University of Chicago Political Science Math Prefresher

Anton Strezhnev

9/12/2022

Table of contents

1 Overview **3**

1.1 Introduction . . . 3

1.2 Course Booklet . . . 3

1.3 Schedule . . . 3

1.4 Software . . . 4

1.5 Acknowledgments . . . 4

2 Sets, Operations, and Functions **5**

2.1 Sets . . . 5

Sets **7**

Sets **8**

2.2 Metric spaces . . . 8

2.3 Operators; Sum and Product notation . . . 9

Operators **12**

Operators **13**

2.4 Introduction to Functions . . . 13

Functions **15**

Functions **16**

2.5 Logarithms and Exponents . . . 16

Logarithms **19**

2.6 Graphing Functions . . . 19

2.7 Solving for Variables and Finding Roots . . . 20

Solving **21**

Roots **22**

3 Limits **23**

Example: The Central Limit Theorem . . . 23

Example: The Law of Large Numbers . . . 24

3.1 Sequences 25

Sequences **26**

3.2 The Limit of a Sequence 27

Ratios **28**

Limits **29**

3.3 Limits of a Function 29

Limits of a function **30**

Limits of a function **31**

Limits of ratios **32**

Limits of a function **33**

3.4 Continuity 33

Continuity **35**

Continuity **36**

4 Calculus **38**

Example: The Mean is a Type of Integral 38

4.1 Derivatives 39

Derivative **40**

Properties of derivatives 41

Power Rule **42**

Derivatives **44**

4.2 Higher-Order Derivatives (Derivatives of Derivatives of Derivatives) 44

Succession of derivatives **46**

4.3 Composite Functions and the Chain Rule 46

Composite functions **47**

Composite Exponent **48**

4.4 Derivatives of natural logs and the exponent 48

Derivative of Exponents/Logs **49**

Derivatives of natural exponential function (e) 49

Derivatives of exponents **51**

Derivatives of logarithms . . . 51

Derivatives of logs **53**

Outline of Proof . . . 53

4.5 Partial Derivatives . . . 54

Partial derivatives **55**

Partial derivatives **56**

4.6 Taylor Series Approximation . . . 56

4.7 The Indefinite Integration . . . 57

Antiderivative **58**

Antiderivative **59**

4.8 Indefinite Integral . . . 59

Graphing **60**

Common Rules of Integration . . . 61

Common Integration **62**

4.9 The Definite Integral: The Area under the Curve . . . 62

The Definite Integral (Riemann) **64**

First Fundamental Theorem of Calculus **65**

Second Fundamental Theorem of Calculus **66**

Definite Integral of a monomial **67**

Indefinite integrals **68**

Common Rules for Definite Integrals . . . 68

Definite integrals **69**

4.10 Integration by Substitution . . . 69

Integration by Substitutiton II **71**

4.11 Integration by Parts . . . 71

Integration by parts **72**

Integration by parts **73**

5 Optimization 74
Example: Meltzer-Richard 74
5.1 Maxima and Minima 76

Plotting a maximum and minimum 78

Maxima and Minima by drawing 80
5.2 Concavity of a Function 80

Concave Function 81

Convex Function 82

Quasiconcave Function 83

Quasiconvex Function 84
Quadratic Forms 84
Definiteness of Quadratic Forms 85
5.3 FOC and SOC 85
First Order Conditions 86

Gradient 87

Critical Point 88
Second Order Conditions 89

Hessian 90

Max and min with two dimensions 91
Definiteness and Concavity 91
5.4 Global Maxima and Minima 92

Optimization 93
5.5 Constrained Optimization 94
Equality Constraints 96

Constrained optimization with two goods and a budget constraint 99
5.6 Inequality Constraints 100

Constrained optimization 103
5.7 Kuhn-Tucker Conditions 104

Kuhn-Tucker with two variables 106

Kuhn-Tucker with logs **108**
5.8 Applications of Quadratic Forms 109

6 Probability Theory **111**
6.1 Counting rules 111

Counting Possibilities **112**

Counting **114**

7 Counting **115**
7.1 Probability 115
Probability Definitions: Formal and Informal 115

Probability **116**

Axioms of Probability **117**
Probability Operations 117

Probability **118**

Probability **119**
7.2 Conditional Probability and Bayes Rule 119

Conditional Probability 1 **120**

Conditional Probability 2 **121**

Multiplicative Law of Probability **122**

Law of total probability **123**

Bayes' Rule **125**

Conditional Probability **126**
7.3 Independence 126

Independence **127**
7.4 Random Variables 128

Random Variable **130**
7.5 Distributions 130

Distribution of a random variable **131**

Total Number of Occurrences **132**
Discrete Random Variables 132

Discrete Random Variable **133**

Probability Mass Function **134**
Continuous Random Variables 135

Continuous Random Variable **136**

Probability density function **137**

Continuous R.V. **138**
7.6 Joint Distributions 138

Discrete, Joint Distributions **140**
7.7 Expectation 140

Expectation of a discrete R.V. **141**

Expectation of a discrete R.V. **142**

Expectation of a continuous R.V. **143**
Expected Value of a Function 143
Properties of Expected Values 143
7.8 Variance and Covariance 144

Covariance **146**

Correlation **147**

Expectation and Variance 1 **148**

Expectation and Variance 2 **149**

Expectation and Variance 3 **150**
7.9 Distributions 150

Binomial Distribution **151**

Binomial distribution **152**

Poisson Distribution **153**

Poisson Distribution **154**

Uniform Distribution **155**

Uniform **156**

Normal Distribution **157**

7.10 Summarizing Observed Events (Data) 157

Sample Covariance and Correlation **160**

7.11 Asymptotic Theory 160

7.11.1 CLT and LLN 160

Central Limit Theorem **161**

Weak Law of Large Numbers (LLN) **162**

7.11.2 Big $\mathcal{O}$ Notation 162

1 Overview

1.1 Introduction

The 2022 UChicago Math Prefresher for incoming Political Science graduate students will be held from September 12-14; September 19-21 and September 23rd. The course is designed as a brief review of math fundamentals – calculus, optimization, probability theory and linear algebra among other topics – as well as an introduction to programming in the R statistical computing language. The course is entirely optional and there are no grades or assignments but we encourage all incoming graduate students to attend if they are able.

1.2 Course Booklet

The course notes for the math and programming sections as well as all practice problems are available on this website and can be accessed by navigating the menus in the sidebar.

1.3 Schedule

The prefresher will run for a total of seven days September 12-14, September 19-21 and September 23rd, with breaks for the APSA conference and the new student orientation. Each day will run from around 9am to 4pm with many breaks in between. We will be meeting in room 407 of Pick Hall.

The morning will focus on math instruction. We will have two one hour sessions from 9:30am - 10:30am and 10:45am-11:45am, with a ~15 minute break in between. These sessions will involve a combination of lectures and working through practice problems.

We will break for lunch from 12:00pm-1:00pm. On September 13th and Spetember 19th, we will have a catered lunch with a faculty member guest. Otherwise, you are free to explore the campus for various lunch options.

The afternoon will focus on coding instruction with lecture/demonstration from 1:30pm-2:45pm. After a short break you will work together on a variety of coding exercises from 3:00-3:30pm. In the last 30 minutes we will regroup to wrap up and discuss any questions on the material.

1.4 Software

As the afternoons of the prefresher will involve instruction in coding, you should be sure to bring a laptop and a charging cable. In addition, prior to the start of the prefresher, please make sure to have installed the following on your computer:

- R (version 4.2.1 or higher)
- RStudio Desktop Open Source License (this is the primary IDE or integrated development environment in which we will be working)
- LaTeX: This is primarily to allow you to generate PDF documents using RMarkdown. We will use the TinyTeX LaTeX distribution which is designed to be minimalist and tailored specifically for R users. After installing R and RStudio, open up an instance of R, install the 'tinytex' package and run the `install_tinytex()` command

```
install.packages('tinytex')
tinytex::install_tinytex()
```

We will also spend some time discussing document preparation and typesetting using LaTeX and Markdown. For the former, we will be using the popular cloud platform Overleaf, which allows for collaborative document editing and streamlines a lot of the irritating parts of typesetting in LaTeX. You should register for an account using your university e-mail as all University of Chicago students and faculty have access to an Overleaf Pro account for free.

You are also welcome to install a LaTeX editor on your local machine to work alongside the TinyTeX distribution or any other TeX distribution that you prefer such as TexMaker

1.5 Acknowledgments

This prefresher draws heavily on the wonderful materials that have been developed by over 20 years of instructors at the Harvard Government Math Prefresher that have been so generously distributed under the GPL 3.0 License. Special thanks to Shiro Kuriwaki, Yon Soo Park, and Connor Jerzak for their efforts in converting the original prefresher materials into the easily distributed Markdown format.

2 Sets, Operations, and Functions

2.1 Sets

Sets are the fundamental building blocks of mathematics. 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. Combining sets with operations, relations, metrics, measures, etc… allows us to define useful mathematical structures. For example, the set of *real numbers* ($\mathbb{R}$) has a notion of *order* as well as defined *operations* of addition and multiplication.

Set : A set is any well defined collection of elements. If x is an element of S, $x \in S$.

Examples:

1. The set of choices available to a player in Rock-Paper-Scissors $\{\text{Rock}, \text{Paper}, \text{Scissors}\}$
2. The set of possible outcomes of a roll of a six-sided die $\{1, 2, 3, 4, 5, 6\}$
3. The set of all natural numbers $\mathbb{N}$
4. The set of all real numbers $\mathbb{R}$

Common mathematical notation relevant to sets:

- $\in$ = “is an element of”; $\notin$ = “is not an element of”
- $\forall$ = “for all” (univeral quantifier)
- $\exists$ = “there exists” (existential quantifier)
- $:$ = “such that”

Subset: If every element of set A is also in set B, then A is a *subset* of B. $A \subseteq B$. If, in addition to being a subset of B, A is not equal to B, A is a *proper subset* $A \subset B$.

Empty Set: a set with no elements. $S = \{\}$. It is denoted by the symbol $\emptyset$.

Cardinality: The cardinality of a set S, typically written $|S|$ is the number of members of S.

Many sets are infinite. For example, $\mathbb{N}$ the set of natural numbers $\mathbb{N} = \{0, 1, 2, 3, 4, \ldots\}$ - Sets with cardinality less than $|\mathbb{N}|$ are *countable* - Sets with the same cardinality as *mathbbN*| are *countably infinite* - Sets with greater cardinality than $|\mathbb{N}|$ are *uncountably infinite* (e.g. the real numbers).

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 2.1.

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 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 2.1.

Sets

Suppose you had a pair of four-sided dice. You sum the results from a single toss.

What is the set of possible outcomes?

Consider subsets $A = \{2, 8\}$ and $B = \{2, 3, 7\}$ of the sample space you found. What is

1. A^c
2. $(A \cup B)^c$

2.2 Metric spaces

A *metric space* is a set that has a notion of *distance* - called a "metric" - defined between any two elements (sometimes referred to as "points").

The distance function $d(x, y)$ defines the distance between element x and element y

- The real numbers $\mathbb{R}$ have a single distance function: $d(x, y) = |x - y|$
- In higher-dimensional real space (e.g. $\mathbb{R}^2$), we can define multiple distance metrics between $x = (x_1, x_2)$ and $y = (y_1, y_2)$
  - "Euclidean" distance: $d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$
  - "Taxicab" distance: $d(x, y) = |x_1 - y_1| + |x_2 - y_2|$
  - Chebyshev distance: $d(x, y) = \max\{|x_1 - y_1| + |x_2 - y_2|\}$
- All of these generalize to $\mathbb{R}^n$

A metric is a function that satisfies the following axioms

1. A distance between a point and itself is zero $d(x, x) = 0$
2. The distance between two points is strictly positive $d(x, y) > 0 \forall x \neq y$
3. Distance from x to y is the same as the distance from y to x ($d(x, y) = d(y, x)$)
4. The "triangle inequality" holds: $d(x, z) \leq d(x, y) + d(y, z)$

Once we have a metric space, we can define some additional useful concepts

Ball: A ball of radius r centered at x_0 is a set that contains all points with a distance less than r from x_0.

Sphere: A sphere of radius r centered at x_0 is the set that contains all points with a distance exactly r from x_0.

Interior Point: The point x is an interior point of the set S if x is in S and if there is some ϵ-ball around x that contains only points in S. The **interior** of S is the collection of all interior points in S. The interior can also be defined as the union of all open sets in S.

- If the set S is circular, the interior points are everything inside of the circle, but not on the circle's rim.
- Example: The interior of the set $\{(x,y) : x^2 + y^2 \leq 4\}$ is $\{(x,y) : x^2 + y^2 < 4\}$.

Boundary Point: The point $\mathbf{x}$ is a boundary point of the set S if every ϵ-ball around $\mathbf{x}$ contains both points that are in S and points that are outside S. The **boundary** is the collection of all boundary points.

- If the set S is circular, the boundary points are everything on the circle's rim.
- Example: The boundary of $\{(x,y) : x^2 + y^2 \leq 4\}$ is $\{(x,y) : x^2 + y^2 = 4\}$.

Open: A set S is open if for each point $\mathbf{x}$ in S, there exists an open ϵ-ball around $\mathbf{x}$ completely contained in S.

- If the set S is circular and open, the points contained within the set get infinitely close to the circle's rim, but do not touch it.
- Example: $\{(x,y) : x^2 + y^2 < 4\}$

Closed: A set S is closed if it contains all of its boundary points.

- Alternatively: A set is closed if its complement is open.
- If the set S is circular and closed, the set contains all points within the rim as well as the rim itself.
- Example: $\{(x,y) : x^2 + y^2 \leq 4\}$
- Note: a set may be neither open nor closed. Example: $\{(x,y) : 2 < x^2 + y^2 \leq 4\}$

2.3 Operators; Sum and Product notation

Addition (+), Subtraction (-), multiplication and division are basic operations of arithmetic. In statistics or calculus, we will often 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.

A few important features of sums:

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

Product:

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

Properties:

- $\prod\limits_{i=1}^{n} cx_i = c^n \prod\limits_{i=1}^{n} x_i$
- $\prod\limits_{i=k}^{n} cx_i = c^{n-k+1} \prod\limits_{i=k}^{n} x_i$
- $\prod\limits_{i=1}^{n} (x_i + y_i) = \text{a total mess}$
- $\prod\limits_{i=1}^{n} c = c^n$

Other Useful Operations

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.

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

Example 2.2.

Operators

1. $\sum_{i=1}^{5} i =$
2. $\prod_{i=1}^{5} i =$
3. $14 \mod 4 =$
4. $4! =$

Exercise 2.2.

Operators

Let $x_1 = 4, x_2 = 3, x_3 = 7, x_4 = 11, x_5 = 2$

1. $\sum\limits_{i=1}^{3} (7)x_i$
2. $\sum\limits_{i=1}^{5} 2$
3. $\prod\limits_{i=3}^{5} (2)x_i$

2.4 Introduction to Functions

A **function** is a mapping, or transformation, 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 from A to B maps every value a in set A such that $f(a) \in B$. Functions can be "many-to-one", where many values or combinations of 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: $\mathbf{R}^1$ is the set of all real numbers extending from $-\infty$ to $+\infty$ — i.e., the real number line. $\mathbf{R}^n$ is an n-dimensional space, where each of the n axes extends from $-\infty$ to $+\infty$.

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

Points in $\mathbf{R}^n$ 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:

- $\mathbf{R}^1$: (3)
- $\mathbf{R}^2$: (-15, 5)

- $\mathbf{R}^3$: (86, 4, 0)

Examples of mapping notation:

Function of one variable: $f : \mathbf{R}^1 \to \mathbf{R}^1$

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

Function of two variables: $f : \mathbf{R}^2 \to \mathbf{R}^1$.

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

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

Example 2.3.

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 $x \in [-\infty, \infty]$, $f : x \to x^2$.

Exercise 2.3.

Functions

For each of the following, state whether they are one-to-one or many-to-one functions.

1. For $x \in [-3, \infty]$, $f : x \to x^2$.
2. For $x \in [0, \infty]$, $f : x \to \sqrt{x}$

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

- **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 $f(X) = \{y : y = f(x), x \in X\}$ Most often used when talking about a function $f : \mathbf{R}^1 \to \mathbf{R}^1$.
- **Image**: same as range, but more often used when talking about a function $f : \mathbf{R}^n \to \mathbf{R}^1$.

Some General 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.

2.5 Logarithms and Exponents

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

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: $\quad y = \log_{10}(x) \iff 10^y = x$. The base 10 logarithm is often simply written as "log(x)" with no base denoted.
2. Base e: $\quad y = \log_e(x) \iff e^y = x$. 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 how to optimize likelihood functions.

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

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:

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

Example 2.4.

Logarithms

Evaluate each of the following logarithms

1. $\log_4(16)$
2. $\log_2(16)$

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

3. $\log_4(x^3y^5)$

Exercise 2.4. Evaluate each of the following logarithms

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

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

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

2.6 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.

2.7 Solving for Variables and Finding Roots

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 2.5.

Solving

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 2.5.

Roots

Solve for x:

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

3 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 3.1 (Central Limit Theorem)**.** *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.*

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathit{Normal}(0, 1)$$

where μ is the mean of X and σ is the standard deviation of X. The arrow is read as "converges in distribution to". $\mathit{Normal}(0,1)$ 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 "→". 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 $n \to \infty$, 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 (Weak) Law of Large Numbers:

Theorem 3.2 ((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 3.1 shows how the sample average from many coin tosses converges to the true value : 0.5.

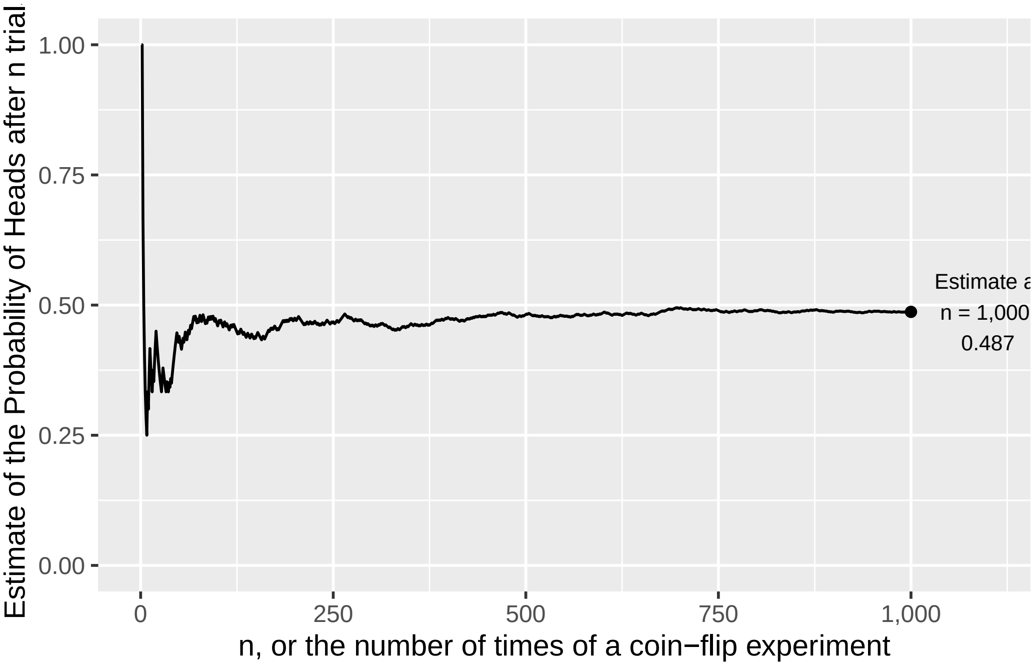


Figure 3.1: As the number of coin tosses goes to infinity, the average probabiity of heads converges to 0.5

3.1 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** $\{x_n\} = \{x_1, x_2, x_3, \ldots, x_n\}$ 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 3.1.

Sequences

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.

Graphing helps you make this point more clearly. See the sequence of $n = 1, ...20$ for each of the three examples in Figure 3.2.

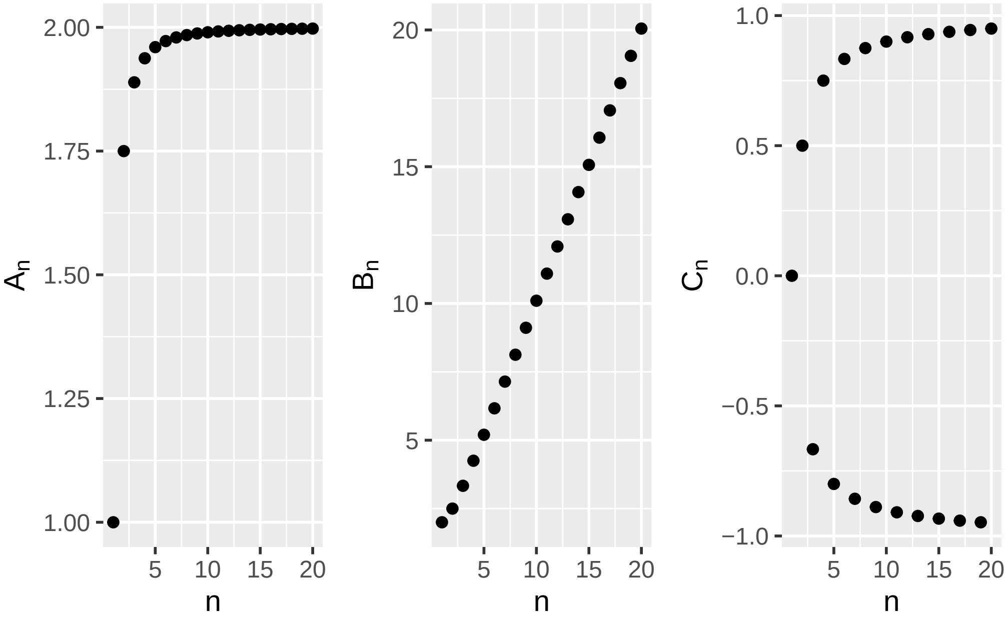


Figure 3.2: Behavior of Some Sequences

3.2 The Limit of a Sequence

The notion of "converging to a limit" is the behavior of the points in Example -@#exm-seqbehav. 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: Limit The sequence $\{y_n\}$ has the limit L, which we write as $\lim\limits_{n\to\infty} y_n = L$, 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 $|y_n| \le K$ 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 3.2.

Ratios

Find the limit of $\lim_{n\to\infty} \frac{n+3}{n}$

Solution. 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 3.1.

Limits

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. $\lim\limits_{n\to\infty} (n^3 - 100n^2)$

3.3 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: $\lim\limits_{x\to c} f(x) = L$. 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 3.1.

Limits of a function

Let $f(x)$ be defined at each point in some open interval containing the point c. Then L equals $\lim\limits_{x \to c} f(x)$ 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 $\lim\limits_{x \to c}$ to exist.

Properties: Let f and g be functions with $\lim\limits_{x \to c} f(x) = k$ and $\lim\limits_{x \to c} g(x) = \ell$.

1. $\lim\limits_{x \to c}[f(x) + g(x)] = \lim\limits_{x \to c} f(x) + \lim\limits_{x \to c} g(x)$
2. $\lim\limits_{x \to c} kf(x) = k \lim\limits_{x \to c} f(x)$
3. $\lim\limits_{x \to c} f(x)g(x) = \left[\lim\limits_{x \to c} f(x)\right] \cdot \left[\lim\limits_{x \to c} g(x)\right]$
4. $\lim\limits_{x \to c} \frac{f(x)}{g(x)} = \frac{\lim_{x \to c} f(x)}{\lim_{x \to c} g(x)}$, provided $\lim\limits_{x \to c} g(x) \neq 0$.

Simple limits of functions can be solved as we did limits of sequences. Just be careful which part of the function is changing.

Example 3.3.

Limits of a function

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,

Example 3.4.

Limits of ratios

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 3.2.

Limits of a function

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?

3.4 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 3.2.

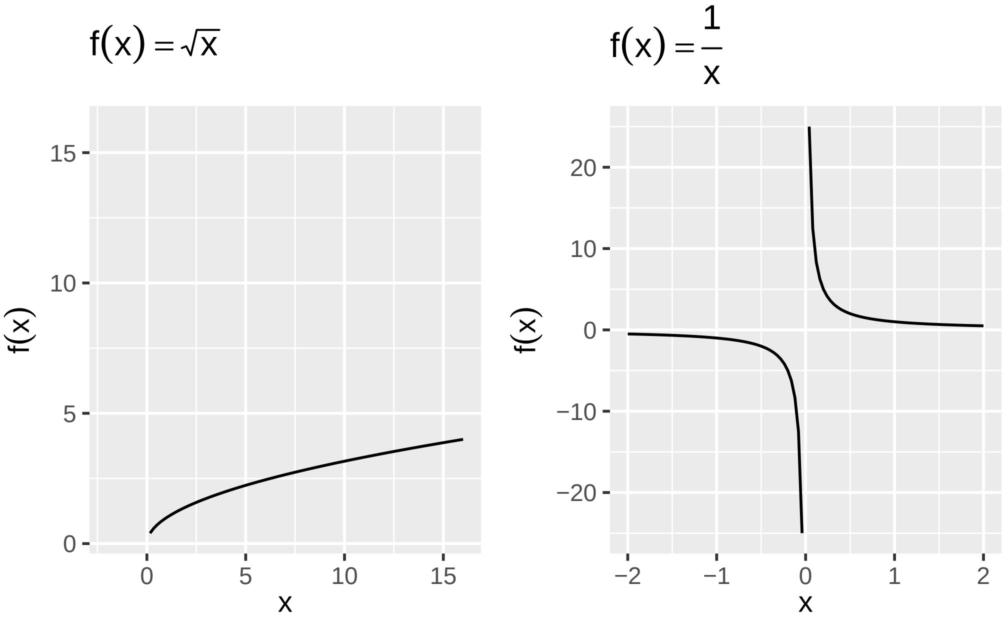


Figure 3.3: Functions which are not defined in some areas

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 3.5.

Continuity

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 3.4, 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.

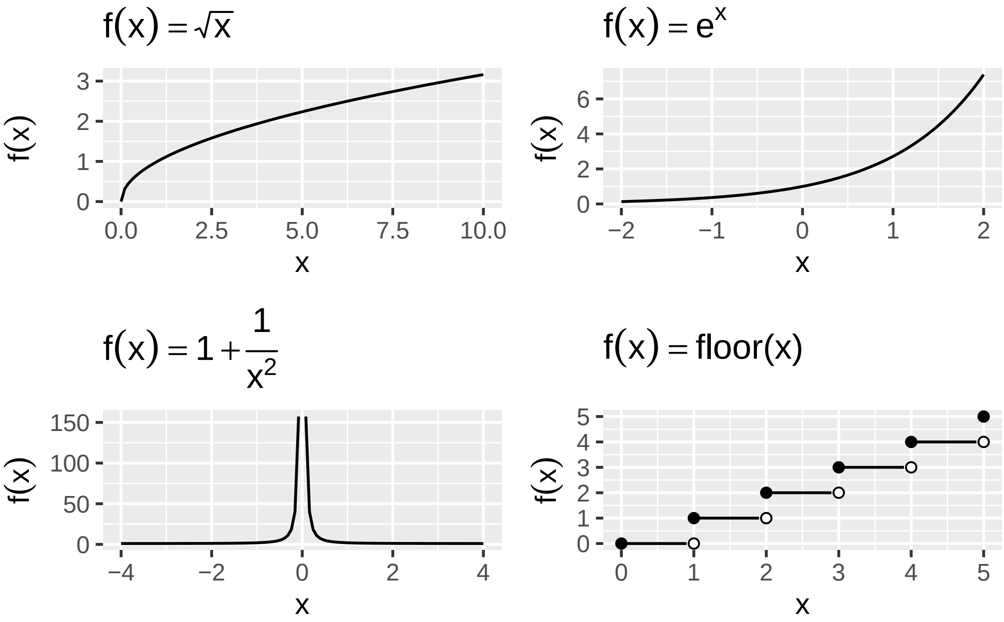


Figure 3.4: Continuous and Discontinuous Functions

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

Let $f(x) = \frac{x^2+2x}{x}$.

1. Graph the function. Is it defined everywhere?
2. What is the functions limit at $x \to 0$?

4 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.

Example: The Mean is a Type of Integral

The average of a quantity is a type of weighted mean, where the potential values are weighted by their likelihood, loosely speaking. The integral is actually a general way to describe this weighted average when there are conceptually an infinite number of potential values.

If X is a continuous random variable, its expected value $E(X)$ – the center of mass – is given by

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

where $f(x)$ is the probability density function of X.

This is a continuous version of the case where X is discrete, in which case

$$E(X) = \sum_{j=1}^{\infty} x_j P(X = x_j)$$

even more concretely, if the potential values of X are finite, then we can write out the expected value as a weighted mean, where the weights is the probability that the value occurs.

$$E(X) = \sum_{x} \quad \left(\underbrace{x}_{\text{value}} \cdot \underbrace{P(X = x)}_{\text{weight, or PMF}} \right)$$

4.1 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 (Section -Chapter 3).

Definition 4.1.

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 -Figure 4.1

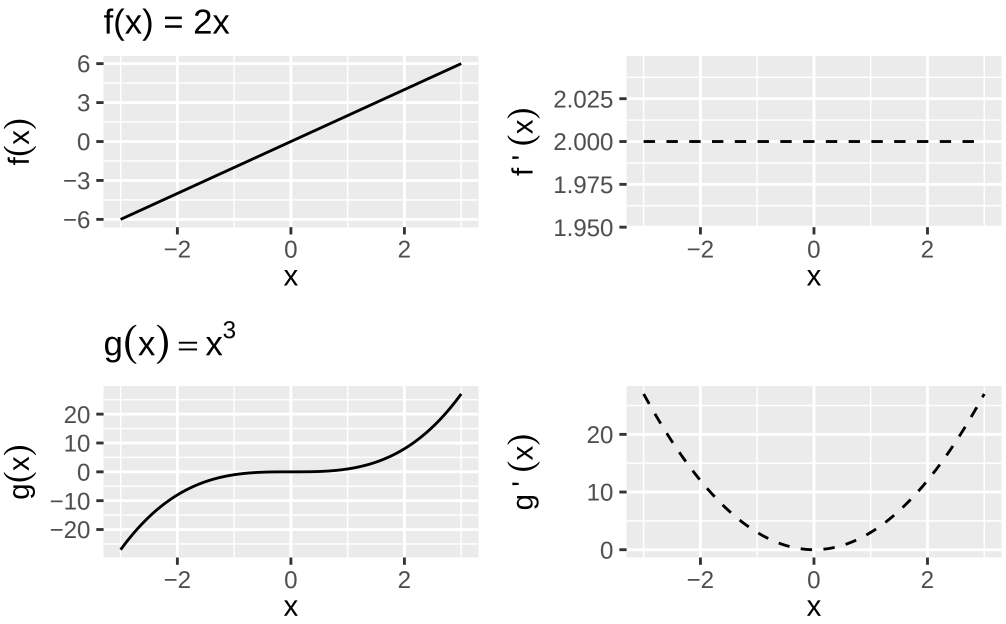


Figure 4.1: The Derivative as a Slope

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:

$$[kf(x)]' = kf'(x)$$

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:

$$\left[x^k\right]' = kx^{k-1}$$

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.

Proposition 4.1.

Power Rule

$$\left[x^k\right]' = kx^{k-1}$$

for any integer k.

Proof. 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

$$\left[x^m\right]' = mx^{m-1}$$

Then, for the case $m + 1$, using the product rule above, we can simplify

$$\begin{aligned}
\left[x^{m+1}\right]' &= [x^m \cdot x]' \\
&= (x^m)' \cdot x + (x^m) \cdot (x)' \\
&= mx^{m-1} \cdot x + x^m \quad \because \text{by previous assumption} \\
&= mx^m + x^m \\
&= (m+1)x^m \\
&= (m+1)x^{(m+1)-1}
\end{aligned}$$

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 4.1.

Derivatives

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. $f(x) = \frac{x^2+1}{x^2-1}$

4.2 Higher-Order Derivatives (Derivatives of Derivatives of 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 4.1.

Succession of derivatives

$$\begin{aligned} f(x) &= x^3 \\ f'(x) &= 3x^2 \\ f''(x) &= 6x \\ f'''(x) &= 6 \\ f''''(x) &= 0 \end{aligned}$$

Earlier, in Section -Section 4.1, 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 (Section @ref(optim)) 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: $f(x) = x^2$, $f(x) = e^x$.

4.3 Composite Functions and 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 4.2.

Composite functions

Let $f(x) = \log x$ for $0 < x < \infty$ and $g(x) = x^2$ for $-\infty < x < \infty$.

Then

$$(f \circ g)(x) = \log x^2, -\infty < x < \infty - \{0\}$$

Also

$$(g \circ f)(x) = [\log x]^2, 0 < x < \infty$$

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 4.3.

Composite Exponent

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,

$$f'(x) = p[g(x)]^{p-1}g'(x)$$

4.4 Derivatives of natural logs and the exponent

Natural logs and exponents (they are inverses of each other; see Section @ref(logexponents)) crop up everywhere in statistics. Their derivative is a special case from the above, but quite elegant.

Theorem 4.1.

Derivative of Exponents/Logs

The functions e^x *and the natural logarithm* $\log(x)$ *are continuous and differentiable in their domains, and their first derivative is*

$$(e^x)' = e^x$$

$$\log(x)' = \frac{1}{x}$$

Also, when these are composite functions, it follows by the generalized power rule that

$$\left(e^{g(x)}\right)' = e^{g(x)} \cdot g'(x)$$

$$\left(\log g(x)\right)' = \frac{g'(x)}{g(x)}, \quad \textit{if} \;\; g(x) > 0$$

Derivatives of natural exponential function (e)

To repeat the main rule in Theorem @ref(thm:derivexplog), the intuition is that

1. Derivative of e^x is itself: $\frac{d}{dx}e^x = e^x$ (See Figure 4.2)
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. $\frac{d}{dx}e^{g(x)} = e^{g(x)}g'(x)$

Example 4.4.

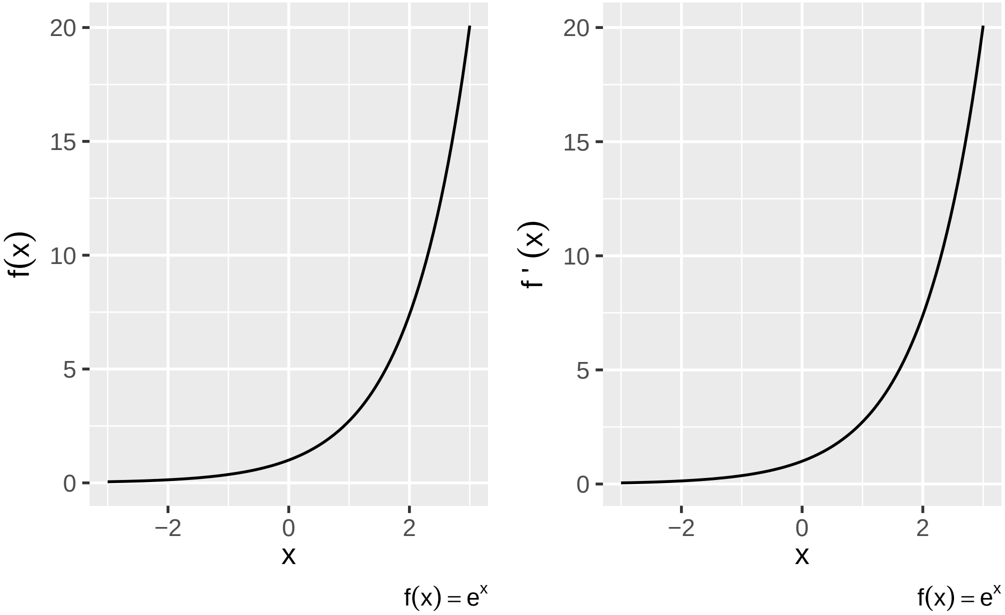


Figure 4.2: Derivative of the Exponential Function

Derivatives of exponents

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 logarithms

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: $\frac{d}{dx} \log x = \frac{1}{x}$ (Figure 4.3)
2. Exponents become multiplicative constants: $\frac{d}{dx} \log x^k = \frac{d}{dx} k \log x = \frac{k}{x}$
3. Chain rule again: $\frac{d}{dx} \log u(x) = \frac{u'(x)}{u(x)}$
4. For any positive base b, $\frac{d}{dx} b^x = (\log b)\,(b^x)$.

Example 4.5.

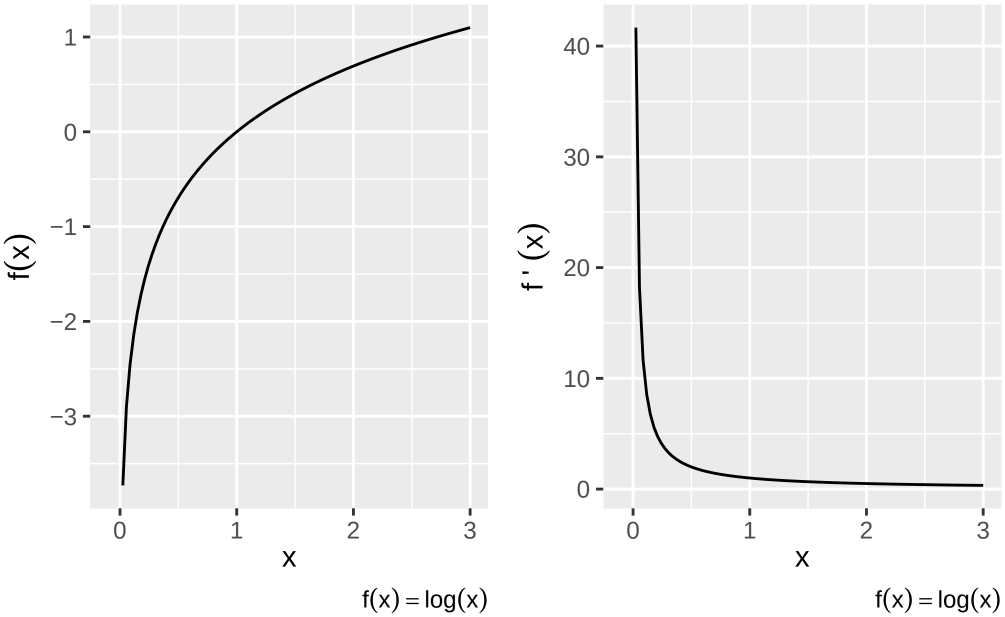


Figure 4.3: Derivative of the Natural Log

Derivatives of logs

Find dy/dx for the following.

1. $f(x) = \log(x^2 + 9)$
2. $f(x) = \log(\log x)$
3. $f(x) = (\log x)^2$
4. $f(x) = \log e^x$

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.

$$(\log_a x)' = \lim_{h \to 0} \frac{1}{h} \log_a \left(1 + \frac{h}{x}\right)$$

Re-express $g = \frac{h}{x}$ and get

$$\begin{aligned}(\log_a x)' &= \frac{1}{x} \lim_{g \to 0} \log_a (1 + g)^{\frac{1}{g}} \\ &= \frac{1}{x} \log_a e\end{aligned}$$

By definition of e. As a special case, when $a = e$, then $(\log x)' = \frac{1}{x}$.

Now let's think about the inverse, taking the derivative of $y = a^x$.

$$\begin{aligned} y &= a^x \\ \Rightarrow \log y &= x \log a \\ \Rightarrow \frac{y'}{y} &= \log a \\ \Rightarrow y' = y \log a \end{aligned}$$

Then in the special case where $a = e$,

$$(e^x)' = (e^x)$$

4.5 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 4.6.

Partial derivatives

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 4.2.

Partial derivatives

Let $f(x, y) = x^3y^4 + e^x - \log y$. What are the following partial derivatives?

$$\begin{aligned}
\frac{\partial f}{\partial x}(x, y) &= \\
\frac{\partial f}{\partial y}(x, y) &= \\
\frac{\partial^2 f}{\partial x^2}(x, y) &= \\
\frac{\partial^2 f}{\partial x \partial y}(x, y) &=
\end{aligned}$$

4.6 Taylor Series 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.

4.7 The Indefinite Integration

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 4.2.

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 4.3.

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.

:::{#def-indefint}

4.8 Indefinite Integral

The indefinite integral of $f(x)$ is written

$$\int f(x)dx$$

and is equal to the antiderivative of f.

Example 4.7.

Graphing

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 4.4 as dotted lines.

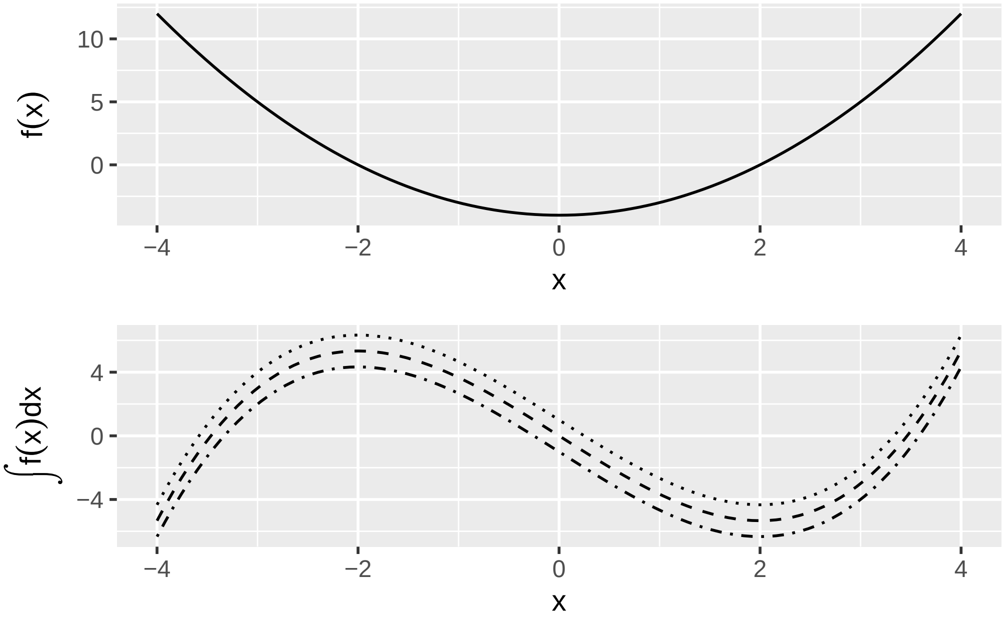


Figure 4.4: The Many Indefinite Integrals of a Function

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.

Common Rules of Integration

Some common rules of integrals follow by virtue of being the inverse of a derivative.

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 $\log(x)$ is one over x, and so: $\int \frac{1}{x}dx = \log x + c$
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: $\int \frac{f'(x)}{f(x)}dx = \log f(x) + c$

Example 4.8.

Common Integration

Simplify the following indefinite integrals:

- (3x^2 dx)
- ((2x+1)dx)
- (e^x e$^{\{\mathrm{e}}$x} dx)

4.9 The Definite Integral: The Area under the Curve

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 $a \leq x \leq b$.

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 4.5 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

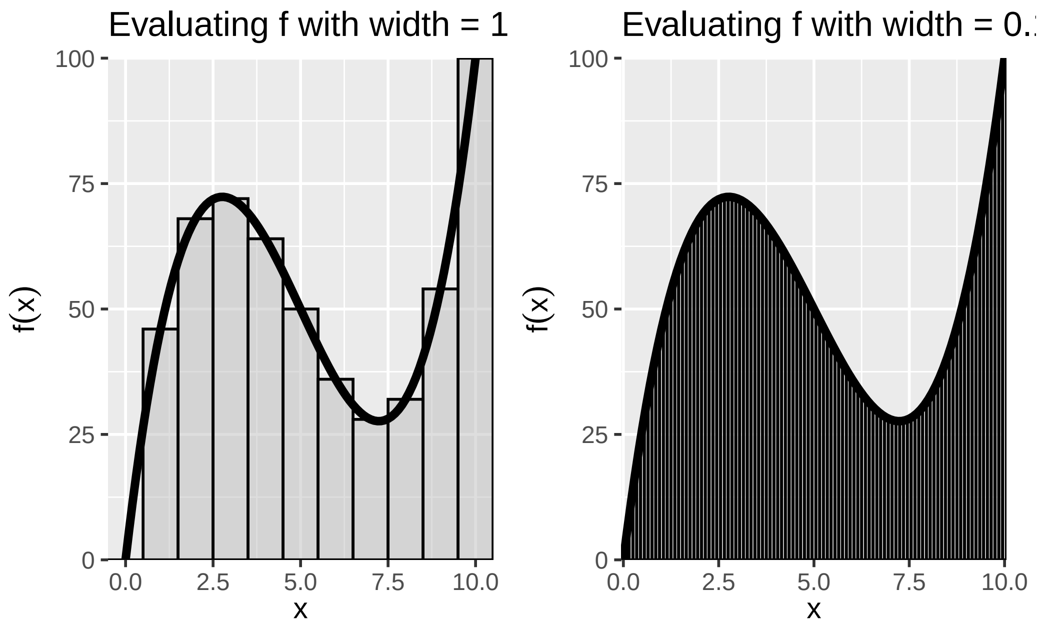


Figure 4.5: The Riemann Integral as a Sum of Evaluations

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 4.3.

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

$$\int\limits_a^b f(x)dx = \lim_{\Delta x \to 0} \sum_{i=1}^{n} f(x_i)\Delta x$$

The most straightforward of a definite integral is the definite integral. That is, we read

$$\int\limits_a^b f(x)dx$$

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 4.2.

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:

$$F(x) = \int\limits_a^x f(t)dt, \quad a \leq x \leq b$$

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.

Theorem 4.3.

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 4.9.

Definite Integral of a monomial

Solve $\int\limits_1^3 3x^2 dx$.

Let $f(x) = 3x^2$.

Exercise 4.4.

Indefinite integrals

What is the value of $\int\limits_{-2}^{2} e^x e^{e^x} dx$?

Common Rules for Definite Integrals

The area-interpretation of the definite integral provides some rules for simplification.

1. There is no area below a point:
$$\int\limits_a^a f(x)dx = 0$$
2. Reversing the limits changes the sign of the integral:
$$\int\limits_a^b f(x)dx = -\int\limits_b^a f(x)dx$$
3. Sums can be separated into their own integrals:
$$\int\limits_a^b [\alpha f(x) + \beta g(x)]dx = \alpha \int\limits_a^b f(x)dx + \beta \int\limits_a^b g(x)dx$$
4. Areas can be combined as long as limits are linked:
$$\int\limits_a^b f(x)dx + \int\limits_b^c f(x)dx = \int\limits_a^c f(x)dx$$

Exercise 4.5.

Definite integrals

Simplify the following definite intergrals.

1. $\int\limits_{1}^{1} 3x^2 dx =$
2. $\int\limits_{0}^{4} (2x+1) dx =$
3. $\int\limits_{-2}^{0} e^x e^{e^x} dx + \int\limits_{0}^{2} e^x e^{e^x} dx =$

4.10 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,

$$\int\limits_a^b g(x)dx = \int\limits_c^d f(u)du = F(d) - F(c)$$

where $c = u(a)$ and $d = u(b)$.

Example 4.10. 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 4.11.

Integration by Substitutiton II

Simplify

$$\int_0^1 \frac{5e^{2x}}{(1+e^{2x})^{1/3}}dx.$$

4.11 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

$$\int_a^b u\frac{dv}{dx}dx = uv|_a^b - \int_a^b v\frac{du}{dx}dx$$

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 4.12.

Integration by parts

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 4.6.

Integration by parts

1. Integrate
$$\int x^n e^{ax} dx$$

2. Integrate
$$\int x^3 e^{-x^2} dx$$

5 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.

Example: Meltzer-Richard

A standard backdrop in comparative political economy, the Meltzer-Richard (1981) model states that redistribution of wealth should be higher in societies where the median income is much smaller than the average income. More to the point, typically income distributions where the median is very different from the average is one of high inequality. In other words, the Meltzer-Richard model says that highly unequal economies will have more re-distribution of wealth. Why is that the case? Here is a simplified example that is not the exact model by Meltzer and Richard[1], but adapted from Persson and Tabellini[2]

We will set the following things about our model human and model democracy.

- Individuals are indexed by i, and the total population is normalized to unity ("1") without loss of generality.
- $U(\cdot)$, u for "utility", is a function that is concave and increasing, and expresses the utility gained from public goods. This tells us that its first derivative is *positive*, and its second derivative is **negative**.
- y_i is the income of person i
- W_i, w for "welfare", is the welfare of person i
- c_i, c for "consumption", is the consumption utility of person i

[1] Allan H. Meltzer and Scott F. Richard. "A Rational Theory of the Size of Government". *Journal of Political Economy* 89:5 (1981), p. 914-927

[2] Adapted from Torsten Persson and Guido Tabellini, *Political Economics: Explaining Economic Policy.* MIT Press.

Also, the government is democratically elected and sets the following redistribution output:

- τ, t for "tax", is a flat tax rate between 0 and 1 that is applied to everyone's income.
- g, "g" for "goods", is the amount of public goods that the government provides.

Suppose an individual's welfare is given by:

$$W_i = c_i + U(g)$$

The consumption good is the person's post-tax income.

$$c_i = (1-\tau)y_i$$

Income varies by person (In the next section we will cover probability, by then we will know that we can express this by saying that y is a random variable with the cumulative distribution function F, i.e. $y \sim F$.). Every distribution has a mean and an median.

- $E(y)$ is the average income of the society.
- med(y) is the **median income** of the society.

What will happen in this economy? What will the tax rate be set too? How much public goods will be provided?

We've skipped ahead of some formal theory results of democracy, but hopefully these are conceptually intuitive. First, if a democracy is competitive, there is no slack in the government's goods, and all tax revenue becomes a public good. So we can go ahead and set the constraint:

$$g = \sum_i \tau y_i P(y_i) = \tau E(y)$$

We can do this trick because of the "normalizes to unity" setting, but this is a general property of the average.

Now given this constraint we can re-write an individual's welfare as

$$\begin{aligned} W_i &= \left(1 - \frac{g}{E(y)}\right) y_i + U(g) \\ &= (E(y) - g)\frac{1}{E(y)} y_i + U(g) \\ &= (E(y) - g)\frac{y_i}{E(y)} + U(g) \end{aligned}$$

When is the individual's welfare maximized, **as a function of the public good**?

$$\frac{d}{dg}W_i = -\frac{y_i}{E(y)} + \frac{d}{dg}U(g)$$

$\frac{d}{dg}W_i = 0$ when $\frac{d}{dg}U(g) = \frac{y_i}{E(y)}$, and so after expressing the derivative as $U_g = \frac{d}{dg}U(g)$ for simplicity,

$$g_i^\star = {U_g}^{-1}\left(\frac{y_i}{E(y)}\right)$$

Now recall that because we assumed concavity, U_g is a negative sloping function whose value is positive. It can be shown that the inverse of such a function is also decreasing. Thus an individual's preferred level of government is determined by a single continuum, the person's income divided by the average income, and the function is **decreasing** in y_i. This is consistent with our intuition that richer people prefer less redistribution.

That was the amount for any given person. The government has to set one value of g, however. So what will that be? Now we will use another result, the median voter theorem. This says that under certain general electoral conditions (single-peaked preferences, two parties, majority rule), the policy winner will be that preferred by the median person in the population. Because the only thing that determines a person's preferred level of government is $y_i/E(y)$, we can presume that the median voter, whose income is $\text{med}(y)$ will prevail in their preferred choice of government. Therefore, we wil see

$$g^\star = {U_g}^{-1}\left(\frac{\text{med}(y)}{E(y)}\right)$$

What does this say about the level of redistribution we observe in an economy? The higher the average income is than the median income, which often (but not always) means *more* inequality, there should be *more* redistribution.

5.1 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 $f'(x) = 2x$

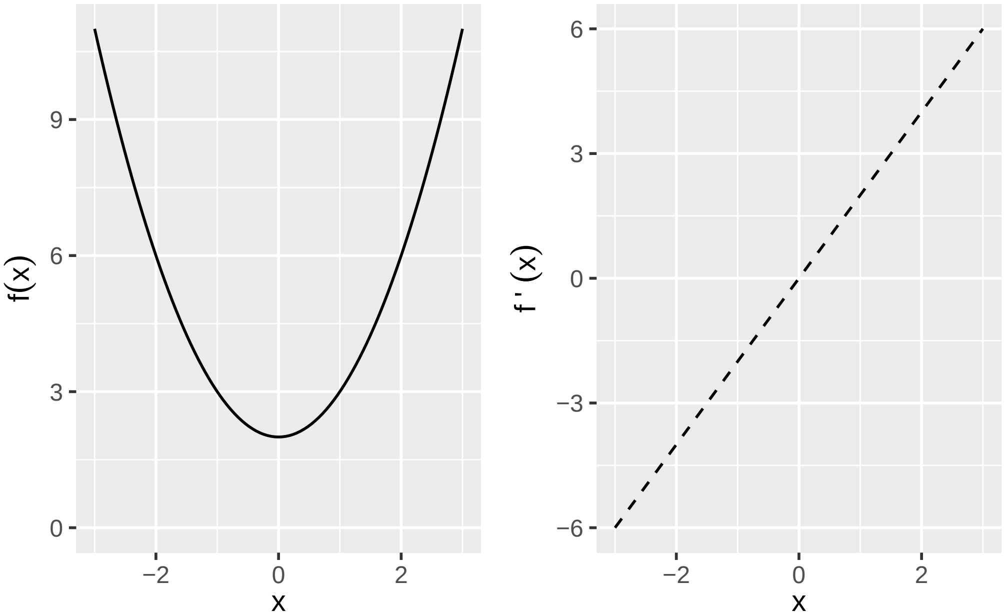


Figure 5.1: Maxima and Minima

Exercise 5.1.

Plotting a maximum 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 down: $f''(x) < 0$
2. Concave up (convex): $f''(x) > 0$

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.
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 5.1.

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. $f(x) = |x^2 - 1|$, $x \in [-2, 2]$

5.2 Concavity of a Function

Concavity helps identify the curvature of a function, $f(x)$, in 2 dimensional space.

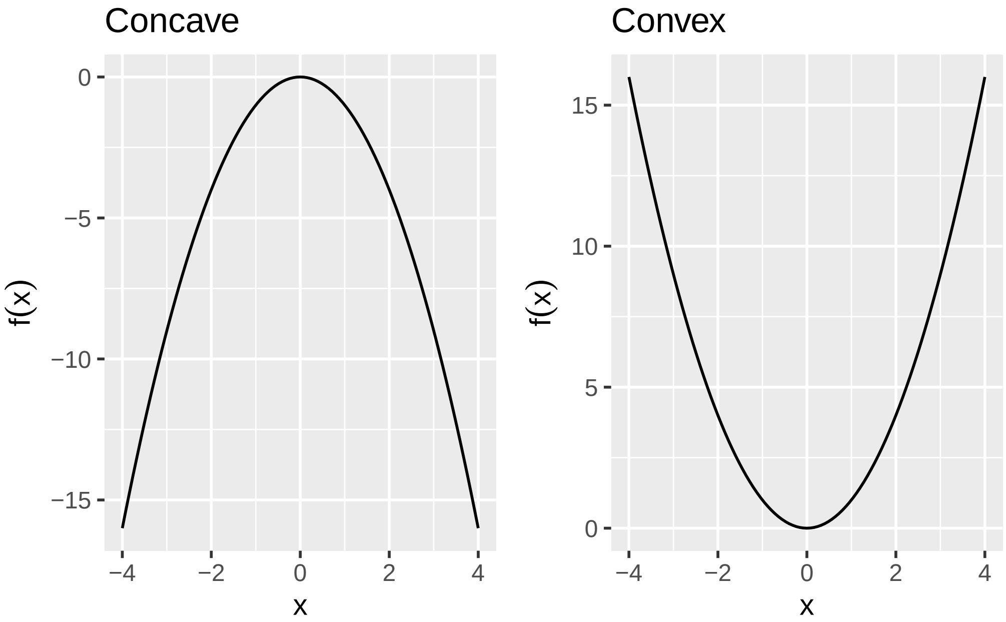
Definition 5.1.

Concave Function

A function f is strictly concave over the set S <u>if</u> $\forall x_1, x_2 \in S$ 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 5.2.

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.

Sometimes, concavity and convexity are strict of a requirement. For most purposes of getting solutions, what we call quasi-concavity is enough.

Definition 5.3.

Quasiconcave Function

A function f is quasiconcave over the set S if $\forall x_1, x_2 \in S$ and $\forall a \in (0,1)$,

$$f(ax_1 + (1-a)x_2) \geq \min(f(x_1), f(x_2))$$

No matter what two points you select, the *lowest* valued point will always be an end point.

Definition 5.4.

Quasiconvex Function

A function f is quasiconvex over the set S if $\forall x_1, x_2 \in S$ and $\forall a \in (0,1)$,

$$f(ax_1 + (1-a)x_2) \leq \max(f(x_1), f(x_2))$$

No matter what two points you select, the *highest* valued point will always be an end point.

Second Derivative Test of Concavity: The second derivative can be used to understand concavity.

If

$$\begin{aligned} f''(x) < 0 &\Rightarrow \text{ Concave} \\ f''(x) > 0 &\Rightarrow \text{ Convex} \end{aligned}$$

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:

$$\begin{aligned} \text{One variable: } & Q(x_1) = a_{11}x_1^2 \\ \text{Two variables: } & Q(x_1, x_2) = a_{11}x_1^2 + a_{12}x_1x_2 + a_{22}x_2^2 \\ \text{N variables: } & Q(x_1, \cdots, x_n) = \sum_{i \leq j} a_{ij}x_ix_j \end{aligned}$$

which can be written in matrix terms:

One variable

$$Q(\mathbf{x}) = x_1^\top a_{11} x_1$$

N variables:

$$\begin{aligned} Q(\mathbf{x}) &= \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix} \begin{pmatrix} a_{11} & \frac{1}{2}a_{12} & \cdots & \frac{1}{2}a_{1n} \\ \frac{1}{2}a_{12} & a_{22} & \cdots & \frac{1}{2}a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2}a_{1n} & \frac{1}{2}a_{2n} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \\ &= \mathbf{x}^\top \mathbf{A} \mathbf{x} \end{aligned}$$

For example, the Quadratic on $\mathbf{R}^2$:

$$\begin{aligned} Q(x_1, x_2) &= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} a_{11} & \frac{1}{2}a_{12} \\ \frac{1}{2}a_{12} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= a_{11}x_1^2 + a_{12}x_1x_2 + a_{22}x_2^2 \end{aligned}$$

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 $\mathbf{x} \neq 0$.

5.3 FOC and SOC

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.

First Order Conditions

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 5.5.

Gradient

Given a function $f(\mathbf{x})$ in n variables, the gradient $\nabla f(\mathbf{x})$ (the greek letter nabla) is a column vector, where the ith element is the partial derivative of $f(\mathbf{x})$ with respect to x_i:

$$\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{pmatrix}$$

Before we know whether a point is a maxima or minima, if it meets the FOC it is a “Critical Point”.

Definition 5.6.

Critical Point

$\mathbf{x}^*$ is a critical point if and only if $\nabla f(\mathbf{x}^*) = 0$. 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.

$$\mathbf{x}^* = \begin{pmatrix} x_1^* \\ x_2^* \\ \vdots \\ x_n^* \end{pmatrix}$$

Example 5.2. 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

$$\begin{aligned} \nabla f(\mathbf{x}) &= \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \end{pmatrix} \\ &= \begin{pmatrix} 2(x_1 - 1) \\ 2x_2 \end{pmatrix} \end{aligned}$$

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

$$\mathbf{x}^* = (1, 0)$$

Second Order Conditions

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 5.7.

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:

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{pmatrix}$$

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 $\implies$ Strict Local Min
2. Hessian is Positive Semidefinite $\forall \mathbf{x} \in B(\mathbf{x}^*, \epsilon)\}$ $\implies$ Local Min
3. Hessian is Negative Definite $\implies$ Strict Local Max
4. Hessian is Negative Semidefinite $\forall \mathbf{x} \in B(\mathbf{x}^*, \epsilon)\}$ $\implies$ Local Max
5. Hessian is Indefinite $\implies$ Saddle Point

Example 5.3.

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

$$\mathbf{H}(\mathbf{x}) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

The Leading principal minors of the Hessian are $M_1 = 2; M_2 = 4$. Now we consider Definiteness. Since both leading principal minors are positive, 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 strict local minimum.

Note: Alternate check of definiteness. Is $\mathbf{H}(\mathbf{x}^*) \geq\leq 0 \quad \forall \quad \mathbf{x} \neq 0$

$$\begin{aligned} \mathbf{x}^\top H(\mathbf{x}^*)\mathbf{x} &= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \\ &= \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \\ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} &= 2x_1^2 + 2x_2^2 \end{aligned}$$

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.

5.4 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: 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 5.2.

Optimization

Given $f(\mathbf{x}) = x_1^3 - x_2^3 + 9x_1x_2$, find any maxima or minima.

1. First order conditions.
   a) Gradient $\nabla f(\mathbf{x}) =$

   b) Critical Points $\mathbf{x}^* =$

2. Second order conditions.
   a) Hessian $\mathbf{H}(\mathbf{x}) =$

   b) Hessian $\mathbf{H}(\mathbf{x_1^*}) =$

   c) Leading principal minors of $\mathbf{H}(\mathbf{x_1^*}) =$

d) Definiteness of $\mathbf{H}(\mathbf{x_1^*})$?

e) Maxima, Minima, or Saddle Point for $\mathbf{x_1^*}$?

f) Hessian $\mathbf{H}(\mathbf{x_2^*}) =$

g) Leading principal minors of $\mathbf{H}(\mathbf{x_2^*}) =$

h) Definiteness of $\mathbf{H}(\mathbf{x_2^*})$?

i) Maxima, Minima, or Saddle Point for $\mathbf{x}_2^*$?

3. Global concavity/convexity.
   a) Is f(x) globally concave/convex?

   b) Are any $\mathbf{x}^*$ global minima or maxima?

5.5 Constrained Optimization

We have already looked at optimizing a function in one or more dimensions over the whole domain of the function. Often, however, we want to find the maximum or minimum of a function over some restricted part of its domain.

ex: Maximizing utility subject to a budget constraint

Types of Constraints: For a function $f(x_1, \dots, x_n)$, there are two types of constraints that can be imposed:

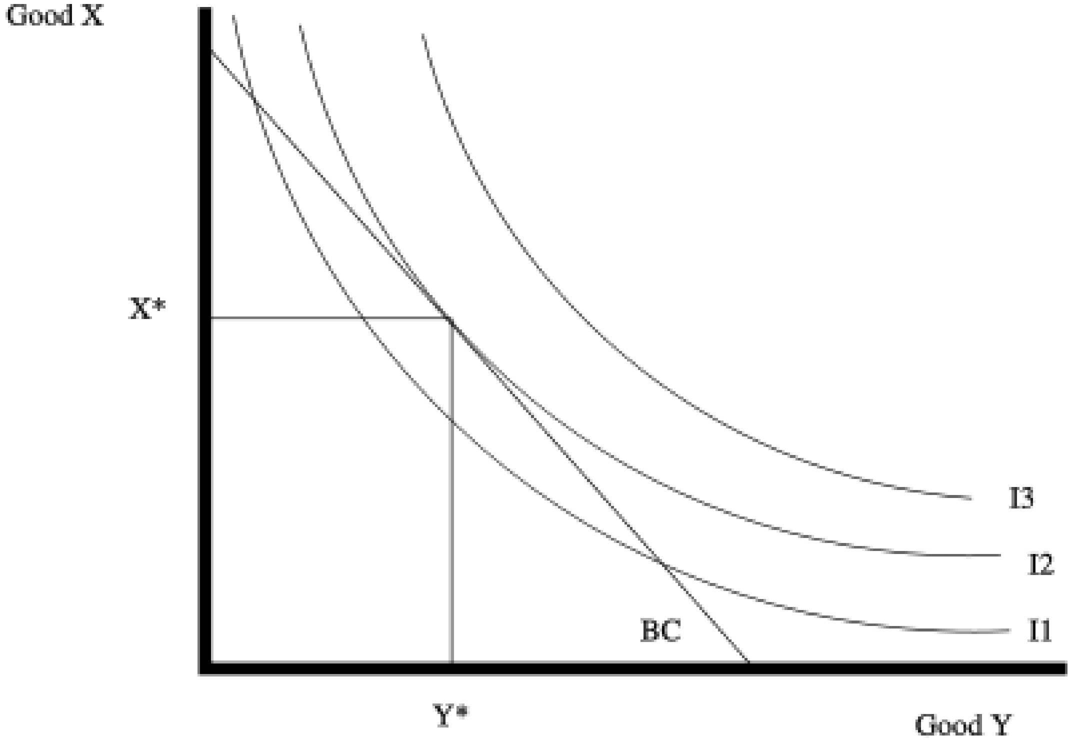


Figure 5.2: A typical Utility Function with a Budget Constraint

1. **Equality constraints:** constraints of the form $c(x_1, \ldots, x_n) = r$. Budget constraints are the classic example of equality constraints in social science.
2. **Inequality constraints:** constraints of the form $c(x_1, \ldots, x_n) \leq r$. These might arise from non-negativity constraints or other threshold effects.

In any constrained optimization problem, the constrained maximum will always be less than or equal to the unconstrained maximum. If the constrained maximum is less than the unconstrained maximum, then the constraint is binding. Essentially, this means that you can treat your constraint as an equality constraint rather than an inequality constraint.

For example, the budget constraint binds when you spend your entire budget. This generally happens because we believe that utility is strictly increasing in consumption, i.e. you always want more so you spend everything you have.

Any number of constraints can be placed on an optimization problem. When working with multiple constraints, always make sure that the set of constraints are not pathological; it must be possible for all of the constraints to be satisfied simultaneously.

Set-up for Constrained Optimization:

$$\max_{x_1,x_2} f(x_1, x_2) \text{ s.t. } c(x_1, x_2)$$

$$\min_{x_1,x_2} f(x_1, x_2) \text{ s.t. } c(x_1, x_2)$$

This tells us to maximize/minimize our function, $f(x_1, x_2)$, with respect to the choice variables, x_1, x_2, subject to the constraint.

Example:

$$\max_{x_1,x_2} f(x_1, x_2) = -(x_1^2 + 2x_2^2) \text{ s.t. } x_1 + x_2 = 4$$

It is easy to see that the *unconstrained* maximum occurs at $(x_1, x_2) = (0, 0)$, but that does not satisfy the constraint. How should we proceed?

Equality Constraints

Equality constraints are the easiest to deal with because we know that the maximum or minimum has to lie on the (intersection of the) constraint(s).

The trick is to change the problem from a constrained optimization problem in n variables to an unconstrained optimization problem in $n + k$ variables, adding *one* variable for *each* equality constraint. We do this using a lagrangian multiplier.

Lagrangian function: The Lagrangian function allows us to combine the function we want to optimize and the constraint function into a single function. Once we have this single function, we can proceed as if this were an *unconstrained* optimization problem.

For each constraint, we must include a **Lagrange multiplier** (λ_i) as an additional variable in the analysis. These terms are the link between the constraint and the Lagrangian function.

Given a *two dimensional* set-up:

$$\max_{x_1,x_2} / \min_{x_1,x_2} f(x_1,x_2) \text{ s.t. } c(x_1,x_2)=a$$

We define the Lagrangian function $L(x_1,x_2,\lambda_1)$ as follows:

$$L(x_1,x_2,\lambda_1)=f(x_1,x_2)-\lambda_1(c(x_1,x_2)-a)$$

More generally, in n *dimensions*:

$$L(x_1,\dots,x_n,\lambda_1,\dots,\lambda_k)=f(x_1,\dots,x_n)-\sum_{i=1}^{k}\lambda_i(c_i(x_1,\dots,x_n)-r_i)$$

Getting the sign right: Note that above we subtract the lagrangian term *and* we subtract the constraint constant from the constraint function. Occasionally, you may see the following alternative form of the Lagrangian, which is *equivalent*:

$$L(x_1,\dots,x_n,\lambda_1,\dots,\lambda_k)=f(x_1,\dots,x_n)+\sum_{i=1}^{k}\lambda_i(r_i-c_i(x_1,\dots,x_n))$$

Here we add the lagrangian term *and* we subtract the constraining function from the constraint constant.

Using the Lagrangian to Find the Critical Points: To find the critical points, we take the partial derivatives of lagrangian function, $L(x_1,\dots,x_n,\lambda_1,\dots,\lambda_k)$, with respect to each of its variables (all choice variables $\mathbf{x}$ *and* all lagrangian multipliers). At a critical point, each of these partial derivatives must be equal to zero, so we obtain a system of $n+k$ equations in $n+k$ unknowns:

$$\begin{aligned}
\frac{\partial L}{\partial x_1} &= \frac{\partial f}{\partial x_1}-\sum_{i=1}^{k}\lambda_i\frac{\partial c_i}{\partial x_1}=0\\
\vdots &= \vdots\\
\frac{\partial L}{\partial x_n} &= \frac{\partial f}{\partial x_n}-\sum_{i=1}^{k}\lambda_i\frac{\partial c_i}{\partial x_n}=0\\
\frac{\partial L}{\partial \lambda_1} &= c_1(x_i,\dots,x_n)-r_1=0\\
\vdots &= \vdots\\
\frac{\partial L}{\partial \lambda_k} &= c_k(x_i,\dots,x_n)-r_k=0
\end{aligned}$$

We can then solve this system of equations, because there are $n + k$ equations and $n + k$ unknowns, to calculate the critical point $(x_1^*, \dots, x_n^*, \lambda_1^*, \dots, \lambda_k^*)$.

Second-order Conditions and Unconstrained Optimization: There may be more than one critical point, i.e. we need to verify that the critical point we find is a maximum/minimum. Similar to unconstrained optimization, we can do this by checking the second-order conditions.

Example 5.4.

Constrained optimization with two goods and a budget constraint

Find the constrained optimization of

$$\max_{x_1,x_2} f(x) = -(x_1^2 + 2x_2^2) \text{ s.t. } x_1 + x_2 = 4$$

Solution. 1. Begin by writing the Lagrangian:

$$L(x_1, x_2, \lambda) = -(x_1^2 + 2x_2^2) - \lambda(x_1 + x_2 - 4)$$

2. Take the partial derivatives and set equal to zero:

$$\begin{aligned}
\frac{\partial L}{\partial x_1} &= -2x_1 - \lambda && = 0 \\
\frac{\partial L}{\partial x_2} &= -4x_2 - \lambda && = 0 \\
\frac{\partial L}{\partial \lambda} &= -(x_1 + x_2 - 4) && = \qquad 0
\end{aligned}$$

3. Solve the system of equations: Using the first two partials, we see that $\lambda = -2x_1$ and $\lambda = -4x_2$ Set these equal to see that $x_1 = 2x_2$. Using the third partial and the above equality, $4 = 2x_2 + x_2$ from which we get

$$x_2^* = 4/3, x_1^* = 8/3, \lambda = -16/3$$

4. Therefore, the only critical point is $x_1^* = \frac{8}{3}$ and $x_2^* = \frac{4}{3}$

5. This gives $f(\frac{8}{3}, \frac{4}{3}) = -\frac{96}{9}$, which is less than the unconstrained optimum $f(0, 0) = 0$

Notice that when we take the partial derivative of L with respect to the Lagrangian multiplier and set it equal to 0, we return exactly our constraint! This is why signs matter.

5.6 Inequality Constraints

Inequality constraints define the boundary of a region over which we seek to optimize the function. This makes inequality constraints more challenging because we do not know if the maximum/minimum lies along one of the constraints (the constraint binds) or in the interior of the region.

We must introduce more variables in order to turn the problem into an unconstrained optimization.

Slack: For each inequality constraint $c_i(x_1, \ldots, x_n) \leq a_i$, we define a slack variable s_i^2 for which the expression $c_i(x_1, \ldots, x_n) \leq a_i - s_i^2$ would hold with equality. These slack variables capture how close the constraint comes to binding. We use s^2 rather than s to ensure that the slack is positive.

Slack is just a way to transform our constraints.

Given a two-dimensional set-up and these edited constraints:

$$\max_{x_1,x_2} / \min_{x_1,x_2} f(x_1, x_2) \text{ s.t. } c(x_1, x_2) \leq a_1$$

Adding in Slack:

$$\max_{x_1,x_2} / \min_{x_1,x_2} f(x_1, x_2) \text{ s.t. } c(x_1, x_2) \leq a_1 - s_1^2$$

We define the Lagrangian function $L(x_1, x_2, \lambda_1, s_1)$ as follows:

$$L(x_1, x_2, \lambda_1, s_1) = f(x_1, x_2) - \lambda_1(c(x_1, x_2) + s_1^2 - a_1)$$

More generally, in n dimensions:

$$L(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_k, s_1, \ldots, s_k) = f(x_1, \ldots, x_n) - \sum_{i=1}^{k} \lambda_i(c_i(x_1, \ldots, x_n) + s_i^2 - a_i)$$

Finding the Critical Points: To find the critical points, we take the partial derivatives of the lagrangian function, $L(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_k, s_1, \ldots, s_k)$, with respect to each of its variables (all choice variables x, all lagrangian multipliers λ, and all slack variables s). At a critical point, *each* of these partial derivatives must be equal to zero, so we obtain a system of $n + 2k$ equations in $n + 2k$ unknowns:

$$
\begin{aligned}
\frac{\partial L}{\partial x_1} &= \frac{\partial f}{\partial x_1} - \sum_{i=1}^{k} \lambda_i \frac{\partial c_i}{\partial x_1} = 0 \\
\vdots &= \vdots \\
\frac{\partial L}{\partial x_n} &= \frac{\partial f}{\partial x_n} - \sum_{i=1}^{k} \lambda_i \frac{\partial c_i}{\partial x_n} = 0 \\
\frac{\partial L}{\partial \lambda_1} &= c_1(x_i, \dots, x_n) + s_1^2 - b_1 = 0 \\
\vdots &= \vdots \\
\frac{\partial L}{\partial \lambda_k} &= c_k(x_i, \dots, x_n) + s_k^2 - b_k = 0 \\
\frac{\partial L}{\partial s_1} &= 2s_1\lambda_1 = 0 \\
\vdots &= \vdots \\
\frac{\partial L}{\partial s_k} &= 2s_k\lambda_k = 0
\end{aligned}
$$

Complementary slackness conditions: The last set of first order conditions of the form $2s_i\lambda_i = 0$ (the partials taken with respect to the slack variables) are known as complementary slackness conditions. These conditions can be satisfied one of three ways:

1. $\lambda_i = 0$ and $s_i \neq 0$: This implies that the slack is positive and thus *the constraint does not bind.*
2. $\lambda_i \neq 0$ and $s_i = 0$: This implies that there is no slack in the constraint and *the constraint does bind.*
3. $\lambda_i = 0$ and $s_i = 0$: In this case, there is no slack but the *constraint binds trivially*, without changing the optimum.

Example: Find the critical points for the following constrained optimization:

$$\max_{x_1, x_2} f(x) = -(x_1^2 + 2x_2^2) \text{ s.t. } x_1 + x_2 \leq 4$$

1. Rewrite with the slack variables:

$$\max_{x_1, x_2} f(x) = -(x_1^2 + 2x_2^2) \text{ s.t. } x_1 + x_2 \leq 4 - s_1^2$$

2. Write the Lagrangian:

$$L(x_1, x_2, \lambda_1, s_1) = -(x_1^2 + 2x_2^2) - \lambda_1(x_1 + x_2 + s_1^2 - 4)$$

3. Take the partial derivatives and set equal to 0:

$$
\begin{aligned}
\frac{\partial L}{\partial x_1} &= -2x_1 - \lambda_1 = 0 \\
\frac{\partial L}{\partial x_2} &= -4x_2 - \lambda_1 = 0 \\
\frac{\partial L}{\partial \lambda_1} &= -(x_1 + x_2 + s_1^2 - 4) = 0 \\
\frac{\partial L}{\partial s_1} &= -2s_1\lambda_1 = 0
\end{aligned}
$$

4. Consider all ways that the complementary slackness conditions are solved:

Hypothesis	s_1	λ_1	x_1	x_2	$f(x_1, x_2)$
$s_1 = 0$ $\lambda_1 = 0$	No solution				
$s_1 \neq 0$ $\lambda_1 = 0$	2	0	0	0	0
$s_1 = 0$ $\lambda_1 \neq 0$	0	$\frac{-16}{3}$	$\frac{8}{3}$	$\frac{4}{3}$	$-\frac{32}{3}$
$s_1 \neq 0$ $\lambda_1 \neq 0$	No solution				

This shows that there are two critical points: $(0, 0)$ and $(\frac{8}{3}, \frac{4}{3})$.

5. Find maximum: Looking at the values of $f(x_1, x_2)$ at the critical points, we see that $f(x_1, x_2)$ is maximized at $x_1^* = 0$ and $x_2^* = 0$.

Exercise 5.3.

Constrained optimization

Example: Find the critical points for the following constrained optimization:

$$\max_{x_1,x_2} f(x) = -(x_1^2 + 2x_2^2) \text{ s.t. } \begin{array}{l} x_1 + x_2 \leq 4 \\ x_1 \geq 0 \\ x_2 \geq 0 \end{array}$$

1. Rewrite with the slack variables:

2. Write the Lagrangian:

3. Take the partial derivatives and set equal to zero:

4. Consider all ways that the complementary slackness conditions are solved:

Hypothesis	s_1	s_2	s_3	λ_1	λ_2	λ_3	x_1	x_2	$f(x_1, x_2)$
$s_1 = s_2 = s_3 = 0$									
$s_1 \neq 0, s_2 = s_3 = 0$									
$s_2 \neq 0, s_1 = s_3 = 0$									
$s_3 \neq 0, s_1 = s_2 = 0$									
$s_1 \neq 0, s_2 \neq 0, s_3 = 0$									
$s_1 \neq 0, s_3 \neq 0, s_2 = 0$									
$s_2 \neq 0, s_3 \neq 0, s_1 = 0$									
$s_1 \neq 0, s_2 \neq 0, s_3 \neq 0$									

5. Find maximum:

5.7 Kuhn-Tucker Conditions

As you can see, this can be a pain. When dealing explicitly with *non-negativity constraints*, this process is simplified by using the Kuhn-Tucker method.

Because the problem of maximizing a function subject to inequality and non-negativity constraints arises frequently in economics, the **Kuhn-Tucker conditions** provides a method that often makes it easier to both calculate the critical points and identify points that are (local) maxima.

Given a *two-dimensional set-up*:

$$\max_{x_1,x_2} / \min_{x_1,x_2} f(x_1,x_2) \text{ s.t. } \begin{array}{l} c(x_1,x_2) \leq a_1 \\ x_1 \geq 0 \\ gx_2 \geq 0 \end{array}$$

We define the Lagrangian function $L(x_1, x_2, \lambda_1)$ the same as if we did not have the non-negativity constraints:

$$L(x_1, x_2, \lambda_2) = f(x_1, x_2) - \lambda_1(c(x_1, x_2) - a_1)$$

More generally, in n dimensions:

$$L(x_1, \dots, x_n, \lambda_1, \dots, \lambda_k) = f(x_1, \dots, x_n) - \sum_{i=1}^{k} \lambda_i (c_i(x_1, \dots, x_n) - a_i)$$

Kuhn-Tucker and Complementary Slackness Conditions: To find the critical points, we first calculate the Kuhn-Tucker conditions by taking the partial derivatives of the lagrangian function, $L(x_1, \dots, x_n, \lambda_1, \dots, \lambda_k)$, with respect to each of its variables (all choice variables x and all lagrangian multipliers λ) and we calculate the *complementary slackness conditions* by multiplying each partial derivative by its respective variable *and* include non-negativity conditions for all variables (choice variables x and lagrangian multipliers λ).

Kuhn-Tucker Conditions

$$\begin{gathered} \frac{\partial L}{\partial x_1} \leq 0, \dots, \frac{\partial L}{\partial x_n} \leq 0 \\ \frac{\partial L}{\partial \lambda_1} \geq 0, \dots, \frac{\partial L}{\partial \lambda_m} \geq 0 \end{gathered}$$

Complementary Slackness Conditions

$$x_1 \frac{\partial L}{\partial x_1} = 0, \ldots, x_n \frac{\partial L}{\partial x_n} = 0$$
$$\lambda_1 \frac{\partial L}{\partial \lambda_1} = 0, \ldots, \lambda_m \frac{\partial L}{\partial \lambda_m} = 0$$

Non-negativity Conditions

$$x_1 \geq 0 \quad \ldots \quad x_n \geq 0$$
$$\lambda_1 \geq 0 \quad \ldots \quad \lambda_m \geq 0$$

Note that some of these conditions are set equal to 0, while others are set as inequalities!

Note also that to minimize the function $f(x_1, \ldots, x_n)$, the simplest thing to do is maximize the function $-f(x_1, \ldots, x_n)$; all of the conditions remain the same after reformulating as a maximization problem.

There are additional assumptions (notably, f(x) is quasi-concave and the constraints are convex) that are sufficient to ensure that a point satisfying the Kuhn-Tucker conditions is a global max; if these assumptions do not hold, you may have to check more than one point.

Finding the Critical Points with Kuhn-Tucker Conditions: Given the above conditions, to find the critical points we solve the above system of equations. To do so, we must check *all* border and interior solutions to see if they satisfy the above conditions.

In a two-dimensional set-up, this means we must check the following cases:

1. $x_1 = 0, x_2 = 0$ Border Solution
2. $x_1 = 0, x_2 \neq 0$ Border Solution
3. $x_1 \neq 0, x_2 = 0$ Border Solution
4. $x_1 \neq 0, x_2 \neq 0$ Interior Solution

Example 5.5.

Kuhn-Tucker with two variables

Solve the following optimization problem with inequality constraints

$$\max_{x_1,x_2} f(x) = -(x_1^2 + 2x_2^2)$$

$$\text{s.t.} \begin{cases} x_1 + x_2* \leq 4 \\ x_1* \geq 0 \\ x_2* \geq 0 \end{cases}$$

1. Write the Lagrangian:

$$L(x_1, x_2, \lambda) = -(x_1^2 + 2x_2^2) - \lambda(x_1 + x_2 - 4)$$

2. Find the First Order Conditions:

Kuhn-Tucker Conditions

$$\begin{aligned} \frac{\partial L}{\partial x_1} &= -2x_1 - \lambda \leq 0 \\ \frac{\partial L}{\partial x_2} &= -4x_2 - \lambda \leq 0 \\ \frac{\partial L}{\partial \lambda} &= -(x_1 + x_2 - 4) \geq 0 \end{aligned}$$

Complementary Slackness Conditions

$$\begin{aligned} x_1 \frac{\partial L}{\partial x_2} &= x_1(-2x_1 - \lambda) = 0 \\ x_2 \frac{\partial L}{\partial x_2} &= x_2(-4x_2 - \lambda) = 0 \\ \lambda \frac{\partial L}{\partial \lambda} &= -\lambda(x_1 + x_2 - 4) = 0 \end{aligned}$$

Non-negativity Conditions

$$\begin{aligned} x_1 &\geq 0 \\ x_2 &\geq 0 \\ \lambda &\geq 0 \end{aligned}$$

3. Consider all border and interior cases:

Hypothesis	λ	x_1	x_2	$f(x_1, x_2)$
$x_1 = 0, x_2 = 0$	0	0	0	0
$x_1 = 0, x_2 \neq 0$	-16	0	4	-32
$x_1 \neq 0, x_2 = 0$	-8	4	0	-16
$x_1 \neq 0, x_2 \neq 0$	$-\frac{16}{3}$	$\frac{8}{3}$	$\frac{4}{3}$	$-\frac{32}{3}$

4. Find Maximum: Three of the critical points violate the requirement that $\lambda \geq 0$, so the point $(0, 0, 0)$ is the maximum.

Exercise 5.4.

Kuhn-Tucker with logs

$$\max_{x_1,x_2} f(x) = \frac{1}{3}\log(x_1+1) + \frac{2}{3}\log(x_2+1) \text{ s.t. } \begin{array}{l} x_1 + 2x_2 \leq 4 \\ x_1 \geq 0 \\ x_2 \geq 0 \end{array}$$

1. Write the Lagrangian:

2. Find the First Order Conditions:
   Kuhn-Tucker Conditions

   Complementary Slackness Conditions

Non-negativity Conditions

3. Consider all border and interior cases:

Hypothesis	λ	x_1	x_2	$f(x_1, x_2)$
$x_1 = 0, x_2 = 0$				
$x_1 = 0, x_2 \neq 0$				
$x_1 \neq 0, x_2 = 0$				
$x_1 \neq 0, x_2 \neq 0$				

4. Find Maximum:

5.8 Applications of Quadratic Forms

Curvature and The Taylor Polynomial as a Quadratic Form: The Hessian is used in a Taylor polynomial approximation to $f(\mathbf{x})$ and provides information about the curvature of $f(\mathbf{x})$ at $\mathbf{x}$ — e.g., which tells us whether a critical point $\mathbf{x}^*$ is a min, max, or saddle point.

1. The second order Taylor polynomial about the critical point $\mathbf{x}^*$ is

$$f(\mathbf{x}^* + \mathbf{h}) = f(\mathbf{x}^*) + \nabla f(\mathbf{x}^*)\mathbf{h} + \frac{1}{2}\mathbf{h}^\top \mathbf{H}(x^*)\mathbf{h} + R(\mathbf{h})$$

2. Since we're looking at a critical point, $\nabla f(\mathbf{x}^*) = 0$; and for small $\mathbf{h}$, $R(\mathbf{h})$ is negligible. Rearranging, we get

$$f(\mathbf{x}^* + \mathbf{h}) - f(\mathbf{x}^*) \approx \frac{1}{2}\mathbf{h}^\top \mathbf{H}(x^*)\mathbf{h}$$

3. The Righthand side here is a quadratic form and we can determine the definiteness of $\mathbf{H}(x^*)$.

6 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.

6.1 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 6.1.

Counting Possibilities

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 6.2.

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 =

7 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?

7.1 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. 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.

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).

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,

Definition 7.1.

Probability

Probability is a function that maps events from a sample space 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 7.2.

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 $A_1, A_2, \ldots$ (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 $P\left(\bigcup\limits_{i=1}^{n} A_i\right) \leq \sum\limits_{i=1}^{n} P(A_i)$.

Example 7.1.

Probability

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. $P\left(\{1, 2\}^C\right) = P(\{3, 4, 5, 6\}) =$
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 7.1.

Probability

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 $P(X = 5 \cup X = 3)^C$?

7.2 Conditional Probability and Bayes Rule

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 7.2.

Conditional Probability 1

Assume A and B occur with the following frequencies:

	A	A^c
B	n_{ab}	n_{a^cb}
B^C	n_{ab^c}	$n_{(ab)^c}$

and let $n_{ab} + n_{a^Cb} + n_{ab^C} + n_{(ab)^C} = N$. 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 7.3.

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 7.3.

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,

$$P(A_1 \cap \cdots \cap A_k) = P(A_k|A_{k-1} \cap \cdots \cap A_1) \times P(A_{k-1}|A_{k-2} \cap \cdots A_1) \times ... \times P(A_2|A_1) \times P(A_1)$$

Sometimes it is easier to calculate these conditional probabilities and sum them than it is to calculate $P(A)$ directly.

Definition 7.4.

Law of total probability

Let S be the sample space of some experiment and let the disjoint k events $B_1, \ldots, B_k$ partition S, such that $P(B_1 \cup \ldots \cup B_k) = P(S) = 1$. If A is some other event in S, then the events $A \cap B_1, A \cap B_2, \ldots, A \cap B_k$ will form a partition of A and we can write A as

$$A = (A \cap B_1) \cup \cdots \cup (A \cap B_k)$$

.

Since the k events are disjoint,

$$\begin{aligned} P(A) &= \sum_{i=1}^{k} P(A \cap B_i) \\ &= \sum_{i=1}^{k} P(B_i) P(A|B_i) \end{aligned}$$

Bayes Rule: Assume that events $B_1, \ldots, B_k$ 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 7.4.

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 7.2.

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?

7.3 Independence

Definition 7.5.

Independence

If the occurrence or nonoccurrence of either events A and B provides no information about 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 $A_1, A_2, \ldots, A_k$ 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, imagine that you own a house with a lawn (a very extreme hypothetical!) Let A be the event that it rained yesterday and B the event that your sprinkler system went off yesterday. Suppose that your sprinkler system is set to randomly go off and so A and B are independent of one another. $P(A \mid B) = P(A)$. But now let C be the event that the grass is wet. The grass can be wet *either* due to the rain or due to the sprinkler. 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.

Let $P(A) = .5$ and $P(B) = .5$.

The marginal probability $P(C)$ is

$$P(C) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = .75$$

The conditional probability $P(A|C)$ is

$$P(A|C) = \frac{P(C|A)P(A)}{P(C)} = \frac{1 \times .5}{.75} = \frac{2}{3}$$

The conditional probability $P(A|B \cap C)$ is

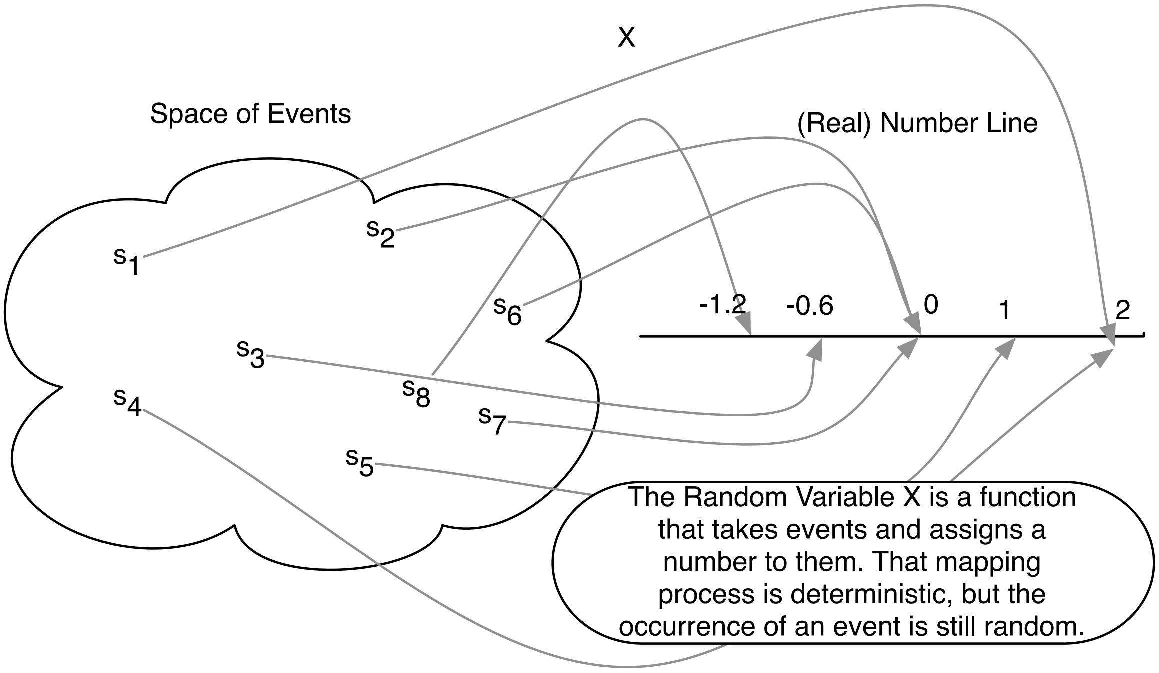
$$P(A|B \cap C) = \frac{P(C \cap B|A)P(A)}{P(C \cap B)} = \frac{P(C \cap B|A)P(A)}{P(C|B)P(B)} \frac{.5 \times .5}{.5} = \frac{1}{2}$$

Intuitively, given that the grass is wet, knowing that it rained yesterday tells us that it is *less* likely that the sprinkler also went off!

7.4 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.

Definition 7.6.



The Random Variable X is a function that takes events and assigns a number to them. That mapping process is deterministic, but the occurrence of an event is still random.

Figure 7.1: The Random Variable as a Real-Valued Function

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 7.1 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?

7.5 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 7.7.

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 7.5.

Total Number of Occurrences

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) and 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 7.8.

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 7.9.

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: 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** , $F(x)$ 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 $P(X > x) = 1 - P(X \leq x)$.

Definition 7.10. For a fair six-sided 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 7.11.

Continuous Random Variable

X is a continuous random variable if there exists a nonnegative 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 7.12.

Probability density function

The function f above is called the probability density function (pdf) of X and must satisfy

$$f(x) \geq 0$$

$$\int\limits_{-\infty}^{\infty} f(x)dx = 1$$

Note also that $P(X = x) = 0$ — i.e., the probability of any point y is zero.

While continuous random variables do not have a PMF (since the PMF would be 0 at every point), the cumulative distribution function is defined in the exact same way. The cumulative distribution gives the probability that Y lies on the interval $(-\infty, y)$ and is defined as

$$F(x) = P(X \leq x) = \int\limits_{-\infty}^{x} f(s)ds$$

We can also make statements about the probability of Y falling in an interval $a \leq y \leq b$.

$$P(a \leq x \leq b) = \int\limits_{a}^{b} f(x)dx$$

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 7.6.

Continuous R.V.

For $f(y) = 1, \quad 0 < y < 1$, find: (1) The CDF $F(y)$ and (2) The probability $P(0.5 < y < 0.75)$.

7.6 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:

$$f(x, y); P((X, Y) \in A) = \iint_A f(x, y) dx dy$$

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 7.3.

Discrete, Joint Distributions

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?

7.7 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 7.13.

Expectation of a discrete R.V.

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 7.7.

Expectation of a discrete R.V.

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 7.8.

Expectation of a continuous R.V.

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:

$$E(XY) = E(X)E(Y)$$

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

7.8 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 7.14. 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 7.9. Given the following PMF:

$$f(x) = \begin{cases} \frac{3!}{x!(3-x)!}(\frac{1}{2})^3 & x = 0, 1, 2, 3 \\ 0 \quad otherwise \end{cases}$$

What is Var(x)?

Hint: First calculate $E(X)$ and $E(X^2)$

Definition 7.15.

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 7.16.

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 7.4.

Expectation and Variance 1

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 7.5.

Expectation and Variance 2

1. Find the expectation and variance

Given the following PDF:

$$f(x) = \begin{cases} \frac{3}{10}(3x - x^2) & 0 \leq x \leq 2 \\ 0 \quad otherwise \end{cases}$$

Exercise 7.6.

Expectation and Variance 3

1. Find the mean and standard deviation of random variable X. The PDF of this X is as follows:

$$f(x) = \begin{cases} \frac{1}{4}x \quad 0 \leq x \leq 2 \\ \frac{1}{4}(4-x) \quad 2 \leq x \leq 4 \\ 0 \quad otherwise \end{cases}$$

2. Next, calculate $P(X < \mu - \sigma)$ Remember, μ is the mean and σ is the standard deviation

7.9 Distributions

A distribution is defined by its cumulative distribution function. There are many common distributions that have useful properties that appear in probability and statistics.

Two *discrete* distributions used often are:

Definition 7.17.

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

$$p(y) = \binom{n}{y} p^y q^{n-y}, \quad y = 0, 1, 2, \dots, n$$

with mean $\mu = E(Y) = np$ and variance $\sigma^2 = \mathrm{Var}(Y) = npq$.

Example 7.10.

Binomial distribution

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. $P(Y = 5) =$ 1. $E(Y) =$ 1. $\mathrm{Var}(Y) =$ 6

Definition 7.18.

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 7.11.

Poisson Distribution

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 7.19.

Uniform Distribution

A random variable Y has a continuous uniform distribution on the interval (α, β) if its density is given by

$$f(y) = \frac{1}{\beta - \alpha}, \quad \alpha \leq y \leq \beta$$

The mean and variance of Y are $E(Y) = \frac{\alpha+\beta}{2}$ and $\mathrm{Var}(Y) = \frac{(\beta-\alpha)^2}{12}$.

Example 7.12.

Uniform

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. $P(Y \leq 2)$
4. $P(Y > 2)$

Definition 7.20.

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 7.2 are various Normal Distributions with the same $\mu = 1$ and two versions of the variance.

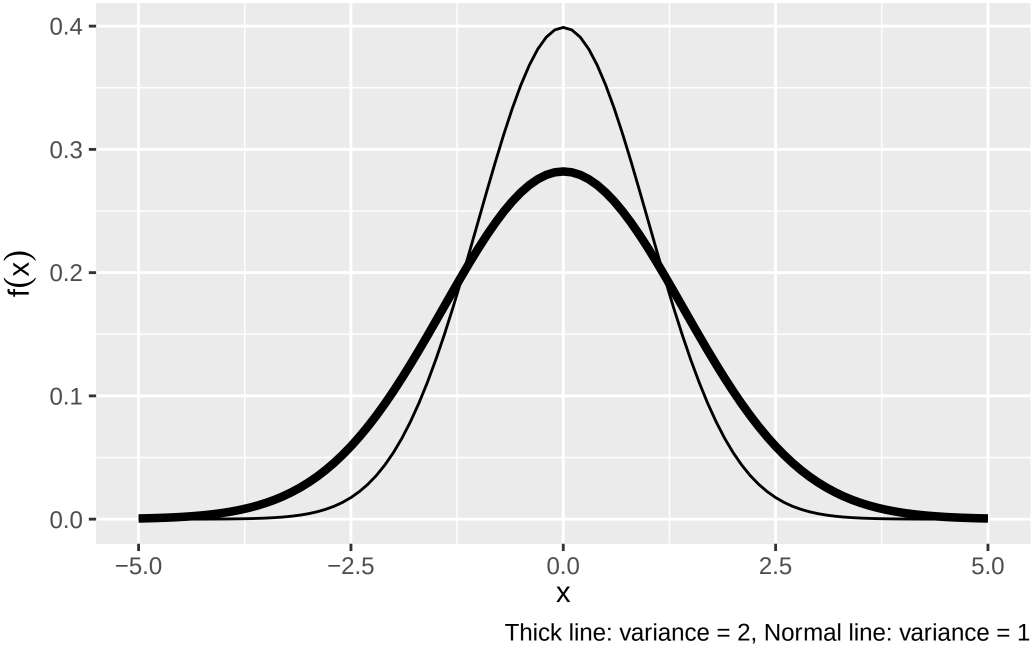


Figure 7.2: Normal Distribution Density

7.10 Summarizing Observed Events (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". We'll spend more time in the intro class talking about this from the standpoint of *estimands*, *estimators* and *estimates*.

Sample mean: This is the most common measure of central tendency, calculated by summing across the observations and dividing by the number of observations.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example:

X	6	3	7	5	5	5	6	4	7	2
Y	1	2	1	2	2	1	2	0	2	0

1. $\bar{x} =$ $\bar{y} =$
2. median(x) = median(y) =
3. $m_x =$ $m_y =$

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.

$$\widehat{\text{Var}}(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Again, this is an *estimator* of the variance of a random variable; we divide by $n-1$ instead of n in order to get an unbiased estimator.

Standard deviation: The sample standard deviation is the square root of the sample variance.

$$\widehat{SD}(X) = \sqrt{\hat{\text{Var}}(X)} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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**: $\hat{\text{Cov}}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

2. **Sample correlation**: $\hat{\text{Corr}} = \frac{\hat{\text{Cov}}(X,Y)}{\sqrt{\hat{\text{Var}}(X)\hat{\text{Var}}(Y)}}$

Example 7.13.

Sample Covariance and Correlation

Example: Using the above table, calculate the sample versions of:

1. $\text{Cov}(X, Y)$
2. $\text{Corr}(X, Y)$

7.11 Asymptotic Theory

In theoretical and applied research, asymptotic arguments are often made. In this section we briefly introduce some of this material.

What are asymptotics? 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.

7.11.1 CLT and LLN

We are now finally ready to revisit, with a bit more precise terms, the two pillars of statistical theory we motivated Section @ref(limitsfun) with.

Theorem 7.1.

Central Limit Theorem

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.

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} Normal(0, 1),$$

Another way to write this as a probability statement is that for all real numbers a,

$$P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq a\right) \rightarrow \Phi(a)$$

as $n \rightarrow \infty$, 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.

Theorem 7.2.

Weak 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.

7.11.2 Big $\mathcal{O}$ Notation

Some of you may encounter "big-OH' '-notation. If f, g are two functions, we say that $f = \mathcal{O}(g)$ if there exists some constant, c, such that $f(n) \leq c \times g(n)$ for large enough n. This notation is useful for simplifying complex problems in game theory, computer science, and statistics.

Example 7.14. What is $\mathcal{O}(5\exp(0.5n) + n^2 + n/2)$? Answer: $\exp(n)$. Why? Because, for large n,

$$\frac{5\exp(0.5n) + n^2 + n/2}{\exp(n)} \leq \frac{c\exp(n)}{\exp(n)} = c.$$

whenever $n > 4$ and where $c = 1$.