunn_full)
```

```
# A tibble: 6 x 59
  respno  ethni~1 murdo~2 isocode region distr~3 townv~4 locat~5 trust~6 trust~7
  <chr>   <chr>   <chr>   <chr>   <chr>  <chr>   <chr>     <dbl>   <dbl>   <dbl>
1 BEN0001 fon     FON     BEN     atlna~ KPOMAS~ TOKPA-~      30       3       3
2 BEN0002 fon     FON     BEN     atlna~ KPOMAS~ TOKPA-~      30       3       3
3 BEN0003 fon     FON     BEN     atlna~ OUIDAH  3ARROND      31       0       0
4 BEN0004 fon     FON     BEN     atlna~ OUIDAH  3ARROND      31       0       0
5 BEN0005 fon     FON     BEN     atlna~ OUIDAH  PAHOU        32       1       1
6 BEN0006 fon     FON     BEN     atlna~ OUIDAH  PAHOU        32       1       1
# ... with 49 more variables: intra_group_trust <dbl>, inter_group_trust <dbl>,
#   trust_local_council <dbl>, ln_export_area <dbl>, export_area <dbl>,
#   export_pop <dbl>, ln_export_pop <dbl>, age <dbl>, age2 <dbl>, male <dbl>,
#   urban_dum <dbl>, occupation <dbl>, religion <dbl>, living_conditions <dbl>,
#   education <dbl>, near_dist <dbl>, distsea <dbl>, loc_murdock_name <chr>,
#   loc_ln_export_area <dbl>, local_council_performance <dbl>,
#   council_listen <dbl>, corrupt_local_council <dbl>, ...
# i Use `colnames()` to see all variable names
```

```r
colnames(nunn_full)
```

```
 [1] "respno"                    "ethnicity"
 [3] "murdock_name"              "isocode"
 [5] "region"                    "district"
 [7] "townvill"                  "location_id"
 [9] "trust_relatives"           "trust_neighbors"
[11] "intra_group_trust"         "inter_group_trust"
[13] "trust_local_council"       "ln_export_area"
[15] "export_area"               "export_pop"
[17] "ln_export_pop"             "age"
[19] "age2"                      "male"
[21] "urban_dum"                 "occupation"
[23] "religion"                  "living_conditions"
[25] "education"                 "near_dist"
[27] "distsea"                   "loc_murdock_name"
[29] "loc_ln_export_area"        "local_council_performance"
[31] "council_listen"            "corrupt_local_council"
[33] "school_present"            "electricity_present"
[35] "piped_water_present"       "sewage_present"
[37] "health_clinic_present"     "district_ethnic_frac"
```

```
[39] "frac_ethnicity_in_district"      "townvill_nonethnic_mean_exports"
[41] "district_nonethnic_mean_exports" "region_nonethnic_mean_exports"
[43] "country_nonethnic_mean_exports"  "murdock_centr_dist_coast"
[45] "centroid_lat"                    "centroid_long"
[47] "explorer_contact"                "railway_contact"
[49] "dist_Saharan_node"               "dist_Saharan_line"
[51] "malaria_ecology"                 "v30"
[53] "v33"                             "fishing"
[55] "exports"                         "ln_exports"
[57] "total_missions_area"             "ln_init_pop_density"
[59] "cities_1400_dum"
```

First, let's consider a small subset of this dataset.

```
nunn <- read_dta("data/input/Nunn_Wantchekon_sample.dta")
```

```
nunn
```

```
# A tibble: 10 x 5
   trust_neighbors exports ln_exports export_area ln_export_area
             <dbl>   <dbl>      <dbl>       <dbl>          <dbl>
 1               3   0.388      0.328     0.00407        0.00406
 2               3   0.631      0.489     0.0971         0.0926
 3               3   0.994      0.690     0.0125         0.0124
 4               0 183.         5.21      1.82           1.04
 5               3   0          0         0              0
 6               2   0          0         0              0
 7               2 666.         6.50     14.0            2.71
 8               0   0.348      0.298     0.00608        0.00606
 9               3   0.435      0.361     0.0383         0.0376
10               3   0          0         0              0
```

7.2 data.frame vs. matricies

This is a `data.frame` object.

```
class(nunn)
```

```
[1] "tbl_df"     "tbl"        "data.frame"
```

But it can be also consider a matrix in the linear algebra sense. What are the dimensions of this matrix?

```
nrow(nunn)
```

```
[1] 10
```

`data.frame`s and matrices have much overlap in `R`, but to explicitly treat an object as a matrix, you'd need to coerce its class. Let's call this matrix `X`.

```
X <- as.matrix(nunn)
```

What is the difference between a `data.frame` and a matrix? A `data.frame` can have columns that are of different types, whereas — in a matrix — all columns must be of the same type (usually either “numeric” or “character”).

You can think of data frames maybe as matrices-plus, because a column can take on characters as well as numbers. As we just saw, this is often useful for real data analyses.

Another way to think about data frames is that it is a type of list. Try the `str()` code below and notice how it is organized in slots. Each slot is a vector. They can be vectors of numbers or characters.

```
# enter this on your console
str(cen10)
```

7.3 Handling matricies in R

You can easily transpose a matrix

```
X
```

```
      trust_neighbors     exports ln_exports  export_area ln_export_area
 [1,]               3   0.3883497  0.3281158  0.004067405    0.004059155
 [2,]               3   0.6311236  0.4892691  0.097059444    0.092633367
 [3,]               3   0.9941893  0.6902376  0.012524694    0.012446908
 [4,]               0 182.5891266  5.2127004  1.824284434    1.038255095
 [5,]               3   0.0000000  0.0000000  0.000000000    0.000000000
 [6,]               2   0.0000000  0.0000000  0.000000000    0.000000000
 [7,]               2 665.9652100  6.5027380 13.975566864    2.706419945
 [8,]               0   0.3476418  0.2983562  0.006082553    0.006064130
 [9,]               3   0.4349871  0.3611559  0.038332380    0.037615947
[10,]               3   0.0000000  0.0000000  0.000000000    0.000000000
```

```
t(X)
```

```
                       [,1]       [,2]       [,3]       [,4] [,5] [,6]
trust_neighbors 3.000000000 3.00000000 3.00000000   0.000000    3    2
exports         0.388349682 0.63112360 0.99418926 182.589127    0    0
ln_exports      0.328115761 0.48926911 0.69023758   5.212700    0    0
export_area     0.004067405 0.09705944 0.01252469   1.824284    0    0
ln_export_area  0.004059155 0.09263337 0.01244691   1.038255    0    0
                      [,7]        [,8]       [,9] [,10]
trust_neighbors   2.000000 0.000000000 3.00000000     3
exports         665.965210 0.347641766 0.43498713     0
ln_exports        6.502738 0.298356235 0.36115587     0
export_area      13.975567 0.006082553 0.03833238     0
ln_export_area    2.706420 0.006064130 0.03761595     0
```

What are the values of all rows in the first column?

```
X[, 1]
```

```
[1] 3 3 3 0 3 2 2 0 3 3
```

What are all the values of "exports"? (i.e. return the whole "exports" column)

```
X[, "exports"]
```

```
 [1]   0.3883497   0.6311236   0.9941893 182.5891266   0.0000000   0.0000000
 [7] 665.9652100   0.3476418   0.4349871   0.0000000
```

What is the first observation (i.e. first row)?

```
X[1, ]
```

```
trust_neighbors         exports      ln_exports     export_area  ln_export_area
    3.000000000     0.388349682     0.328115761     0.004067405     0.004059155
```

What is the value of the first variable of the first observation?

```
X[1, 1]
```

```
trust_neighbors
              3
```

Pause and consider the following problem on your own. What is the following code doing?

```
X[X[, "trust_neighbors"] == 0, "export_area"]
```

```
[1] 1.824284434 0.006082553
```

Why does it give the same output as the following?

```
X[which(X[, "trust_neighbors"] == 0), "export_area"]
```

```
[1] 1.824284434 0.006082553
```

Some more manipulation

```
X + X
```

```
      trust_neighbors      exports ln_exports  export_area ln_export_area
 [1,]               6    0.7766994  0.6562315  0.008134809     0.00811831
 [2,]               6    1.2622472  0.9785382  0.194118887     0.18526673
 [3,]               6    1.9883785  1.3804752  0.025049388     0.02489382
 [4,]               0  365.1782532 10.4254007  3.648568869     2.07651019
 [5,]               6    0.0000000  0.0000000  0.000000000     0.00000000
 [6,]               4    0.0000000  0.0000000  0.000000000     0.00000000
 [7,]               4 1331.9304199 13.0054760 27.951133728     5.41283989
 [8,]               0    0.6952835  0.5967125  0.012165107     0.01212826
 [9,]               6    0.8699743  0.7223117  0.076664761     0.07523189
```

```
[10,]               6    0.0000000  0.0000000  0.000000000     0.00000000
```

```
X - X
```

```
      trust_neighbors exports ln_exports export_area ln_export_area
 [1,]               0       0          0           0              0
 [2,]               0       0          0           0              0
 [3,]               0       0          0           0              0
 [4,]               0       0          0           0              0
 [5,]               0       0          0           0              0
 [6,]               0       0          0           0              0
 [7,]               0       0          0           0              0
 [8,]               0       0          0           0              0
 [9,]               0       0          0           0              0
[10,]               0       0          0           0              0
```

```
t(X) %*% X
```

```
                trust_neighbors    exports ln_exports export_area
trust_neighbors       62.000000   1339.276   18.61181    28.40709
exports             1339.276369 476850.298 5283.76294  9640.42990
ln_exports            18.611811   5283.763   70.50077   100.46202
export_area           28.407085   9640.430  100.46202   198.65558
ln_export_area         5.853106   1992.047   23.08189    39.72847
                ln_export_area
trust_neighbors       5.853106
exports            1992.046502
ln_exports           23.081893
export_area          39.728468
ln_export_area        8.412887
```

```
cbind(X, 1:10)
```

```
      trust_neighbors     exports ln_exports  export_area ln_export_area
 [1,]               3   0.3883497  0.3281158  0.004067405    0.004059155  1
 [2,]               3   0.6311236  0.4892691  0.097059444    0.092633367  2
 [3,]               3   0.9941893  0.6902376  0.012524694    0.012446908  3
 [4,]               0 182.5891266  5.2127004  1.824284434    1.038255095  4
 [5,]               3   0.0000000  0.0000000  0.000000000    0.000000000  5
 [6,]               2   0.0000000  0.0000000  0.000000000    0.000000000  6
 [7,]               2 665.9652100  6.5027380 13.975566864    2.706419945  7
 [8,]               0   0.3476418  0.2983562  0.006082553    0.006064130  8
 [9,]               3   0.4349871  0.3611559  0.038332380    0.037615947  9
[10,]               3   0.0000000  0.0000000  0.000000000    0.000000000 10
```

```
cbind(X, 1)
```

```
      trust_neighbors     exports ln_exports  export_area ln_export_area
 [1,]               3   0.3883497  0.3281158  0.004067405    0.004059155 1
```

```
 [2,]               3   0.6311236  0.4892691  0.097059444    0.092633367 1
 [3,]               3   0.9941893  0.6902376  0.012524694    0.012446908 1
 [4,]               0 182.5891266  5.2127004  1.824284434    1.038255095 1
 [5,]               3   0.0000000  0.0000000  0.000000000    0.000000000 1
 [6,]               2   0.0000000  0.0000000  0.000000000    0.000000000 1
 [7,]               2 665.9652100  6.5027380 13.975566864    2.706419945 1
 [8,]               0   0.3476418  0.2983562  0.006082553    0.006064130 1
 [9,]               3   0.4349871  0.3611559  0.038332380    0.037615947 1
[10,]               3   0.0000000  0.0000000  0.000000000    0.000000000 1
```

```r
colnames(X)
```

```
[1] "trust_neighbors" "exports"         "ln_exports"      "export_area"
[5] "ln_export_area"
```

7.4 Variable Transformations

`exports` is the total number of slaves that were taken from the individual's ethnic group between Africa's four slave trades between 1400-1900.

What is `ln_exports`? The article describes this as the natural log of one plus the `exports`. This is a transformation of one column by a particular function

```r
log(1 + X[, "exports"])
```

```
 [1] 0.3281158 0.4892691 0.6902376 5.2127003 0.0000000 0.0000000 6.5027379
 [8] 0.2983562 0.3611559 0.0000000
```

Question for you: why add the 1?

Verify that this is the same as `X[, "ln_exports"]`

7.5 Linear Combinations

In Table 1 we see "OLS Estimates". These are estimates of OLS coefficients and standard errors. You do not need to know what these are for now, but it doesn't hurt to getting used to seeing them.

Table 1—OLS Estimates of the Determinants of Trust in Neighbors

Dependent variable: Trust of neighbors	Slave exports (thousands) (1)	Exports/ area (2)	Exports/ historical pop (3)	ln (1+ exports) (4)	ln (1+ exports/ area) (5)	ln (1+ exports/ historical pop) (6)
Estimated coefficient	−0.00068	−0.019	−0.531	−0.037	−0.159	−0.743
	[0.00014]	[0.005]	[0.147]	[0.014]	[0.034]	[0.187]
	(0.00015)	(0.005)	(0.147)	(0.014)	(0.034)	(0.187)
	{0.00013}	{0.005}	{0.165}	{0.015}	{0.034}	{0.212}
Individual controls	Yes	Yes	Yes	Yes	Yes	Yes
District controls	Yes	Yes	Yes	Yes	Yes	Yes
Country fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Number of observations	20,027	20,027	17,644	20,027	20,027	17,644
Number of ethnicities	185	185	157	185	185	157
Number of districts	1,257	1,257	1,214	1,257	1,257	1,214
R^2	0.16	0.16	0.15	0.15	0.16	0.15

Notes: The table reports OLS estimates. The unit of observation is an individual. Below each coefficient three standard errors are reported. The first, reported in square brackets, is standard errors adjusted for clustering within ethnic groups. The second, reported in parentheses, is standard errors adjusted for two-way clustering within ethnic groups and within districts. The third, reported in curly brackets, is T. G. Conley (1999) standard errors adjusted for two-dimensional spatial autocorrelation. The standard errors are constructed assuming a window with weights equal to one for observations less than five degrees apart and zero for observations further apart. The individual controls are for age, age squared, a gender indicator variable, five living conditions fixed effects, ten education fixed effects, 18 religion fixed effects, 25 occupation fixed effects, and an indicator for whether the respondent lives in an urban location. The district controls include ethnic fractionalization of each district and the share of the district's population that is the same ethnicity as the respondent.

A very crude way to describe regression is through linear combinations. The simplest linear combination is a one-to-one transformation.

Take the first number in Table 1, which is -0.00068. Now, multiply this by `exports`

```
-0.00068 * X[, "exports"]
```

```
 [1] -0.0002640778 -0.0004291640 -0.0006760487 -0.1241606061  0.0000000000
 [6]  0.0000000000 -0.4528563428 -0.0002363964 -0.0002957912  0.0000000000
```

Now, just one more step. Make a new matrix with just exports and the value 1

```
X2 <- cbind(1, X[, "exports"])
```

name this new column "intercept"

```
colnames(X2)
```

```
NULL
```

```r
colnames(X2) <- c("intercept", "exports")
```

What are the dimensions of the matrix X2?

```r
dim(X2)
```

```
[1] 10  2
```

Now consider a new matrix, called B.

```r
B <- matrix(c(1.62, -0.00068))
```

What are the dimensions of B?

```r
dim(B)
```

```
[1] 2 1
```

What is the product of X2 and B? From the dimensions, can you tell if it will be conformable?

```r
X2 %*% B
```

```
          [,1]
 [1,] 1.619736
 [2,] 1.619571
 [3,] 1.619324
 [4,] 1.495839
 [5,] 1.620000
 [6,] 1.620000
 [7,] 1.167144
 [8,] 1.619764
 [9,] 1.619704
[10,] 1.620000
```

What is this multiplication doing in terms of equations?

7.6 Matrix Basics

Let's take a look at Matrices in the context of R

```r
cen10 <- read_csv("data/input/usc2010_001percent.csv")
head(cen10)
```

```
# A tibble: 6 x 4
  state         sex      age race 
  <chr>         <chr>  <dbl> <chr>
1 New York      Female     8 White
2 Ohio          Male      24 White
```

```
3 Nevada        Male      37 White
4 Michigan      Female    12 White
5 Maryland      Female    18 Black/Negro
6 New Hampshire Male      50 White
```

What is the dimension of this dataframe? What does the number of rows represent? What does the number of columns represent?

```
dim(cen10)
```

```
[1] 30871     4
```

```
nrow(cen10)
```

```
[1] 30871
```

```
ncol(cen10)
```

```
[1] 4
```

What variables does this dataset hold? What kind of information does it have?

```
colnames(cen10)
```

```
[1] "state" "sex"   "age"   "race"
```

We can access column vectors, or vectors that contain values of variables by using the $ sign

```
head(cen10$state)
```

```
[1] "New York"      "Ohio"          "Nevada"        "Michigan"
[5] "Maryland"      "New Hampshire"
```

```
head(cen10$race)
```

```
[1] "White"       "White"       "White"       "White"       "Black/Negro"
[6] "White"
```

We can look at a unique set of variable values by calling the unique function

```
unique(cen10$state)
```

```
 [1] "New York"              "Ohio"                  "Nevada"
 [4] "Michigan"              "Maryland"              "New Hampshire"
 [7] "Iowa"                  "Missouri"              "New Jersey"
[10] "California"            "Texas"                 "Pennsylvania"
[13] "Washington"            "West Virginia"         "Idaho"
[16] "North Carolina"        "Massachusetts"         "Connecticut"
[19] "Arkansas"              "Indiana"               "Wisconsin"
[22] "Maine"                 "Tennessee"             "Minnesota"
```

```
[25] "Florida"              "Oklahoma"             "Montana"
[28] "Georgia"              "Arizona"              "Colorado"
[31] "Virginia"             "Illinois"             "Oregon"
[34] "Kentucky"             "South Carolina"       "Kansas"
[37] "Louisiana"            "Alabama"              "District of Columbia"
[40] "Mississippi"          "Utah"                 "Delaware"
[43] "Nebraska"             "Alaska"               "New Mexico"
[46] "South Dakota"         "Hawaii"               "Vermont"
[49] "Rhode Island"         "Wyoming"              "North Dakota"
```

How many different states are represented (this dataset includes DC as a state)?

```
length(unique(cen10$state))
```

```
[1] 51
```

Matrices are rectangular structures of numbers (they have to be numbers, and they can't be characters).

A cross-tab can be considered a matrix:

```
table(cen10$race, cen10$sex)
```

```
                                  Female  Male
  American Indian or Alaska Native    142   153
  Black/Negro                        2070  1943
  Chinese                             192   162
  Japanese                             51    26
  Other Asian or Pacific Islander     587   542
  Other race, nec                     877   962
  Three or more major races            37    51
  Two major races                     443   426
  White                             11252 10955
```

```
cross_tab <- table(cen10$race, cen10$sex)
dim(cross_tab)
```

```
[1] 9 2
```

```
cross_tab[6, 2]
```

```
[1] 962
```

But a subset of your data – individual values– can be considered a matrix too.

```
# First 20 rows of the entire data
# Below two lines of code do the same thing
cen10[1:20, ]
```

```
# A tibble: 20 x 4
   state          sex      age race
   <chr>          <chr>  <dbl> <chr>
```

```
 1 New York      Female     8 White
 2 Ohio          Male      24 White
 3 Nevada        Male      37 White
 4 Michigan      Female    12 White
 5 Maryland      Female    18 Black/Negro
 6 New Hampshire Male      50 White
 7 Iowa          Female    51 White
 8 Missouri      Female    41 White
 9 New Jersey    Male      62 White
10 California    Male      25 White
11 Texas         Female    23 White
12 Pennsylvania  Female    66 White
13 California    Female    57 White
14 Texas         Female    73 Other race, nec
15 California    Male      43 White
16 Washington    Male      29 White
17 Texas         Male       8 White
18 Missouri      Male      78 White
19 West Virginia Male      10 White
20 Idaho         Female     9 White
```

```
cen10 %>% slice(1:20)
```

```
# A tibble: 20 x 4
   state         sex      age race
   <chr>         <chr>  <dbl> <chr>
 1 New York      Female     8 White
 2 Ohio          Male      24 White
 3 Nevada        Male      37 White
 4 Michigan      Female    12 White
 5 Maryland      Female    18 Black/Negro
 6 New Hampshire Male      50 White
 7 Iowa          Female    51 White
 8 Missouri      Female    41 White
 9 New Jersey    Male      62 White
10 California    Male      25 White
11 Texas         Female    23 White
12 Pennsylvania  Female    66 White
13 California    Female    57 White
14 Texas         Female    73 Other race, nec
15 California    Male      43 White
16 Washington    Male      29 White
17 Texas         Male       8 White
18 Missouri      Male      78 White
19 West Virginia Male      10 White
20 Idaho         Female     9 White
```

```
# Of the first 20 rows of the entire data, look at values of just race and age
# Below two lines of code do the same thing
cen10[1:20, c("race", "age")]
```

```
# A tibble: 20 x 2
   race              age
   <chr>           <dbl>
 1 White               8
 2 White              24
 3 White              37
 4 White              12
 5 Black/Negro        18
 6 White              50
 7 White              51
 8 White              41
 9 White              62
10 White              25
11 White              23
12 White              66
13 White              57
14 Other race, nec    73
15 White              43
16 White              29
17 White               8
18 White              78
19 White              10
20 White               9
```

```
cen10 %>% slice(1:20) %>% select(race, age)
```


```
# A tibble: 20 x 2
   race              age
   <chr>           <dbl>
 1 White               8
 2 White              24
 3 White              37
 4 White              12
 5 Black/Negro        18
 6 White              50
 7 White              51
 8 White              41
 9 White              62
10 White              25
11 White              23
12 White              66
13 White              57
14 Other race, nec    73
15 White              43
16 White              29
17 White               8
18 White              78
19 White              10
20 White               9
```


A vector is a special type of matrix with only one column or only one row

```
# One column
cen10[1:10, c("age")]
```

```
# A tibble: 10 x 1
     age
   <dbl>
 1     8
 2    24
 3    37
 4    12
 5    18
 6    50
 7    51
 8    41
 9    62
10    25
```

```
cen10 %>% slice(1:10) %>% select(c("age"))
```

```
# A tibble: 10 x 1
     age
   <dbl>
 1     8
 2    24
 3    37
 4    12
 5    18
 6    50
 7    51
 8    41
 9    62
10    25
```

```
# One row
cen10[2, ]
```

```
# A tibble: 1 x 4
  state sex     age race
  <chr> <chr> <dbl> <chr>
1 Ohio  Male     24 White
```

```
cen10 %>% slice(2)
```

```
# A tibble: 1 x 4
  state sex     age race
  <chr> <chr> <dbl> <chr>
1 Ohio  Male     24 White
```

What if we want a special subset of the data? For example, what if I only want the records of individuals in California? What if I just want the age and race of individuals in California?

```
# subset for CA rows
ca_subset <- cen10[cen10$state == "California", ]

ca_subset_tidy <- cen10 %>% filter(state == "California")

all_equal(ca_subset, ca_subset_tidy)
```

```
[1] TRUE
```

```
# subset for CA rows and select age and race
ca_subset_age_race <- cen10[cen10$state == "California", c("age", "race")]

ca_subset_age_race_tidy <- cen10 %>% filter(state == "California") %>% select(age, race)

all_equal(ca_subset_age_race, ca_subset_age_race_tidy)
```

```
[1] TRUE
```

Some common operators that can be used to filter or to use as a condition. Remember, you can use the unique function to look at the set of all values a variable holds in the dataset.

```
# all individuals older than 30 and younger than 70
s1 <- cen10[cen10$age > 30 & cen10$age < 70, ]
s2 <- cen10 %>% filter(age > 30 & age < 70)
all_equal(s1, s2)
```

```
[1] TRUE
```

```
# all individuals in either New York or California
s3 <- cen10[cen10$state == "New York" | cen10$state == "California", ]
s4 <- cen10 %>% filter(state == "New York" | state == "California")
all_equal(s3, s4)
```

```
[1] TRUE
```

```
# all individuals in any of the following states: California, Ohio, Nevada, Michigan
s5 <- cen10[cen10$state %in% c("California", "Ohio", "Nevada", "Michigan"), ]
s6 <- cen10 %>% filter(state %in% c("California", "Ohio", "Nevada", "Michigan"))
all_equal(s5, s6)
```

```
[1] TRUE
```

```
# all individuals NOT in any of the following states: California, Ohio, Nevada, Michigan
s7 <- cen10[!(cen10$state %in% c("California", "Ohio", "Nevada", "Michigan")), ]
s8 <- cen10 %>% filter(!state %in% c("California", "Ohio", "Nevada", "Michigan"))
all_equal(s7, s8)
```

```
[1] TRUE
```

Part III

III Calculus

Chapter 8

Limits

Solving limits, i.e. finding out the value of functions as its input moves closer to some value, is important for the social scientist's mathematical toolkit for two related tasks. The first is for the study of calculus, which will be in turn useful to show where certain functions are maximized or minimized. The second is for the study of statistical inference, which is the study of inferring things about things you cannot see by using things you can see.

Example: The Central Limit Theorem

Perhaps the most important theorem in statistics is the Central Limit Theorem,

Theorem 8.1 (Central Limit Theorem (i.i.d. case))**.** *For any series of independent and identically distributed random variables $X_1, X_2, \cdots$, we know the distribution of its sum even if we do not know the distribution of X. The distribution of the sum is a Normal distribution.*

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Normal(0, 1),$$

where μ is the mean of X and σ is the standard deviation of X. The arrow is read as "converges in distribution to". $Normal(0, 1)$ indicates a Normal Distribution with mean 0 and variance 1.

That is, the limit of the distribution of the lefthand side is the distribution of the righthand side.

The sign of a limit is the arrow "$\rightarrow$". Although we have not yet covered probability so we have not described what distributions and random variables are, it is worth foreshadowing the Central Limit Theorem. The Central Limit Theorem is powerful because it gives us a *guarantee* of what would happen if $n \rightarrow \infty$, which in this case means we collected more data.

Example: The Law of Large Numbers

A finding that perhaps rivals the Central Limit Theorem is the Law of Large Numbers:

Theorem 8.2 ((Weak) Law of Large Numbers). *For any draw of identically distributed independent variables with mean μ, the sample average after n draws, $\bar{X}_n$, converges in probability to the true mean as $n \to \infty$:*

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0$$

A shorthand of which is $\bar{X}_n \xrightarrow{p} \mu$, where the arrow is read as "converges in probability to".

Intuitively, the more data, the more accurate is your guess. For example, Figure **??** shows how the sample average from many coin tosses converges to the true value : 0.5.

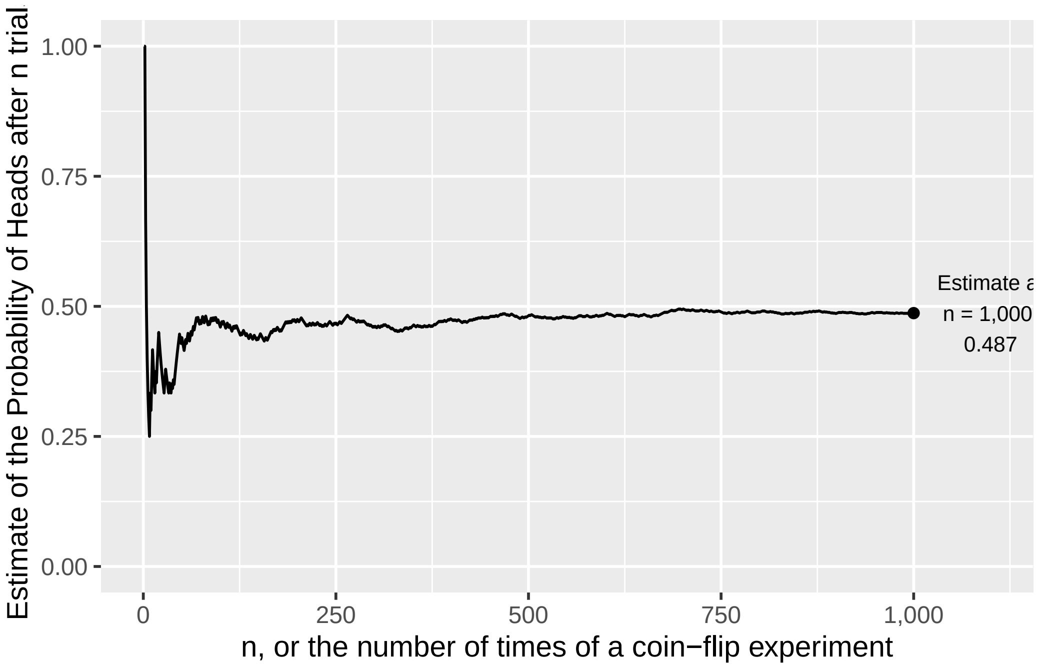


Figure 8.1: As the number of coin tosses goes to infinity, the average probabiity of heads converges to 0.5

8.1 Sequences

We need a couple of steps until we get to limit theorems in probability. First we will introduce a "sequence", then we will think about the limit of a sequence, then we will think about the limit of a *function.*

A **sequence**

$$\{x_n\} = \{x_1, x_2, x_3, \dots, x_n\}$$

is an ordered set of real numbers, where x_1 is the first term in the sequence and y_n is the nth term. Generally, a sequence is infinite, that is it extends to $n = \infty$. We can also write the sequence as

$$\{x_n\}_{n=1}^{\infty}$$

where the subscript and superscript are read together as "from 1 to infinity."

Example 8.1. How do these sequences behave?

1. $\{A_n\} = \left\{2 - \frac{1}{n^2}\right\}$
2. $\{B_n\} = \left\{\frac{n^2+1}{n}\right\}$
3. $\{C_n\} = \left\{(-1)^n\left(1 - \frac{1}{n}\right)\right\}$

We find the sequence by simply "plugging in" the integers into each n. The important thing is to get a sense of how these numbers are going to change. Example 1's numbers seem to come closer and closer to 2, but will it ever surpass 2? Example 2's numbers are also increasing each time, but will it hit a limit? What is the pattern in Example 3? Graphing helps you make this point more clearly. See the sequence of $n = 1, ...20$ for each of the three examples in Figure **??**.

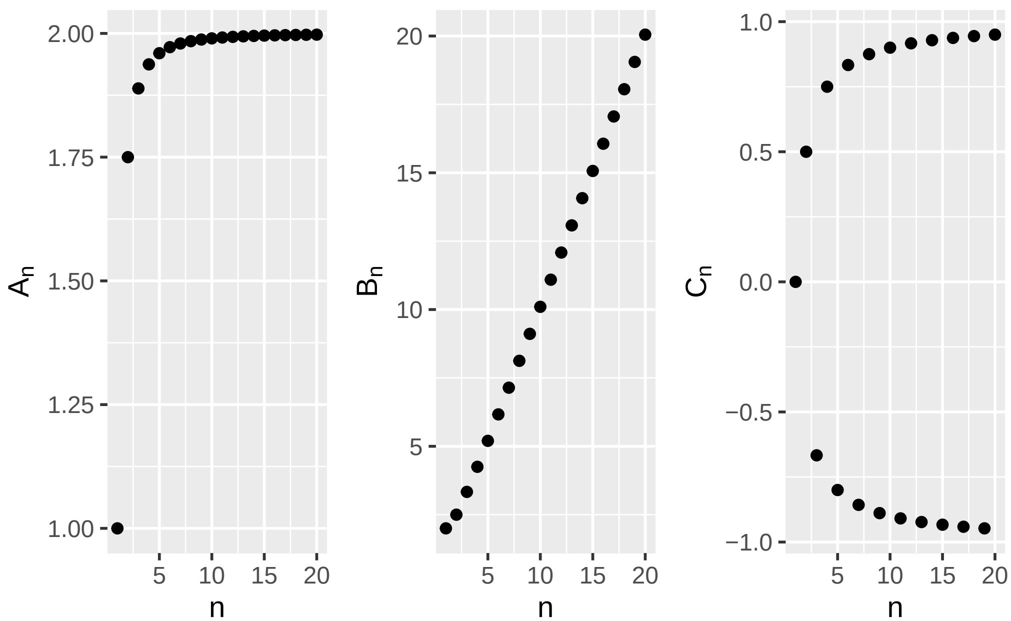


Figure 8.2: Behavior of Some Sequences

8.2 The Limit of a Sequence

The notion of "converging to a limit" is the behavior of the points in Example **??**. In some sense, that's the counterfactual we want to know. What happens as $n \to \infty$?

1. Sequences like 1 above that converge to a limit.
2. Sequences like 2 above that increase without bound.
3. Sequences like 3 above that neither converge nor increase without bound — alternating over the number line.

Definition 8.1 (Limit of a Sequence)**.** The sequence $\{y_n\}$ has the limit L, which we write as

$$\lim_{n\to\infty} y_n = L,$$

if for any $\epsilon > 0$ there is an integer N (which depends on ϵ) with the property that $|y_n - L| < \epsilon$ for each $n > N$. $\{y_n\}$ is said to converge to L. If the above does not hold, then $\{y_n\}$ diverges.

We can also express the behavior of a sequence as bounded or not:

1. Bounded: if $|y_n| \leq K$ for all n
2. Monotonically Increasing: $y_{n+1} > y_n$ for all n
3. Monotonically Decreasing: $y_{n+1} < y_n$ for all n

A limit is *unique*: If $\{y_n\}$ converges, then the limit L is unique.

If a sequence converges, then the sum of such sequences also converges. Let $\lim\limits_{n\to\infty} y_n = y$ and $\lim\limits_{n\to\infty} z_n = z$. Then

1. $\lim\limits_{n\to\infty} [ky_n + \ell z_n] = ky + \ell z$
2. $\lim\limits_{n\to\infty} y_n z_n = yz$
3. $\lim\limits_{n\to\infty} \frac{y_n}{z_n} = \frac{y}{z}$, provided $z \neq 0$

This looks reasonable enough. The harder question, obviously is when the parts of the fraction *don't* converge. If $\lim_{n\to\infty} y_n = \infty$ and $\lim_{n\to\infty} z_n = \infty$, What is $\lim_{n\to\infty} y_n - z_n$? What is $\lim_{n\to\infty} \frac{y_n}{z_n}$?

It is nice for a sequence to converge in limit. We want to know if complex-looking sequences converge or not. The name of the game here is to break that complex sequence up into sums of simple fractions where n only appears in the denominator: $\frac{1}{n}, \frac{1}{n^2}$, and so on. Each of these will converge to 0, because the denominator gets larger and larger. Then, because of the properties above, we can then find the final sequence.

Example 8.2. Find the limit of

$$\lim_{n\to\infty} \frac{n+3}{n}.$$

At first glance, $n+3$ and n both grow to ∞, so it looks like we need to divide infinity by infinity. However, we can express this fraction as a sum, then the limits apply separately:

$$\lim_{n\to\infty} \frac{n+3}{n} = \lim_{n\to\infty} \left(1 + \frac{3}{n}\right) = \underbrace{\lim_{n\to\infty} 1}_{1} + \underbrace{\lim_{n\to\infty} \left(\frac{3}{n}\right)}_{0}$$

so, the limit is actually 1.

After some practice, the key to intuition is whether one part of the fraction grows "faster" than another. If the denominator grows faster to infinity than the numerator, then the fraction will converge to 0, even if the numerator will also increase to infinity. In a sense, limits show how not all infinities are the same.

Exercise 8.1. Find the following limits of sequences, then explain in English the intuition for why that is the case.

1. $\lim\limits_{n\to\infty} \frac{2n}{n^2+1}$
2. $\lim\limits_{n\to\infty} (n^3 - 100n^2)$

8.3 Limits of a Function

We've now covered functions and just covered limits of sequences, so now is the time to combine the two.

A function f is a compact representation of some behavior we care about. Like for sequences, we often want to know if $f(x)$ approaches some number L as its independent variable x moves to some number c (which is usually 0 or $\pm\infty$). If it does, we say that the limit of $f(x)$, as x approaches c, is L: $\lim_{x\to c} f(x) = L$. Unlike a sequence, x is a continuous number, and we can move in decreasing order as well as increasing.

For a limit L to exist, the function $f(x)$ must approach L from both the left (increasing) and the right (decreasing).

Definition 8.2 (Limit of a function)**.** Let $f(x)$ be defined at each point in some open interval containing the point c. Then L equals $\lim_{x\to c} f(x)$ if for any (small positive) number ϵ, there exists a corresponding number $\delta > 0$ such that if $0 < |x - c| < \delta$, then $|f(x) - L| < \epsilon$.

A neat, if subtle result is that $f(x)$ does not necessarily have to be defined at c for $\lim_{x\to c}$ to exist.

Proposition 8.1. *Let f and g be functions with $\lim_{x\to c} f(x) = k$ and $\lim_{x\to c} g(x) = \ell$.*

1. $\lim_{x\to c}[f(x) + g(x)] = \lim_{x\to c} f(x) + \lim_{x\to c} g(x)$
2. $\lim_{x\to c} kf(x) = k \lim_{x\to c} f(x)$
3. $\lim_{x\to c} f(x)g(x) = \left[\lim_{x\to c} f(x)\right] \cdot \left[\lim_{x\to c} g(x)\right]$
4. $\lim_{x\to c} \frac{f(x)}{g(x)} = \frac{\lim_{x\to c} f(x)}{\lim_{x\to c} g(x)}$, *provided* $\lim_{x\to c} g(x) \neq 0$.

Simple limits of functions can be solved as we did limits of sequences. Just be careful which part of the function is changing.

Example 8.3. Find the limit of the following functions.

1. $\lim_{x\to c} k$
2. $\lim_{x\to c} x$
3. $\lim_{x\to 2}(2x - 3)$
4. $\lim_{x\to c} x^n$

Limits can get more complex in roughly two ways. First, the functions may become large polynomials with many moving pieces. Second,the functions may become discontinuous.

The function can be thought of as a more general or "smooth" version of sequences. For example,

Exercise 8.2. Find the limit of

$$\lim_{x\to\infty} \frac{(x^4 + 3x - 99)(2 - x^5)}{(18x^7 + 9x^6 - 3x^2 - 1)(x + 1)}$$

Now, the functions will become a bit more complex:

Exercise 8.3. Solve the following limits of functions

1. $\lim_{x\to 0} |x|$
2. $\lim_{x\to 0} \left(1 + \frac{1}{x^2}\right)$

So there are a few more alternatives about what a limit of a function could be:

1. Right-hand limit: The value approached by $f(x)$ when you move from right to left.

2. Left-hand limit: The value approached by $f(x)$ when you move from left to right.
3. Infinity: The value approached by $f(x)$ as x grows infinitely large. Sometimes this may be a number; sometimes it might be ∞ or $-\infty$.
4. Negative infinity: The value approached by $f(x)$ as x grows infinitely negative. Sometimes this may be a number; sometimes it might be ∞ or $-\infty$.

The distinction between left and right becomes important when the function is not determined for some values of x. What are those cases in the examples below?

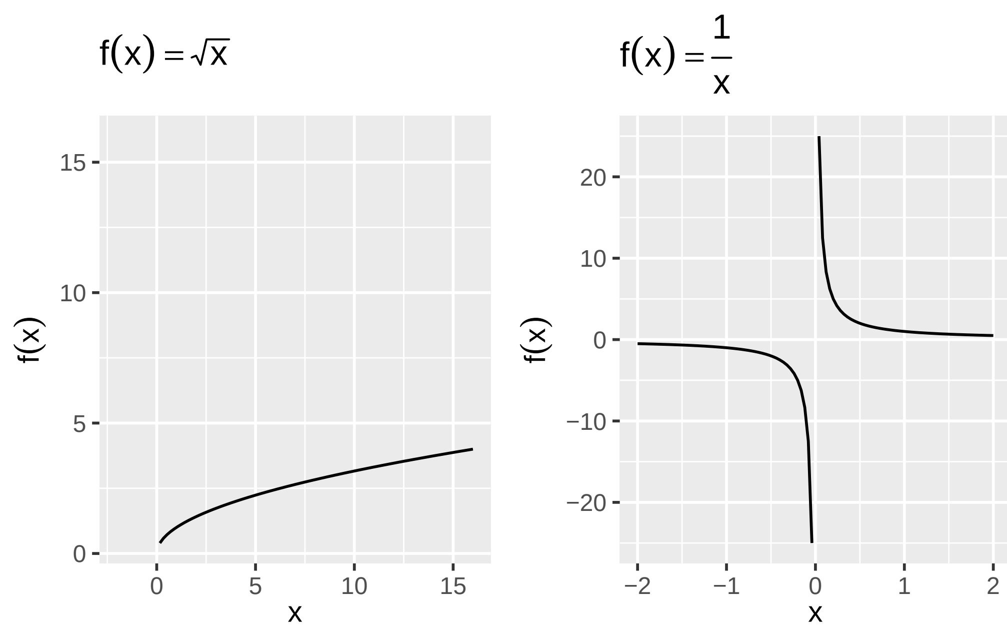


Figure 8.3: Functions which are not defined in some areas

8.4 Continuity

To repeat a finding from the limits of functions: $f(x)$ does not necessarily have to be defined at c for $\lim_{x \to c}$ to exist. Functions that have breaks in their lines are called discontinuous. Functions that have no breaks are called continuous. Continuity is a concept that is more fundamental to, but related to that of "differentiability", which we will cover next in calculus.

Definition 8.3 (Continuity")**.** Suppose that the domain of the function f includes an open interval containing the point c. Then f is continuous at c if $\lim_{x \to c} f(x)$ exists and if $\lim_{x \to c} f(x) = f(c)$. Further, f is continuous on an open interval (a, b) if it is continuous at each point in the interval.

To prove that a function is continuous for all points is beyond this practical introduction to math, but the general intuition can be grasped by graphing.

Example 8.4. For each function, determine if it is continuous or discontinuous.

1. $f(x) = \sqrt{x}$
2. $f(x) = e^x$
3. $f(x) = 1 + \frac{1}{x^2}$

4. $f(x) = \text{floor}(x)$.

The floor is the smaller of the two integers bounding a number. So $\text{floor}(x = 2.999) = 2$, $\text{floor}(x = 2.0001) = 2$, and $\text{floor}(x = 2) = 2$.

Solution. In Figure Figure **??**, we can see that the first two functions are continuous, and the next two are discontinuous. $f(x) = 1 + \frac{1}{x^2}$ is discontinuous at $x = 0$, and $f(x) = \text{floor}(x)$ is discontinuous at each whole number.

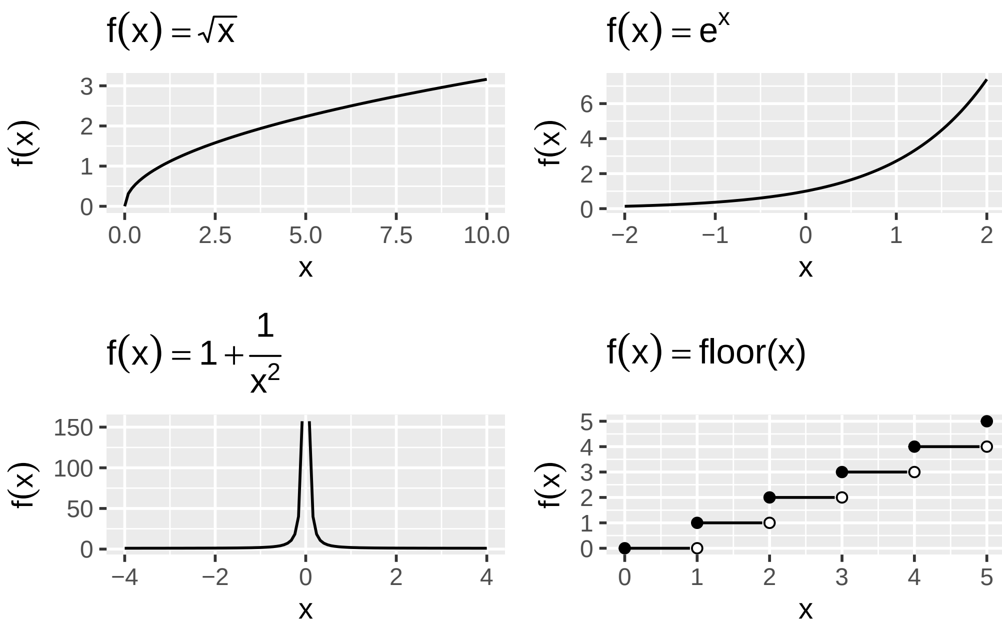


Figure 8.4: Continuous and Discontinuous Functions

Some properties of continuous functions:

1. If f and g are continuous at point c, then $f+g$, $f-g$, $f \cdot g$, $|f|$, and αf are continuous at point c also. f/g is continuous, provided $g(c) \neq 0$.
2. Boundedness: If f is continuous on the closed bounded interval $[a, b]$, then there is a number K such that $|f(x)| \leq K$ for each x in $[a, b]$.
3. Max/Min: If f is continuous on the closed bounded interval $[a, b]$, then f has a maximum and a minimum on $[a, b]$. They may be located at the end points.

Exercise 8.4. Let

$$f(x) = \frac{x^2 + 2x}{x}.$$

1. Graph the function. Is it defined everywhere?
2. What is the functions limit at $x \to 0$?

Answers to Examples

Example **??**

Solution.

1. $\{A_n\} = \left\{2 - \frac{1}{n^2}\right\} = \left\{1, \frac{7}{4}, \frac{17}{9}, \frac{31}{16}, \frac{49}{25}, \ldots\right\} = 2$

2. $\{B_n\} = \left\{\frac{n^2+1}{n}\right\} = \left\{2, \frac{5}{2}, \frac{10}{3}, \frac{17}{4} ...,\right\}$
3. $\{C_n\} = \left\{(-1)^n\left(1-\frac{1}{n}\right)\right\} = \left\{0, \frac{1}{2}, -\frac{2}{3}, \frac{3}{4}, -\frac{4}{5}\right\}$

Exercise **??**
Solution. Plot the function and you'll see the following limits:

1. 0
2. ∞

Example **??**

Solution.

1. k
2. c
3. $\lim_{x\to 2}(2x-3) = 2\lim_{x\to 2} x - 3\lim_{x\to 2} 1 = 1$
4. $\lim_{x\to c} x^n = \lim_{x\to c} x \cdots [\lim_{x\to c} x] = c\cdots c = c^n$

Exercise **??**
Solution. Although this function seems large, the thing our eyes should focus on is where the highest order polynomial remains. That will grow the fastest, so if the highest order term is on the denominator, the fraction will converge to 0, if it is on the numerator it will converge to negative infinity. Previewing the multiplication by hand, we can see that the $-x^9$ on the numerator will be the largest power. So the answer will be $-\infty$. We can also confirm this by writing out fractions:

$$\begin{aligned}
&\lim_{x\to\infty} \frac{\left(1+\frac{3}{x^3}-\frac{99}{4x^4}\right)\left(-\frac{2}{x^5}+1\right)}{\left(1+\frac{9}{18x}-\frac{3}{18x^5}-\frac{1}{18x^7}\right)\left(1+\frac{1}{x}\right)} \\
&\times \frac{x^4}{1} \times -\frac{x^5}{1} \times \frac{1}{18x^7} \times \frac{1}{x} \\
=&1 \times \lim_{-x\to\infty} \frac{x}{18}
\end{aligned}$$

Exercise **??**
Solution. See Figure **??**. We can say $\lim_{x\to 0} f(x) = 2$. Note that we can express

$$f(x) = \begin{cases} x+2 & x \neq 2; \\ \text{undefined} & x = 2 \end{cases}$$

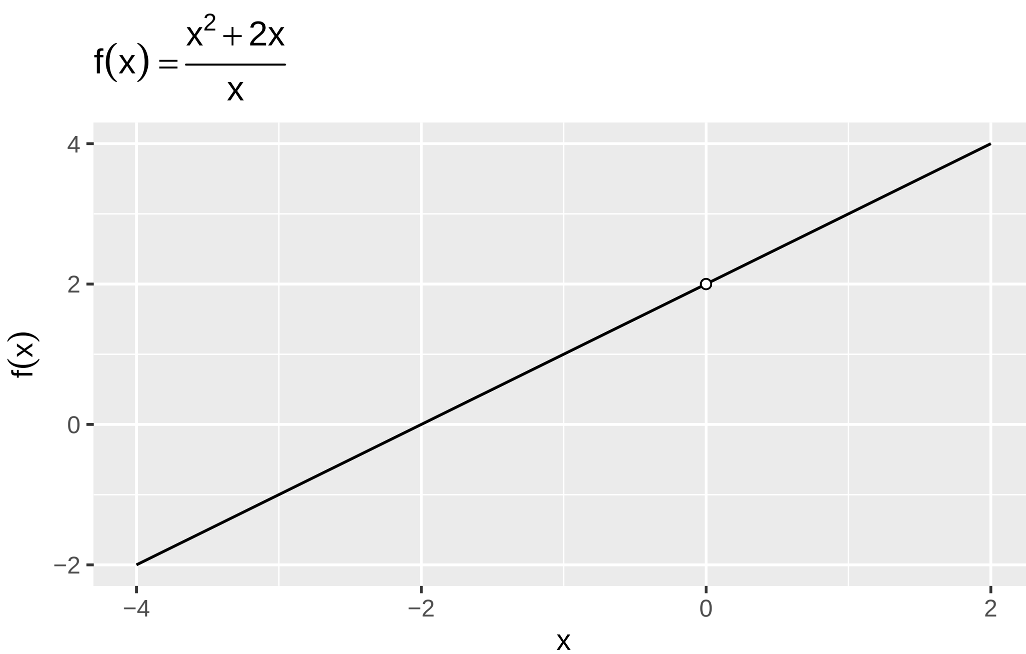


Figure 8.5: A function undedefined at x = 0

Chapter 9

Differential Calculus

Calculus is a fundamental part of any type of statistics exercise. Although you may not be taking derivatives and integral in your daily work as an analyst, calculus undergirds many concepts we use: maximization, expectation, and cumulative probability.

9.1 Derivatives

The derivative of f at x is its rate of change at x: how much $f(x)$ changes with a change in x. The rate of change is a fraction — rise over run — but because not all lines are straight and the rise over run formula will give us different values depending on the range we examine, we need to take a limit.

Definition 9.1 (Derivative)**.** Let f be a function whose domain includes an open interval containing the point x. The derivative of f at x is given by

$$\frac{d}{dx}f(x) = \lim_{h\to 0}\frac{f(x+h)-f(x)}{(x+h)-x} = \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}$$

There are a two main ways to denote a derivate:

- Leibniz Notation: $\frac{d}{dx}(f(x))$
- Prime or Lagrange Notation: $f'(x)$

If $f(x)$ is a straight line, the derivative is the slope. For a curve, the slope changes by the values of x, so the derivative is the slope of the line tangent to the curve at x. See, For example, Figure **??**.

If $f'(x)$ exists at a point x_0, then f is said to be **differentiable** at x_0. That also implies that $f(x)$ is continuous at x_0.

Properties of derivatives

Suppose that f and g are differentiable at x and that α is a constant. Then the functions $f \pm g$, αf, fg, and f/g (provided $g(x) \neq 0$) are also differentiable at x. Additionally,

Constant rule:

$$[kf(x)]' = kf'(x)$$

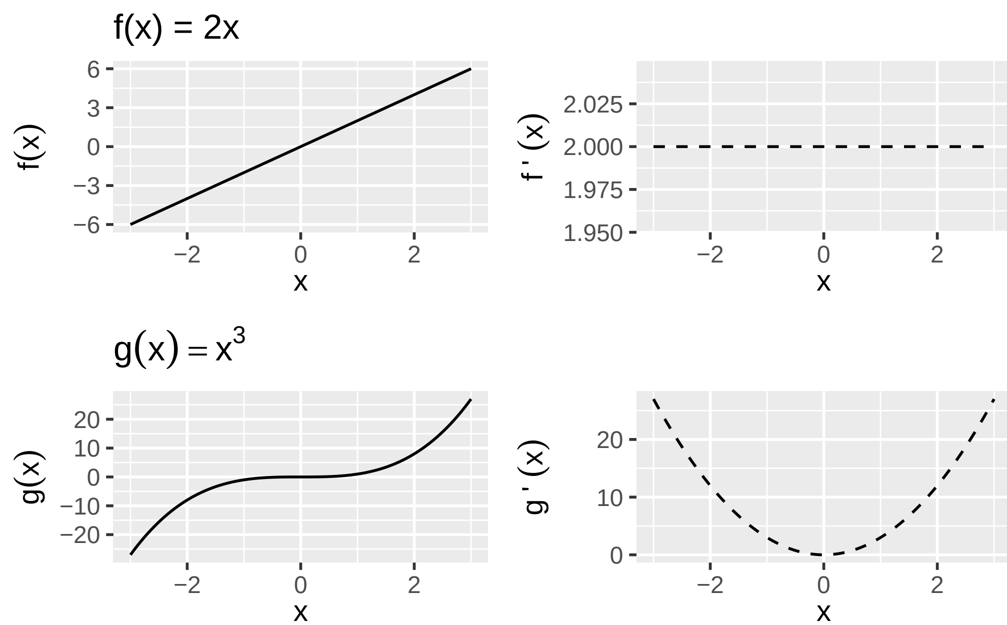


Figure 9.1: The Derivative as a Slope

Sum rule:

$$[f(x) \pm g(x)]' = f'(x) \pm g'(x)$$

With a bit more algebra, we can apply the definition of derivatives to get a formula for of the derivative of a product and a derivative of a quotient.

Product rule:

$$[f(x)g(x)]' = f'(x)g(x) + f(x)g'(x)$$

Quotient rule:

$$[f(x)/g(x)]' = \frac{f'(x)g(x) - f(x)g'(x)}{[g(x)]^2},\ g(x) \neq 0$$

Finally, one way to think of the power of derivatives is that it takes a function a notch down in complexity. The power rule applies to any higher-order function:

Power rule:

$$[x^k]' = kx^{k-1}$$

For any real number k (that is, both whole numbers and fractions). The power rule is proved **by induction**, a neat method of proof used in many fundamental applications to prove that a general statement holds for every possible case, even if there are countably infinite cases. We'll show a simple case where k is an integer here.

Proof. We would like to prove that

$$[x^k]' = kx^{k-1}$$

for any integer k.

First, consider the first case (the base case) of $k = 1$. We can show by the definition of derivatives (setting $f(x) = x^1 = 1$) that

$$[x^1]' = \lim_{h \to 0} \frac{(x+h) - x}{(x+h) - x} = 1.$$

Because 1 is also expressed as $1x^{1-1}$, the statement we want to prove holds for the case $k = 1$.

Now, *assume* that the statement holds for some integer m. That is, assume

$$[x^m]' = mx^{m-1}$$

Then, for the case $m + 1$, using the product rule above, we can simplify

$$\begin{aligned} [x^{m+1}]' &= [x^m \cdot x]' \\ &= (x^m)' \cdot x + (x^m) \cdot (x)' \\ &= mx^{m-1} \cdot x + x^m \quad \because \text{by previous assumption} \\ &= mx^m + x^m \\ &= (m+1)x^m \\ &= (m+1)x^{(m+1)-1} \end{aligned}$$

Therefore, the rule holds for the case $k = m + 1$ once we have assumed it holds for $k = m$. Combined with the first case, this completes proof by induction – we have now proved that the statement holds for all integers $k = 1, 2, 3, \cdots$.

To show that it holds for real fractions as well, we can prove expressing that exponent by a fraction of two integers.

□

These "rules" become apparent by applying the definition of the derivative above to each of the things to be "derived", but these come up so frequently that it is best to repeat until it is muscle memory.

Exercise 9.1. For each of the following functions, find the first-order derivative $f'(x)$.

1. $f(x) = c$
2. $f(x) = x$
3. $f(x) = x^2$
4. $f(x) = x^3$
5. $f(x) = \frac{1}{x^2}$
6. $f(x) = (x^3)(2x^4)$
7. $f(x) = x^4 - x^3 + x^2 - x + 1$
8. $f(x) = (x^2 + 1)(x^3 - 1)$
9. $f(x) = 3x^2 + 2x^{1/3}$
10. $f(x) = \frac{x^2+1}{x^2-1}$

9.2 Higher-Order Derivatives

The first derivative is applying the definition of derivatives on the function, and it can be expressed as

$$f'(x), \quad y', \quad \frac{d}{dx}f(x), \quad \frac{dy}{dx}$$

We can keep applying the differentiation process to functions that are themselves derivatives. The derivative of $f'(x)$ with respect to x, would then be

$$f''(x) = \lim_{h \to 0} \frac{f'(x+h) - f'(x)}{h}$$

and we can therefore call it the **Second derivative:**

$$f''(x), \quad y'', \quad \frac{d^2}{dx^2}f(x), \quad \frac{d^2y}{dx^2}$$

Similarly, the derivative of $f''(x)$ would be called the third derivative and is denoted $f'''(x)$. And by extension, the **nth derivative** is expressed as $\frac{d^n}{dx^n}f(x)$, $\frac{d^ny}{dx^n}$.

Example 9.1.

$$\begin{aligned} f(x) &= x^3 \\ f'(x) &= 3x^2 \\ f''(x) &= 6x \\ f'''(x) &= 6 \\ f''''(x) &= 0 \end{aligned}$$

Earlier, in Section **??**, we said that if a function differentiable at a given point, then it must be continuous. Further, if $f'(x)$ is itself continuous, then $f(x)$ is called continuously differentiable. All of this matters because many of our findings about optimization rely on differentiation, and so we want our function to be differentiable in as many layers. A function that is continuously differentiable infinitly is called "smooth". Some examples: $f(x) = x^2$, $f(x) = e^x$.

9.3 The Chain Rule

As useful as the above rules are, many functions you'll see won't fit neatly in each case immediately. Instead, they will be functions of functions. For example, the difference between $x^2 + 1^2$ and $(x^2 + 1)^2$ may look trivial, but the sum rule can be easily applied to the former, while it's actually not obvious what do with the latter.

Composite functions are formed by substituting one function into another and are denoted by

$$(f \circ g)(x) = f[g(x)].$$

To form $f[g(x)]$, the range of g must be contained (at least in part) within the domain of f. The domain of $f \circ g$ consists of all the points in the domain of g for which $g(x)$ is in the domain of f.

Example 9.2. Let $f(x) = \ln x$ for $0 < x < \infty$ and $g(x) = x^2$ for $-\infty < x < \infty$.

Then

$$(f \circ g)(x) = \ln x^2, -\infty < x < \infty - \{0\}$$

Also

$$(g \circ f)(x) = [\ln x]^2, 0 < x < \infty$$

Notice that $f \circ g$ and $g \circ f$ are not the same functions.

With the notation of composite functions in place, now we can introduce a helpful additional rule that will deal with a derivative of composite functions as a chain of concentric derivatives.

Chain Rule:

Let $y = (f \circ g)(x) = f[g(x)]$. The derivative of y with respect to x is

$$\frac{d}{dx}\{f[g(x)]\} = f'[g(x)]g'(x)$$

We can read this as: "the derivative of the composite function y is the derivative of f evaluated at $g(x)$, times the derivative of g."

The chain rule can be thought of as the derivative of the "outside" times the derivative of the "inside", remembering that the derivative of the outside function is evaluated at the value of the inside function.

- The chain rule can also be written as
$$\frac{dy}{dx} = \frac{dy}{dg(x)}\frac{dg(x)}{dx}$$
  This expression does not imply that the $dg(x)$'s cancel out, as in fractions. They are part of the derivative notation and you can't separate them out or cancel them.)

Example 9.3. Find $f'(x)$ for $f(x) = (3x^2 + 5x - 7)^6$.

The direct use of a chain rule is when the exponent of is itself a function, so the power rule could not have applied generaly:

Generalized Power Rule:

If $f(x) = [g(x)]^p$ for any rational number p,

$$f'(x) = p[g(x)]^{p-1}g'(x)$$

9.4 Derivatives of logs and exponents

Natural logs and exponents (they are inverses of each other; see Prerequisites) crop up everywhere in statistics. Their derivative is a special case from the above, but quite elegant.

Theorem 9.1. *The functions e^x and the natural logarithm $\ln(x)$ are continuous and differentiable in their domains, and their first derivate is*

$$(e^x)' = e^x$$

$$\ln(x)' = \frac{1}{x}$$

Also, when these are composite functions, it follows by the generalized power rule that

$$\left(e^{g(x)}\right)' = e^{g(x)} \cdot g'(x)$$

$$(\ln g(x))' = \frac{g'(x)}{g(x)}, \quad \text{if} \ \ g(x) > 0$$

We will relegate the proofs to small excerpts.

Derivatives of exponents

To repeat the main rule in Theorem **??**, the intuition is that

1. Derivative of e^x is itself: $\frac{d}{dx}e^x = e^x$ (See Figure **??**)
2. Same thing if there were a constant in front: $\frac{d}{dx}\alpha e^x = \alpha e^x$
3. Same thing no matter how many derivatives there are in front: $\frac{d^n}{dx^n}\alpha e^x = \alpha e^x$
4. Chain Rule: When the exponent is a function of x, remember to take derivative of that function and add to product. $\frac{d}{dx}e^{g(x)} = e^{g(x)}g'(x)$

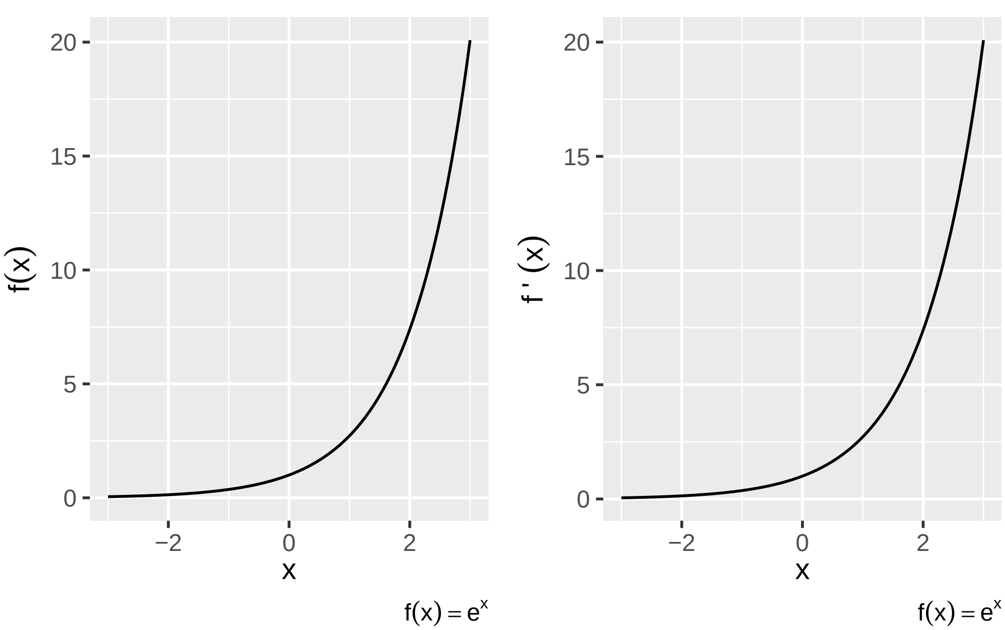


Figure 9.2: Derivative of the Exponential Function

Example 9.4. Find the derivative for the following.

1. $f(x) = e^{-3x}$
2. $f(x) = e^{x^2}$
3. $f(x) = (x-1)e^x$

Derivatives of logs

The natural log is the mirror image of the natural exponent and has mirroring properties, again, to repeat the theorem,

1. log prime x is one over x (Figure **??**):

$$\frac{d}{dx}\ln x = \frac{1}{x}$$

2. Exponents become multiplicative constants:

$$\frac{d}{dx}\ln x^k = \frac{d}{dx}k\ln x = \frac{k}{x}$$

3. Chain rule again:

$$\frac{d}{dx}\ln u(x) = \frac{u'(x)}{u(x)}$$

4. For any positive base b,

$$\frac{d}{dx}b^x = (\ln b)\,(b^x)$$

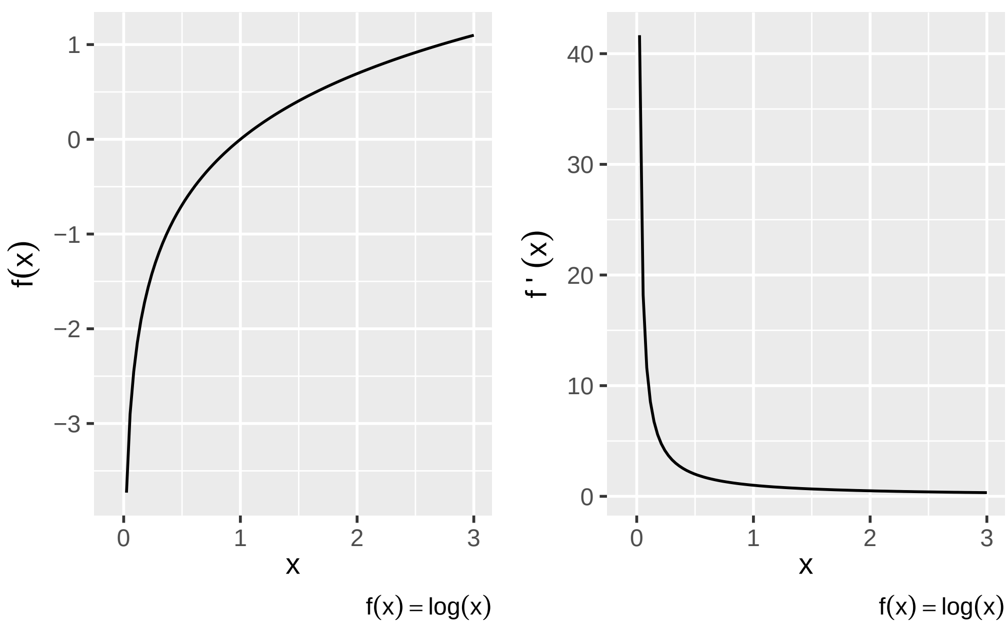


Figure 9.3: Derivative of the Natural Log

Example 9.5. Find dy/dx for the following.

1. $f(x) = \ln(x^2 + 9)$
2. $f(x) = \ln(\ln x)$
3. $f(x) = (\ln x)^2$
4. $f(x) = \ln e^x$

Outline of Proof

We actually show the derivative of the log first, and then the derivative of the exponential naturally follows.

The general derivative of the log at any base a is solvable by the definition of derivatives.

$$(\ln_a x)' = \lim_{h \to 0} \frac{1}{h} \ln_a \left(1 + \frac{h}{x}\right)$$

Re-express $g = \frac{h}{x}$ and get

$$\begin{aligned}(\ln_a x)' &= \frac{1}{x} \lim_{g \to 0} \ln_a (1 + g)^{\frac{1}{g}} \\ &= \frac{1}{x} \ln_a e\end{aligned}$$

By definition of e. As a special case, when $a = e$, then $(\ln x)' = \frac{1}{x}$.

Now let's think about the inverse, taking the derivative of $y = a^x$.

$$\begin{aligned}y &= a^x \\ \Rightarrow \ln y &= x \ln a \\ \Rightarrow \frac{y'}{y} &= \ln a \\ \Rightarrow y' = y \ln a\end{aligned}$$

Then in the special case where $a = e$,

$$(e^x)' = (e^x)$$

9.5 Partial Derivatives

What happens when there's more than variable that is changing?

> If you can do ordinary derivatives, you can do partial derivatives: just hold all the other input variables constant except for the one you're differentiating with respect to. (Joe Blitzstein's Math Notes)

Suppose we have a function f now of two (or more) variables and we want to determine the rate of change relative to one of the variables. To do so, we would find its partial derivative, which is defined similar to the derivative of a function of one variable.

Partial Derivative: Let f be a function of the variables $(x_1, \dots, x_n)$. The partial derivative of f with respect to x_i is

$$\frac{\partial f}{\partial x_i}(x_1, \dots, x_n) = \lim_{h \to 0} \frac{f(x_1, \dots, x_i + h, \dots, x_n) - f(x_1, \dots, x_i, \dots, x_n)}{h}$$

Only the ith variable changes — the others are treated as constants.

We can take higher-order partial derivatives, like we did with functions of a single variable, except now the higher-order partials can be with respect to multiple variables.

Example 9.6. Notice that you can take partials with regard to different variables.

Suppose $f(x, y) = x^2 + y^2$. Then

$$\begin{aligned}
\frac{\partial f}{\partial x}(x, y) &= \\
\frac{\partial f}{\partial y}(x, y) &= \\
\frac{\partial^2 f}{\partial x^2}(x, y) &= \\
\frac{\partial^2 f}{\partial x \partial y}(x, y) &=
\end{aligned}$$

Exercise 9.2. Let $f(x, y) = x^3 y^4 + e^x - \ln y$. What are the following partial derivaitves?

$$\begin{aligned}
\frac{\partial f}{\partial x}(x, y) &= \\
\frac{\partial f}{\partial y}(x, y) &= \\
\frac{\partial^2 f}{\partial x^2}(x, y) &= \\
\frac{\partial^2 f}{\partial x \partial y}(x, y) &=
\end{aligned}$$

9.6 Taylor Approximation

A common form of approximation used in statistics involves derivatives. A Taylor series is a way to represent common functions as infinite series (a sum of infinite elements) of the function's derivatives at some point a.

For example, Taylor series are very helpful in representing nonlinear (read: difficult) functions as linear (read: manageable) functions. One can thus **approximate** functions by using lower-order, finite series known as **Taylor polynomials**. If $a = 0$, the series is called a Maclaurin series.

Specifically, a Taylor series of a real or complex function $f(x)$ that is infinitely differentiable in the neighborhood of point a is:

$$\begin{aligned} f(x) &= f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \cdots \\ &= \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n \end{aligned}$$

Taylor Approximation: We can often approximate the curvature of a function $f(x)$ at point a using a 2nd order Taylor polynomial around point a:

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + R_2$$

R_2 is the remainder (R for remainder, 2 for the fact that we took two derivatives) and often treated as negligible, giving us:

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{f''(a)}{2}(x-a)^2$$

The more derivatives that are added, the smaller the remainder R and the more accurate the approximation. Proofs involving limits guarantee that the remainder converges to 0 as the order of derivation increases.

Chapter 10

Optimization

To optimize, we use derivatives and calculus. Optimization is to find the maximum or minimum of a functon, and to find what value of an input gives that extremum. This has obvious uses in engineering. Many tools in the statistical toolkit use optimization. One of the most common ways of estimating a model is through "Maximum Likelihood Estimation", done via optimizing a function (the likelihood).

Optimization also comes up in Economics, Formal Theory, and Political Economy all the time. A go-to model of human behavior is that they optimize a certain utility function. Humans are not pure utility maximizers, of course, but nuanced models of optimization – for example, adding constraints and adding uncertainty – will prove to be quite useful.

10.1 Maxima and Minima

The first derivative, $f'(x)$, quantifies the slope of a function. Therefore, it can be used to check whether the function $f(x)$ at the point x is increasing or decreasing at x.

1. **Increasing:** $f'(x) > 0$
2. **Decreasing:** $f'(x) < 0$
3. **Neither increasing nor decreasing**: $f'(x) = 0$ i.e. a maximum, minimum, or saddle point

So for example, $f(x) = x^2 + 2$ and $f'(x) = 2x$

Exercise 10.1 (Plotting a mazimum and minimum)**.** Plot $f(x) = x^3 + x^2 + 2$, plot its derivative, and identifiy where the derivative is zero. Is there a maximum or minimum?

The second derivative $f''(x)$ identifies whether the function $f(x)$ at the point x is

1. Concave / concave down: $f''(x) < 0$
2. Convex / Concave up: $f''(x) > 0$

Maximum (Minimum): x_0 is a **local maximum (minimum)** if $f(x_0) > f(x)$ $(f(x_0) < f(x))$ for all x within some open interval containing x_0. x_0 is a **global maximum (minimum)** if $f(x_0) > f(x)$ $(f(x_0) < f(x))$ for all x in the domain of f.

Given the function f defined over domain D, all of the following are defined as **critical points**:

1. Any interior point of D where $f'(x) = 0$.
2. Any interior point of D where $f'(x)$ does not exist.

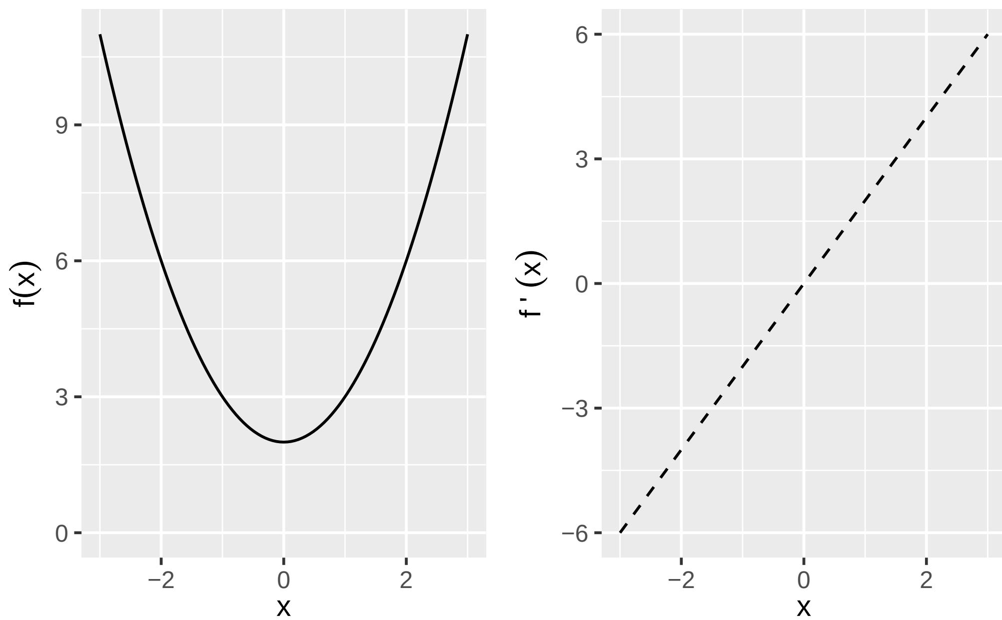


Figure 10.1: Maxima and Minima

3. Any endpoint that is in D.

The maxima and minima will be a subset of the critical points.

Second Derivative Test of Maxima/Minima: We can use the second derivative to tell us whether a point is a maximum or minimum of $f(x)$.

1. Local Maximum: $f'(x) = 0$ and $f''(x) < 0$
2. Local Minimum: $f'(x) = 0$ and $f''(x) > 0$
3. Need more info: $f'(x) = 0$ and $f''(x) = 0$

Global Maxima and Minima Sometimes no global max or min exists — e.g., $f(x)$ not bounded above or below. However, there are three situations where we can fairly easily identify global max or min.

1. **Functions with only one critical point.** If x_0 is a local max or min of f and it is the only critical point, then it is the global max or min.
2. **Globally concave up or concave down functions.** If $f''(x)$ is never zero, then there is at most one critical point. That critical point is a global maximum if $f'' < 0$ and a global minimum if $f'' > 0$.
3. **Functions over closed and bounded intervals** must have both a global maximum and a global minimum.

Example 10.1 (Maxima and Minima by drawing)**.** Find any critical points and identify whether they are a max, min, or saddle point:

1. $f(x) = x^2 + 2$
2. $f(x) = x^3 + 2$
3. $f(x) = |x^2 - 1|$, $x \in [-2, 2]$

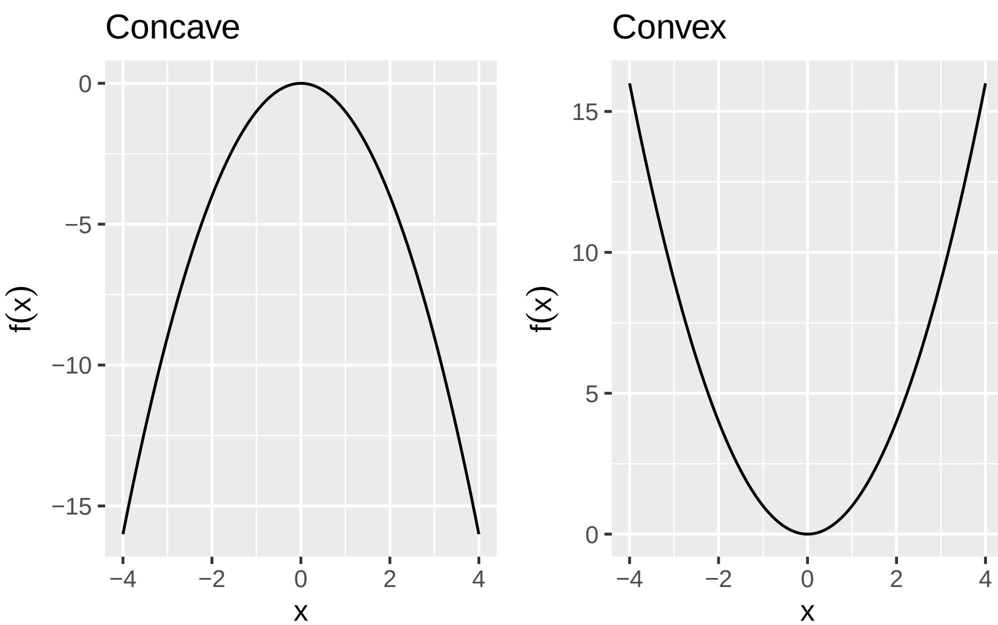
10.2 Concavity of a Function

Concavity helps identify the curvature of a function, $f(x)$, in 2 dimensional space.

Definition 10.1 (Concave Function)**.** A function f is strictly concave over the set S if $\forall x_1, x_2 \in S$ and $\forall a \in (0, 1)$,

$$f(ax_1 + (1-a)x_2) > af(x_1) + (1-a)f(x_2)$$

Any line connecting two points on a concave function will lie *below* the function.



Definition 10.2 (Convex Function)**.** Convex: A function f is strictly convex over the set S if $\forall x_1, x_2 \in S$ and $\forall a \in (0, 1)$,

$$f(ax_1 + (1-a)x_2) < af(x_1) + (1-a)f(x_2)$$

Any line connecting two points on a convex function will lie above the function.

Second Derivative Test of Concavity: The second derivative can be used to understand concavity.

If

$$\begin{cases} f''(x) < 0 & \Rightarrow \quad \text{Concave} \\ f''(x) > 0 & \Rightarrow \quad \text{Convex} \end{cases}$$

Quadratic Forms

Quadratic forms is shorthand for a way to summarize a function. This is important for finding concavity because

1. Approximates local curvature around a point — e.g., used to identify max vs min vs saddle point.
2. They are simple to express even in n dimensions:
3. Have a matrix representation.

Quadratic Form: A polynomial where each term is a monomial of degree 2 in any number of variables:

$$\begin{aligned}
\text{One variable: } & Q(x_1) = a_{11}x_1^2 \\
\text{Two variables: } & Q(x_1, x_2) = a_{11}x_1^2 + a_{12}x_1x_2 + a_{22}x_2^2 \\
\text{N variables: } & Q(x_1, \cdots, x_n) = \sum_{i \le j} a_{ij}x_ix_j
\end{aligned}$$

which can be written in matrix terms:

One variable

$$Q(\mathbf{x}) = x_1^\top a_{11} x_1$$

N variables:

$$\begin{aligned}
Q(\mathbf{x}) &= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} a_{11} & \frac{1}{2}a_{12} & \cdots & \frac{1}{2}a_{1n} \\ \frac{1}{2}a_{12} & a_{22} & \cdots & \frac{1}{2}a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}a_{1n} & \frac{1}{2}a_{2n} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \\
&= \mathbf{x}^\top \mathbf{A} \mathbf{x}
\end{aligned}$$

For example, the Quadratic on $\mathbb{R}^2$:

$$\begin{aligned}
Q(x_1, x_2) &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a_{11} & \frac{1}{2}a_{12} \\ \frac{1}{2}a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\
&= a_{11}x_1^2 + a_{12}x_1x_2 + a_{22}x_2^2
\end{aligned}$$

Definiteness of Quadratic Forms

When the function $f(\mathbf{x})$ has more than two inputs, determining whether it has a maxima and minima (remember, functions may have many inputs but they have only one output) is a bit more tedious. Definiteness helps identify the curvature of a function, $Q(\mathbf{x})$, in n dimensional space.

Definiteness: By definition, a quadratic form always takes on the value of zero when $x = 0$, $Q(\mathbf{x}) = 0$ at $\mathbf{x} = 0$. The definiteness of the matrix $\mathbf{A}$ is determined by whether the quadratic form $Q(\mathbf{x}) = \mathbf{x}^\top \mathbf{A}\mathbf{x}$ is greater than zero, less than zero, or sometimes both over all $\mathbf{x} \neq 0$.

10.3 Gradient and FOC

We can see from a graphical representation that if a point is a local maxima or minima, it must meet certain conditions regarding its derivative. These are so commonly used that we refer these to "First Order Conditions" (FOCs) and "Second Order Conditions" (SOCs) in the economic tradition.

When we examined functions of one variable x, we found critical points by taking the first derivative, setting it to zero, and solving for x. For functions of n variables, the critical points are found in much the same way, except now we set the partial derivatives equal to zero. Note: We will only consider critical points on the interior of a function's domain.

In a derivative, we only took the derivative with respect to one variable at a time. When we take the derivative separately with respect to all variables in the elements of $\mathbf{x}$ and then express the result as a vector, we use the term Gradient and Hessian.

Definition 10.3 (Gradient)**.** Given a function $f(\mathbf{x})$ in n variables, the gradient $\nabla f(\mathbf{x})$ (the greek letter nabla) is a row vector, where the ith element is the partial derivative of $f(\mathbf{x})$ with respect to x_i:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} & \frac{\partial f(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix}$$

The gradient points in the direction of the steepest rate of increase at each point $\mathbf{x}$.

Before we know whether a point is a maxima or minima, if it meets the FOC it is a "Critical Point".

Definition 10.4 (Critical Point)**.** $\mathbf{x}^*$ is a critical point if and only if $\nabla f(\mathbf{x}^*) = \mathbf{0}$ (the vector of zeros). If the partial derivative of f(x) with respect to x^* is 0, then $\mathbf{x}^*$ is a critical point. To solve for $\mathbf{x}^*$, find the gradient, set each element equal to 0, and solve the system of equations.

$$\mathbf{x}^* = \begin{bmatrix} x_1^* \\ x_2^* \\ \vdots \\ x_n^* \end{bmatrix}$$

Example 10.2. Example: Given a function $f(\mathbf{x}) = (x_1 - 1)^2 + x_2^2 + 1$, find the (1) Gradient and (2) Critical point of $f(\mathbf{x})$.
Solution. Gradient

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} & \frac{\partial f(\mathbf{x})}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2(x_1 - 1) & 2x_2 \end{bmatrix}$$

Critical Point $\mathbf{x}^*$:

$$\begin{aligned} \frac{\partial f(\mathbf{x})}{\partial x_1} &= 2(x_1 - 1) = 0 && \Rightarrow x_1^* = 1 \\ \frac{\partial f(\mathbf{x})}{\partial x_2} &= 2x_2 = 0 && \Rightarrow x_2^* = 0 \end{aligned}$$

So

$$\mathbf{x}^* = (1, 0)$$

10.4 Hessian and SOC

When we found a critical point for a function of one variable, we used the second derivative as a indicator of the curvature at the point in order to determine whether the point was a min, max, or saddle (second derivative test of concavity). For functions of n variables, we use *second order partial derivatives* as an indicator of curvature.

Definition 10.5 (Hessian)**.** Given a function $f(\mathbf{x})$ in n variables, the hessian $\mathbf{H}(\mathbf{x})$ is an $n \times n$ matrix, where the (i, j)th element is the second order partial derivative of $f(\mathbf{x})$ with respect to x_i and x_j:

$$\mathbf{H}(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix}$$

Note that the hessian will be a symmetric matrix because $\frac{\partial f(\mathbf{x})}{\partial x_1 \partial x_2} = \frac{\partial f(\mathbf{x})}{\partial x_2 \partial x_1}$.

Also note that given that $f(\mathbf{x})$ is of quadratic form, each element of the hessian will be a constant.

These definitions will be employed when we determine the **Second Order Conditions** of a function:

Given a function $f(\mathbf{x})$ and a point $\mathbf{x}^*$ such that $\nabla f(\mathbf{x}^*) = 0$,

1. Hessian is Positive Definite around $\mathbf{x}^*$ $\implies$ Local Min
2. Hessian is Negative Definite around $\mathbf{x}^*$ $\implies$ Local Max
3. Hessian is Indefinite around $\mathbf{x}^*$ $\implies$ Saddle Point

Furthermore, there's an easier way to check whether a 2×2 matrix is positive or negative definite.

Definiteness of 2 $\times$ 2 Matrix: For a 2×2 matrix $$A=

$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$

$$

1. If $\det(\mathbf{A}) > 0$ and $A_{11} > 0$, then $\mathbf{A}$ is positive definite
2. If $\det(\mathbf{A}) > 0$ and $A_{11} < 0$, then $\mathbf{A}$ is negative definite
3. If $\det(\mathbf{A}) < 0$, then $\mathbf{A}$ is indefinite

Example 10.3 (Max and min with two dimensions)**.** We found that the only critical point of $f(\mathbf{x}) = (x_1 - 1)^2 + x_2^2 + 1$ is at $\mathbf{x}^* = (1, 0)$. Is it a min, max, or saddle point?
Solution. The Hessian is

$$\mathbf{H}(\mathbf{x}) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Since $\det(\mathbf{H}(x)) = 4 > 0$ and $H_{11} = 2 > 0$, the Hessian is positive definite.

Maxima, Minima, or Saddle Point? Since the Hessian is positive definite and the gradient equals 0, $x^\star = (1, 0)$ is a local minimum.

Note: Alternate check of definiteness. Is $\mathbf{H}(\mathbf{x}^*) \geq\leq 0 \quad \forall \quad \mathbf{x} \neq 0$

$$\begin{aligned} \mathbf{x}^\top H(\mathbf{x}^*)\mathbf{x} &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \\ \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= 2x_1^2 + 2x_2^2 \end{aligned}$$

For any $\mathbf{x} \neq 0$, $2(x_1^2 + x_2^2) > 0$, so the Hessian is positive definite and $\mathbf{x}^*$ is a strict local minimum.

Definiteness and Concavity

Although definiteness helps us to understand the curvature of an n-dimensional function, it does not necessarily tell us whether the function is globally concave or convex.

We need to know whether a function is globally concave or convex to determine whether a critical point is a global min or max. We can use the definiteness of the Hessian to determine whether a function is globally concave or convex:

1. Hessian is Positive Semidefinite $\forall \mathbf{x} \implies$ Globally Convex
2. Hessian is Negative Semidefinite $\forall \mathbf{x} \implies$ Globally Concave

Notice that the definiteness conditions must be satisfied over the entire domain.

10.5 Global Maxima and Minima

Global Max/Min Conditions: Given a function $f(\mathbf{x})$ and a point $\mathbf{x}^*$ such that $\nabla f(\mathbf{x}^*) = 0$,

1. $f(\mathbf{x})$ Globally Convex $\implies$ Global Min
2. $f(\mathbf{x})$ Globally Concave $\implies$ Global Max

Note that showing that $\mathbf{H}(\mathbf{x}^*)$ is negative semidefinite is not enough to guarantee $\mathbf{x}^*$ is a local max. However, showing that $\mathbf{H}(\mathbf{x})$ is negative semidefinite for all $\mathbf{x}$ guarantees that x^* is a global max. (The same goes for positive semidefinite and minima.)

Example 10.4. Take $f_1(x) = x^4$ and $f_2(x) = -x^4$.

- Both have $x = 0$ as a critical point.

- Unfortunately, $f_1''(0) = 0$ and $f_2''(0) = 0$, so we can't tell whether $x = 0$ is a min or max for either. However, $f_1''(x) = 12x^2$ and $f_2''(x) = -12x^2$.

- For all x, $f_1''(x) \geq 0$ and $f_2''(x) \leq 0$ — i.e., $f_1(x)$ is globally convex and $f_2(x)$ is globally concave.
- So $x = 0$ is a global min of $f_1(x)$ and a global max of $f_2(x)$.

Exercise 10.2. Given $f(\mathbf{x}) = x_1^3 - x_2^3 + 9x_1x_2$, find any maxima or minima.

Solution.

1. First order conditions
   (1) Gradient $\nabla f(\mathbf{x})$

$$\begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 3x_1^2 + 9x_2 \\ -3x_2^2 + 9x_1 \end{bmatrix}$$

   (2) Critical Points $\mathbf{x}^*$
   - Set the gradient equal to zero and solve for x_1 and x_2. We have two equations and two unknowns. Solving for x_1 and x_2, we get two critical points: $\mathbf{x}_1^* = (0, 0)$ and $\mathbf{x}_2^* = (3, -3)$.
   - $3x_1^2 + 9x_2 = 0 \quad \Rightarrow \quad 9x_2 = -3x_1^2 \quad \Rightarrow \quad x_2 = -\frac{1}{3}x_1^2$
   - $-3x_2^2 + 9x_1 = 0 \quad \Rightarrow \quad -3(-\frac{1}{3}x_1^2)^2 + 9x_1 = 0$
   - $\Rightarrow \quad -\frac{1}{3}x_1^4 + 9x_1 = 0 \quad \Rightarrow \quad x_1^3 = 27x_1 \quad \Rightarrow \quad x_1 = 3$
   - $3(3)^2 + 9x_2 = 0 \quad \Rightarrow \quad x_2 = -3$
2. Second order conditions.

(1) Hessian $\mathbf{H}(\mathbf{x})$

$$\begin{bmatrix} 6x_1 & 9 \\ 9 & -6x_2 \end{bmatrix}$$

(2) Hessian $\mathbf{H}(\mathbf{x_1^*})$

$$\begin{bmatrix} 0 & 9 \\ 9 & 0 \end{bmatrix}$$

(3) Leading principal minors of $\mathbf{H}(\mathbf{x_1^*})$

$$M_1 = 0; M_2 = -81$$

(4) Definiteness of $\mathbf{H}(\mathbf{x_1^*})$?
- $\mathbf{H}(\mathbf{x_1^*})$ is indefinite

(5) Maxima, Minima, or Saddle Point for $\mathbf{x_1^*}$?
- Since $\mathbf{H}(\mathbf{x_1^*})$ is indefinite, $\mathbf{x}_1^* = (0, 0)$ is a saddle point.

(6) Hessian

$$\mathbf{H}(\mathbf{x_2^*}) = \begin{bmatrix} 18 & 9 \\ 9 & 18 \end{bmatrix}$$

(7) Leading principal minors of $\mathbf{H}(\mathbf{x_2^*})$

$$M_1 = 18; M_2 = 243$$

(8) Definiteness of $\mathbf{H}(\mathbf{x_2^*})$?
- $\mathbf{H}(\mathbf{x_2^*})$ is positive definite

(9) Maxima, Minima, or Saddle Point for $\mathbf{x}_2^*$?
- Since $\mathbf{H}(\mathbf{x_2^*})$ is positive definite, $\mathbf{x}_1^* = (3, -3)$ is a strict local minimum

3. Global concavity/convexity.

(1) Is f(x) globally concave/convex?
- No. In evaluating the Hessians for $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ we saw that the Hessian is not positive semidefinite at x = (0,0).

(2) Are any $\mathbf{x}^*$ global minima or maxima?
- No. Since the function is not globally concave/convex, we can't infer that $\mathbf{x}_2^* = (3, -3)$ is a global minimum. In fact, if we set $x_1 = 0$, the $f(\mathbf{x}) = -x_2^3$, which will go to $-\infty$ as $x_2 \to \infty$.

Chapter 11

Integral Calculus

11.1 The Indefinite Integral

So far, we've been interested in finding the derivative $f = F'$ of a function F. However, sometimes we're interested in exactly the reverse: finding the function F for which f is its derivative. We refer to F as the antiderivative of f.

Definition 11.1 (Antiderivative)**.** The antiverivative of a function $f(x)$ is a differentiable function F whose derivative is f.

$$F' = f.$$

Another way to describe is through the inverse formula. Let DF be the derivative of F. And let $DF(x)$ be the derivative of F evaluated at x. Then the antiderivative is denoted by D^{-1} (i.e., the inverse derivative). If $DF = f$, then $F = D^{-1}f$.

This definition bolsters the main takeaway about integrals and derivatives: They are inverses of each other.

Exercise 11.1 (Antiderivative)**.** Find the antiderivative of the following:

1. $f(x) = \frac{1}{x^2}$
2. $f(x) = 3e^{3x}$

We know from derivatives how to manipulate F to get f. But how do you express the procedure to manipulate f to get F? For that, we need a new symbol, which we will call indefinite integration.

Definition 11.2 (Indefinite Integral)**.** The indefinite integral of $f(x)$ is written

$$\int f(x)dx$$

and is equal to the antiderivative of f.

Example 11.1. Draw the function $f(x)$ and its indefinite integral, $\int f(x)dx$

$$f(x) = (x^2 - 4)$$

Solution. The Indefinite Integral of the function $f(x) = (x^2 - 4)$ can, for example, be $F(x) = \frac{1}{3}x^3 - 4x$. But it can also be $F(x) = \frac{1}{3}x^3 - 4x + 1$, because the constant 1 disappears when taking the derivative.

Some of these functions are plotted in the bottom panel of Figure **??** as dotted lines.

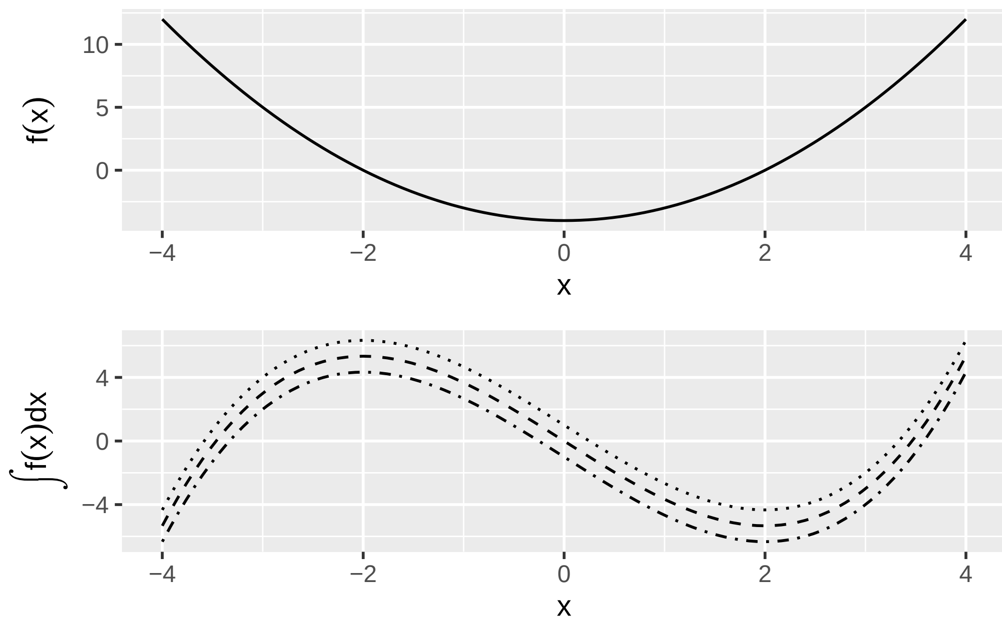


Figure 11.1: The Many Indefinite Integrals of a Function

Notice from these examples that while there is only a single derivative for any function, there are multiple antiderivatives: one for any arbitrary constant c. c just shifts the curve up or down on the y-axis. If more information is present about the antiderivative — e.g., that it passes through a particular point — then we can solve for a specific value of c.

Properties of Integration

Some useful properties of integrals follow by virtue of being the inverse of a derivative.

11.1.1 Properties of Integration

1. Constants are allowed to slip out: $\int af(x)dx = a\int f(x)dx$
2. Integration of the sum is sum of integrations: $\int[f(x) + g(x)]dx = \int f(x)dx + \int g(x)dx$
3. Reverse Power-rule: $\int x^n dx = \frac{1}{n+1}x^{n+1} + c$
4. Exponents are still exponents: $\int e^x dx = e^x + c$
5. Recall the derivative of $\ln(x)$ is one over x, and so: $\int \frac{1}{x}dx = \ln x + c$
6. Reverse chain-rule: $\int e^{f(x)}f'(x)dx = e^{f(x)} + c$
7. More generally: $\int[f(x)]^n f'(x)dx = \frac{1}{n+1}[f(x)]^{n+1} + c$
8. Remember the derivative of a log of a function: $\int \frac{f'(x)}{f(x)}dx = \ln f(x) + c$

Example 11.2 (Common Integration)**.** Simplify the following indefinite integrals:

- $\int 3x^2 dx$
- $\int (2x+1) dx$
- $\int e^x e^{e^x} dx$

11.2 The Definite Integral

If there is a indefinite integral, there *must* be a definite integral. Indeed there is, but the notion of definite integrals comes from a different objective: finding the are a under a function. We will find, perhaps remarkably, that the formula we find to get the sum turns out to be expressible by the anti-derivative.

Suppose we want to determine the area $A(R)$ of a region R defined by a curve $f(x)$ and some interval $a \leq x \leq b$.

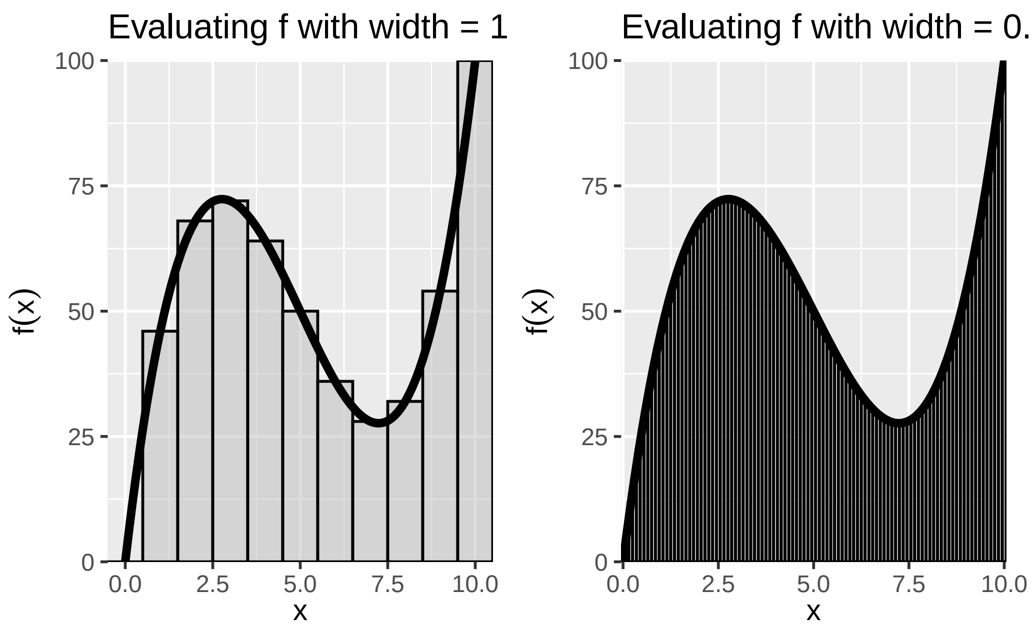


Figure 11.2: The Riemann Integral as a Sum of Evaluations

One way to calculate the area would be to divide the interval $a \leq x \leq b$ into n subintervals of length Δx and then approximate the region with a series of rectangles, where the base of each rectangle is Δx and the height is $f(x)$ at the midpoint of that interval. $A(R)$ would then be approximated by the area of the union of the rectangles, which is given by

$$S(f, \Delta x) = \sum_{i=1}^{n} f(x_i)\Delta x$$

and is called a **Riemann sum**.

As we decrease the size of the subintervals Δx, making the rectangles "thinner," we would expect our approximation of the area of the region to become closer to the true area. This allows us to express the area as a limit of a series:

$$A(R) = \lim_{\Delta x \to 0} \sum_{i=1}^{n} f(x_i)\Delta x$$

Figure **??** shows that illustration. The curve depicted is $f(x) = -15(x-5)+(x-5)^3+50$. We want approximate the area under the curve between the x values of 0 and 10. We can do this in blocks of arbitrary width, where the sum of rectangles (the area of which is width times $f(x)$ evaluated at the midpoint of the bar) shows the Riemann Sum. As the width of the bars Δx becomes smaller, the better the estimate of $A(R)$.

This is how we define the "Definite" Integral:

Definition 11.3 (The Definite Integral (Riemann))**.** If for a given function f the Riemann sum approaches a limit as $\Delta x \to 0$, then that limit is called the Riemann integral of f from a to b. We express this with the $\int$, symbol, and write

$$\int_a^b f(x)dx = \lim_{\Delta x \to 0} \sum_{i=1}^{n} f(x_i)\Delta x$$

The most straightforward of a definite integral is the definite integral. That is, we read

$$\int_a^b f(x)dx$$

as the definite integral of f from a to b and we defined as the area under the "curve" $f(x)$ from point $x = a$ to $x = b$.

The fundamental theorem of calculus shows us that this sum is, in fact, the antiderivative.

Theorem 11.1 (First Fundamental Theorem of Calculus)**.** *Let the function f be bounded on $[a, b]$ and continuous on (a, b). Then, suggestively, use the symbol $F(x)$ to denote the definite integral from a to x*

$$F(x) = \int_a^x f(t)dt, \quad a \leq x \leq b$$

Then $F(x)$ has a derivative at each point in (a, b) and

$$F'(x) = f(x), \quad a < x < b$$

That is, the definite integral function of f is *the one of the antiderivatives of some f.*

This is again a long way of saying that that differentiation is the inverse of integration. But now, we've covered definite integrals.

The second theorem gives us a simple way of computing a definite integral as a function of indefinite integrals.

11.2.1 Second Fundamental Theorem of Calculus

Let the function f be bounded on $[a, b]$ and continuous on (a, b). Let F be any function that is continuous on $[a, b]$ such that $F'(x) = f(x)$ on (a, b). Then

$$\int_a^b f(x)dx = F(b) - F(a)$$

So the procedure to calculate a simple definite integral $\int_a^b f(x)dx$ is then

1. Find the indefinite integral $F(x)$.
2. Evaluate $F(b) - F(a)$.

Example 11.3 (Definite Integral of a monomial)**.** Solve $\int_1^3 3x^2 dx$. Let $f(x) = 3x^2$.

Exercise 11.2. What is the value of $\int_{-2}^{2} e^x e^{e^x} dx$?

Properties for Definite Integrals

The area-interpretation of the definite integral provides some useful properties for definite integrals

1. There is no area below a point:
$$\int_a^a f(x)dx = 0$$
2. Reversing the limits changes the sign of the integral:
$$\int_a^b f(x)dx = -\int_b^a f(x)dx$$
3. Sums can be separated into their own integrals:
$$\int_a^b [\alpha f(x) + \beta g(x)]dx = \alpha \int_a^b f(x)dx + \beta \int_a^b g(x)dx$$
4. Areas can be combined as long as limits are linked:
$$\int_a^b f(x)dx + \int_b^c f(x)dx = \int_a^c f(x)dx$$

Exercise 11.3. Simplify the following definite intergrals.

1. $\int_1^1 3x^2 dx =$
2. $\int_0^4 (2x+1)dx =$
3. $\int_{-2}^{0} e^x e^{e^x} dx + \int_0^2 e^x e^{e^x} dx =$

11.3 Integration by Substitution

From the second fundamental theorem of calculus, we now that a quick way to get a definite integral is to first find the indefinite integral, and then just plug in the bounds.

Sometimes the integrand (the thing that we are trying to take an integral of) doesn't appear integrable using common rules and antiderivatives. A method one might try is **integration by substitution**, which is related to the Chain Rule.

Suppose we want to find the indefinite integral

$$\int g(x)dx$$

but $g(x)$ is complex and none of the formulas we have seen so far seem to apply immediately. The trick is to come up with a *new* function $u(x)$ such that

$$g(x) = f[u(x)]u'(x).$$

Why does an introduction of yet another function end of simplifying things? Let's refer to the antiderivative of f as F. Then the chain rule tells us that

$$\frac{d}{dx}F[u(x)] = f[u(x)]u'(x)$$

. So, $F[u(x)]$ is the antiderivative of g. We can then write

$$\int g(x)dx = \int f[u(x)]u'(x)dx = \int \frac{d}{dx}F[u(x)]dx = F[u(x)] + c$$

To summarize, the procedure to determine the indefinite integral $\int g(x)dx$ by the method of substitution:

1. Identify some part of $g(x)$ that might be simplified by substituting in a single variable u (which will then be a function of x).
2. Determine if $g(x)dx$ can be reformulated in terms of u and du.
3. Solve the indefinite integral.
4. Substitute back in for x

Substitution can also be used to calculate a definite integral. Using the same procedure as above,

$$\int_a^b g(x)dx = \int_c^d f(u)du = F(d) - F(c)$$

where $c = u(a)$ and $d = u(b)$.

Example 11.4 (Integration by Substitution I)**.** Solve the indefinite integral

$$\int x^2\sqrt{x+1}dx.$$

For the above problem, we could have also used the substitution $u = \sqrt{x+1}$. Then $x = u^2 - 1$ and $dx = 2udu$. Substituting these in, we get

$$\int x^2\sqrt{x+1}dx = \int (u^2-1)^2 u2udu$$

which when expanded is again a polynomial and gives the same result as above.

Another case in which integration by substitution is is useful is with a fraction.

Example 11.5 (Integration by Substitutiton II)**.** Simplify

$$\int_0^1 \frac{5e^{2x}}{(1+e^{2x})^{1/3}}dx.$$

11.4 Integration by Parts

Another useful integration technique is **integration by parts**, which is related to the Product Rule of differentiation. The product rule states that

$$\frac{d}{dx}(uv) = u\frac{dv}{dx} + v\frac{du}{dx}$$

Integrating this and rearranging, we get

$$\int u\frac{dv}{dx}dx = uv - \int v\frac{du}{dx}dx$$

or

$$\int u(x)v'(x)dx = u(x)v(x) - \int v(x)u'(x)dx$$

More easily remembered with the mnemonic "Ultraviolet Voodoo":

$$\int udv = uv - \int vdu$$

where $du = u'(x)dx$ and $dv = v'(x)dx$.

For definite integrals, this is simply

$$\int_a^b u\frac{dv}{dx}dx = uv\big|_a^b - \int_a^b v\frac{du}{dx}dx$$

Our goal here is to find expressions for u and dv that, when substituted into the above equation, yield an expression that's more easily evaluated.

Example 11.6 (Integration by Parts I)**.** Simplify the following integrals. These seemingly obscure forms of integrals come up often when integrating distributions.

$$\int xe^{ax}dx$$

Solution. Let $u = x$ and $\frac{dv}{dx} = e^{ax}$. Then $du = dx$ and $v = (1/a)e^{ax}$. Substituting this into the integration by parts formula, we obtain

$$\begin{aligned}\int xe^{ax}dx &= uv - \int vdu \\ &= x\left(\frac{1}{a}e^{ax}\right) - \int \frac{1}{a}e^{ax}dx \\ &= \frac{1}{a}xe^{ax} - \frac{1}{a^2}e^{ax} + c\end{aligned}$$

Exercise 11.4 (Integration by Parts II). 1. Integrate

$$\int x^n e^{ax} dx$$

2. Integrate

$$\int x^3 e^{-x^2} dx$$

Answers to Examples and Exercises

Exercise **??**

Solution.

1. $f'(x) = 0$
2. $f'(x) = 1$
3. $f'(x) = 2x^3$
4. $f'(x) = 3x^2$
5. $f'(x) = -2x^{-3}$
6. $f'(x) = 14x^6$
7. $f'(x) = 4x^3 - 3x^2 + 2x - 1$
8. $f'(x) = 5x^4 + 3x^2 - 2x$
9. $f'(x) = 6x + \frac{2}{3}x^{\frac{-2}{3}}$
10. $f'(x) = \frac{-4x}{x^4-2x^2+1}$

Example **??**
Solution. For convenience, define $f(z) = z^6$ and $z = g(x) = 3x^2 + 5x - 7$. Then, $y = f[g(x)]$ and

$$\begin{aligned}\frac{d}{dx}y &= f'(z)g'(x) \\ &= 6(3x^2 + 5x - 7)^5(6x + 5)\end{aligned}$$

Example **??**

Solution.

1. Let $u(x) = -3x$. Then $u'(x) = -3$ and $f'(x) = -3e^{-3x}$.
2. Let $u(x) = x^2$. Then $u'(x) = 2x$ and $f'(x) = 2xe^{x^2}$.

Example **??**

Solution.

1. Let $u(x) = x^2 + 9$. Then $u'(x) = 2x$ and

$$\frac{dy}{dx} = \frac{u'(x)}{u(x)} = \frac{2x}{(x^2 + 9)}$$

2. Let $u(x) = \ln x$. Then $u'(x) = 1/x$ and $\frac{dy}{dx} = \frac{1}{(x \ln x)}$.

3. Use the generalized power rule.

$$\frac{dy}{dx} = \frac{(2\ln x)}{x}$$

4. We know that $\ln e^x = x$ and that $dx/dx = 1$, but we can double check. Let $u(x) = e^x$. Then $u'(x) = e^x$ and $\frac{dy}{dx} = \frac{u'(x)}{u(x)} = \frac{e^x}{e^x} = 1$.

Example **??**
Solution. What is $F(x)$? From the power rule, recognize $\frac{d}{dx}x^3 = 3x^2$ so

$$\begin{aligned} F(x) &= x^3 \\ \int_1^3 f(x)dx &= F(x=3) - F(x-1) \\ &= 3^3 - 1^3 \\ &= 26 \end{aligned}$$

Example **??**
Solution. The problem here is the $\sqrt{x+1}$ term. However, if the integrand had $\sqrt{x}$ times some polynomial, then we'd be in business. Let's try $u = x+1$. Then $x = u-1$ and $dx = du$. Substituting these into the above equation, we get

$$\begin{aligned} \int x^2\sqrt{x+1}dx &= \int (u-1)^2\sqrt{u}du \\ &= \int (u^2 - 2u + 1)u^{1/2}du \\ &= \int (u^{5/2} - 2u^{3/2} + u^{1/2})du \end{aligned}$$

We can easily integrate this, since it is just a polynomial. Doing so and substituting $u = x+1$ back in, we get

$$\int x^2\sqrt{x+1}dx = 2(x+1)^{3/2}\left[\frac{1}{7}(x+1)^2 - \frac{2}{5}(x+1) + \frac{1}{3}\right] + c$$

Example **??**
Solution. When an expression is raised to a power, it is often helpful to use this expression as the basis for a substitution. So, let $u = 1 + e^{2x}$. Then $du = 2e^{2x}dx$ and we can set $5e^{2x}dx = 5du/2$. Additionally, $u = 2$ when $x = 0$ and $u = 1 + e^2$ when $x = 1$. Substituting all of this in, we get

$$
\begin{aligned}
\int_0^1 \frac{5e^{2x}}{(1+e^{2x})^{1/3}}dx &= \frac{5}{2}\int_2^{1+e^2} \frac{du}{u^{1/3}} \\
&= \frac{5}{2}\int_2^{1+e^2} u^{-1/3}du \\
&= \frac{15}{4}u^{2/3}\Big|_2^{1+e^2} \\
&= 9.53
\end{aligned}
$$

Exercise **??**

Solution.

1.

$$\int x^n e^{ax}dx$$

As in the first problem, let

$$u = x^n, dv = e^{ax}dx$$

Then $du = nx^{n-1}dx$ and $v = (1/a)e^{ax}$.

Substituting these into the integration by parts formula gives

$$
\begin{aligned}
\int x^n e^{ax}dx &= uv - \int vdu \\
&= x^n\left(\frac{1}{a}e^{ax}\right) - \int \frac{1}{a}e^{ax}nx^{n-1}dx \\
&= \frac{1}{a}x^n e^{ax} - \frac{n}{a}\int x^{n-1}e^{ax}dx
\end{aligned}
$$

Notice that we now have an integral similar to the previous one, but with x^{n-1} instead of x^n.

For a given n, we would repeat the integration by parts procedure until the integrand was directly integratable — e.g., when the integral became $\int e^{ax}dx$.

2.

$$\int x^3 e^{-x^2}dx$$

We could, as before, choose $u = x^3$ and $dv = e^{-x^2}dx$. But we can't then find v — i.e., integrating $e^{-x^2}dx$ isn't possible. Instead, notice that

$$\frac{d}{dx}e^{-x^2} = -2xe^{-x^2},$$

which can be factored out of the original integrand

$$\int x^3 e^{-x^2}dx = \int x^2(xe^{-x^2})dx.$$

We can then let $u = x^2$ and $dv = xe^{-x^2}dx$. The$du = 2xdx$ and $v = -\frac{1}{2}e^{-x^2}$. Substituting these in, we have

$$\begin{aligned}
\int x^3 e^{-x^2} dx &= uv - \int v du \\
&= x^2 \left(-\frac{1}{2}e^{-x^2}\right) - \int \left(-\frac{1}{2}e^{-x^2}\right) 2xdx \\
&= -\frac{1}{2}x^2 e^{-x^2} + \int xe^{-x^2} dx \\
&= -\frac{1}{2}x^2 e^{-x^2} - \frac{1}{2}e^{-x^2} + c
\end{aligned}$$

Part IV

IV Probability

Chapter 12

Probability Theory

Probability and Inferences are mirror images of each other, and both are integral to social science. Probability quantifies uncertainty, which is important because many things in the social world are at first uncertain. Inference is then the study of how to learn about facts you don't observe from facts you do observe.

12.1 Counting Rules

Probability in high school is usually really about combinatorics: the probability of event A is the number of ways in which A can occur divided by the number of all other possibilities. This is a very simplified version of probability, which we can call the "counting definition of probability", essentially because each possible event to count is often equally likely and discrete. But it is still good to review the underlying rules here.

Fundamental Theorem of Counting: If an object has j different characteristics that are independent of each other, and each characteristic i has n_i ways of being expressed, then there are $\prod_{i=1}^{j} n_i$ possible unique objects.

Example 12.1. Suppose we are given a stack of cards. Cards can be either red or black and can take on any of 13 values. There is only one of each color-number combination. In this case,

1. $j =$
2. $n_{\text{color}} =$
3. $n_{\text{number}} =$
4. Number of Outcomes =

We often need to count the number of ways to choose a subset from some set of possibilities. The number of outcomes depends on two characteristics of the process: does the order matter and is replacement allowed?

It is useful to think of any problem concretely, e.g. through a **sampling table**: If there are n objects which are numbered 1 to n and we select $k < n$ of them, how many different outcomes are possible?

If the order in which a given object is selected matters, selecting 4 numbered objects in the following order (1, 3, 7, 2) and selecting the same four objects but in a different order such as (7, 2, 1, 3) will be counted as different outcomes.

If replacement is allowed, there are always the same n objects to select from. However, if replacement is not allowed, there is always one less option than the previous round when making a selection. For example, if replacement is not allowed and I am selecting 3 elements from the following set {1, 2, 3, 4, 5, 6}, I will have 6 options at first, 5 options as I make my second selection, and 4 options as I make my third.

1. So if ***order matters*** AND we are sampling ***with replacement***, the number of different outcomes is n^k.
2. If ***order matters*** AND we are sampling ***without replacement***, the number of different outcomes is $n(n-1)(n-2)...(n-k+1) = \frac{n!}{(n-k)!}$.
3. If ***order doesn't matter*** AND we are sampling ***without replacement***, the number of different outcomes is $\binom{n}{k} = \frac{n!}{(n-k)!k!}$.

Expression $\binom{n}{k}$ is read as "n choose k" and denotes $\frac{n!}{(n-k)!k!}$. Also, note that $0! = 1$.

Example 12.2 (Counting)**.** There are five balls numbered from 1 through 5 in a jar. Three balls are chosen. How many possible choices are there?

1. Ordered, with replacement =
2. Ordered, without replacement =
3. Unordered, without replacement =

Exercise 12.1 (Counting)**.** Four cards are selected from a deck of 52 cards. Once a card has been drawn, it is not reshuffled back into the deck. Moreover, we care only about the complete hand that we get (i.e. we care about the set of selected cards, not the sequence in which it was drawn). How many possible outcomes are there?

12.2 Sets

Probability is about quantifying the uncertainty of events. *Sets* (set theory) are the mathematical way we choose to formalize those events. Events are not inherently numerical: the onset of war or the stock market crashing is not inherently a number. Sets can define such events, and we wrap math around so that we have a transparent language to communicate about those events. *Measure theory* might sound mysterious or hard, but it is also just a mathematical way to quantify things like length, volume, and mass. Probability can be thought of as a particular application of measure theory where we want to quantify the measure of a set.

Set : A set is any well defined collection of elements. If x is an element of S, $x \in S$.

Sample Space (S): A set or collection of all possible outcomes from some process. Outcomes in the set can be discrete elements (countable) or points along a continuous interval (uncountable).

Examples:

1. Discrete: the numbers on a die, whether a vote cast is republican or democrat.
2. Continuous: GNP, arms spending, age.

Event: Any collection of possible outcomes of an experiment. Any subset of the full set of possibilities, including the full set itself. Event A $\subset$ S.

Empty Set: a set with no elements. $S = \{\}$. It is denoted by the symbol $\emptyset$.

Set operations:

1. **Union**: The union of two sets A and B, $A \cup B$, is the set containing all of the elements in A or B.

$$A_1 \cup A_2 \cup \cdots \cup A_n = \bigcup_{i=1}^{n} A_i$$

2. **Intersection**: The intersection of sets A and B, $A \cap B$, is the set containing all of the elements in both A and B.
$$A_1 \cap A_2 \cap \cdots \cap A_n = \bigcap_{i=1}^{n} A_i$$
3. **Complement**: If set A is a subset of S, then the complement of A, denoted A^C, is the set containing all of the elements in S that are not in A.

Properties of set operations:

- **Commutative**: $A \cup B = B \cup A$; $A \cap B = B \cap A$
- **Associative**: $A \cup (B \cup C) = (A \cup B) \cup C$; $A \cap (B \cap C) = (A \cap B) \cap C$
- **Distributive**: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$; $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- **de Morgan's laws**: $(A \cup B)^C = A^C \cap B^C$; $(A \cap B)^C = A^C \cup B^C$
- **Disjointness**: Sets are disjoint when they do not intersect, such that $A \cap B = \emptyset$. A collection of sets is pairwise disjoint (**mutually exclusive**) if, for all $i \neq j$, $A_i \cap A_j = \emptyset$. A collection of sets form a partition of set S if they are pairwise disjoint and they cover set S, such that $\bigcup_{i=1}^{k} A_i = S$.

Example 12.3 (Sets)**.** Let set A be $\{1, 2, 3, 4\}$, B be $\{3, 4, 5, 6\}$, and C be $\{5, 6, 7, 8\}$. Sets A, B, and C are all subsets of the sample space S which is $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$

Write out the following sets:

1. $A \cup B$
2. $C \cap B$
3. B^c
4. $A \cap (B \cup C)$

Exercise 12.2. Suppose you had a pair of four-sided dice. You sum the results from a single toss.

What is the set of possible outcomes (i.e. the sample space)?

Consider subsets A $\{2, 8\}$ and B $\{2,3,7\}$ of the sample space you found. What is

1. A^c
2. $(A \cup B)^c$

12.3 Probability

Probability Definitions: Formal and Informal

Many things in the world are uncertain. In everyday speech, we say that we are *uncertain* about the outcome of random events. Probability is a formal model of uncertainty which provides a measure of uncertainty governed by a particular set of rules (Figure **??**). A different model of uncertainty would, of course, have a set of rules different from anything we discuss here. Our focus on probability is justified because it has proven to be a particularly useful model of uncertainty.

Probability Distribution Function: a mapping of each event in the sample space S to the real numbers that satisfy the following three axioms (also called Kolmogorov's Axioms).

Formally,

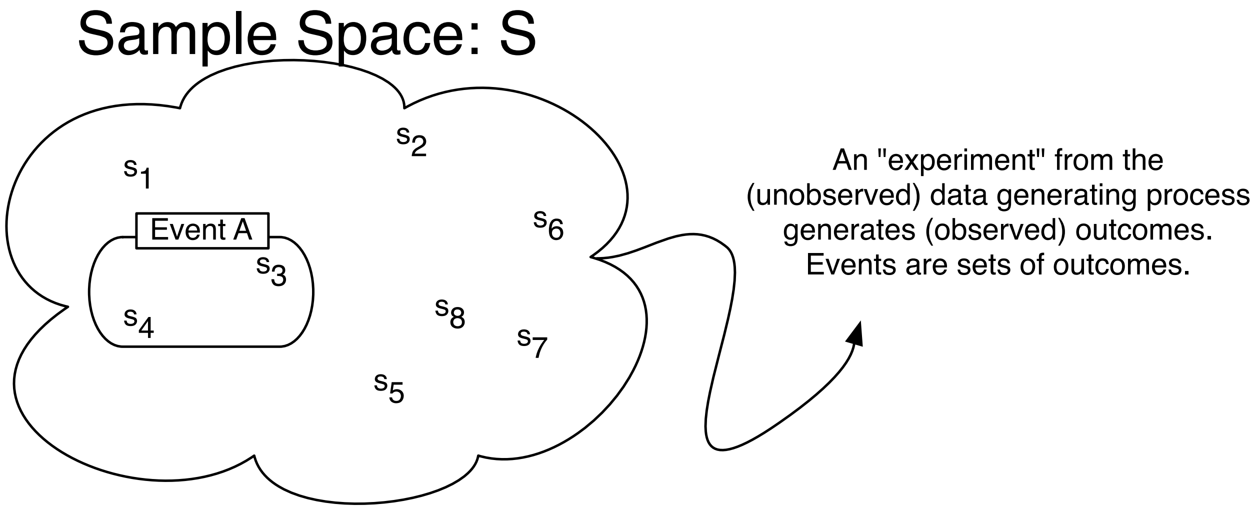


Figure 12.1: Probablity as a Measure[1]

Definition 12.1 (Probability)**.** Probability is a function that maps events to a real number, obeying the axioms of probability.

The axioms of probability make sure that the separate events add up in terms of probability, and – for standardization purposes – that they add up to 1.

Definition 12.2 (Axioms of Probability)**.**
1. For any event A, $P(A) \geq 0$.
2. $P(S) = 1$
3. The Countable Additivity Axiom: For any sequence of *disjoint* (mutually exclusive) events $A_1, A_2, ...$ (of which there may be infinitely many),

$$P\left(\bigcup_{i=1}^{k} A_i\right) = \sum_{i=1}^{k} P(A_i)$$

The last axiom is an extension of a union to infinite sets. When there are only two events in the space, it boils down to:

$$P(A_1 \cup A_2) = P(A_1) + P(A_2) \quad \text{for disjoint } A_1, A_2$$

Probability Operations

Using these three axioms, we can define all of the common rules of probability.

1. $P(\emptyset) = 0$
2. For any event A, $0 \leq P(A) \leq 1$
3. $P(A^C) = 1 - P(A)$
4. If $A \subset B$ (A is a subset of B), then $P(A) \leq P(B)$
5. For *any* two events A and B, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

6. Boole's Inequality: For any sequence of n events (which need not be disjoint) $A_1, A_2, \ldots, A_n$, then

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i)$$

Example 12.4. Assume we have an evenly-balanced, six-sided die.

Then,

1. Sample space $S =$
2. $P(1) = \cdots = P(6) =$
3. $P(\emptyset) = P(7) =$
4. $P(\{1,3,5\}) =$
5. $P(\{1,2\}^C) = P(\{3,4,5,6\}) =$
6. Let $A = \{1,2,3,4,5\} \subset S$. Then $P(A) = 5/6 < P(S) =$
7. Let $A = \{1,2,3\}$ and $B = \{2,4,6\}$. Then $A \cup B$? $A \cap B$? $P(A \cup B)$?

Exercise 12.3. Suppose you had a pair of four-sided dice. You sum the results from a single toss. Let us call this sum, or the outcome, X.

1. What is $P(X = 5)$, $P(X = 3)$, $P(X = 6)$?
2. What is $P(X = 5 \cup X = 3)^C$?

12.4 Conditional Probability

Conditional Probability: The conditional probability $P(A|B)$ of an event A is the probability of A, given that another event B has occurred. Conditional probability allows for the inclusion of other information into the calculation of the probability of an event. It is calculated as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Note that conditional probabilities are probabilities and must also follow the Kolmagorov axioms of probability.

Example 12.5 (Conditional Probability)**.** Assume A and B occur with the following frequencies:

	A	A^c
B	n_{ab}	$n_{a^c b}$
B^C	n_{ab^c}	$n_{(ab)^c}$

and let $n_{ab} + n_{a^C b} + n_{ab^C} + n_{(ab)^C} = N$. Then

1. $P(A) =$
2. $P(B) =$
3. $P(A \cap B) =$
4. $P(A|B) = \frac{P(A \cap B)}{P(B)} =$
5. $P(B|A) = \frac{P(A \cap B)}{P(A)} =$

Example 12.6 (Conditional Probability 2)**.** A six-sided die is rolled. What is the probability of a 1, given the outcome is an odd number?

You could rearrange the fraction to highlight how a joint probability could be expressed as the product of a conditional probability.

Definition 12.3 (Multiplicative Law of Probability)**.** The probability of the intersection of two events A and B is

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$$

which follows directly from the definition of conditional probability. More generally,

$$\begin{aligned} P(A_1 \cap \dots \cap A_k) = & P(A_k | A_{k-1} \cap \dots \cap A_1) \\ & \times P(A_{k-1} | A_{k-2} \cap \dots A_1) \\ & \vdots \\ & \times P(A_2 | A_1) \\ & \times P(A_1) \end{aligned}$$

Sometimes it is easier to calculate these conditional probabilities and sum them than it is to calculate $P(A)$ directly.

Definition 12.4 (Law of Total Probability)**.** Let S be the sample space of some experiment and let the disjoint k events $B_1, \dots, B_k$ partition S, such that $P(B_1 \cup \dots \cup B_k) = P(S) = 1$. If A is some other event in S, then the events $A \cap B_1, A \cap B_2, \dots, A \cap B_k$ will form a partition of A and we can write A as

$$A = (A \cap B_1) \cup \dots \cup (A \cap B_k)$$

.

Since the k events are disjoint,

$$\begin{aligned} P(A) &= \sum_{i=1}^{k} P(A \cap B_i) \\ &= \sum_{i=1}^{k} P(B_i) P(A|B_i) \end{aligned}$$

12.5 Bayes Rule

Bayes Rule: Assume that events $B_1, \dots, B_k$ form a partition of the space S. Then by the Law of Total Probability

$$P(B_j|A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(B_j)P(A|B_j)}{\sum\limits_{i=1}^{k} P(B_i)P(A|B_i)}$$

If there are only two states of B, then this is just

$$P(B_1|A) = \frac{P(B_1)P(A|B_1)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2)}$$

Bayes' rule determines the posterior probability of a state $P(B_j|A)$ by calculating the probability $P(A \cap B_j)$ that both the event A and the state B_j will occur and dividing it by the probability that the event will occur regardless of the state (by summing across all B_i). The states could be something like Normal/Defective, Healthy/Diseased, Republican/Democrat/Independent, etc. The event on which one conditions could be something like a sampling from a batch of components, a test for a disease, or a question about a policy position.

Prior and Posterior Probabilities: Above, $P(B_1)$ is often called the prior probability, since it's the probability of B_1 before anything else is known. $P(B_1|A)$ is called the posterior probability, since it's the probability after other information is taken into account.

Example 12.7 (Bayes' Rule)**.** In a given town, 40% of the voters are Democrat and 60% are Republican. The president's budget is supported by 50% of the Democrats and 90% of the Republicans. If a randomly (equally likely) selected voter is found to support the president's budget, what is the probability that they are a Democrat?

Exercise 12.4 (Conditional Probability)**.** Assume that 2% of the population of the U.S. are members of some extremist militia group. We develop a survey that positively classifies someone as being a member of a militia group given that they are a member 95% of the time and negatively classifies someone as not being a member of a militia group given that they are not a member 97% of the time. What is the probability that someone positively classified as being a member of a militia group is actually a militia member?

12.6 Independence

Definition 12.5 (Independence)**.** If the occurrence or nonoccurrence of either events A and B have no effect on the occurrence or nonoccurrence of the other, then A and B are independent.

If A and B are independent, then

1. $P(A|B) = P(A)$
2. $P(B|A) = P(B)$
3. $P(A \cap B) = P(A)P(B)$
4. More generally than the above, $P(\bigcap_{i=1}^{k} A_i) = \prod_{i=1}^{K} P(A_i)$

Are mutually exclusive events independent of each other?

No. If A and B are mutually exclusive, then they cannot happen simultaneously. If we know that A occurred, then we know that B couldn't have occurred. Because of this, A and B aren't independent.

Pairwise Independence: A set of more than two events $A_1, A_2, \dots, A_k$ is pairwise independent if $P(A_i \cap A_j) = P(A_i)P(A_j)$, $\forall i \neq j$. Note that this does **not** necessarily imply joint independence.

Conditional Independence: If A and B are independent once you know the occurrence of a third event C, then A and B are conditionally independent (conditional on C):

1. $P(A|B \cap C) = P(A|C)$
2. $P(B|A \cap C) = P(B|C)$
3. $P(A \cap B|C) = P(A|C)P(B|C)$

Just because two events are conditionally independent does not mean that they are independent. Actually it is hard to think of real-world things that are "unconditionally" independent. That's why it's always important to ask about a finding: What was it conditioned on? For example, suppose that a graduate school admission decisions are done by only one professor, who picks a group of 50 bright students and flips a coin for each student to generate a class of about 25 students. Then the the probability that two students get accepted are conditionally independent, because they are determined by two separate coin tosses. However, this does not mean that their admittance is not completely independent. Knowing that student A got in gives us information about whether student B got in, if we think that the professor originally picked her pool of 50 students by merit.

Perhaps more counter-intuitively: If two events are already independent, then it might seem that no amount of "conditioning" will make them dependent. But this is not always so. For example[2], suppose I only get a call from two people, Alice and Bob. Let A be the event that Alice calls, and B be the event that Bob calls. Alice and Bob do not communicate, so $P(A \mid B) = P(A)$. But now let C be the event that your phone rings. For conditional independence to hold here, then $P(A \mid C)$ must be equal to $P(A \mid B \cap C)$. But this is not true – $A \mid C$ may or may not be true, but $P(A \mid B \cap C)$ certainly is true.

12.7 Random Variables

Most questions in the social sciences involve events, rather than numbers per se. To analyze and reason about events quantitatively, we need a way of mapping events to numbers. A random variable does exactly that.

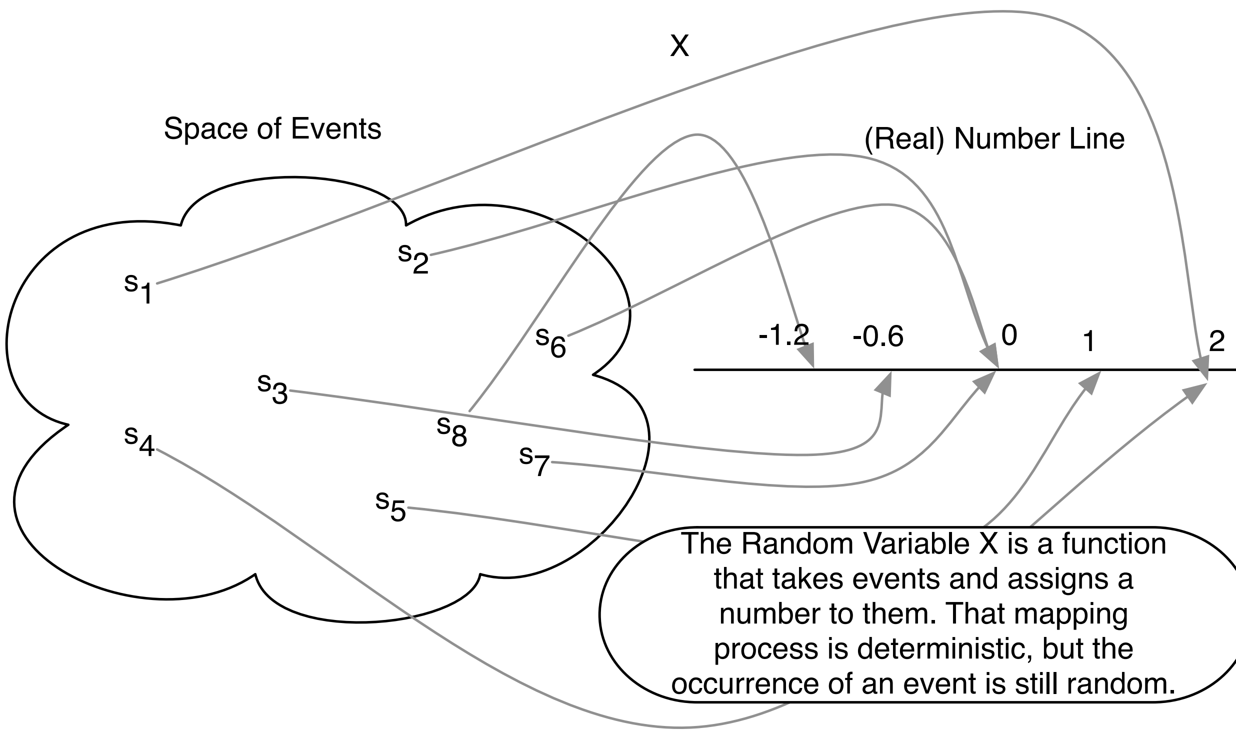


Figure 12.2: The Random Variable as a Real-Valued Function

[2]Example taken from Blitzstein and Hwang, Example 2.5.10

Definition 12.6 (Random Variable)**.** A random variable is a measurable function X that maps from the sample space S to the set of real numbers R. It assigns a real number to every outcome $s \in S$.

Figure **??** shows a image of the function. It might seem strange to define a random variable as a function – which is neither random nor variable. The randomness comes from the realization of an event from the sample space s.

Randomness means that the outcome of some experiment is not deterministic, i.e. there is some probability $(0 < P(A) < 1)$ that the event will occur.

The support of a random variable is all values for which there is a positive probability of occurrence.

Example: Flip a fair coin two times. What is the sample space?

A random variable must map events to the real line. For example, let a random variable X be the number of heads. The event (H, H) gets mapped to 2 $X(s) = 2$, and the events $\{(H, T), (T, H)\}$ gets mapped to 1 $(X(s) = 1)$, the event (T, T) gets mapped to 0 $(X(s) = 0)$.

What are other possible random variables?

12.8 Distributions

We now have two main concepts in this section – probability and random variables. Given a sample space S and the same experiment, both probability and random variables take events as their inputs. But they output different things (probabilities measure the "size" of events, random variables give a number in a way that the analyst chose to define the random variable). How do the two concepts relate?

The concept of distributions is the natural bridge between these two concepts.

Definition 12.7 (Distribution of a random variable)**.** A distribution of a random variable is a function that specifies the probabilities of all events associated with that random variable. There are several types of distributions: A probability mass function for a discrete random variable and probability density function for a continuous random variable.

Notice how the definition of distributions combines two ideas of random variables and probabilities of events. First, the distribution considers a random variable, call it X. X can take a number of possible numeric values.

Example 12.8. Consider three binary outcomes, one for each patient recovering from a disease: R_i denotes the event in which patient i $(i = 1, 2, 3)$ recovers from a disease. R_1, R_2, and R_3. How would we represent the total number of people who end up recovering from the disease?
Solution. Define the random variable X be the total number of people (out of three) who recover from the disease. Random variables are functions, that take as an input a set of events (in the sample space S) and deterministically assigns them to a number of the analyst's choice.

Recall that with each of these numerical values there is a class of *events*. In the previous example,

- For $X = 3$ there is one outcome (R_1, R_2, R_3)
- For $X = 1$ there are multiple $\{(R_1, R_2^c, R_3^c), (R_1^c, R_2, R_3^c), (R_1^c, R_2^c, R_3), \}$

Now, the thing to notice here is that each of these events naturally come with a probability associated with them. That is, $P(R_1, R_2, R_3)$ is a number from 0 to 1, as is $P(R_1, R_2^c, R_3^c)$. These all have probabilities because they are in the sample space S. The function that tells us these probabilities that are associated with a numerical value of a random variable is called a distribution.

In other words, a random variable X *induces a probability distribution* P (sometimes written P_X to emphasize that the probability density is about the r.v. X)

Discrete Random Variables

The formal definition of a random variable is easier to given by separating out two cases: discrete random variables when the numeric summaries of the events are discrete, and continuous random variables when they are continuous.

Definition 12.8 (Discrete Random Variable)**.** X is a discrete random variable if it can assume only a finite or countably infinite number of distinct values. Examples: number of wars per year, heads or tails.

The distribution of a discrete r.v. is a PMF:

Definition 12.9 (Probability Mass Function)**.** For a discrete random variable X, the probability mass function (Also referred to simply as the "probability distribution.") (PMF), $p(x) = P(X = x)$, assigns probabilities to a countable number of distinct x values such that

1. $0 \leq p(x) \leq 1$
2. $\sum\limits_{y} p(x) = 1$

Example 12.9. For a fair six-sided die, there is an equal probability of rolling any number. Since there are six sides, the probability mass function is then $p(y) = 1/6$ for $y = 1, \ldots, 6$, 0 otherwise.

In a discrete random variable, **cumulative distribution function** (Also referred to simply as the "cumulative distribution" or previously as the "distribution function"), $F(x)$ or $P(X \leq x)$, is the probability that X is less than or equal to some value x, or

$$P(X \leq x) = \sum_{i \leq x} p(i)$$

Properties a CDF must satisfy:

1. $F(x)$ is non-decreasing in x.
2. $\lim\limits_{x \to -\infty} F(x) = 0$ and $\lim\limits_{x \to \infty} F(x) = 1$
3. $F(x)$ is right-continuous.

Note that $P(X > x) = 1 - P(X \leq x)$.

Example 12.10. For a fair die with its value as Y, What are the following?

1. $P(Y \leq 1)$
2. $P(Y \leq 3)$
3. $P(Y \leq 6)$

Continuous Random Variables

We also have a similar definition for *continuous* random variables.

Definition 12.10 (Continuous Random Variable)**.** X is a continuous random variable if there exists a non-negative function $f(x)$ defined for all real $x \in (-\infty, \infty)$, such that for any interval A, $P(X \in A) = \int\limits_{A} f(x)dx$.

Examples: age, income, GNP, temperature.

Definition 12.11 (Probability Density Function)**.** The function f above is called the probability density function (pdf) of X and must satisfy

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Note also that $P(X = x) = 0$ — i.e., the probability of any point y is zero.

For both discrete and continuous random variables, we have a unifying concept of another measure: the cumulative distribution:

Definition 12.12 (Cumulative Distribution Function)**.** Because the probability that a continuous random variable will assume any particular value is zero, we can only make statements about the probability of a continuous random variable being within an interval. The cumulative distribution gives the probability that Y lies on the interval $(-\infty, y)$ and is defined as

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(s)ds.$$

Note that $F(x)$ has similar properties with continuous distributions as it does with discrete - non-decreasing, continuous (not just right-continuous), and $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

We can also make statements about the probability of Y falling in an interval $a \leq y \leq b$.

$$P(a \leq x \leq b) = \int_{a}^{b} f(x)dx$$

The PDF and CDF are linked by the integral: The CDF of the integral of the PDF:

$$f(x) = F'(x) = \frac{dF(x)}{dx}$$

Example 12.11. For $f(y) = 1, \quad 0 < y < 1$, find: (1) The CDF $F(y)$ and (2) The probability $P(0.5 < y < 0.75)$.

Answers to Examples and Exercises

Answer to Example **??**:

1. $5 \times 5 \times 5 = 125$
2. $5 \times 4 \times 3 = 60$
3. $\binom{5}{3} = \frac{5!}{(5-3)!3!} = \frac{5 \times 4}{2 \times 1} = 10$

Answer to Exercise **??**:

1. $\binom{52}{4} = \frac{52!}{(52-4)!4!} = 270725$

Answer to Example **??**:

1. $\{1, 2, 3, 4, 5, 6\}$
2. $\{5, 6\}$
3. $\{1, 2, 7, 8, 9, 10\}$
4. $\{3, 4\}$

Answer to Exercise **??**:

Sample Space: $\{2, 3, 4, 5, 6, 7, 8\}$

1. $\{3, 4, 5, 6, 7\}$
2. $\{4, 5, 6\}$

Answer to Example **??**:

1. $1, 2, 3, 4, 5, 6$
2. $\frac{1}{6}$
3. 0
4. $\frac{1}{2}$
5. $\frac{4}{6} = \frac{2}{3}$
6. 1
7. $A \cup B = \{1, 2, 3, 4, 6\}$, $A \cap B = \{2\}$, $\frac{5}{6}$

Answer to Exercise **??**:

1. $P(X = 5) = \frac{4}{16}$, $P(X = 3) = \frac{2}{16}$, $P(X = 6) = \frac{3}{16}$
2. What is $P(X = 5 \cup X = 3)^C = \frac{10}{16}$?

Answer to Example **??**:

1. $\frac{n_{ab}+n_{ab^c}}{N}$
2. $\frac{n_{ab}+n_{a^cb}}{N}$
3. $\frac{n_{ab}}{N}$
4. $\frac{\frac{n_{ab}}{N}}{\frac{n_{ab}+n_{a^cb}}{N}} = \frac{n_{ab}}{n_{ab}+n_{a^cb}}$
5. $\frac{\frac{n_{ab}}{N}}{\frac{n_{ab}+n_{ab^c}}{N}} = \frac{n_{ab}}{n_{ab}+n_{ab^c}}$

Answer to Example **??**:

$P(1|Odd) = \frac{P(1 \cap Odd)}{P(Odd)} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$

Answer to Example **??**:

We are given that

$$P(D) = .4, P(D^c) = .6, P(S|D) = .5, P(S|D^c) = .9$$

Using this, Bayes' Law and the Law of Total Probability, we know:

$$P(D|S) = \frac{P(D)P(S|D)}{P(D)P(S|D) + P(D^c)P(S|D^c)}$$

$$P(D|S) = \frac{.4 \times .5}{.4 \times .5 + .6 \times .9} = .27$$

Answer to Exercise **??**:

We are given that

$$P(M) = .02, P(C|M) = .95, P(C^c|M^c) = .97$$

$$P(M|C) = \frac{P(C|M)P(M)}{P(C)}$$

$$= \frac{P(C|M)P(M)}{P(C|M)P(M) + P(C|M^c)P(M^c)}$$

$$= \frac{P(C|M)P(M)}{P(C|M)P(M) + [1 - P(C^c|M^c)]P(M^c)}$$

$$= \frac{.95 \times .02}{.95 \times .02 + .03 \times .98} = .38$$

Chapter 13

Summarizing Distributions

13.1 Expectation

We often want to summarize some characteristics of the distribution of a random variable. The most important summary is the expectation (or expected value, or mean), in which the possible values of a random variable are weighted by their probabilities.

Definition 13.1 (Expectation of a Discrete Random Variable)**.** The expected value of a discrete random variable Y is

$$E(Y) = \sum_{y} yP(Y = y) = \sum_{y} yp(y)$$

In words, it is the weighted average of all possible values of Y, weighted by the probability that y occurs. It is not necessarily the number we would expect Y to take on, but the average value of Y after a large number of repetitions of an experiment.

Example 13.1. What is the expectation of a fair, six-sided die?

Expectation of a Continuous Random Variable: The expected value of a continuous random variable is similar in concept to that of the discrete random variable, except that instead of summing using probabilities as weights, we integrate using the density to weight. Hence, the expected value of the continuous variable Y is defined by

$$E(Y) = \int_{y} yf(y)dy$$

Example 13.2 (Expectation of a Continuous Random Variable)**.** Find $E(Y)$ for $f(y) = \frac{1}{1.5}, \quad 0 < y < 1.5$.

Expected Value of a Function

Remember: An Expected Value is a type of weighted average. We can extend this to composite functions. For random variable Y, If Y is Discrete with PMF $p(y)$,

$$E[g(Y)] = \sum_y g(y)p(y)$$

If Y is Continuous with PDF $f(y)$,

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)dy$$

Properties of Expected Values

Dealing with Expectations is easier when the thing inside is a sum. The intuition behind this that Expectation is an integral, which is a type of sum.

1. Expectation of a constant is a constant
$$E(c) = c$$
2. Constants come out
$$E(cg(Y)) = cE(g(Y))$$
3. Expectation is Linear
$$E(g(Y_1) + \cdots + g(Y_n)) = E(g(Y_1)) + \cdots + E(g(Y_n)),$$
regardless of independence
4. Expected Value of Expected Values:
$$E(E(Y)) = E(Y)$$
(because the expected value of a random variable is a constant)

Finally, if X and Y are independent, even products are easy:

$$X \text{ and } Y \text{areindependent } \Rightarrow E(XY) = E(X)E(Y)$$

Conditional Expectation: With joint distributions, we are often interested in the expected value of a variable Y if we could hold the other variable X fixed. This is the conditional expectation of Y given $X = x$:

1. Y discrete: $E(Y|X = x) = \sum_y y p_{Y|X}(y|x)$
2. Y continuous: $E(Y|X = x) = \int_y y f_{Y|X}(y|x)dy$

The conditional expectation is often used for prediction when one knows the value of X but not Y

13.2 Variance and Covariance

We can also look at other summaries of the distribution, which build on the idea of taking expectations. Variance tells us about the "spread" of the distribution; it is the expected value of the squared deviations from the mean of the distribution. The standard deviation is simply the square root of the variance.

Definition 13.2 (Variance)**.** The Variance of a Random Variable Y is

$$\text{Var}(Y) = E[(Y - E(Y))^2] = E(Y^2) - [E(Y)]^2$$

The Standard Deviation is the square root of the variance :

$$SD(Y) = \sigma_Y = \sqrt{\text{Var}(Y)}$$

Example 13.3. Given the following PMF:

$$f(x) = \begin{cases} \frac{3!}{x!(3-x)!}(\frac{1}{2})^3 & x = 0, 1, 2, 3 \\ 0 & otherwise \end{cases}$$

What is $\text{Var}(x)$?

Hint: First calculate $E(X)$ and $E(X^2)$

Definition 13.3 (Covariance)**.** The covariance measures the degree to which two random variables vary together; if the covariance between X and Y is positive, X tends to be larger than its mean when Y is larger than its mean.

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

We can also write this as

$$\begin{aligned} \text{Cov}(X, Y) &= E\left(XY - XE(Y) - E(X)Y + E(X)E(Y)\right) \\ &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

The covariance of a variable with itself is the variance of that variable.

The Covariance is unfortunately hard to interpret in magnitude. The correlation is a standardized version of the covariance, and always ranges from -1 to 1.

Definition 13.4 (Correlation)**.** The correlation coefficient is the covariance divided by the standard deviations of X and Y. It is a unitless measure and always takes on values in the interval $[-1, 1]$.

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{SD(X)SD(Y)}$$

Properties of Variance and Covariance:

1. $\text{Var}(c) = 0$
2. $\text{Var}(cY) = c^2\text{Var}(Y)$
3. $\text{Cov}(Y, Y) = \text{Var}(Y)$
4. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$

5. $\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$
6. $\text{Cov}(X + a, Y) = \text{Cov}(X, Y)$
7. $\text{Cov}(X + Z, Y + W) = \text{Cov}(X, Y) + \text{Cov}(X, W) + \text{Cov}(Z, Y) + \text{Cov}(Z, W)$
8. $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$

Exercise 13.1 (Expectation and Variance)**.** Suppose we have a PMF with the following characteristics:

$$\begin{aligned} P(X=-2) &= \frac{1}{5} \\ P(X=-1) &= \frac{1}{6} \\ P(X=0) &= \frac{1}{5} \\ P(X=1) &= \frac{1}{15} \\ P(X=2) &= \frac{11}{30} \end{aligned}$$

1. Calculate the expected value of X

Define the random variable $Y = X^2$.

2. Calculate the expected value of Y. (Hint: It would help to derive the PMF of Y first in order to calculate the expected value of Y in a straightforward way)
3. Calculate the variance of X.

Exercise 13.2. Given the following PDF:

$$f(x) = \begin{cases} \frac{3}{10}(3x - x^2) & 0 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

Find the expectation and variance of X.

Exercise 13.3. Find the mean and standard deviation of random variable X. The PDF of this X is as follows:

$$f(x) = \begin{cases} \frac{1}{4}x & 0 \le x \le 2 \\ \frac{1}{4}(4 - x) & 2 \le x \le 4 \\ 0 & \text{otherwise} \end{cases}$$

Next, calculate $P(X < \mu - \sigma)$ Remember, μ is the mean and σ is the standard deviation.

13.3 Common Distributions

Two *discrete* distributions used often are:

Definition 13.5 (Binomial Distribution)**.** Y is distributed binomial if it represents the number of "successes" observed in n independent, identical "trials," where the probability of success in any trial is p and the probability of failure is $q = 1 - p$.

For any particular sequence of y successes and $n-y$ failures, the probability of obtaining that sequence is $p^y q^{n-y}$ (by the multiplicative law and independence). However, there are $\binom{n}{y} = \frac{n!}{(n-y)!y!}$ ways of obtaining a sequence with y successes and $n-y$ failures. So the binomial distribution is given by

$$p(y) = \binom{n}{y} p^y q^{n-y}, \quad y = 0, 1, 2, \ldots, n$$

with mean $\mu = E(Y) = np$ and variance $\sigma^2 = \mathrm{Var}(Y) = npq$.

Example 13.4. Republicans vote for Democrat-sponsored bills 2% of the time. What is the probability that out of 10 Republicans questioned, half voted for a particular Democrat-sponsored bill? What is the mean number of Republicans voting for Democrat-sponsored bills? The variance? 1. $P(Y = 5) =$ 2. $E(Y) =$ 3. $\mathrm{Var}(Y) = 6$

Definition 13.6 (Poisson Distribution)**.** A random variable Y has a Poisson distribution if

$$P(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}, \quad y = 0, 1, 2, \ldots, \quad \lambda > 0$$

The Poisson has the unusual feature that its expectation equals its variance: $E(Y) = \mathrm{Var}(Y) = \lambda$. The Poisson distribution is often used to model rare event counts: counts of the number of events that occur during some unit of time. λ is often called the "arrival rate."

Example 13.5. Border disputes occur between two countries through a Poisson Distribution, at a rate of 2 per month. What is the probability of 0, 2, and less than 5 disputes occurring in a month?

Two *continuous* distributions used often are:

Definition 13.7 (Uniform Distribution)**.** A random variable Y has a continuous uniform distribution on the interval (α, β) if its density is given by

$$f(y) = \frac{1}{\beta - \alpha}, \quad \alpha \le y \le \beta$$

The mean and variance of Y are $E(Y) = \frac{\alpha+\beta}{2}$ and $\mathrm{Var}(Y) = \frac{(\beta-\alpha)^2}{12}$.

Example 13.6. For Y uniformly distributed over $(1, 3)$, what are the following probabilities?

1. $P(Y = 2)$
2. Its density evaluated at 2, or $f(2)$
3. $P(Y \le 2)$
4. $P(Y > 2)$

Definition 13.8 (Normal Distribution)**.** A random variable Y is normally distributed with mean $E(Y) = \mu$ and variance $\mathrm{Var}(Y) = \sigma^2$ if its density is

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

See Figure Figure **??** are various Normal Distributions with the same $\mu = 1$ and two versions of the variance.

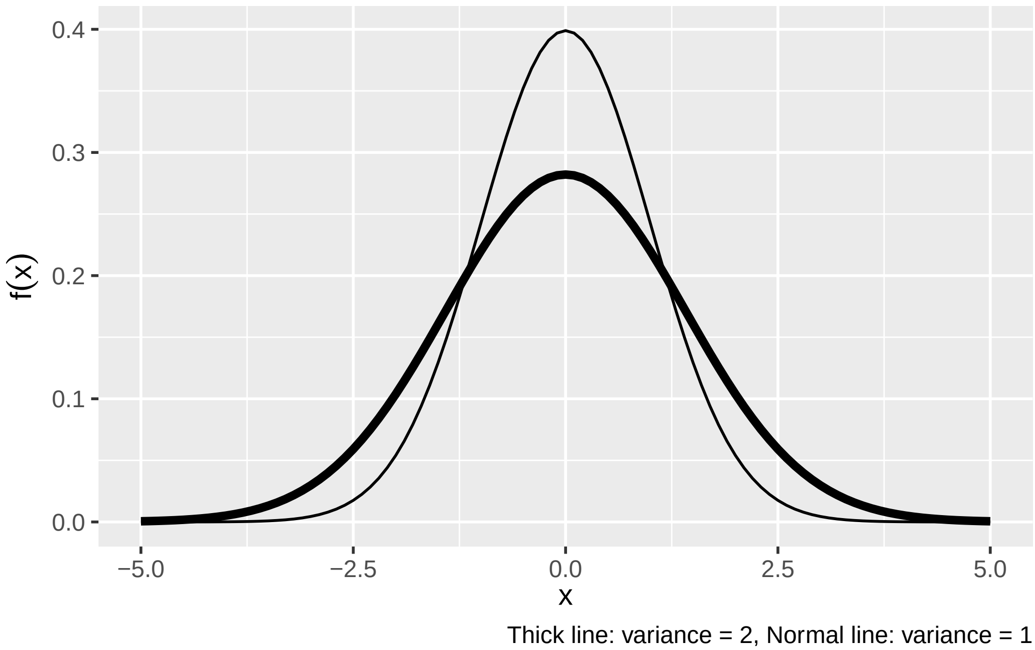


Figure 13.1: Normal Distribution Density

13.4 Joint Distributions

Often, we are interested in two or more random variables defined on the same sample space. The distribution of these variables is called a joint distribution. Joint distributions can be made up of any combination of discrete and continuous random variables.

Joint Probability Distribution: If both X and Y are random variable, their joint probability mass/density function assigns probabilities to each pair of outcomes

Discrete:

$$p(x, y) = P(X = x, Y = y)$$

such that $p(x, y) \in [0, 1]$ and

$$\sum \sum p(x, y) = 1$$

Continuous:

$$f(x, y); P((X, Y) \in A) = \int\int_A f(x, y) dx dy$$

s.t. $f(x, y) \geq 0$ and

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$$

If X and Y are independent, then $P(X = x, Y = y) = P(X = x)P(Y = y)$ and $f(x, y) = f(x)f(y)$

Marginal Probability Distribution: probability distribution of only one of the two variables (ignoring information about the other variable), we can obtain the marginal distribution by summing/integrating across the variable that we don't care about:

- Discrete: $p_X(x) = \sum_i p(x, y_i)$
- Continuous: $f_X(x) = \int_{-\infty}^{\infty} f(x, y)dy$

Conditional Probability Distribution: probability distribution for one variable, holding the other variable fixed. Recalling from the previous lecture that $P(A|B) = \frac{P(A \cap B)}{P(B)}$, we can write the conditional distribution as

- Discrete: $p_{Y|X}(y|x) = \frac{p(x,y)}{p_X(x)}, \quad p_X(x) > 0$
- Continuous: $f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}, \quad f_X(x) > 0$

Exercise 13.4. Suppose we are interested in the outcomes of flipping a coin and rolling a 6-sided die at the same time. The sample space for this process contains 12 elements:

$$\{(H, 1), (H, 2), (H, 3), (H, 4), (H, 5), (H, 6), (T, 1), (T, 2), (T, 3), (T, 4), (T, 5), (T, 6)\}$$

We can define two random variables X and Y such that $X = 1$ if heads and $X = 0$ if tails, while Y equals the number on the die.

We can then make statements about the joint distribution of X and Y. What are the following?

1. $P(X = x)$
2. $P(Y = y)$
3. $P(X = x, Y = y)$
4. $P(X = x|Y = y)$
5. Are X and Y independent?

Answers to Examples and Exercises

Answer to Example Example **??**:

$E(Y) = 7/2$

We would never expect the result of a rolled die to be 7/2, but that would be the average over a large number of rolls of the die.

Answer to Example **??**

0.75

Answer to Example **??**:

$E(X) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{3}{2}$

Since there is a 1 to 1 mapping from X to $X^2 : E(X^2) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 4 \times \frac{3}{8} + 9 \times \frac{1}{8} = \frac{24}{8} = 3$

$$\begin{aligned}
\text{Var}(x) &= E(X^2) - E(x)^2 \\
&= 3 - (\frac{3}{2})^2 \\
&= \frac{3}{4}
\end{aligned}$$

Answer to Exercise **??**:

1. $E(X) = -2(\frac{1}{5}) + -1(\frac{1}{6}) + 0(\frac{1}{5}) + 1(\frac{1}{15}) + 2(\frac{11}{30}) = \frac{7}{30}$
2. $E(Y) = 0(\frac{1}{5}) + 1(\frac{7}{30}) + 4(\frac{17}{30}) = \frac{5}{2}$
3.

$$\begin{aligned}\mathrm{Var}(X) &= E[X^2] - E[X]^2 \\ &= E(Y) - E(X)^2 \\ &= \frac{5}{2} - \frac{7}{30}^2 \approx 2.45\end{aligned}$$

Answer to Exercise **??**:

1. expectation $= \frac{6}{5}$, variance $= \frac{6}{25}$

Answer to Exercise **??**:

1. mean $= 2$, standard deviation $= \sqrt{\frac{2}{3}}$
2. $\frac{1}{8}(2 - \sqrt{\frac{2}{3}})^2$

Chapter 14

Learning from Data

14.1 Summarizing Data

So far, we've talked about distributions in a theoretical sense, looking at different properties of random variables. We don't observe random variables; we observe realizations of the random variable. These realizations of events are roughly equivalent to what we mean by "data".

Sample mean: This is the most common measure of central tendency, calculated by summing across the observations and dividing by the number of observations.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

The sample mean is an *estimate* of the expected value of a distribution.

Dispersion: We also typically want to know how spread out the data are relative to the center of the observed distribution. There are several ways to measure dispersion.

Sample variance: The sample variance is the sum of the squared deviations from the sample mean, divided by the number of observations minus 1.

$$\hat{\mathrm{Var}}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Again, this is an *estimate* of the variance of a random variable; we divide by $n-1$ instead of n in order to get an unbiased estimate.

Standard deviation: The sample standard deviation is the square root of the sample variance.

$$\hat{SD}(X) = \sqrt{\hat{\mathrm{Var}}(X)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Covariance and Correlation: Both of these quantities measure the degree to which two variables vary together, and are estimates of the covariance and correlation of two random variables as defined above.

1. **Sample covariance**: $\hat{\mathrm{Cov}}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

2. **Sample correlation**: $\hat{\text{Corr}} = \frac{\hat{\text{Cov}}(X,Y)}{\sqrt{\hat{\text{Var}}(X)\hat{\text{Var}}(Y)}}$

Example 14.1. Example: Using the above table, calculate the sample versions of:

1. $\text{Cov}(X, Y)$
2. $\text{Corr}(X, Y)$

14.2 Law of Large Numbers

In probability theory, asymptotic analysis is the study of limiting behavior. By limiting behavior, we mean the behavior of some random process as the number of observations gets larger and larger.

Why is this important? We rarely know the true process governing the events we see in the social world. It is helpful to understand how such unknown processes theoretically must behave and asymptotic theory helps us do this.

Theorem 14.1 (Law of Large Numbers (LLN)). *For any draw of independent random variables with the same mean μ, the sample average after n draws, $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + ... + X_n)$, converges in probability to the expected value of X, μ as $n \to \infty$:*

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0$$

A shorthand of which is $\bar{X}_n \xrightarrow{p} \mu$, where the arrow is read as "converges in probability to" as $n \to \infty$. In other words, $P(\lim_{n\to\infty} \bar{X}_n = \mu) = 1$. This is an important motivation for the widespread use of the sample mean, as well as the intuition link between averages and expected values.

More precisely this version of the LLN is called the *weak* law of large numbers because it leaves open the possibility that $|\bar{X}_n - \mu| > \varepsilon$ occurs many times. The *strong* law of large numbers states that, under a few more conditions, the probability that the limit of the sample average is the true mean is 1 (and other possibilities occur with probability 0), but the difference is rarely consequential in practice.

The Strong Law of Large Numbers holds so long as the expected value exists; no other assumptions are needed. However, the rate of convergence will differ greatly depending on the distribution underlying the observed data. When extreme observations occur often (i.e. kurtosis is large), the rate of convergence is much slower. Cf. The distribution of financial returns.

14.3 Central Limit Theorem

We are now finally ready to revisit, with a bit more precise terms, the two pillars of statistical theory we motivated Probability with.

Theorem 14.2 (Central Limit Theorem (i.i.d. case)). *Let $\{X_n\} = \{X_1, X_2, ...\}$ be a sequence of i.i.d. random variables with finite mean (μ) and variance (σ^2). Then, the sample mean $\bar{X}_n = \frac{X_1+X_2+\cdots+X_n}{n}$ increasingly converges into a Normal distribution.*

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathit{Normal}(0, 1),$$

Another way to write this as a probability statement is that for all real numbers a,

$$P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq a\right) \to \Phi(a)$$

as $n \to \infty$, where

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx$$

is the CDF of a Normal distribution with mean 0 and variance 1.

This result means that, as n grows, the distribution of the sample mean $\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ is approximately normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, i.e.,

$$\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

The standard deviation of $\bar{X}_n$ (which is roughly a measure of the precision of $\bar{X}_n$ as an estimator of μ) decreases at the rate $1/\sqrt{n}$, so, for example, to increase its precision by 10 (i.e., to get one more digit right), one needs to collect $10^2 = 100$ times more units of data.

Intuitively, this result also justifies that whenever a lot of small, independent processes somehow combine together to form the realized observations, practitioners often feel comfortable assuming Normality.

Warmup Questions Solutions

Linear Algebra

Vectors

Define the vectors $u = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$, $v = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}$, and the scalar $c = 2$.

1. $u + v = \begin{pmatrix} 5 \\ 7 \\ 9 \end{pmatrix}$
2. $cv = \begin{pmatrix} 8 \\ 10 \\ 12 \end{pmatrix}$
3. $u \cdot v = 1(4) + 2(5) + 3(6) = 32$

Are the following sets of vectors linearly independent?

1. $u = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, $v = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$

$\rightsquigarrow$ No:

$$2u = \begin{pmatrix} 2 \\ 4 \end{pmatrix}, v = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$$

so infinitely many linear combinations of u and v that amount to 0 exist.

2. $u = \begin{pmatrix} 1 \\ 2 \\ 5 \end{pmatrix}$, $v = \begin{pmatrix} 3 \\ 7 \\ 9 \end{pmatrix}$

$\rightsquigarrow$ Yes: we cannot find linear combination of these two vectors that would amount to zero.

3. $a = \begin{pmatrix} 2 \\ -1 \\ 1 \end{pmatrix}$, $b = \begin{pmatrix} 3 \\ -4 \\ -2 \end{pmatrix}$, $c = \begin{pmatrix} 5 \\ -10 \\ -8 \end{pmatrix}$

$\rightsquigarrow$ No: After playing around with some numbers, we can find that

$$-2a = \begin{pmatrix} -4 \\ 2 \\ -2 \end{pmatrix}, 3b = \begin{pmatrix} 9 \\ -12 \\ -6 \end{pmatrix}, -1c = \begin{pmatrix} -5 \\ 10 \\ 8 \end{pmatrix}$$

So

$$-2a + 3b - c = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

i.e., a linear combination of these three vectors that would amount to zero exists.

Matrices

$$\mathbf{A} = \begin{pmatrix} 7 & 5 & 1 \\ 11 & 9 & 3 \\ 2 & 14 & 21 \\ 4 & 1 & 5 \end{pmatrix}$$

What is the dimensionality of matrix $\mathbf{A}$? 4×3

What is the element a_{23} of $\mathbf{A}$? 3

Given that

$$\mathbf{B} = \begin{pmatrix} 1 & 2 & 8 \\ 3 & 9 & 11 \\ 4 & 7 & 5 \\ 5 & 1 & 9 \end{pmatrix}$$

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} 8 & 7 & 9 \\ 14 & 18 & 14 \\ 6 & 21 & 26 \\ 9 & 2 & 14 \end{pmatrix}$$

Given that

$$\mathbf{C} = \begin{pmatrix} 1 & 2 & 8 \\ 3 & 9 & 11 \\ 4 & 7 & 5 \end{pmatrix}$$

$$\mathbf{A} + \mathbf{C} = \text{No solution, matrices non-conformable}$$

Given that

$$c = 2$$

$$c\mathbf{A} = \begin{pmatrix} 14 & 10 & 2 \\ 22 & 18 & 6 \\ 4 & 28 & 42 \\ 8 & 2 & 10 \end{pmatrix}$$

Operations

Summation

Simplify the following

1. $\sum_{i=1}^{3} i = 1 + 2 + 3 = 6$

2. $\sum_{k=1}^{3}(3k + 2) = 3\sum_{k=1}^{3} k + \sum_{k=1}^{3} 2 = 3 \times 6 + 3 \times 2 = 24$

3. $\sum_{i=1}^{4}(3k + i + 2) = 3\sum_{i=1}^{4} k + \sum_{i=1}^{4} i + \sum_{i=1}^{4} 2 = 12k + 10 + 8 = 12k + 18$

Products

1. $\prod_{i=1}^{3} i = 1 \cdot 2 \cdot 3 = 6$

2. $\prod_{k=1}^{3}(3k + 2) = (3 + 2) \cdot (6 + 2) \cdot (9 + 2) = 440$

Logs and exponents

Simplify the following

1. $4^2 = 16$
2. $4^2 2^3 = 2^{2\cdot 2} 2^3 = 2^{4+3} = 128$
3. $\log_{10} 100 = \log_{10} 10^2 = 2$
4. $\log_2 4 = \log_2 2^2 = 2$
5. when log is the natural log, $\log e = \log_e e^1 = 1$
6. when a, b, c are each constants, $e^a e^b e^c = e^{a+b+c}$,
7. $\log 0 =$ undefined – no exponentiation of anything will result in a 0.
8. $e^0 = 1$ – any number raised to the 0 is always 1.
9. $e^1 = e$ – any number raised to the 1 is always itself
10. $\log e^2 = \log_e e^2 = 2$

Limits

Find the limit of the following.

1. $\lim_{x\to 2}(x - 1) = 1$
2. $\lim_{x\to 2} \frac{(x-2)(x-1)}{(x-2)} = 1$, though note that the original function $\frac{(x-2)(x-1)}{(x-2)}$ would have been undefined at $x = 2$ because of a divide by zero problem; otherwise it would have been equal to $x - 1$.
3. $\lim_{x\to 2} \frac{x^2-3x+2}{x-2} = 1$, same as above.

Calculus

For each of the following functions $f(x)$, find the derivative $f'(x)$ or $\frac{d}{dx} f(x)$

1. $f(x) = c$, $f'(x) = 0$

2. $f(x) = x$, $f'(x) = 1$
3. $f(x) = x^2$, $f'(x) = 2x$
4. $f(x) = x^3$, $f'(x) = 3x^2$
5. $f(x) = 3x^2 + 2x^{1/3}$, $f'(x) = 6x + \frac{2}{3}x^{-2/3}$
6. $f(x) = (x^3)(2x^4)$, $f'(x) = \frac{d}{dx}2x^7 = 14x^6$

Optimization

For each of the followng functions $f(x)$, does a maximum and minimum exist in the domain $x \in \mathbf{R}$? If so, for what are those values and for which values of x?

1. $f(x) = x \rightsquigarrow$ neither exists.
2. $f(x) = x^2 \rightsquigarrow$ a minimum $f(x) = 0$ exists at $x = 0$, but not a maximum.
3. $f(x) = -(x-2)^2 \rightsquigarrow$ a maximum $f(x) = 0$ exists at $x = 2$, but not a minimum.

If you are stuck, please try sketching out a picture of each of the functions.

Probability

1. If there are 12 cards, numbered 1 to 12, and 4 cards are chosen, $\binom{12}{4} = \frac{12 \cdot 11 \cdot 10 \cdot 9}{4!} = 495$ possible hands exist (unordered, without replacement) .

2. Let $A = \{1, 3, 5, 7, 8\}$ and $B = \{2, 4, 7, 8, 12, 13\}$. Then $A \cup B = \{1, 2, 3, 4, 5, 7, 8, 12, 13\}$, $A \cap B = \{7, 8\}$? If A is a subset of the Sample Space $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, then the complement $A^C = \{2, 4, 6, 9, 10\}$

3. If we roll two fair dice, what is the probability that their sum would be 11? $\rightsquigarrow \frac{1}{18}$

4. If we roll two fair dice, what is the probability that their sum would be 12? $\rightsquigarrow \frac{1}{36}$. There are two independent dice, so $6^2 = 36$ options in total. While the previous question had two possibilities for a sum of 11 (5,6 and 6,5), there is only one possibility out of 36 for a sum of 12 (6,6).