A First Course in Probability and Statistics

A First Course in Probability and Statistics

David Goldsman, Ph.D.
Professor
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

Paul Goldsman, Ph.D.

v. 201127.241225

For inquiries, please contact David Goldsman at: sman@gatech.edu

A hard copy of the text can be ordered here:
https://www.lulu.com/search?page=1&q=goldsman&pageSize=10&adult_audience_rating=00

Preface

Probability and Statistics are complementary disciplines concerned with the analysis of randomness in the real world. *Probability*, roughly speaking, describes the random variation in systems, while *Statistics* uses sample data to draw general conclusions about a population.

This book was written with first-year Masters students in engineering and other technical disciplines in mind, but it could also be used for advanced undergraduates or practitioners. Most of the content could be completed in one solid semester, or alternatively in a more leisurely two-course sequence.

Organization

The fun starts in Chapter 1, with an introduction to *basic probability concepts*. Our emphasis is on applications in science and engineering, with the goal of enhancing modeling and analysis skills for a variety of real-world problems.

We have endeavored to keep the text as self-contained as possible, and so we begin by reviewing concepts from set theory and calculus; and we even provide some cheerleading on how to prove things (because we'll be doing some of that in the book). We then discuss fundamental probability axioms that serve as the basis for all of our subsequent work — this discussion will give us the tools to study elementary probability counting rules, including permutations and combinations. We illustrate these concepts with various cool applications, including poker probabilities and wedding invitations gone horribly wrong!

The next stops on our tour are the concepts of *independence* and *conditional probability*, which concern how the probabilities of different events are related to each other, and how new information can be used to update probabilities. The chapter culminates with a discussion of *Bayes Rule* and how it is used to update probabilities when certain information becomes available.

Chapter 2 introduces the concepts of *discrete* and *continuous* random variables. A discrete random variable is a random construct that can take on a countable set of values (such as a subset of the integers), for instance: *How many customers are likely to arrive in the next hour?* A continuous random variable takes values from an uncountable set (such as a continuous interval), for instance: *How long will a lightbulb last?* We'll discuss various properties of random variables, including expected value, variance, and moment generating functions. This leads us to a discussion of functions of random variables, which have important applications in computer simulation and many other areas.

Chapter 3 generalizes our work to *bivariate* (two-dimensional) random variables that may be dependent on each other in some way, for example, a person's height and weight. This chapter provides methodology that complements the one-dimensional case from the previous chapter, while allowing us to extract marginal (one-dimensional) and conditional information from the bivariate versions. This will prepare us for the important concepts of independence and correlation, which we'll discuss in some detail.

Chapter 4 takes us on an excursion of important *probability distributions*, including the Bernoulli, binomial, geometric, Poisson, uniform, exponential, and normal distributions. Particular attention is paid to the normal distribution, because it leads us to the *Central Limit Theorem* (the most-important mathematical result in the universe, actually), which enables us to do wonderful things, such as make probability calculations for arbitrary averages and sums of random variables.

Chapter 5 discusses elementary *descriptive statistics* and *point estimation methods*, including unbiased estimation, maximum likelihood estimation, and the method of moments. We also describe the t, χ^2, and F sampling distributions, which will prove useful in the upcoming statistical applications.

Chapter 6 is concerned with *confidence intervals*, which allow us to make probabilistic statements such as: "Based on the sample of observations we collected, we are 95% sure that the unknown mean lies between A and B," and "We are 95% sure that Candidate Smith's popularity is $52\% \pm 3\%$." We formulate and interpret confidence intervals for a large variety of probability distributions.

Hypothesis testing, the subject of Chapter 7, tells us how to pose hypotheses and test their validity in a statistically rigorous way. For instance: "Does a new drug result in a higher cure rate than the old drug," or "Is the mean tensile strength of item A greater than that of item B?" We discuss the types of errors that can occur with hypothesis testing, and how to design tests to mitigate those errors, and then we formulate and interpret hypothesis tests for a variety of probability distributions.

Computing Resources

We'll be doing a bit of computation in support of the material, and the good news is that statistical software is ubiquitous these days. Elementary analyses can certainly be carried out in Excel, but specialty products such as Minitab, JMP, and R (the latter being freeware) can all be used in conjunction with the text. In fact, we'll provide occasional examples showing how to implement these technologies. Happily, all of these packages are easy to use and only require light programming experience.

Our Approach

Our goal is to gently get readers started on the wonderful topics of Probability and Statistics. Even if your math is a little rusty, we try to give you all the math background you'll need in order to work through the necessary concepts. Some of the topic areas are a little more theoretical than others, and are usually designated as "Honors" items. In any case, we'll try to warn you about these ahead of time, so that you can be properly wary — e.g., §§3.4, 3.7, 4.3.4, and 4.5.2. We have plenty of homework exercises to give you lots of enjoyment, as well as a series of mini-assessments, practice exams, and other resources available on the book's website.

We regard the text as a living document. To this end, we will provide periodic updates. We are human, so we make missteaks, and we encourage you to look around for any goofs and let us know when you find them. As a reward, you shall have our eternal gratitude, and maybe even an "It's Time for t (distribution)" collectible mug.

Dave Goldsman, Atlanta, GA (contact: sman@gatech.edu)
Paul Goldsman, Syracuse, NY
December 18, 2020

Contents

Chapter 1

Probability Basics

This chapter gets us up and running on the basics of probability. We'll cover the following topics with the goal of introducing fundamental tools that will be useful for the remainder of the course.

1.1 Introduction and Motivational Examples

Since this is a course about randomness, let's start with a few examples of everyday phenomena that are chock full of randomness.

- Will my next coin flip be heads or tails?
- How much snow will fall tomorrow?
- Will IBM make a profit this year?
- Should I buy a call or put option?
- Can I win at blackjack?

- How cost-effective is a new drug?
- Which horse will win the Kentucky Derby?
- Will my plane arrive at its destination on time?
- How many times will a political candidate lie during his speech?
- Will a certain gun control law save lives?
- Considering the current traffic situation in The Land of Oz (where I work), what's the best route back to Kansas today?
- If a circuit is experiencing random electrical fluctuations, what is its average resistance? [1]

On the other hand, there are other things that don't seem to exhibit any randomness at all, for instance,

- An assembly line churns out exactly 20 items per hour. After t hours, we have $20t$ items.
- Galileo drops a ball from the Leaning Tower of Pisa from height h_0. After t seconds, its height is $h(t) = h_0 - 16t^2$.
- Deposit \$1000 in a continuously compounding 2% checking account. At time t, it's worth exactly $\$1000e^{0.02t}$.

Or maybe they do — most real-world processes involve at least some degree of randomness and may have to be analyzed using a *probabilistic model*. Let's take a closer look at the previous examples.

- What if the assembly line is subject to random breakdowns? How might that affect production capacity?
- What if Galileo gives the ball a little push by mistake when he drops it?
- What if the bank defaults?

Do randomness and uncertainty genuinely exist in the world? Practically speaking, there are many observable systems that appear to be random — perhaps our knowledge of the underlying rules that govern the systems is incomplete; maybe the true parameters are unknown, unknowable or immeasurable; maybe there are chaotic effects or quantum effects. In many such cases, where we are not able to create an acceptable deterministic model, we can still create a useful probabilistic model. Here are some motivational examples to get us going.

Example: *The Birthday Problem.* Sometimes the results from probability theory can be quite unexpected. What is the probability that two persons in a class (or in any group) have the same birthday?

Assuming that birthdays are distributed evenly across the year,[2] most people would be surprised to learn that:

[1] There's no place like Ohm.

[2] ...and assuming no weirdos born on Feb. 29

- If there are just *23* people in the room, the odds are better than 50–50 that there will be a match.
- If there are *50* people, the probability is about 97%!

Example: *Monopoly.* In the long run, you have the highest probability of landing on Illinois Ave.

Example: *Poker.* Pick five cards from a standard deck of cards. Then,

- The probability that you'll have exactly two pairs is about 0.0475.
- The probability of a full house is about 0.00144.
- The probability of a flush is about 0.00198.

Example: *Stock Market.* Monkeys randomly selecting stocks should be able to outperform most market analysts![3]

Example: A couple has two children and at least one is a boy. What's the probability that *both* are boys? Clearly, the (equally likely) possibilities are GG, BG, GB, and BB. Eliminate GG since we know there's at least one boy. Then the probability that both are boys is 1/3.

Example: From *Ask Marilyn*[4]. Suppose that you are a contestant on the old TV game show *Let's Make a Deal!* The host Monty Hall shows you three doors: A, B, and C. Behind one of the doors is a car; behind the other two are goats. You pick door A. Monty opens door B and reveals a goat. He then offers you a chance to switch to door C. What should you do? Answer: You switch! Stay tuned to see why!

Example: Vietnam War Draft Lottery — not as "fair" as you might think!

Example: What soft drink is the most popular? Well, those of us living in Atlanta know the answer to that one!

Example: Why do some election polls get it so wrong?

Example: How do they do real-time updates of win probabilities as a basketball game progresses?

Example: How can you simulate randomness on a computer, and what can you use it for?

Example: How can you tell if the quality of an item produced by your manufacturing plant has started to decline?

[3] www.telegraph.co.uk/finance/personalfinance/investing/9971683/Monkeys-beat-the-stock-market.html

[4] This is the so-called "Monty Hall" problem.

Let's start with simple, one-sentence **working definitions** of probability and statistics.

Probability— A methodology that describes the random variation in systems. (We'll spend about 50% of our time on this.)

Statistics — A branch of applied mathematics that uses sample data to draw general conclusions about the population from which a sample was taken. (50% of our time.)

Before launching into probability, we'll start with a little bootcamp to review some background material.

1.2 Math Bootcamp

We will review some basic set theory in §1.2.1, calculus in §1.2.2, and then the art of proving things in §1.2.3. This material has been included in order to keep the book more-or-less self-contained. Readers can skip this section if the memories are fresh.

1.2.1 The Joy of Sets

Definition: A **set** is a collection of distinct objects that are likely to be related in some way. Members of a set are called **elements**.

Here is some standard notation that you may have seen in grade school.

- Capital letters for sets (e.g., A, X, Ω, etc.).
- Small letters for elements of a set (a, x, ω, etc.).
- $\in$ denotes set membership, e.g., $x \in X$ means that x is an element of the set X.
- $\notin$ for non-membership, e.g., $x \notin X$.
- Ω often denotes the "universal" set (i.e., the set of everything of interest).
- $\emptyset$ is the **empty set** (aka **null set**), i.e., the set consisting of nothing — no elements at all.

Examples/Notation:

- $A = \{1, 2, 3, 4, 5\}$, i.e., the set of integers from 1 to 5. The order of the elements in sets doesn't matter, so you could also write $A = \{5, 4, 3, 2, 1\}$, and the world would be perfectly okay with that.
- $B = \{x \,|\, -1 \leq x \leq 1\}$, i.e., the set of all x, *such that* x is between -1 and 1, inclusive.

- $C = \{x \,|\, x^2 - 3x + 2 = 0\} = \{1, 2\}$. (Either is fine.)
- $D = \{\text{great songs by Justin Bieber}\} = \emptyset$. (He literally doesn't have any!)
- $E = 2^C \equiv$ **power set** of $C \equiv$ set of all subsets of $C = \{\emptyset, \{1\}, \{2\}, C\}$.
- $\mathbb{R}$ is the set of all real numbers. Note that $\sqrt{-1} \notin \mathbb{R}$.
- $\mathbb{R}^2$ is the set of all ordered pairs of real numbers.
- $\mathbb{Q}$ is the set of all rational numbers. Note that $3.14159 \in \mathbb{Q}$, but $\pi \notin \mathbb{Q}$.

Definition: If every element of set S is an element of set T, then S is called a **subset** of T, which we denote by $S \subseteq T$. Of course, if it's also the case that $T \subseteq S$, then $S = T$.

Examples:

- {a, e, i, o, u} is a subset of the alphabet.
- $\mathbb{Q} \subseteq \mathbb{R}$.
- For any set S and the universal set Ω, we have $\emptyset \subseteq S \subseteq S \subseteq \Omega$. Thus, the empty set is a subset of any other set; and any set is a subset of itself, as well as a subset of Ω.
- Transitivity of subsets: If $X \subseteq Y$ and $Y \subseteq Z$, then $X \subseteq Z$.

Definitions: The **cardinality** of a set S is the number of elements in S, and is denoted by $|S|$. If $|S| < \infty$, then S is **finite**. For example, $S = \{3, 4\}$ is finite, since $|S| = 2$.

The set of natural numbers $\mathbb{N} \equiv \{0, 1, 2, \ldots\}$ is **countably infinite**; the infinite cardinality $|\mathbb{N}|$ is sometimes denoted by $\aleph_0$ ("aleph-naught"). A set that has the same cardinality as a subset of $\mathbb{N}$ is called **countable**. Countable sets can be finite or infinite.

Any set of real numbers defining an interval, e.g., $\{x \,|\, 0 \leq x \leq 1\}$, is **uncountably infinite**; and this cardinality is denoted by $\aleph_1$.

Remarks: Some facts about infinity are surprising, for example: (i) The set of rational numbers $\mathbb{Q}$ has cardinality $\aleph_0$ — the same cardinality as that of the integers! (ii) $\aleph_1$ is a "larger infinity" than $\aleph_0$; in fact, we sometimes write $\aleph_1 = 2^{\aleph_0}$. (iii) The father of set theory was Georg Cantor; his work was so controversial at the time that he suffered severe health issues for many years.

More Definitions:

- The **complement**[5] of a set X is the set of all elements that are not in X, i.e., $\bar{X} \equiv \{x \,|\, x \notin X\}$.

[5] "You're one fine set, X!"

- The **intersection** of sets X and Y is the set of all elements that are in *both* sets, i.e., $X \cap Y \equiv \{z \,|\, z \in X \text{ and } z \in Y\}$. If $X \cap Y = \emptyset$ (i.e., no common elements), then X and Y are **disjoint** or **mutually exclusive** sets.
- The **union** of sets X and Y is the set of all elements contained in either or both sets, i.e., $X \cup Y \equiv \{z \,|\, z \in X \text{ and/or } z \in Y\}$.
- The **minus** operation for sets X and Y is $X - Y \equiv X \cap \bar{Y}$, i.e., the set one gets when one removes from X anything that also happens to be in Y.
- The **symmetric difference** or **XOR** ("exclusive or") between two sets is everything in X or Y but not both,

$$\begin{aligned} X \,\Delta\, Y &\equiv (X - Y) \cup (Y - X) \\ &= (X \cap \bar{Y}) \cup (\bar{X} \cap Y) \\ &= (X \cup Y) - (X \cap Y). \end{aligned}$$

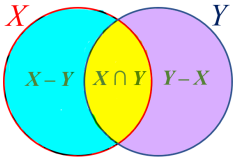
Example: Consider the following universal set of funny symbols,

$$\Omega = \{※, \infty, ☺, \heartsuit, \nabla, \Delta, \$, £, ¥\},$$

along with subsets $X =$ types of money $= \{\$, £, ¥\}$, and $Y = \{\heartsuit, \nabla, \Delta, \$\}$.

- $\bar{X} = \{※, \infty, ☺, \heartsuit, \nabla, \Delta\}$.
- $X \cap Y = \{\$\}$.
- $X \cup Y = \{\heartsuit, \nabla, \Delta, \$, £, ¥\}$.
- $X - Y = \{£, ¥\}$.
- $X \,\Delta\, Y = \{\heartsuit, \nabla, \Delta, £, ¥\}$. □

Remark: One can illustrate the various set operations via *Venn diagrams*. Here's what the minus and intersection operations look like.



Some Laws Involving Set Operations:

- Complements: $X \cup \bar{X} = \Omega$, $X \cap \bar{X} = \emptyset$, and $\bar{\bar{X}} = X$.
- Commutative: $X \cap Y = Y \cap X$, and $X \cup Y = Y \cup X$.
- Associative: $X \cap (Y \cap Z) = (X \cap Y) \cap Z = X \cap Y \cap Z$, and $X \cup (Y \cup Z) = (X \cup Y) \cup Z = X \cup Y \cup Z$.

- Distributive: $X \cap (Y \cup Z) = (X \cap Y) \cup (X \cap Z)$, and $X \cup (Y \cap Z) = (X \cup Y) \cap (X \cup Z)$.
- DeMorgan's: $\overline{X \cap Y} = \bar{X} \cup \bar{Y}$, and $\overline{X \cup Y} = \bar{X} \cap \bar{Y}$.

Proof: The proofs of these results can be carried out using Venn diagrams or other elementary methods. To illustrate, we'll do the second half of DeMorgan.

$$\begin{aligned} z \in \overline{X \cup Y} &\Leftrightarrow z \notin X \cup Y \\ &\Leftrightarrow z \notin X \text{ and } z \notin Y \\ &\Leftrightarrow z \in \bar{X} \text{ and } z \in \bar{Y} \\ &\Leftrightarrow z \in \bar{X} \cap \bar{Y}. \quad \square \end{aligned}$$

Example: Suppose that $\Omega = \{1, 2, \ldots, 10\}$, X is the set of prime numbers (≤ 10), and $Y = \{1, 2, 3, 4, 6\}$. Let's verify that DeMorgan works (we'll start at the ends and meet in the middle):

$$\begin{aligned} \overline{X \cup Y} &= \overline{\{2,3,5,7\} \cup \{1,2,3,4,6\}} \\ &= \overline{\{1,2,3,4,5,6,7\}} \\ &= \{8,9,10\} \\ &= \{1,4,6,8,9,10\} \cap \{5,7,8,9,10\} \\ &= \overline{\{2,3,5,7\}} \cap \overline{\{1,2,3,4,6\}} \\ &= \bar{X} \cap \bar{Y}. \quad \square \end{aligned}$$

1.2.2 Calculus Primer

In this bootcamp, we'll begin with some fundamentals, and then proceed to discussions on derivatives, integration, and other items of interest.

1.2.2.1 The Basics

First of all, let's suppose that $f(x)$ is a **function** that maps values of x from a certain **domain** X to a certain **range** Y. We denote this by the shorthand $f : X \to Y$.

Example: If $f(x) = x^2$, then the function takes x-values from the real line $\mathbb{R}$ (domain X) to the nonnegative portion of the real line $\mathbb{R}^+$ (range Y).

Definition: We say that $f(x)$ is a **continuous** function if, for any x_0 and $x \in X$, we have $\lim_{x \to x_0} f(x) = f(x_0)$, where "lim" denotes a **limit**, and $f(x)$ is assumed to exist for all $x \in X$.

Example: The function $f(x) = 3x^2$ is continuous for all x. The function $f(x) = \lfloor x \rfloor$ (round down to the nearest integer, e.g., $\lfloor 3.4 \rfloor = 3$) has a "jump" discontinuity at any integer x. $\square$

Definition: The **inverse** of a function $f : X \to Y$ is (informally) the "reverse" mapping $g : Y \to X$, such that $f(x) = y$ if and only if $g(y) = x$, for all appropriate

x and y. The inverse of f is often written as f^{-1}, and is especially useful if $f(x)$ is a strictly increasing or strictly decreasing function. Note that $f^{-1}(f(x)) = x$.

Examples: If $f(x) = x^3$, then we have $f^{-1}(y) = y^{1/3}$. If $f(x) = e^x$, then $f^{-1}(y) = \ell\mathrm{n}(y)$. But sadly, $f(x) = x^2$ *doesn't* have a unique inverse. □

1.2.2.2 Differentiation

Definition: If $f(x)$ is continuous, then it is **differentiable** (has a **derivative**) if

$$\frac{d}{dx}f(x) \equiv f'(x) \equiv \lim_{h\to 0}\frac{f(x+h)-f(x)}{h}$$

exists and is well-defined for any given x. Think of the derivative as the slope of the function.

Example: Some well-known derivatives are:

$$\begin{aligned}
[x^k]' &= kx^{k-1}\\
[e^x]' &= e^x\\
[\sin(x)]' &= \cos(x)\\
[\cos(x)]' &= -\sin(x)\\
[\ell\mathrm{n}(x)]' &= \frac{1}{x}\\
[\arctan(x)]' &= \frac{1}{1+x^2}. \quad \square
\end{aligned}$$

Theorem: Some well-known properties of derivatives are:

$$\begin{aligned}
[af(x)+b]' &= af'(x)\\
[f(x)+g(x)]' &= f'(x)+g'(x)\\
[f(x)g(x)]' &= f'(x)g(x)+f(x)g'(x) \quad \textbf{(product rule)}\\
\left[\frac{f(x)}{g(x)}\right]' &= \frac{g(x)f'(x)-f(x)g'(x)}{g^2(x)} \quad \textbf{(quotient rule)}\\
[f(g(x))]' &= f'(g(x))g'(x) \quad \textbf{(chain rule for compositions)}.
\end{aligned}$$

Example: Suppose that $f(x) = x^2$, and $g(x) = \ell\mathrm{n}(x)$. Then

$$\begin{aligned}
[f(x)g(x)]' &= f'(x)g(x)+f(x)g'(x)\\
&= 2x\ell\mathrm{n}(x)+x^2(1/x)\\
&= 2x\ell\mathrm{n}(x)+x,
\end{aligned}$$

$$\begin{aligned}
\left[\frac{f(x)}{g(x)}\right]' &= \frac{g(x)f'(x)-f(x)g'(x)}{g^2(x)}\\
&= \frac{[\ell\mathrm{n}(x)]2x-x^2(1/x)}{\ell\mathrm{n}^2(x)}\\
&= \frac{2x\ell\mathrm{n}(x)-x}{\ell\mathrm{n}^2(x)}, \quad \text{and}
\end{aligned}$$

$$[f(g(x))]' \;=\; \left[(g(x))^2\right]' \;=\; 2g(x)g'(x) \;=\; \frac{2\ell\mathrm{n}(x)}{x}. \quad \square$$

Remarks: The second derivative $f''(x) \equiv \frac{d}{dx}f'(x)$ and is the "slope of the slope." If $f(x)$ is "position," then $f'(x)$ can be regarded as "velocity," and $f''(x)$ as "acceleration." The minimum or maximum of $f(x)$ can only occur when the slope of $f(x)$ is 0, i.e., only when $f'(x) = 0$, say at the **critical point** $x = x_0$. Exception: Check the endpoints of your interval of interest as well. Then if $f''(x_0) < 0$, you get a max; if $f''(x_0) > 0$, you get a min; and if $f''(x_0) = 0$, you get a **point of inflection**.

Example: Find the value of x that minimizes $f(x) = e^{2x} + e^{-x}$. The minimum can only occur when $f'(x) = 2e^{2x} - e^{-x} = 0$. After a little algebra, we find that this occurs at $x_0 = -(1/3)\ell\mathrm{n}(2) \doteq -0.231$. It's also easy to show that $f''(x) > 0$ for all x, so x_0 yields a minimum. $\square$

1.2.2.3 Integration

Definition: The function $F(x)$ having derivative $f(x)$ is called the **antiderivative** (or **indefinite integral**). It is denoted by $F(x) = \int f(x)\,dx$.

Fundamental Theorem of Calculus: If $f(x)$ is continuous, then the area under the curve for $x \in [a, b]$ is denoted and given by the **definite integral**[6]

$$\int_a^b f(x)\,dx \;\equiv\; F(x)\Big|_a^b \;\equiv\; F(b) - F(a).$$

Example: Some well-known indefinite integrals are:

$$\int x^k\,dx \;=\; \frac{x^{k+1}}{k+1} + C, \quad \text{for } k \neq -1$$

$$\int \frac{dx}{x} \;=\; \ell\mathrm{n}|x| + C$$

$$\int e^x\,dx \;=\; e^x + C$$

$$\int \cos(x)\,dx \;=\; \sin(x) + C$$

$$\int \frac{dx}{1+x^2} \;=\; \arctan(x) + C,$$

where C is an arbitrary constant. $\square$

Example: It is easy to see that

$$\int \frac{d\,\mathrm{cabin}}{\mathrm{cabin}} \;=\; \ell\mathrm{n}|\mathrm{cabin}| + C \;=\; \text{houseboat}. \quad \ddot{\smile}$$

[6] "I'm *really* an integral!"

Theorem: Some well-known properties of definite integrals are:

$$\int_a^a f(x)\,dx \;=\; 0$$

$$\int_a^b f(x)\,dx \;=\; -\int_b^a f(x)\,dx$$

$$\int_a^b f(x)\,dx \;=\; \int_a^c f(x)\,dx + \int_c^b f(x)\,dx.$$

Theorem: Some other properties of general integrals are:

$$\int [f(x)+g(x)]\,dx \;=\; \int f(x)\,dx + \int g(x)\,dx$$

$$\int f(x)g'(x)\,dx \;=\; f(x)g(x) - \int g(x)f'(x)\,dx \quad \text{(integration by parts)}$$

$$\int f(g(x))g'(x)\,dx \;=\; \int f(u)\,du \quad \text{(substitution rule)}.$$

Example: To demonstrate integration by parts on a definite integral, let $f(x) = x$ and $g'(x) = e^{2x}$, so that $g(x) = e^{2x}/2$. Then

$$\begin{aligned}
\int_0^1 xe^{2x}\,dx &= \int_0^1 f(x)g'(x)\,dx \\
&= f(x)g(x)\Big|_0^1 - \int_0^1 g(x)f'(x)\,dx \quad \text{(parts)} \\
&= \frac{xe^{2x}}{2}\Big|_0^1 - \int_0^1 \frac{e^{2x}}{2}\,dx \\
&= \frac{e^2}{2} - \frac{e^{2x}}{4}\Big|_0^1 \\
&= \frac{e^2+1}{4}. \quad \square
\end{aligned}$$

1.2.2.4 Series

Definition: Derivatives of arbitrary order k can be written as $f^{(k)}(x)$ or $\frac{d^k}{dx^k}f(x)$. By convention, $f^{(0)}(x) = f(x)$.

The **Taylor series expansion** of $f(x)$ about a point a is given by

$$f(x) \;=\; \sum_{k=0}^{\infty} \frac{f^{(k)}(a)(x-a)^k}{k!}.$$

The **Maclaurin series** is simply Taylor expanded around $a = 0$.

Example: Here are some famous Maclaurin series:

$$\sin(x) = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k+1}}{(2k+1)!}$$

$$\cos(x) = \sum_{k=0}^{\infty} \frac{(-1)^k x^{2k}}{(2k)!}$$

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}.$$

Example: And while we're at it, here are some miscellaneous sums that you should know:

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}$$

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6}$$

$$\sum_{k=0}^{\infty} p^k = \frac{1}{1-p} \text{ (for } -1 < p < 1).$$

1.2.2.5 Going to the Hospital

Theorem: Occasionally, we run into trouble when taking indeterminate ratios of the form $0/0$ or ∞/∞. In such cases, **L'Hôpital's Rule** is useful:

If $\lim_{x\to a} f(x)$ and $\lim_{x\to a} g(x)$ both go to 0 or both go to ∞, then

$$\lim_{x\to a} \frac{f(x)}{g(x)} = \lim_{x\to a} \frac{f'(x)}{g'(x)}.$$

Example: L'Hôpital shows that

$$\lim_{x\to 0} \frac{\sin(x)}{x} = \lim_{x\to 0} \frac{\cos(x)}{1} = 1. \quad \square$$

1.2.2.6 Double Integration

We'll have occasion to calculate several double integrals. Whereas our usual (single) integrals get us the area under a curve, double integrals represent the *volume* under a three-dimensional function.

Example: The volume under $f(x, y) = 8xy$, over the region $0 < x < y < 1$, is given by

$$\int_0^1 \int_0^y f(x,y)\, dx\, dy = \int_0^1 \int_0^y 8xy\, dx\, dy = \int_0^1 4y^3\, dy = 1. \quad \square$$

We can usually swap the order of integration to get the same answer (a manipulation known as "Fubini magic"):

$$\int_0^1 \int_x^1 8xy \, dy \, dx \;=\; \int_0^1 4x(1-x^2) \, dx \;=\; 1. \quad \square$$

Some double integration problems can be solved more easily via a transformation from the (x, y)-plane to **polar coordinates** (r, θ), in which we set $x = r\cos(\theta)$ and $y = r\sin(\theta)$. Then it can be shown that

$$\iint_A f(x, y) \, dx \, dy \;=\; \iint_B f(r\cos(\theta), r\sin(\theta)) \, r \, dr \, d\theta,$$

for appropriate regions of integration A and B.

Example: We can use polar coordinates and Fubini to calculate the otherwise challenging quantity,

$$\begin{aligned}
\left(\int_{\mathbb{R}} e^{-x^2/2} \, dx \right)^2 &= \int_{\mathbb{R}} \int_{\mathbb{R}} e^{-(x^2+y^2)/2} \, dx \, dy \\
&= \int_0^{2\pi} \int_0^{\infty} e^{-r^2(\cos^2(\theta)+\sin^2(\theta))/2} \, r \, dr \, d\theta \\
&= \int_0^{\infty} r \, e^{-r^2/2} \int_0^{2\pi} d\theta \, dr \;=\; 2\pi. \quad \square
\end{aligned}$$

1.2.2.7 Saved by Zero! (How to Solve for a Root)

Suppose that we want to solve some equation $g(r) = 0$ for a root $r^\star$, where $g(r)$ is a nicely behaved continuous function. We can use:

- trial-and-error or some sort of linear search — that's for losers! ☹
- **bisection method** — pretty fast! ☺
- **Newton's method** — really fast! 😲

First, a useful theorem...

Intermediate Value Theorem (IVT): If $g(\ell)g(u) < 0$, then there is a zero (root) $r^\star \in [\ell, u]$. In other words, if (i) $g(\ell) < 0$ and $g(u) > 0$, or (ii) $g(\ell) > 0$ and $g(u) < 0$, then $g(r)$ crosses 0 somewhere between ℓ and u.

Bisection uses the IVT to hone in on a zero via sequential bisecting:

- Initialization: Find lower and upper bounds $\ell_0 < u_0$, such that $g(\ell_0)g(u_0) < 0$. Then the IVT implies that $r^\star \in [\ell_0, u_0]$.
- For $j = 0, 1, 2, \ldots,$
  - Let the midpoint of the current interval be $r_j \leftarrow (\ell_j + u_j)/2$.

- If $g(r_j)$ is sufficiently close to 0, or the interval width $u_j - \ell_j$ is sufficiently small, or your iteration budget is exceeded, then set $r^\star \leftarrow r_j$ and STOP.
- If the sign of $g(r_j)$ matches that of $g(\ell_j)$, this means that $r^\star \in [r_j, u_j]$; so set $\ell_{j+1} \leftarrow r_j$ and $u_{j+1} \leftarrow u_j$. Otherwise, $r^\star \in [\ell_j, r_j]$; so set $\ell_{j+1} \leftarrow \ell_j$ and $u_{j+1} \leftarrow r_j$.

Each iteration of the algorithm chops the search area in half and therefore converges to $r^\star$ pretty quickly.

Example: Use bisection to find $\sqrt{2}$, by solving $g(x) = x^2 - 2 = 0$. In order to initialize the bisection algorithm, we note that $g(1) = -1$ and $g(2) = 2$. So there's a root in $[1, 2]$ just itching to be found!

step	ℓ_j	$g(\ell_j)$	u_j	$g(u_j)$	r_j	$g(r_j)$
0	1	-1	2	2	1.5	0.25
1	1	-1	1.5	0.25	1.25	-0.4375
2	1.25	-0.4375	1.5	0.25	1.375	-0.1094
3	1.375	-0.1094	1.5	0.25	1.4375	0.0664
4	1.375	-0.1094	1.4375	0.0664	1.40625	-0.0225
$\vdots$						

You can see that r_j seems to be converging to $r^\star = \sqrt{2} \doteq 1.4142$. □

Newton's method. This technique uses information about both the function g and its derivative g', in order to aim the search in the right direction more efficiently. As a result, it's usually a lot faster than bisection. Here's a reasonable implementation.

1. Initialize r_0 as some first guess of the root. Set $j \leftarrow 0$.
2. Update $r_{j+1} \leftarrow r_j - g(r_j)/g'(r_j)$.
3. If $|g(r_{j+1})|$ or $|r_{j+1} - r_j|$ or your budget is suitably small, then STOP and set $r^\star \leftarrow r_{j+1}$. Otherwise, let $j \leftarrow j + 1$ and go back to Step 2.

Example: Use Newton to find $\sqrt{2}$ by solving $g(x) = x^2 - 2 = 0$. To do so, note that

$$r_{j+1} \leftarrow r_j - \frac{g(r_j)}{g'(r_j)} \;=\; r_{j+1} \leftarrow r_j - \frac{r_j^2 - 2}{2r_j} \;=\; \frac{r_j^2 + 2}{2r_j}.$$

If $r_0 = 1$, then we find that $r_1 = 3/2$, $r_2 = 17/12 \doteq 1.4167$, $r_3 = 577/408 \doteq 1.4142$, ... That's fast convergence! ☺

1.2.3 Proving Things

We do a number of proofs in this book (some of which you can feel free to skip over). The current subsection presents a small list of useful proof techniques that you are likely to encounter in your travels. If you enjoy this topic, a terrific reference is Polya's classic text [6].

§1.2.3.1 It's Completely Obvious

Sometimes things are so patently clear that it's probably okay to just assume that they are true — so you don't have to "prove" them at all! For example, $1 + 1 = 2$, $a^2 + b^2 = c^2$ for a right triangle, $x^2 \geq 0$, Justin Bieber is annoying, etc. Most of the time, it's perfectly fine to assume that all is fine and dandy with such results and to proceed with our lives uninterrupted.

Nevertheless, you should always be on the lookout for trouble. After all, $1 + 1 = 10$ in base 2 (also see Chapter 5 of Enderton [2] to see just how much care it takes to prove that $1 + 1 = 2$ rigorously); $a^2 + b^2 \neq c^2$ on a sphere; $i^2 = -1 < 0$, where $i \equiv \sqrt{-1}$; and Justin Bieber — well, actually, he *is* annoying! More to the point, we will encounter a number of extremely non-intuitive examples later in this very chapter, so you may have to watch your step.

The bottom line is that you really shouldn't worry about most of the obvious stuff — though it never hurts to be just a bit paranoid, because maybe they *are* out to get you!

§1.2.3.2 Intimidation (aka, "You'd Better Believe Me!")

We've all heard people tell us "It's right because I say it's right!" That also happens in mathematics. For instance, you should know the following from grade school, and don't you dare ask me to prove it!

Binomial Theorem:

$$(x + y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}, \tag{1.1}$$

where the binomial coefficient is $\binom{n}{k} \equiv \frac{n!}{k!\,(n-k)!}$, for $k = 1, 2, \ldots, n$.

We may sneak in this questionable proof technique every once in a while, but we'll always warn you so that you can proceed with the proper mindset. Just because we're awesome doesn't mean that you always have to believe us — but we will try not to lead you astray!

§1.2.3.3 Slippery Proofs

Some questionable proofs don't rise to the level of intimidation, but they can start you on a slippery slope. For instance, you may see phrases like "without going into details" or "under certain conditions" or "trust me. . . ." We are occasionally guilty of such proofiness, often for the sake of clarity of exposition. For instance, in order to avoid clogging up an otherwise clean proof, we have been known to make illegal interchanges such as $\frac{d}{dt} \sum_{k=1}^{\infty} f_n(t) = \sum_{k=1}^{\infty} \frac{d}{dt} f_n(t)$, which are not necessarily true due to that infinite sum. The good news is that we will let you know when we pull such a move.

§1.2.3.4 Complete Enumeration

Until now, we have discussed some shadowy proof techniques. Now we'll finally start commenting on more-rigorous methodologies.

You can occasionally prove a result simply by enumerating every single possibility and showing that the result is true for each. This approach is probably fine for smaller problems, but can quickly become impossible as the size of the problem grows.

Example: There are exactly eight prime numbers less than 20. To establish this, we can tediously write out the relevant prime factorizations. □

number	1	2	3	4	5	6	7	8	9	10
factors	??	☺	☺	2^2	☺	$2 \cdot 3$	☺	2^3	3^2	$2 \cdot 5$
number	11	12	13	14	15	16	17	18	19	20
factors	☺	$2 \cdot 3 \cdot 4$	☺	$2 \cdot 7$	$3 \cdot 5$	2^4	☺	$2 \cdot 3^2$	☺	$2^2 \cdot 5$

Example: The Traveling Salesman Problem challenges you to find the minimum-distance itinerary that goes through every city in the set $\{1, 2, \ldots, n\}$ in any order you please, so long as you start and stop at the same city and don't visit any city (other than your starting point) more than once. Assuming the distance from city i to city j is the same as that from j to i, then it is easy to show (especially after you finish §1.5.2) that you have to check the distances of $(n-1)!/2$ itineraries. One might be tempted to enumerate all of the possibilities, but this becomes impossibly tedious, even for modest n (see Cook [1]). □

§1.2.3.5 Grind It Out

A family relative of complete enumeration is that of grinding out an answer, usually after significant blood, sweat, and tears involving vehicles such as calculus and algebra. Examples of this approach appear as Honors results here and there in the text, e.g., our proof of the general Principle of Inclusion-Exclusion in §1.3.4.

There is really nothing to be ashamed of in this workmanlike approach, though it may be possible to find elegant short cuts that save on the effort.

§1.2.3.6 Special Case of Something Bigger

It often happens that you already have the tools you need right in front of you — just like Dorothy in The Land of Oz! Indeed, you can save a lot of work by spotting the right special cases of known results.

Example: Surprising fact:

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n.$$

Proof: By the Binomial Theorem (1.1),

$$2^n \;=\; (1+1)^n \;=\; \sum_{k=0}^{n} \binom{n}{k} 1^k \, 1^{n-k}. \quad \square$$

§1.2.3.7 Counterexample

In order for a statement to be true, it has be true *all of the time*. If you can find just one case, when the statement doesn't hold then it's false. Counterexamples are nice ways to settle true/false arguments. Once you find one, you can declare the statement to be false and walk away. Of course, a counterexample isn't a proof (it's just the opposite), but it saves you a lot of effort if you're trying to find a proof for a false result.

Example: We can disprove the crazy claim that all of Justin Bieber's songs are great, simply by finding a single terrible song . . . which is quite easy to do. ※

Example: How about the claim that if p is a prime number, then so is $2^p - 1$? On the surface, this looks pretty reasonable. After all, $2^2 - 1 = 3$ is prime, and so are $2^3 - 1 = 7$, $2^5 - 1 = 31$, and $2^7 - 1 = 127$. But, sadly, $2^{11} - 1 = 2047 = 23 \cdot 89$, disproving my claim. ※

On the other hand, it turns out that the largest known prime number (as of late 2018) indeed has the form $2^p - 1$, where $p = 282{,}589{,}933$. So now you can impress your friends!

§1.2.3.8 Contradiction

Proof by contradiction is a wonderful tool to have in your wheelhouse. Suppose you want to prove claim A. Contradiction asks you to instead proceed under the (incorrect) assumption that A is *false*. If that assumption results in a fundamental contradiction, then the assumption that A is false must have been incorrect; therefore, the only remaining alternative is that A must be true! Let's illustrate this clever trick with a couple of famous examples.

Example: There are an infinite number of prime numbers. To prove this, suppose there are only a *finite* number n of primes, say, $p_1, p_2, \ldots, p_n$. Now let's consider the number $P = 1 + \prod_{k=1}^{n} p_k$, i.e., one plus the product of all of our n primes. Note that none of $p_1, p_2, \ldots, p_n$ are factors of P (because of the $+1$ we've added to the product). Therefore, P must be prime — but this would be our $(n+1)^{\text{st}}$ prime, which is a contradiction of assumption of only n primes. So there are an infinite number of primes. ※

Example: $\sqrt{2}$ is irrational. To prove this, suppose that it's rational. This means (by definition of rationals) that there exist integers p and q *with no common factors* such that $\sqrt{2} = p/q$, or, equivalently, $p^2 = 2q^2$. This, in turn, implies that p^2 is even; and this means that p must be even. Thus, p^2 is actually divisible by 4, and so we can write $p^2 = 4k$ for some integer k. Thus, $4k = 2q^2$, so that $q^2 = 2k$. This now

implies q^2 and then q are even. So we now have p and q even. But this contradicts the requirement that p and q have no common factors! So our assumption that $\sqrt{2}$ is rational is false. So $\sqrt{2}$ is irrational. ※

§1.2.3.9 Induction

Induction is a powerful technique that comes into play when we want to prove statements of the form $\mathcal{S}(n)$ for all integers $n \geq 1$. The idea is to

1. Establish the result for a "base case," say $\mathcal{S}(1)$.
2. Assume the truth of $\mathcal{S}(n)$.
3. Show that the truth of $\mathcal{S}(n)$ implies the truth of $\mathcal{S}(n+1)$.

Putting these steps together, we will have

$$\mathcal{S}(0) \;\Rightarrow\; \mathcal{S}(1) \;\Rightarrow\; \mathcal{S}(2) \;\Rightarrow\; \cdots \Rightarrow\; \mathcal{S}(n) \;\Rightarrow\; \mathcal{S}(n+1) \;\Rightarrow\; \mathcal{S}(n+2) \;\Rightarrow\; \cdots,$$

i.e., $\mathcal{S}(n)$ is true for *all* $n \geq 1$. A couple of examples will illustrate how this works.

Example: We'll use induction to prove the well-known algebra result from §1.2.2.4 that $\sum_{k=1}^{n} k^2 = n(n+1)(2n+1)/6$.

Base Case ($n = 1$): $\sum_{k=1}^{1} k^2 = 1 = 1(1+1)(2(1)+1)/6$. ☺

Inductive Step: Assume $\sum_{k=1}^{n} k^2 = n(n+1)(2n+1)/6$ is true.

Now the "$n+1$" step:

$$\begin{aligned}
\sum_{k=1}^{n+1} k^2 &= \sum_{k=1}^{n} k^2 + (n+1)^2 \\
&= \frac{n(n+1)(2n+1)}{6} + (n+1)^2 \quad \text{(inductive step)} \\
&= \frac{(n+1)\big[n(2n+1) + 6(n+1)\big]}{6} \\
&= \frac{(n+1)\big((n+1)+1\big)\,\big(2(n+1)+1\big)}{6}. \quad \square
\end{aligned}$$

Example: Here's a proof by induction of the Binomial Theorem (1.1).

Base Case ($n = 0$): Noting that $0! = 1$, we have

$$\sum_{k=0}^{0} \binom{0}{k} x^k y^{0-k} \;=\; \binom{0}{0} x^0 y^{0-0} \;=\; 1 \;=\; (x+y)^0. \quad ☺$$

Inductive Step: Assume Equation (1.1) is true.

We will be done when we prove the "$n+1$" case, i.e.,

$$(x+y)^{n+1} \;=\; \sum_{k=0}^{n+1} \binom{n+1}{k} x^k y^{n+1-k}.$$

Before we jump into the fire, note that

$$\begin{aligned}
\binom{n+1}{k} &= \frac{(n+1)!}{k!\,(n+1-k)!} \\
&= \frac{n!}{(k-1)!\,(n-k)!}\left[\frac{n+1}{k\,(n+1-k)}\right] \\
&= \frac{n!}{(k-1)!\,(n-k)!}\left[\frac{1}{n+1-k}+\frac{1}{k}\right] \\
&= \frac{n!}{(k-1)!\,(n+1-k)!}+\frac{n!}{k!\,(n-k)!} \\
&= \binom{n}{k-1}+\binom{n}{k}.
\end{aligned}$$

Since we need $k \geq 1$ for the $\binom{n}{k-1}$ term and $k \leq n$ for the $\binom{n}{k}$ term, we have

$$\begin{aligned}
\sum_{k=0}^{n+1}\binom{n+1}{k}x^k y^{n+1-k} &= \sum_{k=1}^{n+1}\binom{n}{k-1}x^k y^{n+1-k}+\sum_{k=0}^{n}\binom{n}{k}x^k y^{n+1-k} \\
&= x\sum_{k=0}^{n}\binom{n}{k}x^k y^{n-k}+y\sum_{k=0}^{n+1}\binom{n}{k}x^k y^{n-k} \\
&= x(x+y)^n+y(x+y)^n \;=\; (x+y)^{n+1}. \quad \square
\end{aligned}$$

1.3 Experiments and Probability Spaces

In this section, we'll finally start talking about random experiments, how we can represent the possible outcomes from these experiments, and then some of the issues involving the probabilities associated with the outcomes.

First of all, here are a few examples of random experiments that one could conduct.

- Toss a coin and observe the outcome: heads (H) or tails (T).
- Toss a coin twice and observe the sequence of H's and T's.
- Toss a six-sided die and observe the outcome.
- What is the GPA of a random university student?
- How long will a lightbulb last?
- Ask 10 people if they prefer Coke or Pepsi.

In order to analyze experiments such as these, we need a common language to describe the "players." The notion of the *probability space* of an experiment fits the bill, where we speak of the constituent *sample space*, *events* and *probability function*.

Definition: The **probability space** associated with an experiment is a triple consisting of the following components.

1. A **sample space**, which is the set of all possible outcomes of the experiment. It is usually denoted by S or Ω.

2. A set $\mathcal{F}$ of events, where each **event** is itself a subset of Ω. We can informally think of $\mathcal{F}$ as the set of all subsets of Ω (i.e., the power set of Ω); see Problem 2.

3. A **probability function** $\mathrm{P} : \mathcal{F} \to [0, 1]$ that assigns probabilities to events.

Let's look at each of these components in a bit more detail.

1.3.1 Sample Space

Examples: Here are sample spaces for each of the experiments described above.

- Toss a coin: $\Omega = \{\mathrm{H}, \mathrm{T}\}$.
- Toss a coin twice: $\Omega = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$.
- Toss a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- University student GPA: $\Omega = \{0, 0.1, 0.2, \ldots, 3.9, 4.0\}$.
- Lightbulb life: $\Omega = \{x \,|\, x \geq 0\}$.
- Prefer Coke over Pepsi: $\Omega = \{0, 1, 2, \ldots, 10\}$.

Remark: A sample space doesn't have to be uniquely defined — it depends on what we're interested in. For instance, an alternate sample space for the experiment in which we toss a coin twice could be $\Omega' = \{0, 1, 2\}$, which could be interpreted as the number of H's that we observe.

1.3.2 Events

An event is simply a set of possible outcomes. Thus,

- Any subset of the sample space Ω is an event.
- The empty set $\emptyset$ is an event of Ω ("none of the possible outcomes of the experiment are observed").
- The sample space Ω is itself an event ("something from the sample space happens").
- If $A \subseteq \Omega$ is an event, then $\bar{A}$ is the complementary (opposite) event ("A doesn't happen").
- If A and B are events, then $A \cup B$ is an event ("A or B or both happen").
- If A and B are events, then $A \cap B$ is an event ("A and B both happen").

Example: Toss an eight-sided Dungeons and Dragons die. $\Omega = \{1, 2, \ldots, 8\}$. If A is the event "an odd number occurs," then $A = \{1, 3, 5, 7\}$, i.e., when the die is tossed, we get 1 or 3 or 5 or 7. Moreover, $\bar{A} = \{2, 4, 6, 8\}$, i.e., the complementary event that an even number occurs.

Example: Toss three coins.

$$\begin{aligned} A &= \text{"exactly one T was observed"} = \{\text{HHT}, \text{HTH}, \text{THH}\} \\ B &= \text{"no T's observed"} = \{\text{HHH}\} \\ C &= \text{"first coin is H"} = \{\text{HHH}, \text{HHT}, \text{HTH}, \text{HTT}\} \end{aligned}$$

Then,

$$\begin{aligned} A \cup B &= \text{"at most one T observed"} = \{\text{HHT}, \text{HTH}, \text{THH}, \text{HHH}\} \\ A \cap C &= \{\text{HHT}, \text{HTH}\}. \quad \square \end{aligned}$$

1.3.3 What is Probability?

We all have an intuitive idea of what probability is. In this subsection, we'll try to formalize the concept to some extent.

We can regard probability as a function that maps an event A from the sample space Ω to the interval $[0, 1]$. (Clearly, the probability of an event cannot be less than 0 or greater than 1.) The probability that A occurs is denoted by $\text{P}(A)$.

Example: If we toss a fair coin, then of course $\Omega = \{\text{H}, \text{T}\}$. What is the probability that H will come up (i.e., that the event $A = \{\text{H}\}$ occurs)? We all know that the answer is $\text{P}(\{\text{H}\}) = \text{P}(\text{H}) = 1/2$ (we drop the annoying braces when the meaning is obvious).

What does this mean?

The *frequentist view* of probability says that if the experiment were to be repeated n times, where n is very large, then we would expect about $1/2$ of the tosses to be H's:

$$\frac{\text{Total number of H's out of } n \text{ tosses}}{n} \doteq 1/2. \quad \square$$

Example: Toss a fair die. $\Omega = \{1, 2, 3, 4, 5, 6\}$, where each individual outcome has probability $1/6$. Then $\text{P}(1 \text{ or } 2) = 1/3$.

More-Formal Definition (see, e.g., Meyer [5]): The **probability** of a generic event $A \subseteq \Omega$ is a function that adheres to the following *axioms*:

(1) $0 \leq \text{P}(A) \leq 1$ (probabilities are *always* between 0 and 1).

(2) $\text{P}(\Omega) = 1$ (probability of *some* outcome from Ω is 1).

Example: Toss a die. $\text{P}(\Omega) = \text{P}(1, 2, 3, 4, 5, 6) = 1$.

(3) If A and B are *disjoint* events, i.e., $A \cap B = \emptyset$, then $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$.

Example: $\mathrm{P}(1 \text{ or } 2) = \mathrm{P}(1) + \mathrm{P}(2) = 1/6 + 1/6 = 1/3$.

Remark: After noting that $\Omega = A \cup \bar{A}$, this axiom immediately implies that the complementary probability $\mathrm{P}(\bar{A}) = 1 - \mathrm{P}(A)$.

Example: The probability that it'll rain tomorrow is 1 minus the probability that it won't rain.

Remark: In particular, the probability of *no* outcome from Ω is (intuitively) $\mathrm{P}(\emptyset) = 1 - \mathrm{P}(\Omega) = 0$.

Remark/Example: But the converse is *false*: $\mathrm{P}(A) = 0$ does *not* imply $A = \emptyset$. As a counterexample, pick a random number between 0 and 1. Later on, we'll show why any particular outcome actually has probability 0! ※

(4) Suppose $A_1, A_2, \ldots$ is a sequence of *disjoint* events (i.e., $A_i \cap A_j = \emptyset$, for $i \neq j$). Then

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(A_i).$$

Example: Toss a fair die until a 3 appears for the first time. We define the *disjoint* events $A_i =$ "3 appears for the first time on toss i", for $i = 1, 2, \ldots$, i.e.,

$$A_1 = \{3\},\ A_2 = \{\bar{3}3\},\ A_3 = \{\bar{3}\bar{3}3\}, \ldots.$$

Then Axioms (2) and (4) imply that

$$1 = \mathrm{P}(\Omega) = \mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(A_i).$$

Later on, we'll learn that $\mathrm{P}(A_i) = 5^{i-1}/6^i$, $i = 1, 2, \ldots$. □

Remark: The finite version for $A_1, A_2, \ldots, A_n$ follows as a special case.

The axioms will allow us to build all of the technology we need in order to discuss probability theory.

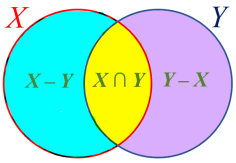
1.3.4 Some Examples Involving Unions

The next results concern the probability of the union of several events.

Theorem: For *any* two events A and B (not necessarily disjoint),

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B).$$

Proof: This follows by an easy Venn diagram argument. (Subtract $\mathrm{P}(A \cap B)$ to avoid double-counting.)



Remark: Axiom (3) is a "special case" of this theorem, with $A \cap B = \emptyset$.

Example: Suppose there is a...

40% chance of colder weather
10% chance of rain *and* colder weather
80% chance of rain *or* colder weather.

Then the chance of rain is

$$\mathrm{P}(R) \;=\; \mathrm{P}(R \cup C) - \mathrm{P}(C) + \mathrm{P}(R \cap C) \;=\; 0.8 - 0.4 + 0.1 \;=\; 0.5. \quad \square$$

Theorem: For any three events A, B, and C,

$$\begin{aligned} \mathrm{P}(A \cup B \cup C) \;=\;& \mathrm{P}(A) + \mathrm{P}(B) + \mathrm{P}(C) \\ & -\mathrm{P}(A \cap B) - \mathrm{P}(A \cap C) - \mathrm{P}(B \cap C) \\ & +\mathrm{P}(A \cap B \cap C). \end{aligned}$$

Proof: You can try an informal proof via Venn diagrams, but be careful about double- and triple-counting events. Here's a proof that builds on the previous theorem.

$$\begin{aligned} \mathrm{P}\big((A \cup B) \cup C\big) \;&=\; \mathrm{P}(A \cup B) + \mathrm{P}(C) - \mathrm{P}\big((A \cup B) \cap C\big) \\ &=\; \mathrm{P}(A \cup B) + \mathrm{P}(C) - \mathrm{P}\big((A \cap C) \cup (B \cap C)\big) \\ &=\; \mathrm{P}(A \cup B) + \mathrm{P}(C) \\ &\qquad -\big[\mathrm{P}(A \cap C) + \mathrm{P}(B \cap C) - \mathrm{P}\big((A \cap C) \cap (B \cap C)\big)\big] \\ &=\; \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B) + \mathrm{P}(C) \\ &\qquad -\mathrm{P}(A \cap C) - \mathrm{P}(B \cap C) + \mathrm{P}(A \cap B \cap C). \quad \square \end{aligned}$$

Example: In a certain population of music aficionados, 60% love The **B**eatles, 50% like The Rolling **S**tones, 40% adore The **Z**ombies, 40% enjoy The Beatles *and* The Stones, 30% The Beatles *and* The Zoms, 35% The Stones *and* The Zoms, and 30% *all three*. What's the probability that a random person isn't a fan of any of the three groups?

Solution: First, we'll calculate the probability that a random person likes at least one of the groups,

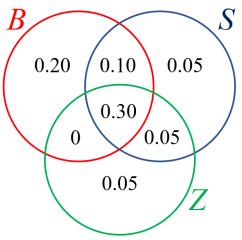
$$\begin{aligned} \mathrm{P}(B \cup S \cup Z) \;&=\; \mathrm{P}(B) + \mathrm{P}(S) + \mathrm{P}(Z) - \mathrm{P}(B \cap S) - \mathrm{P}(B \cap Z) - \mathrm{P}(S \cap Z) \\ &\qquad +\mathrm{P}(B \cap S \cap Z) \\ &=\; 0.60 + 0.50 + 0.40 - 0.40 - 0.30 - 0.35 + 0.30 \;=\; 0.75. \end{aligned}$$

Then the desired probability that the person doesn't like any group is

$$\mathrm{P}\big(\overline{B \cup S \cup Z}\big) \;=\; 1 - \mathrm{P}(B \cup S \cup Z) \;=\; 0.25. \quad \square$$

Now find the probability that a random person will like precisely two of the three groups.

To solve this problem, we can use a Venn diagram argument, starting from the center (since $\mathrm{P}(B \cap S \cap Z) = 0.30$) and building out.



Then we have

$$\begin{aligned} &\mathrm{P}(\text{only } B \text{ and } S) + \mathrm{P}(\text{only } B \text{ and } Z) + \mathrm{P}(\text{only } S \text{ and } Z) \\ &\quad = \mathrm{P}(B \cap S \cap \bar{Z}) + \mathrm{P}(B \cap \bar{S} \cap Z) + \mathrm{P}(\bar{B} \cap S \cap Z) \\ &\quad = 0.10 + 0 + 0.05 \;=\; 0.15. \quad \square \end{aligned}$$

Here's the general result.

Theorem: The **Principle of Inclusion-Exclusion** for the union of n events:

$$\begin{aligned} &\mathrm{P}(A_1 \cup A_2 \cup \cdots \cup A_n) \\ &\quad = \sum_{i=1}^{n} \mathrm{P}(A_i) - \sum\sum_{i<j} \mathrm{P}(A_i \cap A_j) + \sum\sum\sum_{i<j<k} \mathrm{P}(A_i \cap A_j \cap A_k) \\ &\qquad - \cdots + (-1)^{n-1} \mathrm{P}(A_1 \cap A_2 \cap \cdots \cap A_n). \end{aligned} \tag{1.2}$$

Remark: You "include" all of the 'single" events, "exclude" the double events, include the triple events, etc. Obviously, the previous two theorems are special cases.

Honors Proof

We'll proceed by induction. We define $P_i \equiv \mathrm{P}(A_i)$, $P_{ij} \equiv \mathrm{P}(A_i \cap A_j)$, $P_{ijk} \equiv \mathrm{P}(A_i \cap A_j \cap A_k)$, etc., for all $i, j, k, \ldots$.

Also, let $B_i \equiv A_i \cap A_{n+1}$, for all i, so that $\mathrm{P}(B_i) = P_{i,n+1}$, $\mathrm{P}(B_i \cap B_j) = P_{ij,n+1}$, $\mathrm{P}(B_i \cap B_j \cap B_k) = P_{ijk,n+1}$, etc.

And let $\mathcal{A}_n \equiv \bigcup_{i=1}^{n} A_i$ and $\mathcal{B}_n \equiv \bigcup_{i=1}^{n} B_i$, for $n \geq 1$.

Finally, we can get going! The induction's *base case* is trivial (just $\mathrm{P}(A_1) = \mathrm{P}(A_1)$), and then the *induction step* assumes

$$\mathrm{P}(\mathcal{A}_n) = \sum_{i=1}^{n} P_i - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P_{ij} + \sum_{i=1}^{n-2} \sum_{j=i+1}^{n-1} \sum_{k=j+1}^{n} P_{ijk} - \cdots + (-1)^{n-1} P_{12\cdots n}.$$

The induction will be complete when we establish the result for the "$n+1$"

case. To this end,

$$\begin{aligned} \mathrm{P}(\mathcal{A}_{n+1}) &= \mathrm{P}\big(\mathcal{A}_n \cup A_{n+1}\big) \\ &= \mathrm{P}(A_{n+1}) + \mathrm{P}(\mathcal{A}_n) - \mathrm{P}\big(\mathcal{A}_n \cap A_{n+1}\big) \\ &= P_{n+1} + \mathrm{P}(\mathcal{A}_n) - \mathrm{P}(\mathcal{B}_n). \end{aligned}$$

By the induction step for $\mathrm{P}(\mathcal{A}_n)$ and then again for $\mathrm{P}(\mathcal{B}_n)$, we have

$$\begin{aligned} &\mathrm{P}(\mathcal{A}_{n+1}) = \\ &\quad \left[\sum_{i=1}^{n+1} P_i - \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} P_{ij} + \sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}\sum_{k=j+1}^{n} P_{ijk} - \cdots + (-1)^{n-1} P_{12\cdots n}\right] \\ &\quad - \left[\sum_{i=1}^{n} P_{i,n+1} - \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} P_{ij,n+1} - \cdots + (-1)^{n-1} P_{12\cdots n+1}\right]. \end{aligned}$$

After the smoke clears from the algebra, we obtain

$$\mathrm{P}(\mathcal{A}_{n+1}) = \sum_{i=1}^{n+1} P_i - \sum_{i=1}^{n}\sum_{j=i+1}^{n+1} P_{ij} + \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=j+1}^{n+1} P_{ijk} - \cdots + (-1)^{n} P_{12\cdots n+1}.$$

Whew! □

1.4 Finite Sample Spaces

Suppose a sample space Ω is **finite**, say $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$. Finite sample spaces often allow us to calculate the probabilities of certain events more efficiently.

To illustrate, let $A \subseteq \Omega$ be any event. Then $\mathrm{P}(A) = \sum_{\omega_i \in A} \mathrm{P}(\omega_i)$.

Example: You have two red cards, one blue card, and one yellow card. Pick one card at random, where "**at random**" means that each of the four cards has the same probability (1/4) of being chosen.

The sample space $\Omega = \{\text{red, blue, yellow}\} = \{\omega_1, \omega_2, \omega_3\}$.

$\mathrm{P}(\omega_1) = 1/2$, $\mathrm{P}(\omega_2) = 1/4$, $\mathrm{P}(\omega_3) = 1/4$.

Then, $\mathrm{P}(\text{red or yellow}) = \mathrm{P}(\omega_1) + \mathrm{P}(\omega_3) = 3/4$. □

Definition: A **simple sample space** (SSS) is a finite sample space in which all outcomes are *equally likely*. In fact, for a SSS, all outcomes have probability $1/|\Omega|$.

Remark: In the above example, Ω is *not* simple, since $\mathrm{P}(\omega_1) \neq \mathrm{P}(\omega_2)$.

Example: Toss two fair coins.

$\Omega = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$ is a SSS (all probabilities are 1/4).

$\Omega' = \{0, 1, 2\}$ (number of H's) is *not* a SSS. Why not? □

The next theorem establishes the ease of calculating certain probabilities when we have a SSS.

Theorem: For any event A in a SSS Ω,

$$\mathrm{P}(A) \;=\; \frac{|A|}{|\Omega|} \;=\; \frac{\#\text{ of elements in } A}{\#\text{ of elements in } \Omega}.$$

Example: Toss a die. Let $A = \{1, 2, 4, 6\}$. Each outcome has probability 1/6, so $\mathrm{P}(A) = 4/6$. □

Example: Let's now toss a pair of dice. Here are the possible results (each ordered pair with probability 1/36):

$$\begin{array}{cccc} (1,1) & (1,2) & \cdots & (1,6) \\ (2,1) & (2,2) & \cdots & (2,6) \\ & \vdots & & \\ (6,1) & (6,2) & \cdots & (6,6) \end{array}$$

From this SSS, we easily obtain the following results on the sum of the two tosses.

Sum of tosses	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

For instance, $\mathrm{P}(\text{Sum} = 4) = \mathrm{P}((1,3)) + \mathrm{P}((2,2)) + \mathrm{P}((3,1)) = 3/36$. □

With this material in mind, we can now move on to more-complicated counting problems.

1.5 Counting Techniques

Idea: Count the elements in events from a SSS in order to calculate certain probabilities efficiently. We'll look at various helpful rules/techniques, including (i) some intuitive baby examples, (ii) **permutations**, and (iii) **combinations.**

1.5.1 Baby Examples

We present a few examples that are intuitively obvious — just to illustrate some simple counting rules and the resulting probabilities.

Example: Suppose that you can make choice A in n_{A} ways, and choice B in n_{B} ways. If only one choice can be made, you have $n_{\mathrm{A}} + n_{\mathrm{B}}$ possible ways of doing so. For instance, you go to Starbucks and decide to have either a muffin (blueberry or oatmeal) or a bagel (sesame, plain, salt, garlic), but not both. You have $2 + 4 = 6$ choices in total. □

Example: Suppose there are n_{AB} ways to go from City A to City B, and n_{BC} ways to go from City B to City C. Then you can go from A to C (via B) in $n_{\text{AB}} \times n_{\text{BC}}$ ways. For instance, if you can walk, bike, or drive from A to B, and if you can drive, fly, take a boat, or take a train from B to C, then you have 12 possible itineraries from A to B to C. □

Example: Roll two dice. How many outcomes? (Assume $(3,2) \neq (2,3)$.) Answer is $6 \times 6 = 36$. □

Example: Toss n dice. There are 6^n possible outcomes. □

Example: Flip three coins. $2 \times 2 \times 2 = 8$ possible outcomes,

$$\big\{\text{HHH}, \text{HHT}, \text{HTH}, \text{HTT}, \text{THH}, \text{THT}, \text{TTH}, \text{TTT}\big\}.$$

What's the probability that the third coin is a head? Answer (by looking at the list) $= 4/8$. □

Remark: In the previous examples, we implicitly specified that *selection order* was significant (e.g., $(3,2) \neq (2,3)$). In addition, when we are selecting items randomly from a finite set, it is important to specify whether we are *selecting with or without replacement*. Selecting *with replacement* means that the selected element is returned to the set before the next selection, and so the same element could be chosen more than once. Selecting *without replacement* means that each selection removes one element from the set, and so each item can be selected at most one time.

Example: Select two cards from a deck ***without replacement*** and ***care about order*** (i.e., $(Q\spadesuit, 7\clubsuit) \neq (7\clubsuit, Q\spadesuit)$). How many ways can we do this? It's easy to see that the answer is

$$(\#\text{ of ways to choose } 1^{\text{st}} \text{ card}) \cdot (\#\text{ of ways to choose } 2^{\text{nd}} \text{ (after } 1^{\text{st}})),$$

which is simply $52 \cdot 51 = 2652$. □

Example: Given a box containing 10 socks — two red and eight black — pick two without replacement.

- What is the probability that both of the selected socks are red? Let A be the event that both are red. Then
$$\text{P}(A) \;=\; \frac{\#\text{ of ways to pick two reds}}{\#\text{ of ways to pick two socks}} \;=\; \frac{2 \cdot 1}{10 \cdot 9} \;=\; \frac{1}{45}. \quad \square$$
- What is the probability that both socks are black? Let B be the event that both are black. Then $\text{P}(B) = \frac{8 \cdot 7}{10 \cdot 9} = \frac{28}{45}$. □
- What is the probability that one sock is red and the other is black? Let C be the event that one is of each color. Since A and B are disjoint, we have
$$\text{P}(C) \;=\; 1 - \text{P}(\bar{C}) \;=\; 1 - \text{P}(A \cup B) \;=\; 1 - \text{P}(A) - \text{P}(B) \;=\; \frac{16}{45}. \quad \square$$

1.5.2 Permutations

Now the fun begins. Suppose we want to count the number of ways to choose r out of n objects ***with regard to order.*** Let's start with a special case.

Definition: An arrangement of n different symbols in a ***definite order*** is a **permutation** of the n symbols.

Example: How many ways can the numbers $\{1, 2, 3\}$ be arranged?

Answer: There are $3! = 6$ ways. This is easy to verify by just listing all the possibilities: 123, 132, 213, 231, 312, 321. □

In most practical examples, the number of elements involved is too large to make complete enumeration a viable strategy, so we need a more-general method.

Example: How many different ways can the numbers $\{1, 2, \ldots, n\}$ be arranged? Equivalently, how many words can you spell from n different letters, using each exactly once? Or how ways can you schedule a series of n jobs? Or visit an itinerary of n cities? Answer: A little reflection shows that this is just an example of sequential sampling without replacement:

$$\begin{gathered}(\text{visit } 1^{\text{st}} \text{ city})\,(\text{then visit } 2^{\text{nd}} \text{ city}) \cdots (\text{then visit last city}) \\ = n(n-1)(n-2)\cdots 2 \cdot 1 \;=\; n!.\end{gathered}$$

Thus, if you want to visit 9 cities, you can do them in $9! = 362880$ orders. □

Definition: An **r-tuple** is an ordered list of cardinality r (i.e., it has r elements). For instance, $(5, 6, 1, 3)$ is a 4-tuple.

Definition/Theorem: The number of r-tuples we can make from n different symbols (*each used at most once*) is called the **number of permutations of n things taken r-at-a-time**, and

$$P_{n,r} \;\equiv\; \frac{n!}{(n-r)!}.$$

(We treated the special case of $r = n$ above, resulting in $P_{n,n} = n!/0! = n!$.)

Proof: The number of words of size r that you can make from n distinct letters (using each at most once) is

$$\begin{aligned} P_{n,r} &= (\text{choose } 1^{\text{st}} \text{ letter})(\text{choose } 2^{\text{nd}}) \cdots (\text{choose } r^{\text{th}}) \\ &= n(n-1)(n-2)\cdots(n-r+1) \\ &= \frac{n(n-1)\cdots(n-r+1)(n-r)\cdots 2 \cdot 1}{(n-r)\cdots 2 \cdot 1} \\ &= \frac{n!}{(n-r)!}. \quad \square \end{aligned}$$

Example: How many two-digit numbers can you make from $\{1, 2, 3, 4, 5\}$?

Answer: $P_{5,2} = 5!/(5-2)! = 20$. Let's list 'em:

12	13	14	15	21	23	24	25	31	32
34	35	41	42	43	45	51	52	53	54

How many three-digit numbers? Answer: $P_{5,3} = 5!/2! = 60$. (Feel free to list them if you really have nothing better to do.) □

Example: There are eight candidates running for three positions (President, VP, and Treasurer) of the Gilligan's Island Appreciation Society (GIAS). How many ways can those executive positions be selected? Answer: $P_{8,3} = 8!/(8-3)! = 336$ ways. □

Example: How many of these 336 ways have Jones as President?

Method 1: The first three positions look like: (Jones,?,?). This is equivalent to choosing two candidates from the remaining seven. So $P_{7,2} = 7!/(7-2)! = 42$ ways.

Method 2: Another way to solve the problem is simply by recognizing that each of the eight candidates is "equally likely" to be President. Thus, the number of ways that Pres. Jones is equal to $336/8 = 42$. □

Example: How many six-digit license plates can be made from the numbers $\{1, 2, \ldots, 9\}$...

- With no repetitions? (For instance, 123465 is okay, but 133354 isn't okay.) $P_{9,6} = 9!/3! = 60480$.
- Allowing repetitions? (Anything's okay.) $9 \times 9 \times \cdots \times 9 = 9^6 = 531441$.
- Containing repetitions? $531441 - 60480 = 470961$. □

1.5.3 Combinations

We might also wish to count the number of ways to choose r out of n objects ***without regard to order.*** In this case we ask: How many different subsets of size r can be chosen from a set of n objects? (Recall that the order of the members of a set has no significance.)

Example: How many subsets of $\{1, 2, 3\}$ contain exactly two elements? (Order isn't important.)

Answer: Three subsets — $\{1, 2\}$, $\{1, 3\}$, $\{2, 3\}$.

Definition: The number of subsets with r elements out of a set with n elements is called the **number of combinations of n things taken r-at-a-time**.

Notation: $\binom{n}{r}$ or $C_{n,r}$ (read as "n choose r"). These are also called **binomial coefficients**, and have already made an appearance in the Binomial Theorem, Equation (1.1). We will see below that $C_{n,r} = \frac{n!}{r!(n-r)!}$.

Differences between permutations and combinations:

- Combinations are not concerned with the order of the elements, i.e., $\{a, b, c\} = \{b, a, c\}$.
- Permutations *are* concerned with the order of the elements, i.e., $(a, b, c) \neq (b, a, c)$.
- The number of permutations of n things taken r-at-a-time is always as least as large as the number of combinations. In fact, ...

Choosing a permutation is the same as first choosing a combination of r objects out of n and *then* putting the r elements in order, i.e.,

$$\frac{n!}{(n-r)!} = \binom{n}{r} r!.$$

So,

$$C_{n,r} \equiv \binom{n}{r} \equiv \frac{n!}{(n-r)!r!}.$$

The following results should all be intuitive:

$$\binom{n}{r} = \binom{n}{n-r}, \quad \binom{n}{0} = \binom{n}{n} = 1, \quad \binom{n}{1} = \binom{n}{n-1} = n.$$

In fact, you may recognize these binomial coefficients in the context of **Pascal's Triangle**, which you may have seen way back in grade school.

n	$\binom{n}{r},\ r = 0, 1, \ldots, n$
0	1
1	1 1
2	1 2 1
3	1 3 3 1
4	1 4 6 4 1
5	1 5 10 10 5 1
6	1 6 15 20 15 6 1

Example: An office volleyball team has 14 members. How many ways can the coach choose the starting six players? (Order doesn't matter.)

$$\binom{14}{6} = \frac{14!}{6!\,8!} = 3003. \quad \square$$

Example: Smith is one of the players on the team. How many of the 3003 starting line-ups include her?

$$\binom{13}{5} = \frac{13!}{5!\,8!} = 1287$$

(Smith gets one of the six positions for free. There are now five left to be filled by the remaining 13 players.) □

Example: Suppose that you are the proud owner of seven identical red shoes and five identical blue shoes. What is the total number of possible arrangements of your collection? Here is one such arrangement:

$$\text{R} \quad \text{B} \quad \text{R} \quad \text{R} \quad \text{B} \quad \text{B} \quad \text{R} \quad \text{R} \quad \text{R} \quad \text{B} \quad \text{R} \quad \text{B}.$$

Answer: You can think of this as the number of ways to put seven red shoes into 12 slots (or, equivalently, the number of ways to put five blue shoes into 12 slots), which is simply equal to $\binom{12}{7}$. □

1.6 Counting Applications

In this section we'll discuss a potpourri of applications of counting techniques. Here's what's coming up.

1.6.1 Hypergeometric Distribution

Suppose you have a collection consisting of a objects of type 1 and b objects of type 2. From this collection of $a+b$ objects, you will select a total of n objects ***without replacement***. What is the probability that k of the objects selected will be of type 1, where $k = \max\{0, n-b\}, \ldots, \min\{a, n\}$.

We have

$$\begin{aligned} &\text{P}(k \text{ type 1's are picked}) \\ &= \frac{(\#\text{ ways to choose } k \text{ type 1's from } a)(\text{choose } n-k \text{ type 2's from } b)}{\#\text{ ways to choose } n \text{ out of } a+b} \\ &= \frac{\binom{a}{k}\binom{b}{n-k}}{\binom{a+b}{n}}. \quad \square \end{aligned}$$

The number of type 1's chosen is said to have the **hypergeometric distribution**. We'll have a very thorough discussion on "distributions" later.

Example: Suppose that your sock drawer contains 15 red socks and 10 blue socks, for a total of 25 socks. Randomly pick seven, without replacement. What is the probability that you pick three reds and four blues? Answer:

$$\text{P(exactly three reds are picked)} \;=\; \frac{\binom{15}{3}\binom{10}{4}}{\binom{25}{7}} \;=\; 0.1988. \quad \square$$

Example: Starting from the same sock drawer (containing its full contingent of 25 socks), suppose that immediately upon awakening (and before you have had your morning coffee), you blindly pick two socks without replacement. What is your probability of picking a matching pair? Answer:

$$\begin{aligned}
&\text{P(matching pair)} \\
&\quad = \quad \text{P(exactly two reds are picked)} + \text{P(exactly two blues are picked)} \\
&\quad = \quad \frac{\binom{15}{2}\binom{10}{0}}{\binom{25}{2}} + \frac{\binom{15}{0}\binom{10}{2}}{\binom{25}{2}} \;=\; \frac{\binom{15}{2}+\binom{10}{2}}{\binom{25}{2}} \;=\; 0.5. \quad \square
\end{aligned}$$

1.6.2 Binomial Distribution

Let's begin with the same collection of objects as in §1.6.1, consisting of a objects of type 1 and b objects of type 2. But now select n objects ***with replacement*** from the $a + b$. Then,

$$\begin{aligned}
&\text{P}(k \text{ type 1's are picked}) \\
&\quad = \quad (\#\text{ of ways to choose } k \text{ type 1's and } n-k \text{ type 2's}) \\
&\qquad\quad \times \text{P(a particular selection of } k \text{ type 1's and } n-k \text{ type 2's)} \\
&\quad = \quad \binom{n}{k}\text{P(choose } k \text{ type 1's in a row, then } n-k \text{ type 2's in a row)} \\
&\quad = \quad \binom{n}{k}\Big(\frac{a}{a+b}\Big)^k\Big(\frac{b}{a+b}\Big)^{n-k}, \quad k = 0, 1, \ldots, n. \quad \square
\end{aligned}$$

The number of type 1's chosen is said to have the **binomial distribution**.

Example: 25 socks in a box, with 15 red and 10 blue. Pick seven with replacement.

$$\text{P(exactly three reds are picked)} \;=\; \binom{7}{3}\Big(\frac{15}{25}\Big)^3\Big(\frac{10}{25}\Big)^{7-3} \;=\; 0.1936. \quad \square$$

Be sure to compare this answer with that of the analogous hypergeometric example. We'll discuss the binomial distribution in great detail in upcoming chapters.

1.6.3 Multinomial Coefficients

In this subsection, we'll expand a bit on the concepts of the hypergeometric and binomial distributions by adding more categories. Previously, we had on hand a objects of type 1 and b objects of type 2, from which we sampled n objects. Here

we consider the more-general case in which we have a_i objects of type i, where $i = 1, 2, \ldots, c$, so that we now have c categories instead of just two. Let $A = \sum_{i=1}^{c} a_i$ denote the total number of objects of all types, and suppose we take a sample of n of them. We are interested in obtaining the probability of selecting k_i objects of types $i = 1, 2, \ldots, c$, where $n = \sum_{i=1}^{c} k_i$, and where the k_i's are all "legal," e.g., $k_i \leq a_i$.

§1.6.3.1 Sampling without Replacement

In the case of sampling *without replacement*, a straightforward generalization of the hypergeometric distribution reveals that

$$\mathrm{P}\big(\text{Select } k_i \text{ type } i\text{'s, } i = 1, 2, \ldots, c\big) \;=\; \frac{\binom{a_1}{k_1}\binom{a_2}{k_2}\cdots\binom{a_c}{k_c}}{\binom{A}{n}} \;=\; \frac{\prod_{i=1}^{c}\binom{a_i}{k_i}}{\binom{A}{n}}.$$

Example: Suppose that your sock drawer contains 15 red socks, 10 blue socks, and 12 green socks, for a total of 37 socks. Randomly pick nine, without replacement. What is the probability that you pick three reds, four blues, and two greens? Answer:

$$\mathrm{P}(3 \text{ R's, } 4 \text{ B's, } 2 \text{ G's}) \;=\; \frac{\binom{15}{3}\binom{10}{4}\binom{12}{2}}{\binom{37}{9}} \;=\; 0.0507. \quad \square$$

§1.6.3.2 Sampling with Replacement

In the case of sampling *with replacement*, a similarly straightforward generalization of the binomial distribution yields the **multinomial distribution**,

$$\mathrm{P}\big(\text{Select } k_i \text{ type } i\text{'s, } i = 1, 2, \ldots, c\big) \;=\; \binom{n}{k_1, k_2, \ldots, k_c} \prod_{i=1}^{c} (a_i/A)^{k_i},$$

where

$$\binom{n}{k_1, k_2, \ldots, k_c} \;\equiv\; \frac{n!}{k_1!\, k_2! \cdots k_c!}$$

is what is known as a **multinomial coefficient**, which represents the number of ways you can choose the k_i objects of types $i = 1, 2, \ldots, c$ out of the total of n objects. Of course, the binomial coefficient corresponds to the case $c = 2$.

Example: How many different 11-digit numbers can be formed from the digits 1,2,2,3,3,3,3,4,4,4,4? This is simply a multinomial coefficient,

$$\frac{\#\text{ of permutations of 11 digits}}{(\#\text{ 1's})!\,(\#\text{ 2's})!\,(\#\text{ 3's})!\,(\#\text{ 3's})!} \;=\; \frac{11!}{1!\,2!\,4!\,4!} \;=\; 34{,}650. \quad \square$$

Example: Again, consider that sock drawer with 15 red socks, 10 blue socks, and 12 green socks, for a total of 37 socks. Randomly pick nine, with replacement. Now what's the probability that you pick three reds, four blues, and two greens? Answer:

$$\mathrm{P}(3 \text{ R's, } 4 \text{ B's, } 2 \text{ G's}) \;=\; \binom{9}{3, 4, 2}(15/37)^3(10/37)^4(12/37)^2 \;=\; 0.0471. \quad \square$$

1.6.4 Permutations vs. Combinations

By now you may be a bit confused about when to think in terms of permutations and when to think in terms of combinations. Actually, there are often a number of methods that you can use to solve a particular problem. The following is an example of a problem that can be solved by approaching it in terms of either permutations or combinations. Both methods will give you the correct answer, but (in this case, anyhow) one approach ends up being much simpler.

Example: Suppose you have four red marbles and two white marbles. For some reason (known only to yourself), you want to randomly line them up in a row. Find:

(a) P(The two end marbles are both W).

(b) P(The two end marbles aren't both W).

(c) P(The two W's are side-by-side).

Method 1 (*using permutations*): Let the sample space

$$\Omega \;=\; \{\text{every random ordering of the six marbles}\}.$$

(a) Define event A: The two end marbles are W, i.e., WRRRRW. Note that

$$\begin{aligned} |A| \;&=\; (\#\text{ of ways to permute the 2 W's in the end slots}) \\ &\qquad \times (\#\text{ of ways to permute the 4 R's in the middle slots}) \\ &=\; 2! \times 4! \;=\; 48. \end{aligned}$$

This implies that

$$\mathrm{P}(A) \;=\; \frac{|A|}{|\Omega|} \;=\; \frac{48}{720} \;=\; \frac{1}{15}. \quad \square$$

(b) The event "the end marbles are not both white" is just the complement of A. Thus, $\mathrm{P}(\bar{A}) = 1 - \mathrm{P}(A) = 14/15$. $\square$

(c) Define event B: The two W's are side-by-side, i.e., WWRRRR or RWWRRR or $\cdots$ or RRRRWW. Then

$$\begin{aligned} |B| \;&=\; (\#\text{ of ways to select a pair of side-by-side slots for 2 W's}) \\ &\qquad \times (\#\text{ of ways to insert W's into a pair of slots}) \\ &\qquad \times (\#\text{ of ways to insert R's into the remaining slots}) \\ &=\; 5 \times 2! \times 4! \;=\; 240. \end{aligned}$$

Thus,

$$\mathrm{P}(B) \;=\; \frac{|B|}{|\Omega|} \;=\; \frac{240}{720} \;=\; \frac{1}{3}. \quad \square$$

But — the above method took too much time! Here's an easier way...

Method 2 (*using combinations*): In this approach, we begin by redefining the sample space as follows:

$$\Omega \;=\; \{\text{possible } \textit{pairs} \text{ of slots that the white marbles can occupy}\}.$$

Clearly, $|\Omega| = \binom{6}{2} = 15$.

(a) Since the W's must occupy the end slots in order for A to occur, then

$$|A| = 1 \Rightarrow \mathrm{P}(A) = |A|/|\Omega| = 1/15. \quad \square$$

(b) $\mathrm{P}(\bar{A}) = 14/15.$ $\square$

(c) $|B| = 5 \Rightarrow \mathrm{P}(B) = 5/15 = 1/3.$ (That was much nicer!) $\square$

1.6.5 The Birthday Problem

The "Birthday Problem" is a classic example traditionally presented in introductory probability classes. The results may surprise you!

Suppose there are n people in a room. Find the probability that at least two of them have the same birthday. To simplify the analysis, we ignore anyone born on February 29, and we assume that all 365 days have equal probability.

The (simple) sample space is

$$\Omega = \{(x_1, \ldots, x_n) : x_i = 1, 2, \ldots, 365, \forall i\},$$

where x_i is person i's birthday, and note that $|\Omega| = (365)^n$.

Let A be the event "all birthdays are different." Then the size of A is just the number of permutations of n different days, which is simply

$$|A| = P_{365,n} = (365)(364)\cdots(365 - n + 1).$$

Thus, we have

$$\begin{aligned} \mathrm{P}(A) &= \frac{(365)(364)\cdots(365-n+1)}{(365)^n} \\ &= 1 \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365-n+1}{365}. \end{aligned}$$

But we want

$$\mathrm{P}(\bar{A}) = 1 - \left(1 \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365-n+1}{365}\right).$$

Remarks:

- When $n = 366$, $\mathrm{P}(\bar{A}) = 1$.
- For $\mathrm{P}(\bar{A})$ to be $> 1/2$, n must be ≥ 23. In other words, a class of just 23 students has a better than even chance of having two students with the same birthday. (Surprising!)
- When $n = 50$, $\mathrm{P}(\bar{A}) = 0.97$.

1.6.6 The Envelope Problem

A group of n people receives n envelopes with their names on them — *but someone has completely mixed up the envelopes!* Find the probability that at least one person will receive the proper envelope. (There are lots of variations to this story that are mathematically equivalent.)

Let A_i signify that person i receives the correct envelope. Then we obviously want $\mathrm{P}(A_1 \cup A_2 \cup \cdots \cup A_n)$.

Recall the general Principle of Inclusion-Exclusion, Equation (1.2), which states that

$$\begin{aligned}&\mathrm{P}(A_1 \cup A_2 \cup \cdots \cup A_n)\\ &= \sum_{i=1}^{n} \mathrm{P}(A_i) - \sum\sum_{i<j} \mathrm{P}(A_i \cap A_j) + \sum\sum\sum_{i<j<k} \mathrm{P}(A_i \cap A_j \cap A_k)\\ &\qquad - \cdots + (-1)^{n-1} \mathrm{P}(A_1 \cap A_2 \cap \cdots \cap A_n).\end{aligned}$$

Since all of the $\mathrm{P}(A_i)$'s are the same, all of the $\mathrm{P}(A_i \cap A_j)$'s are the same, etc., we have

$$\begin{aligned}&\mathrm{P}(A_1 \cup A_2 \cup \cdots \cup A_n)\\ &= n\mathrm{P}(A_1) - \binom{n}{2}\mathrm{P}(A_1 \cap A_2) + \binom{n}{3}\mathrm{P}(A_1 \cap A_2 \cap A_3)\\ &\qquad - \cdots + (-1)^{n-1} \mathrm{P}(A_1 \cap A_2 \cap \cdots \cap A_n).\end{aligned}$$

Finally, $\mathrm{P}(A_1) = 1/n$, $\mathrm{P}(A_1 \cap A_2) = 1/(n(n-1))$, etc. (why?), implying that

$$\begin{aligned}&\mathrm{P}(A_1 \cup A_2 \cup \cdots \cup A_n)\\ &= \frac{n}{n} - \binom{n}{2}\frac{1}{n} \cdot \frac{1}{n-1} + \binom{n}{3}\frac{1}{n} \cdot \frac{1}{n-1} \cdot \frac{1}{n-2} - \cdots + (-1)^{n-1}\frac{1}{n!}\\ &= 1 - \frac{1}{2!} + \frac{1}{3!} - \cdots + (-1)^{n-1}\frac{1}{n!} \doteq 1 - \frac{1}{e} \doteq 0.6321. \quad \square\end{aligned}$$

Example: If there are just $n = 4$ envelopes, then

$$\mathrm{P}(A_1 \cup A_2 \cup A_3 \cup A_4) = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} = 0.625,$$

which is right on the asymptotic money! $\square$

1.6.7 Poker Problems

Gambling problems were important motivators in the early development of probability theory. We'll look at several examples related to the game of poker.

First of all, we draw a "hand" of five cards at random from a standard deck of 52 cards. The sample space Ω is the set of all possible hands, so that the number of possible hands is $|\Omega| = \binom{52}{5} = 2{,}598{,}960$.

Here is some standard terminology (not that we're advocating gambling, but if you aren't familiar with poker, take a few minutes to learn the basics):

ranks $= 2, 3, \ldots, 10,$ Jack (11), Queen (12), King (13), Ace (14, sometimes 1)
suits $= \clubsuit$ (clubs), $\diamondsuit$ (diamonds), $\heartsuit$ (hearts), $\spadesuit$ (spades).

We'll now calculate the probabilities of obtaining various special hands.

(a) **One pair**, e.g., the hand $K\diamondsuit, A\clubsuit, K\spadesuit, 7\heartsuit, 6\diamondsuit$ sports precisely one pair of kings, but nothing better (such as two pairs or a three-of-a-kind, described below).

We can pick the rank of the pair (a K here) in 13 ways, and then the suits of the two kings (here $\diamondsuit$ and $\spadesuit$) in $\binom{4}{2} = 6$ ways.

We now have to deal with the remaining three cards (here $A\clubsuit, 7\heartsuit, 6\diamondsuit$). The restrictions are that no more K's are allowed, and that no other pairs (or a three-of-a-kind) are allowed. Thus, we must choose those cards from three different ranks (from the remaining 12 non-K's). Those three ranks can be chosen in $\binom{12}{3}$ ways, and then the suits of the corresponding three cards can be chosen in $4 \cdot 4 \cdot 4 = 4^3$ ways.

Putting all of this together, we get

$$\text{P(one pair)} = \frac{13\binom{4}{2}\binom{12}{3}4^3}{\binom{52}{5}} = \frac{1{,}098{,}240}{2{,}598{,}960} \doteq 0.4226. \quad \square$$

(b) **Two pairs**, e.g., $K\diamondsuit, 6\clubsuit, K\spadesuit, 7\heartsuit, 6\diamondsuit$. This hand has exactly two pairs (K's and 6's) and a junk card (but does *not* include a full house, which is coming up next.)

Since one pair is as good as the other, the order in which we pick the ranks of the pairs doesn't matter. Thus, we can pick the ranks of the two pairs (K and 6 here) in $\binom{13}{2}$ ways.

Then we select the suits of the two kings (here $\diamondsuit$ and $\spadesuit$) in $\binom{4}{2} = 6$ ways, and the suits of the two 6's (here $\clubsuit$ and $\diamondsuit$) in $\binom{4}{2} = 6$ ways.

The remaining card (here $7\heartsuit$) can be chosen from anything except for K's and 6's, which can happen in 44 ways.

Putting all of this together, we get

$$\text{P(two pairs)} = \frac{\binom{13}{2}\binom{4}{2}\binom{4}{2}44}{\binom{52}{5}} = \frac{123{,}552}{2{,}598{,}960} \doteq 0.0475. \quad \square$$

(c) **Full house** (one pair + one three-of-a-kind), e.g., $K\diamondsuit, 6\clubsuit, K\spadesuit, 6\heartsuit, 6\diamondsuit$. Unlike the previous example involving two pairs, we note that *order matters* when selecting the ranks of the pair and the triple (because a pair and a triple are different). In fact, this selection can be accomplished in $P_{13,2} = 13 \cdot 12$ ways.

Once we have the rank of the three-of-a-kind, we can select the corresponding suits in $\binom{4}{3}$ ways; and, similarly, $\binom{4}{2}$ ways for the pair. Then

$$\text{P(full house)} = \frac{P_{13,2}\binom{4}{3}\binom{4}{2}}{\binom{52}{5}} = \frac{3744}{2{,}598{,}960} \doteq 0.00144. \quad \square$$

(d) **Straight** (five ranks in a row), e.g., 3♢, 4♣, 5♠, 6♡, 7♢. We begin by selecting a starting point for the straight $(A, 2, 3, \ldots, 10$, where convention allows the Ace to pull double duty as a "1"); this can be done in 10 ways. And then we simply choose a suit for each card in the straight, which can be done in 4^5 ways. Thus,

$$\text{P(straight)} \;=\; \frac{10 \cdot 4^5}{2{,}598{,}960} \;\doteq\; 0.00394. \quad \square$$

(e) **Flush** (all five cards from same suit), e.g., 3♢, 7♢, 9♢, J♢, A♢. We start by selecting a suit for the flush (4 ways), and then five cards from that suit, which can be achieved in $\binom{13}{5}$ ways. Then

$$\text{P(flush)} \;=\; \frac{5148}{2{,}598{,}960} \;\doteq\; 0.00198. \quad \square$$

(f) **Straight flush**, e.g., 3♢, 4♢, 5♢, 6♢, 7♢. Select a starting point for the straight (10 ways), and then a suit (4 ways). This yields

$$\text{P(straight flush)} \;=\; \frac{40}{2{,}598{,}960} \;\doteq\; 0.0000154.$$

That's a really tiny probability! What The Flush?! $\square$

Remark: There are unlimited combinatorics problems involving games like poker. Can you do bridge problems? Yahtzee?

Example: **Yahtzee!** Toss five six-sided dice. What's the probability that you'll see a full house, e.g., 2,5,5,2,5? This is actually easier to analyze than the corresponding poker problem. Select the numbers that will correspond to the pair and triple (order matters), which can be done in $P_{6,2} = 6 \cdot 5$ ways. Then choose the positions in the group of five tosses that the pair occupies, which can be done in $\binom{5}{2}$ ways; and finish by choosing the remaining slots for the triple, which can be done in $\binom{3}{3}$ ways. Thus,

$$\text{P(Yahtzee full house)} \;=\; \frac{P_{6,2}\binom{5}{2}\binom{3}{3}}{6^5} \;=\; \frac{300}{7776} \;\doteq\; 0.03858. \quad \square$$

For more Yahtzee fun, see

`https://datagenetics.com/blog/january42012/index.html.`

1.7 Conditional Probability and Independence

Idea: We can often update the probability of an event as we obtain more information. Sometimes the probability will change, and sometimes not. In this section, we'll explore this issue and come up with lots of interesting findings.

1.7.1 Conditional Probability

Example: If A is the event that a person weighs at least 150 pounds, then $\mathrm{P}(A)$ is certainly related to the person's height, e.g., if B is the event that the person is at least 6 feet tall vs. B being the event that the person is < 5 feet tall. □

Example: For a standard die toss, define the events $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 4, 5\}$. Then $\mathrm{P}(A) = 1/2$ and $\mathrm{P}(B) = 5/6$.

Suppose we *know* that B occurs (so that there is no way that a "6" can come up). Then the probability that A occurs given that B occurs is

$$\mathrm{P}(A|B) \;=\; \frac{2}{5} \;=\; \frac{|A \cap B|}{|B|},$$

where the notation "$A|B$" can be read as the event "A given B." □

Thus, the probability of A depends on the information that you have! The information that B occurs allows us to regard B as a new, restricted sample space. So, assuming we have a *simple sample space* Ω, then

$$\mathrm{P}(A|B) \;=\; \frac{|A \cap B|}{|B|} \;=\; \frac{|A \cap B|/|\Omega|}{|B|/|\Omega|} \;=\; \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}.$$

Finally, here's the (more-general) definition of conditional probability that doesn't rely on the need for a SSS.

Definition: If $\mathrm{P}(B) > 0$, the **conditional probability of A given B** is

$$\mathrm{P}(A|B) \;\equiv\; \mathrm{P}(A \cap B)/\mathrm{P}(B).$$

Remarks: If A and B are disjoint, then $\mathrm{P}(A|B) = 0$. (If B occurs, there's no chance that A can also occur.)

What happens if $\mathrm{P}(B) = 0$? Don't worry! In this case, it makes no sense to consider $\mathrm{P}(A|B)$.

Example: Toss two dice and observe the sum. Define the events A: odd sum, i.e., $\{3, 5, 7, 9, 11\}$, and B: $\{2, 3\}$. Then

$$\begin{aligned}
\mathrm{P}(A) &= \mathrm{P}(3) + \mathrm{P}(5) + \cdots + \mathrm{P}(11) \;=\; \frac{2}{36} + \frac{4}{36} + \cdots + \frac{2}{36} \;=\; \frac{1}{2}, \\
\mathrm{P}(B) &= \mathrm{P}(2) + \mathrm{P}(3) \;=\; \frac{1}{36} + \frac{2}{36} \;=\; \frac{1}{12}, \quad \text{and} \\
\mathrm{P}(A|B) &= \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \;=\; \frac{\mathrm{P}(3)}{\mathrm{P}(B)} \;=\; \frac{2/36}{1/12} \;=\; \frac{2}{3}.
\end{aligned}$$

Thus, in light of the information provided by B, we see that $\mathrm{P}(A) = 1/2$ increases to $\mathrm{P}(A|B) = 2/3$. □

Example: Suppose your sock drawer is down to four white socks and eight red socks. (No doubt you have often found yourself in a not dissimilar situation.) Select two socks *without replacement*. Define the following events:

A: 1st sock is white.

B: 2nd sock is white.

C: Both socks are white ($= A \cap B$).

Let's find $\mathrm{P}(C)$ and $\mathrm{P}(B)$, the latter of which will turn out to be intuitively trivial (but we'll do it carefully anyway). First of all, by definition of conditional,

$$\mathrm{P}(C) \;=\; \mathrm{P}(A \cap B) \;=\; \mathrm{P}(A)\mathrm{P}(B|A) \;=\; \frac{4}{12} \cdot \frac{3}{11} \;=\; \frac{1}{11}.$$

Moreover, it is easy to see that $B = (A \cap B) \cup (\bar{A} \cap B)$, where the two components of the union are disjoint. So,

$$\begin{aligned} \mathrm{P}(B) &= \mathrm{P}(A \cap B) + \mathrm{P}(\bar{A} \cap B) \\ &= \mathrm{P}(A)\mathrm{P}(B|A) + \mathrm{P}(\bar{A})\mathrm{P}(B|\bar{A}) \\ &= \frac{4}{12} \cdot \frac{3}{11} + \frac{8}{12} \cdot \frac{4}{11} \;=\; \frac{1}{3}. \quad \square \end{aligned}$$

Could you have gotten this result without thinking about it? (Answer: Yes, and with a little practice, you'll get very good at it!)

Example: A couple has two children, at least one of whom is a boy. What is the probability that *both* are boys?[7]

The sample space and relevant events are:

- $\Omega = \{\mathrm{GG}, \mathrm{GB}, \mathrm{BG}, \mathrm{BB}\}$ (where, e.g., "BG" means "boy then girl").
- C: Both are boys $= \{\mathrm{BB}\}$.
- D: At least one is a boy $= \{\mathrm{GB}, \mathrm{BG}, \mathrm{BB}\}$.

We need to calculate

$$\mathrm{P}(C|D) \;=\; \frac{\mathrm{P}(C \cap D)}{\mathrm{P}(D)} \;=\; \frac{\mathrm{P}(\mathrm{BB})}{\mathrm{P}(\mathrm{GB}, \mathrm{BG}, \mathrm{BB})} \;=\; \frac{1/4}{3/4} \;=\; 1/3.$$

(My intuition from years ago when I was a Mathlete in high school was 1/2 — the *wrong* answer! The problem is that we don't know whether D means the first or second child.) $\square$

As you get more information, you can make some even more-surprising findings...

Honors Example

A couple has two kids and at least one is a boy ***born on a Tuesday.*** Assuming that any kid has a 1/14 chance of any gender / day-of-week combination, what's the probability that *both* of the couple's kids are boys?

[7] This is an interesting and tricky problem that has led to many lively discussions; intuition can sometimes fool you!

Let the events B_x [G_x] = Boy [Girl] born on day x, $x = 1, 2, \ldots, 7$ (e.g., $x = 3$ is Tuesday). A viable sample space comprising ordered pairs of the kids is:

$$\Omega \;=\; \{(\mathrm{G}_x, \mathrm{G}_y), (\mathrm{G}_x, \mathrm{B}_y), (\mathrm{B}_x, \mathrm{G}_y), (\mathrm{B}_x, \mathrm{B}_y),\; x, y = 1, 2, \ldots, 7\},$$

so that $|\Omega| = 4 \times 49 = 196$. Meanwhile, define the following events:

- C: Both are boys (with at least one born on a Tuesday)

  $$= \{(\mathrm{B}_x, \mathrm{B}_3),\; x = 1, 2, \ldots, 7\} \cup \{(\mathrm{B}_3, \mathrm{B}_y),\; y = 1, 2, \ldots, 7\}.$$

  Note that $|C| = 13$ (to avoid double counting $(\mathrm{B}_3, \mathrm{B}_3)$).

- D: There is at least one boy born on a Tuesday

  $$= C \cup \{(\mathrm{G}_x, \mathrm{B}_3), (\mathrm{B}_3, \mathrm{G}_y),\; x, y = 1, 2, \ldots, 7\}.$$

  So $|D| = 27$ (list 'em out if you don't believe me).

Then

$$\mathrm{P}(C|D) \;=\; \frac{\mathrm{P}(C \cap D)}{\mathrm{P}(D)} \;=\; \frac{\mathrm{P}(C)}{\mathrm{P}(D)} \;=\; \frac{13/196}{27/196} \;=\; \mathbf{13/27}.$$

Do you have any intuition as to why this answer is so much closer to 1/2 compared to the answer of 1/3 from the previous example? □

Properties of Conditional Probability — Conditional probabilities are simply probabilities of events that are on a restricted sample space. Thus, conditional probabilities have properties that are analogous to the Axioms of Probability from §1.3.3, which we list here for completeness.

(1′) $0 \le \mathrm{P}(A \,|\, C) \le 1$.

(2′) $\mathrm{P}(\Omega \,|\, C) = 1$.

(3′) If A and B are disjoint, then $\mathrm{P}(A \cup B \,|\, C) = \mathrm{P}(A \,|\, C) + \mathrm{P}(B \,|\, C)$.

(4′) If $A_1, A_2, \ldots$ are all disjoint, then $\mathrm{P}\big(\bigcup_{i=1}^{\infty} A_i \,\big|\, C\big) = \sum_{i=1}^{\infty} \mathrm{P}(A_i \,|\, C)$.

1.7.2 Independence

We now discuss the concept of independence,[8] and how it's related to conditional probability.

The short story is that any unrelated events are independent. For example,

A: It rains on Mars tomorrow and B: A coin lands on heads.

Definition: A and B are **independent** iff $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$.

[8] Remember this section on July 4th!

Example: If P(A) = 0.2, P(B) = 0.5, and P(A and B) = P(A)P(B) = 0.1, then A and B are independent. □

Remark: If P(A) = 0, then A is independent of any other event.

Remark: Events don't have to be physically unrelated to be independent.

Example: Toss a die, and define $A = \{2, 4, 6\}$ and $B = \{1, 2, 3, 4\}$. Then $A \cap B = \{2, 4\}$, so that P(A) = 1/2, P(B) = 2/3, and P($A \cap B$) = 1/3.
Then P(A)P(B) = 1/3 = P($A \cap B$) $\Rightarrow A, B$ independent. □

The next theorem provides a more-natural interpretation of independence.

Theorem: Suppose P(B) > 0. Then A and B are independent $\Leftrightarrow$ P($A|B$) = P(A).

Proof: By the definitions of independence and conditional probability,

$$\begin{aligned} A, B \text{ independent} &\Leftrightarrow \mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B) \\ &\Leftrightarrow \mathrm{P}(A \cap B)/\mathrm{P}(B) = \mathrm{P}(A) \\ &\Leftrightarrow \mathrm{P}(A|B) = \mathrm{P}(A). \quad \square \end{aligned}$$

Remark: The theorem says that if A and B are independent, then the probability of A does not depend on whether or not B occurs.

Important Remark: Don't confuse independence with disjointness! In fact, the following theorem suggests that independence and disjointness are almost opposites. Intuitively, if A and B are disjoint and A occurs, then you have *information* that B cannot occur — so A and B can't be independent!

Theorem: If P(A) and P(B) > 0, then A and B cannot be independent and disjoint at the same time.

Proof: Suppose A, B are disjoint (i.e., $A \cap B = \emptyset$). Then P($A \cap B$) = 0 < P(A)P(B). Thus, A, B are not independent. Similarly, independence implies that A and B are not disjoint. □

The following definition extends the concept of independence to more than two events.

Definition: A, B, C are independent $\Leftrightarrow$

(a) P($A \cap B \cap C$) = P(A)P(B)P(C) *and*

(b) All *pairs* are independent:

$$\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B),\ \mathrm{P}(A \cap C) = \mathrm{P}(A)\mathrm{P}(C),\ \text{and } \mathrm{P}(B \cap C) = \mathrm{P}(B)\mathrm{P}(C).$$

Remark: Careful! Note that condition (a) by itself isn't enough.

Example: Let $\Omega = \{1, 2, \ldots, 8\}$, where each element has probability 1/8. Define the events $A = \{1, 2, 3, 4\}$, $B = \{1, 5, 6, 7\}$, and $C = \{1, 2, 3, 8\}$.

(a) $A \cap B \cap C = \{1\}$. Thus, $\mathrm{P}(A \cap B \cap C) = \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C) = 1/8$, so (a) is satisfied. However, (b) is *not*...

(b) $A \cap B = \{1\}$. $\mathrm{P}(A \cap B) = 1/8 \neq 1/4 = \mathrm{P}(A)\mathrm{P}(B)$. So, A and B are not independent! ※

Stay vigilant, because, as the following example shows, (b) by itself isn't enough either!

Example: Let $\Omega = \{1, 2, 3, 4\}$ (each element w.p. 1/4). Define the events $A = \{1, 2\}$, $B = \{1, 3\}$, and $C = \{1, 4\}$.

(b) $\mathrm{P}(A \cap B) = 1/4 = \mathrm{P}(A)\mathrm{P}(B)$. Same deal with A, C and B, C. So (b) is okay. But (a) *isn't*...

(a) $\mathrm{P}(A \cap B \cap C) = 1/4 \neq 1/8 = \mathrm{P}(A)\mathrm{P}(B)\mathrm{P}(C)$. ※

General Definition: We can extend the concept of independence to more than two events. In fact, $A_1, A_2 \ldots, A_n$ are independent $\Leftrightarrow$

$$\mathrm{P}(A_1 \cap A_2 \cap \cdots \cap A_n) \;=\; \mathrm{P}(A_1)\mathrm{P}(A_2)\cdots\mathrm{P}(A_n)$$

and the A_i's comprising *any* subset of $\{A_1, A_2 \ldots, A_n\}$ are independent.

Definition: If n trials of an experiment are performed such that the outcome of one trial is independent of the outcomes of the other trials, then they are said to be **independent trials**.

Example: Flip three coins independently.

(a) $\mathrm{P}(1^{\text{st}}$ coin is H$) = 1/2$. Don't worry about the other two coins since they're independent of the 1^{st}.

(b) $\mathrm{P}(1^{\text{st}}$ coin H, 3^{rd} T$) = \mathrm{P}(1^{\text{st}}$ coin H$)\mathrm{P}(3^{\text{rd}}$ T$) = 1/4$. □

Remark: For independent trials, just multiply the individual probabilities.

Example: Flip a coin infinitely many times (each flip is independent of the others).

$$\begin{aligned} p_n &\equiv \mathrm{P}(1^{\text{st}} \text{ H on } n^{\text{th}} \text{ trial}) \\ &= \mathrm{P}(\underbrace{\mathrm{T}\,\mathrm{T}\cdots\mathrm{T}}_{n-1}\mathrm{H}) \\ &= \underbrace{\mathrm{P}(\mathrm{T})\mathrm{P}(\mathrm{T})\cdots\mathrm{P}(\mathrm{T})}_{n-1}\mathrm{P}(\mathrm{H}) \;=\; 1/2^n, \end{aligned}$$

so that

$$\mathrm{P}(\text{H eventually}) \;=\; \sum_{n=1}^{\infty} p_n \;=\; \sum_{n=1}^{\infty} 2^{-n} \;=\; 1. \quad \square$$

1.8 Bayes Theorem

Bayes Theorem allows us to adjust the probability of events in light of certain information that becomes available. It's particularly useful when we are able to divide the sample space into distinct pieces (i.e., a partition). The ensuing discussion partially follows the presentation in Meyer [5].

Definition: A **partition** of a sample space Ω splits the sample space into disjoint, all-encompassing subsets. Specifically, the events $A_1, A_2, \ldots, A_n$ form a partition of Ω if:

(a) $A_1, A_2, \ldots, A_n$ are disjoint.

(b) $\bigcup_{i=1}^{n} A_i = \Omega$.

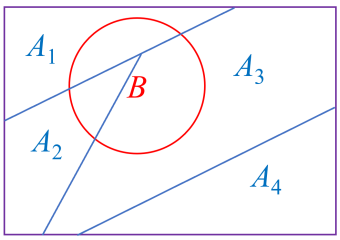
Remark: When an experiment is performed, *exactly one* of the A_i's occurs.

Example: A and $\bar{A}$ form a partition, for any event A

Example: "Vowels" and "consonants" form a partition of the letters (if you pretend that only a,e,i,o,u are vowels).

Remark: It's often convenient to choose all of the A_i's such that $\mathrm{P}(A_i) > 0$, but this is not actually a requirement.

Suppose that $A_1, A_2, \ldots, A_n$ form a partition of a sample space Ω, and B is some arbitrary event. Then (see the figure) $B = \bigcup_{i=1}^{n}(A_i \cap B)$.



So, if $A_1, A_2, \ldots, A_n$ is a partition, then we can decompose B into pieces of the A_i's,

$$\begin{aligned}
\mathrm{P}(B) &= \mathrm{P}\left(\bigcup_{i=1}^{n}(A_i \cap B)\right) \\
&= \sum_{i=1}^{n} \mathrm{P}(A_i \cap B) \quad \text{(since } A_1, A_2, \ldots, A_n \text{ are disjoint)} \\
&= \sum_{i=1}^{n} \mathrm{P}(A_i)\mathrm{P}(B|A_i) \quad \text{(definition of conditional probability).}
\end{aligned}$$

This is the **Law of Total Probability** — It's not only a good idea, it's The Law!

Example: $\mathrm{P}(B) = \mathrm{P}(A)\mathrm{P}(B|A) + \mathrm{P}(\bar{A})\mathrm{P}(B|\bar{A})$, which we saw in one of the examples from §1.7.

Example: Suppose we have 10 students from Clever University and 20 students from Gifted University taking a test. Clever students have a 95% chance of passing the test, but Gifted students somehow only have a 50% chance of passing. Pick a student at random, and determine the probability that he/she passes.

By the Law of Total Probability,

$$\begin{aligned} \mathrm{P}(\text{passes}) &= \mathrm{P}(\mathrm{C})\mathrm{P}(\text{passes} \,|\, \mathrm{C}) + \mathrm{P}(\mathrm{G})\mathrm{P}(\text{passes} \,|\, \mathrm{G}) \\ &= (1/3)(0.95) + (2/3)(0.5) \;=\; 0.65. \quad \square \end{aligned}$$

And here is an important immediate consequence of the Law of Total Probability, which we'll state and prove at the same time (because we can).

Bayes Theorem: If $A_1, A_2, \ldots, A_n$ form a partition of the sample space Ω, and B is any event, then

$$\mathrm{P}(A_j|B) \;=\; \frac{\mathrm{P}(A_j \cap B)}{\mathrm{P}(B)} \;=\; \frac{\mathrm{P}(A_j)\mathrm{P}(B|A_j)}{\sum_{i=1}^{n} \mathrm{P}(A_i)\mathrm{P}(B|A_i)}.$$

- The $\mathrm{P}(A_j)$'s are called **prior** probabilities ("before B").
- The $\mathrm{P}(A_j|B)$'s are called **posterior** probabilities ("after B").
- The $\mathrm{P}(A_j|B)$'s add up to 1, as you can see from the right-hand side of the equation. That's why we have the funny-looking denominator.
- The result — dating back to 1763 — is named after Rev. Thomas Bayes.

Example: There are two boxes of sockses. Box A contains one red sock and one blue sock. Box B contains two reds. Your friend picks a box randomly and then shows you a random sock from it, which turns out to be red. What's the probability that it's from Box A?

Let R denote the event that a red sock was drawn. The relevant partition is $\{A, B\}$.

$$\begin{aligned} \mathrm{P}(A|R) &= \frac{\mathrm{P}(A)\mathrm{P}(R|A)}{\mathrm{P}(A)\mathrm{P}(R|A) + \mathrm{P}(B)\mathrm{P}(R|B)} \\ &= \frac{(0.5)(0.5)}{(0.5)(0.5) + (0.5)(1.0)} \;=\; 1/3. \end{aligned}$$

Notice how the probability of Box A went from 0.5 (prior) to 1/3 (posterior) — it was certainly influenced by the information we received. $\square$

Example: Two political candidates are at a debate. Candidate A is asked 60% of the questions, and candidate B is asked just 40% (for some reason). A is likely to make a dumb answer 20% of the time, and B makes a dumb answer a whopping

50% of the time. One of the candidates is asked a question and makes a dumb answer. What's the probability that it was A?

(We'd expect the probability to be $< 60\%$, since dumb answers favor B.)

Let D denote the event that the person makes a dumb answer. The relevant partition is $\{A, B\}$. Then,

$$\begin{aligned} \mathrm{P}(A|D) &= \frac{\mathrm{P}(A)\mathrm{P}(D|A)}{\mathrm{P}(A)\mathrm{P}(D|A) + \mathrm{P}(B)\mathrm{P}(D|B)} \\ &= \frac{(0.6)(0.2)}{(0.6)(0.2) + (0.4)(0.5)} = 0.375. \quad \square \end{aligned}$$

Example: You are a contestant on the old TV show *Let's Make a Deal!* Behind one of three doors is a car; behind the other two are goats. You pick Door number 1. Monty Hall opens Door 2 and reveals a goat. Monty offers you a chance to switch to Door 3. What should you do? (Note: If the car had actually been behind your Door 1, Monty would've randomly chosen to show you Door 2 or 3.)

By Bayes, we have

$$\begin{aligned} &\mathrm{P}(\text{Car behind Door 1} \,|\, \text{Monty shows Door 2}) \\ &= \frac{\mathrm{P}(\text{Monty shows Door 2} \,|\, \text{Car behind 1})\mathrm{P}(\text{Car behind 1})}{\sum_{i=1}^{3} \mathrm{P}(\text{Monty shows Door 2} \,|\, \text{Car behind } i)\mathrm{P}(\text{Car behind } i)} \\ &= \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = 1/3. \end{aligned}$$

On the other hand,

$$\begin{aligned} &\mathrm{P}(\text{Car behind Door 3} \,|\, \text{Monty shows Door 2}) \\ &= \frac{\mathrm{P}(\text{Monty shows Door 2} \,|\, \text{Car behind 3})\mathrm{P}(\text{Car behind 3})}{\sum_{i=1}^{3} \mathrm{P}(\text{Monty shows Door 2} \,|\, \text{Car behind } i)\mathrm{P}(\text{Car behind } i)} \\ &= \frac{1 \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = 2/3. \end{aligned}$$

Thus, the prudent action is to switch to Door 3! $\square$

If you don't quite believe this, you aren't alone. But consider what you would do if there were 1,000 doors and Monty revealed 998 of them — of course you would switch from your door to the remaining one! ☺

1.9 Exercises

1. (§1.2) Suppose we define the following line segments: $U = [0, 2]$, $A = [0.5, 1]$, and $B = [0.5, 1.5]$. What are $\bar{\bar{A}}$, $\overline{A \cup B}$, and $A \cup \bar{B}$?

2. (§1.2) The set of all subsets of a set S is called the **power set** of S, and is denoted by 2^S. For instance, if $S = \{a, b\}$, then $2^S = \big\{\emptyset, \{a\}, \{b\}, S\big\}$. If $|S| = n$, find $|2^S|$.

3. (§1.2) Prove one of DeMorgan's Laws: $\overline{A \cap B} = \bar{A} \cup \bar{B}$. You can use Venn diagrams or argue mathematically.

4. (§1.2) If $f(x) = \ell\text{n}(2x - 3)$, find the derivative $f'(x)$.

5. (§1.2) If $f(x) = \cos(1/x)$, find the derivative $f'(x)$.

6. (§1.2) If $f(x) = \sin(\ell\text{n}(x))$, find the derivative $f'(x)$.

7. (§1.2) Find $\int_0^1 (2x + 1)^2 \, dx$.

8. (§1.2) Find $\int_1^2 e^{2x} \, dx$.

9. (§1.2) Find $\int_1^2 \ell\text{n}(x) \, dx$.

10. (§1.2) Calculate the limit
$$\lim_{x \to 0} \frac{e^x - 1}{\sin(x)}.$$
(Hint: If this problem makes you sick, you'll have to go to the. . . ?)

11. (§1.2) Calculate the limit
$$\lim_{x \to 0} \frac{\sin(x) - x}{x}.$$

12. (§1.2) Use induction to prove the well-known result $\sum_{k=1}^{n} k = n(n + 1)/2$.

13. (§1.3) A box contains four marbles (one red, two greens, and one purple).
    (a) Consider an experiment that consists of taking one marble from the box, then replacing it in the box, and then drawing a second marble from the box. What is the sample space?
    (b) Repeat the above when the second marble is drawn *without* replacing the first marble.

14. (§1.3) Let A, B, and C be three events, and suppose that the sample space $\Omega = A \cup B \cup C$.
    (a) Find an expression for the event that B and C occur, but not A.
    (b) Find an expression for the event that only B occurs.

15. (§1.3) As if you have nothing better to do, toss a die 6,000 times. (I suppose you could do this in Excel or some other software.) How many times do each of the numbers come up? Approximately how many would you expect?

16. (§1.3) Fun exercise! Toss a coin until you see the sequence HT. How many tries did it take? (E.g., the sequence "TTHHHT" would take six tosses before you stop.) Now toss a coin until you see two H's in a row. How long did that take? (E.g., the sequence TTHTTTHH" would take eight tosses.) Thoughts?

17. (§1.3) Let's look at the preferences of the membership of a travel club for couples. It turns out that 60% of the folks adore **A**ndorra, 60% love **L**as Vegas, and 70% are partial for **P**aris. Further, 30% like Andorra and Las Vegas, 40% Andorra and Paris, 50% Las Vegas and Paris, and 20% all three.

(a) James and Joyce are a random couple in the club. What's the probability that they enjoy going to at least one of the three destinations?

(b) Ricky and Nelson are another random couple in the club. They're travellin' men, and they've made a lot of stops all over the world. Find the probability that they'll enjoy precisely two of the three destinations.

18. (§1.3) Prove Bonferroni's inequality: $\mathrm{P}(A \cap B) \geq \mathrm{P}(A) + \mathrm{P}(B) - 1$.

19. (§1.4) Suppose I flip a coin three times. What's the probability that I get exactly two heads and one tail?

20. (§1.5) How many five-letter words can be formed from the alphabet if we require

(a) The second and fourth letters to be vowels (a, e, i, o, u)?

(b) At least one vowel?

21. (§1.5) TRUE or FALSE? $P_{n,k} = \binom{n}{k} k!$ for all appropriate (n, k).

22. (§1.5) Consider a baseball team with exactly nine players.

(a) How many ways can you fill the first four positions the batting order?

(b) How many of these ways have Smith batting first?

23. (§1.5) Suppose that 12 clowns are trying to get into a small car that will only accommodate a maximum of seven clowns. How many possible choices of seven clowns are there?

24. (§1.5) Write a computer program in your favorite language to calculate combinations. Demonstrate your program on $C_{100,50}$.

25. (§1.6) Three dice are tossed. What is the probability that the same number appears on exactly two of the three dice?

26. (§1.6) How many ways can you arrange the letters in "SYZYGY"? (Do you know what this word means?)

27. (§1.6) A bridge hand contains 13 cards from a standard deck. Find the probability that a bridge hand will contain...

(a) Exactly three kings.

(b) All 13 cards of the same suit.

28. (§1.6) A six-sided die is thrown seven times. Find

(a) P('3' comes up at least once).

(b) P(each face appears at least once).

29. (§1.6) The planet Glubnor has 50-day years.

(a) Suppose there are two Glubnorians in the room. What's the probability that they'll have the same birthday?

(b) Suppose there are three Glubnorians in the room. (They're big, so the room is getting crowded.) What's the probability that at least two of them have the same birthday?

30. (§1.6) Wedding invitations! We have four invitation cards and accompanying envelopes. But oops — we've randomly mixed the cards and the envelopes! What's the probability that we'll get at least one correct match?

31. (§1.6) Choose six cards from a standard deck. What's the probability that you will get three pairs?

32. (§1.6) Yahtzee! Toss five dice.

    (a) What's the probability that you'll observe a so-called "large straight" of length five? (E.g., such a straight arises from the toss 3,2,6,4,5. And note that you can only have straights starting at 1 or 2.)

    (b) What's the probability that you'll see exactly two pairs and a junk toss? (Note that this precludes the possibility of a full house. Also, perhaps surprisingly, this doesn't get you any points in Yahtzee!)

33. (§1.7) You have 20 marbles — 12 reds and 8 blues. Suppose you choose two of these marbles randomly (assuming that you don't lose your marbles). What's the probability that...

    (a) Both are red?

    (b) One is red and one is blue?

34. (§1.7) Prove: If $\mathrm{P}(A|B) > \mathrm{P}(A)$, then $\mathrm{P}(B|A) > \mathrm{P}(B)$.

35. (§1.7) Suppose A and B are independent, $\mathrm{P}(A) = 0.4$, and $\mathrm{P}(A \cup B) = 0.6$. Find $\mathrm{P}(B)$.

36. (§1.7) Suppose $\mathrm{P}(A) = 0.4$, $\mathrm{P}(A \cup B) = 0.7$, and $\mathrm{P}(B) = x$.

    (a) For what choice of x are A and B disjoint?

    (b) Independent?

37. (§1.7) If $\mathrm{P}(A) = \mathrm{P}(B) = \mathrm{P}(C) = 0.6$, and A, B, and C are independent, find the probability that *exactly one* of A, B, and C occurs.

38. (§1.7) Prove: C and D are independent $\Leftrightarrow$ C and $\bar{D}$ are independent.

39. (§1.8) Consider two boxes. Box I contains one blue marble and one white marble. Box II contains two blues and one white. A box is selected at random and a marble is drawn at random from the selected box.

    (a) Find P(the marble is blue).

    (b) What is the probability that the marble was selected from Box I, given that the marble is white?

40. (§1.8) I have a fair coin and a two-headed coin.

(a) I select one at random, and when I flip it, it comes up heads. What's the probability that the coin is fair?

(b) I flip the same coin, and again it comes up heads. Same question.

(c) I flip the coin eight times, and I get HHHHHHHT. Same question.

41. (§1.8) The city of Springfield has police with terrific judgment — so much so that

$$\text{P(Any defendant who is brought to trial is actually guilty)} \;=\; 0.99.$$

And Springfield's court system is pretty good, too. In fact, in any trial,

$$\text{P(The jury frees the defendant if he's actually innocent)} \;=\; 0.95$$

and

$$\text{P(The jury convicts if the defendant is actually guilty)} \;=\; 0.95.$$

Find P(Defendant is innocent|Jury sets him free).

42. (§1.8) A tough problem! Three prisoners A, B, and C are informed by their jailer that one of them has been chosen at random to be executed, and the other two are to be freed. Prisoner A asks the jailer to tell him privately which of his fellow prisoners will be set free (with the assumption that the jailer will randomize between B and C if A is the unfortunate soul). A claims that there would be no harm in divulging this information, since he already knows that at least one of the two will go free. The jailer refuses by arguing that if A knew, then A's probability of being executed would rise from 1/3 to 1/2 (i.e., there would only be two prisoners left). Who is correct?

43. Mathemusical Bonus: Use Bayes Theorem and Bill Withers to simplify the following expression.

$$\frac{\text{P(she's gone}\,|\,\text{no sunshine)P(no sunshine}\big)}{\text{P(she's gone}\,|\,\text{no sunshine)P(no sunshine)} + \text{P(she's gone}\,|\,\text{sunshine)P(sunshine)}}.$$

Chapter 2

Random Variables

This chapter introduces the important concept of random variables, which are useful in the calculation of probabilities, as well as measures of centrality and variation.

2.1 Introduction and Definitions

Definition: A **random variable (RV)** X is a function from the sample space to the real line, $X : \Omega \to \mathbb{R}$.

Example: Flip two coins. The sample space is $\Omega = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$. Suppose X is the RV corresponding to the number of H's. Then (suppressing the extraneous "$\{\cdot\}$" notation),

$$X(\text{TT}) \;=\; 0, \quad X(\text{HT}) \;=\; X(\text{TH}) \;=\; 1, \quad X(\text{HH}) \;=\; 2.$$

This results in

$$\text{P}(X = 0) \;=\; \frac{1}{4}, \quad \text{P}(X = 1) \;=\; \frac{1}{2}, \quad \text{P}(X = 2) \;=\; \frac{1}{4}. \quad \square$$

Notation: Capital letters like X, Y, Z, U, V, W usually represent RV's. Small letters like x, y, z, u, v, w usually represent particular values of the RV's. Thus, you can speak of quantities such as $\mathrm{P}(X = x)$.

Example: Let X be the sum of two dice rolls. The sample space is the set of all the ways to roll two dice. Then, e.g., $(4, 6)$ is an outcome from the sample space, and of course $X((4, 6)) = 10$. In addition, we can calculate the probability of each possible outcome:

$$\mathrm{P}(X = x) = \begin{cases} 1/36 & \text{if } x = 2 \\ 2/36 & \text{if } x = 3 \\ \vdots & \\ 6/36 & \text{if } x = 7 \\ \vdots & \\ 1/36 & \text{if } x = 12 \\ 0 & \text{otherwise.} \end{cases} \qquad \square$$

Example: Flip a coin.

$$X \equiv \begin{cases} 0 & \text{if T} \\ 1 & \text{if H} \end{cases}$$

Example: Roll a die.

$$Y \equiv \begin{cases} 0 & \text{if } \{1, 2, 3\} \\ 1 & \text{if } \{4, 5, 6\} \end{cases}$$

For all intents and purposes, the RV's X and Y are the same, since $\mathrm{P}(X = 0) = \mathrm{P}(Y = 0) = \frac{1}{2}$, and $\mathrm{P}(X = 1) = \mathrm{P}(Y = 1) = \frac{1}{2}$. $\square$

Example: Select a real number at random between 0 and 1. This experiment has an *infinite* number of "equally likely" outcomes.

Conclusion: P(we choose the *individual point* x) $= \mathrm{P}(X = x) = 0$, believe it or not!

But note that $\mathrm{P}(X \leq 0.65) = 0.65$, and $\mathrm{P}(X \in [0.3, 0.7]) = 0.4$. In fact, if A is any *interval* in [0,1], then $\mathrm{P}(X \in A)$ is the length of A. $\square$

Definition: If the number of possible values of a random variable X is finite or countably infinite, then X is a **discrete** random variable. Otherwise,...

A **continuous** random variable is one with probability 0 at every point.

Example: Flip a coin — result is H or T. Discrete.

Example: Pick a point at random in $[0, 1]$. Continuous.

Example: The amount of time you wait in a line is either 0 (with positive probability) or some positive real number — a *combined* discrete-continuous random

variable!

2.2 Discrete Random Variables

Definition: If X is a discrete random variable, its **probability mass function (pmf)** is $f(x) \equiv \mathrm{P}(X = x)$, for all x.

Note that if X is discrete, then $0 \le f(x) \le 1$, for all x, and $\sum_x f(x) = 1$.

Example: Flip two coins. Let X be the number of heads.

$$f(x) = \begin{cases} 1/4 & \text{if } x = 0 \text{ or } 2 \\ 1/2 & \text{if } x = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Example/Definition: Suppose X can equal $1, 2, \ldots, n$, each with probability $1/n$, i.e., $f(i) = 1/n$, $i = 1, 2, \ldots, n$. Then we say that X has the **discrete uniform**$\{1, 2, \ldots, n\}$ **distribution**.

Example: A discrete RV can have any values. For instance, let X denote the possible profits from an inventory policy, where $f(-5.1) = 0.2$ (lose money), $f(1.3) = 0.5$ (break even), and $f(11) = 0.3$ (big bucks).

Example/Definition: Let X denote the number of "successes" from n independent trials such that the probability of success at each trial is p ($0 \le p \le 1$). Then X has the **binomial distribution**, with parameters n and p. The trials are referred to as **Bernoulli trials**, named after Jakob Bernoulli (1655–1705).

Notation: $X \sim \mathrm{Bin}(n, p)$.

Example: Roll a die three independent times. Find P(exactly two 6's).

We interpret each toss as a Bernoulli trial in which a 6 is a success, and anything else (1,2,3,4,5) is a failure. All three trials are independent, and P(success) = 1/6 doesn't change from trial to trial.

Let $X =$ the number of 6's. Then $X \sim \mathrm{Bin}(3, \frac{1}{6})$.

Theorem: If $X \sim \mathrm{Bin}(n, p)$, then the probability of k successes in n trials is

$$\mathrm{P}(X = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \ldots, n, \quad \text{with } q = 1 - p.$$

Proof: Consider the particular sequence of successes and failures:

$$\underbrace{\mathrm{SS}\cdots\mathrm{S}}_{k \text{ successes}} \; \underbrace{\mathrm{FF}\cdots\mathrm{F}}_{n-k \text{ fails}} \quad (\text{probability} = p^k q^{n-k}).$$

The number of ways to arrange the sequence is $\binom{n}{k}$. Done. □

Example (cont'd): Back to the dice example, where $X \sim \text{Bin}(3, \frac{1}{6})$, and we want P(exactly two 6's). We have $n = 3$, $k = 2$, $p = 1/6$, and $q = 5/6$. Then

$$\text{P}(X = 2) \;=\; \binom{3}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^1 \;=\; \frac{15}{216}.$$

In fact, we can get the entire pmf,

k	0	1	2	3
$\text{P}(X = k)$	$\frac{125}{216}$	$\frac{75}{216}$	$\frac{15}{216}$	$\frac{1}{216}$

. □

Example: Roll two dice and get the sum. Repeat this experiment 12 times. Find P(Sum will be 7 or 11 exactly three times).

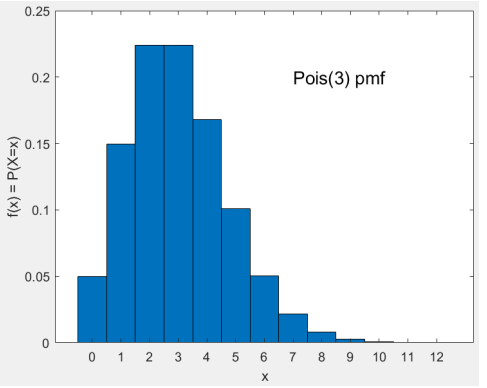
Let $X =$ the number of times we get 7 or 11. Then

$$\text{P}(7 \text{ or } 11) \;=\; \text{P}(7) + \text{P}(11) \;=\; \frac{6}{36} + \frac{2}{36} \;=\; \frac{2}{9}.$$

So $X \sim \text{Bin}(12, 2/9)$, and then

$$\text{P}(X = 3) \;=\; \binom{12}{3}\left(\frac{2}{9}\right)^3\left(\frac{7}{9}\right)^9. \quad \square$$

Definition: If $\text{P}(X = k) = e^{-\lambda}\lambda^k/k!$, $k = 0, 1, 2, \ldots$, and $\lambda > 0$, we say that X has the **Poisson distribution** with parameter λ, so named after Siméon Denis Poisson (1781–1840).



Notation: $X \sim \text{Pois}(\lambda)$.

Example: Suppose the number of raisins in a cup of cookie dough is Pois(10). Find the probability that a cup of dough has at least four raisins.

$$\begin{aligned}
\text{P}(X \geq 4) &= 1 - \text{P}(X = 0, 1, 2, 3) \\
&= 1 - e^{-10}\left(\frac{10^0}{0!} + \frac{10^1}{1!} + \frac{10^2}{2!} + \frac{10^3}{3!}\right) \\
&= 0.9897. \quad \square
\end{aligned}$$

2.3 Continuous Random Variables

Example: Pick a point X randomly between 0 and 1, and define the continuous function

$$f(x) \;=\; \begin{cases} 1 & \text{if } 0 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Intuitively, if $0 \le a \le b \le 1$, then

$$\mathrm{P}(a < X < b) \;=\; \text{the "area" under } f(x), \text{ from } a \text{ to } b \;=\; b - a.$$

Definition: Suppose X is a continuous RV. The magic function $f(x)$ is the **probability density function (pdf)** if

- $\int_{\mathbb{R}} f(x)\,dx = 1$ (the area under $f(x)$ is 1),
- $f(x) \ge 0$, $\forall x$ (the function is always non-negative), and
- If $A \subseteq \mathbb{R}$, then $\mathrm{P}(X \in A) = \int_A f(x)\,dx$ (the probability that X is in A).

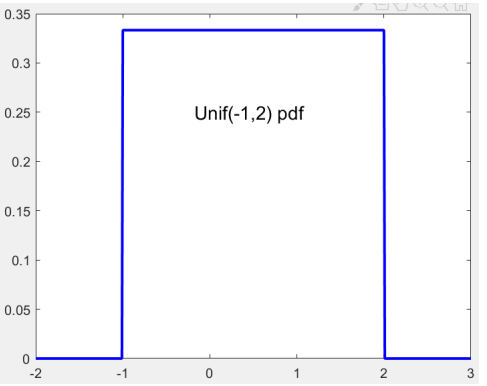
Remarks: If X is a continuous RV, then $\mathrm{P}(a < X < b) = \int_a^b f(x)\,dx$. As a special case, any individual point has probability *zero*, i.e., $\mathrm{P}(X = a) = \int_a^a f(x)\,dx = 0$. This implies that $\mathrm{P}(a < X < b) = \mathrm{P}(a \le X \le b)$, so we can be a little loose with the inequalities for the continuous case.

Remark: Note that $f(x)$ denotes both the pmf (**discrete** case) and pdf (**continuous** case) — but they are *different*:

- If X is *discrete*, then $f(x) = \mathrm{P}(X = x)$ and we must have $0 \le f(x) \le 1$.
- If X is *continuous*, then
  - $f(x)$ *isn't a probability*, but it's used to *calculate* probabilities.
  - Instead, think of $f(x)\,dx$ as $\doteq \mathrm{P}(x < X < x + dx)$.
  - Must have $f(x) \ge 0$ (and possibly > 1).
  - Calculate the probability of an event A by integrating, $\int_A f(x)\,dx$.

Example: If X is "equally likely" to be anywhere between a and b, then X has the **uniform distribution** on (a, b).

$$f(x) \;=\; \begin{cases} \frac{1}{b-a} & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$$



Notation: $X \sim \text{Unif}(a, b)$.

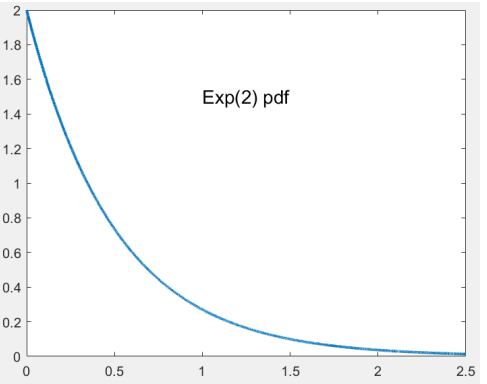
Remark: Note that $\int_{\mathbb{R}} f(x)\,dx = \int_a^b \frac{1}{b-a}\,dx = 1$ (as desired).

Example: If $X \sim \text{Unif}(-2, 8)$, then

$$\text{P}(-1 < X < 6) \;=\; \int_{-1}^{6} \frac{1}{8-(-2)}\,dx \;=\; 0.7. \quad \square$$

Example: X has the **exponential distribution** with parameter $\lambda > 0$ if it has pdf

$$f(x) \;=\; \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$



Notation: $X \sim \text{Exp}(\lambda)$.

Remark: $\int_{\mathbb{R}} f(x)\,dx = \int_0^\infty \lambda e^{-\lambda x}\,dx = 1$ (as desired).

Example: Suppose $X \sim \text{Exp}(1)$. Then

$\text{P}(X \leq 3) \;=\; \int_0^3 e^{-x}\,dx \;=\; 1 - e^{-3}$.

$\text{P}(X \geq 5) \;=\; \int_5^\infty e^{-x}\,dx \;=\; e^{-5}$.

$$\mathrm{P}(2 \le X < 4) \;=\; \mathrm{P}(2 \le X \le 4) \;=\; \int_2^4 e^{-x}\,dx \;=\; e^{-2} - e^{-4}.$$
$$\mathrm{P}(X = 3) \;=\; \int_3^3 e^{-x}\,dx \;=\; 0. \quad \square$$

Example: Suppose X is a continuous RV with pdf

$$f(x) \;=\; \begin{cases} cx^2 & \text{if } 0 < x < 2 \\ 0 & \text{otherwise.} \end{cases}$$

First of all, let's find c. Noting that the pdf must integrate to 1, we have

$$1 \;=\; \int_{\mathbb{R}} f(x)\,dx \;=\; \int_0^2 cx^2\,dx \;=\; 8c/3,$$

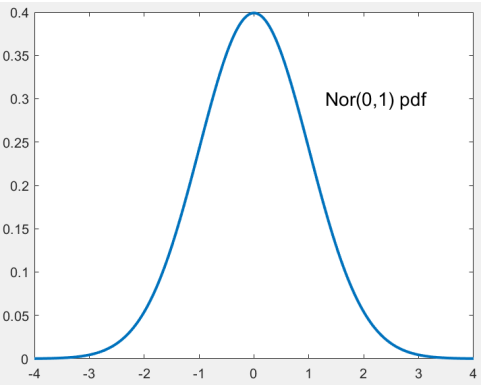
so that $c = 3/8$. Now we can calculate any reasonable probabilities, e.g.,

$$\mathrm{P}(0 < X < 1) \;=\; \int_0^1 \frac{3}{8}x^2\,dx \;=\; 1/8.$$

And more-complicated ones, e.g.,

$$\begin{aligned} &\mathrm{P}\Big(0 < X < 1 \,\Big|\, 1/2 < X < 3/2\Big) \\ &= \frac{\mathrm{P}(0 < X < 1 \text{ and } 1/2 < X < 3/2)}{\mathrm{P}(1/2 < X < 3/2)} \\ &= \frac{\mathrm{P}(1/2 < X < 1)}{\mathrm{P}(1/2 < X < 3/2)} \\ &= \frac{\int_{1/2}^1 \frac{3}{8}x^2\,dx}{\int_{1/2}^{3/2} \frac{3}{8}x^2\,dx} \;=\; 7/26. \quad \square \end{aligned}$$

Example: X has the **standard normal distribution** if its pdf is $\phi(x) \equiv \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, for all $x \in \mathbb{R}$. This is the famous "bell curve" distribution.



Notation: $X \sim \mathrm{Nor}(0, 1)$.

2.4 Cumulative Distribution Functions

Definition: For any random variable X (discrete or continuous), the **cumulative distribution function (cdf)** is $F(x) \equiv P(X \leq x)$, $\forall x \in \mathbb{R}$.

For X discrete,

$$F(x) = \sum_{y|y\leq x} f(y) = \sum_{y|y\leq x} \mathrm{P}(X = y).$$

For X continuous,

$$F(x) = \int_{-\infty}^{x} f(y)\, dy.$$

Discrete cdf's

Example: Flip a coin twice. Let X = number of H's. Then,

$$X = \begin{cases} 0 \text{ or } 2 & \text{w.p. } 1/4 \\ 1 & \text{w.p. } 1/2, \end{cases}$$

and the cdf is the following *step function*:

$$F(x) = \mathrm{P}(X \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ 1/4 & \text{if } 0 \leq x < 1 \\ 3/4 & \text{if } 1 \leq x < 2 \\ 1 & \text{if } x \geq 2. \quad \square \end{cases}$$

Remark: **Warning!** For *discrete* RV's, you have to be careful about "$\leq$" vs. "$<$" at the endpoints of the intervals (where the step function jumps).

Continuous cdf's

Theorem: If X is a *continuous* random variable, then $f(x) = F'(x)$ (assuming the derivative exists).

Proof: $F'(x) = \frac{d}{dx} \int_{-\infty}^{x} f(t)\, dt = f(x)$, which follows by the Fundamental Theorem of Calculus (§1.2.2.3). $\square$

Example: $X \sim \mathrm{Unif}(0, 1)$. The pdf and cdf, respectively, are

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } 0 < x < 1 \\ 1 & \text{if } x \geq 1. \quad \square \end{cases}$$

Example: $X \sim \mathrm{Exp}(\lambda)$. Recall that the pdf is $f(x) = \lambda e^{-\lambda x}$, $x > 0$. Then the cdf is

$$F(x) = \int_{-\infty}^{x} f(t)\, dt = \begin{cases} 0 & \text{if } x \leq 0 \\ 1 - e^{-\lambda x} & \text{if } x > 0. \end{cases}$$

We can use the cdf to find the **median** of X, that is, the point m such that

$$0.5 \;=\; \mathrm{P}(X \le m) \;=\; F(m) \;=\; 1 - e^{-\lambda m}.$$

Solving, we obtain $m = (1/\lambda)\ell\mathrm{n}(2)$. $\square$

Example: If $X \sim \mathrm{Nor}(0,1)$ with pdf $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, for all $x \in \mathbb{R}$, then the cdf is $\Phi(x) \equiv \int_{-\infty}^{x} \phi(t)\, dt$, which has no closed form, but is widely available in tables (e.g., our Table B.1 in the Appendix) or via ubiquitous software. We shall visit this distribution over and over again in the sequel.

Properties of all cdf's: Any cdf has the following intuitive properties that we will use throughout the book.

- $F(x) = \mathrm{P}(X \le x)$ is *non-decreasing* in x, i.e., $a < b$ implies that $F(a) \le F(b)$. This certainly makes sense in light of the fact that $F(x)$ is the totality of the probabilities up to the point x.
- $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

Theorem: $\mathrm{P}(X > x) = 1 - F(x)$.

Proof: By complements, $\mathrm{P}(X > x) = 1 - \mathrm{P}(X \le x) = 1 - F(x)$. $\square$

Theorem: $a < b \;\Rightarrow\; \mathrm{P}(a < X \le b) = F(b) - F(a)$.

Proof: Since $a < b$, we have

$$\begin{aligned} \mathrm{P}(a < X \le b) &= \mathrm{P}(X > a \;\cap\; X \le b) \\ &= \mathrm{P}(X > a) + \mathrm{P}(X \le b) - \mathrm{P}(X > a \;\cup\; X \le b) \\ &= 1 - F(a) + F(b) - 1. \quad \square \end{aligned}$$

2.5 Great Expectations

We start out in §2.5.1 by defining the *expected value* (aka the *mean*) of a random variable, which is a measure of a RV's "central tendency." We then introduce the very general *Law of the Unconscious Statistician* in §2.5.2, and use this to define other RV measures such as the *variance* and *moments*. We'll end the discussion with some useful approximations for the expected value and variance of complicated functions of X in §2.5.3.

2.5.1 Expected Value

Definition: The **mean** or **expected value** or **average** of a RV X is

$$\mu \;\equiv\; \mathrm{E}[X] \;\equiv\; \begin{cases} \sum_x x f(x) & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} x f(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

The mean gives an indication of an RV's *central tendency*. It can be thought of as a weighted average of the possible x's, where the weights are given by $f(x)$.

Example: The discrete RV X has the **Bernoulli distribution**, with parameter p, if it has the very simple pmf $\mathrm{P}(X = 0) = 1 - p$ and $\mathrm{P}(X = 1) = p$. We think of $X = 1$ as a "success" and $X = 0$ as a "failure." And a Bern(p) RV is the same as a Bin$(1, p)$. Then

$$\mathrm{E}[X] \;=\; \sum_x x f(x) \;=\; (0 \cdot (1-p)) + (1 \cdot p) \;=\; p. \quad \square$$

Example: Toss a die. Then $X = 1, 2, \ldots, 6$, each with probability $1/6$, and

$$\mathrm{E}[X] \;=\; \sum_x x f(x) \;=\; 1 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} \;=\; 3.5. \quad \square$$

Remark: The above couple examples show that $\mathrm{E}[X]$ does not have to equal one of the potential values of the RV — nothing to worry about!

Definition: We say that X has the **geometric distribution**, with parameter p, if X is defined as the number of Bern(p) trials until you obtain your first success. (For example, FFFFS would give $X = 5$.) Then X has pmf

$$f(x) \;=\; (1-p)^{x-1} p, \; x = 1, 2, \ldots.$$

Notation: $X \sim \mathrm{Geom}(p)$.

Extended Example: Here's an application that requires me to admit that I'm not a very good basketball player. Suppose I take independent foul shots, but the probability of making any particular shot is only 0.4. What's the probability that it'll take me at least three tries to make a successful shot?

Answer: The number of tries until my first success is $X \sim \mathrm{Geom}(0.4)$. Thus,

$$\begin{aligned} \mathrm{P}(X \geq 3) &= 1 - \mathrm{P}(X \leq 2) \\ &= 1 - \mathrm{P}(X = 1) - \mathrm{P}(X = 2) \\ &= 1 - 0.4 - (0.6)(0.4) \;=\; 0.36. \quad \square \end{aligned}$$

Now, let's find the expected value of $X \sim \mathrm{Geom}(p)$. Letting $q = 1 - p$, we have

$$\begin{aligned} \mathrm{E}[X] &= \sum_x x f(x) \;=\; \sum_{x=1}^{\infty} x q^{x-1} p \\ &= p \sum_{x=1}^{\infty} \frac{d}{dq} q^x \;=\; p \, \frac{d}{dq} \sum_{x=1}^{\infty} q^x \quad \text{(carefully swap derivative and sum)} \\ &= p \, \frac{d}{dq} \frac{q}{1-q} \quad \text{(geometric sum)} \\ &= p \left[\frac{(1-q) - q(-1)}{(1-q)^2} \right] \;=\; 1/p. \quad \square \end{aligned}$$

Thus, owing to my questionable basketball skills, it will take, on average, $\mathrm{E}[X] = 1/p = 1/0.4 = 2.5$ shots before I make my first bucket. □

Example: $X \sim \mathrm{Exp}(\lambda)$. $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$. Then

$$\begin{aligned}
\mathrm{E}[X] &= \int_{\mathbb{R}} x f(x)\, dx \\
&= \int_0^\infty x\lambda e^{-\lambda x}\, dx \\
&= -xe^{-\lambda x}\Big|_0^\infty - \int_0^\infty (-e^{-\lambda x})\, dx \quad \text{(by parts)} \\
&= \int_0^\infty e^{-\lambda x}\, dx \quad \text{(L'Hôpital rules!)} \\
&= 1/\lambda. \quad \square
\end{aligned}$$

Remark: Strictly speaking, there is a technical requirement that must hold in order for the mean to exist. Namely, we require $\sum_x |x| f(x) < \infty$ or $\int_{\mathbb{R}} |x| f(x)\, dx < \infty$ if X is, respectively, discrete or continuous. So, for example, if X has the **Cauchy distribution** with pdf $f(x) = \frac{1}{\pi(1+x^2)}$ for $x \in \mathbb{R}$, then $\int_{\mathbb{R}} |x| f(x) = \infty$, so that X does not have a mean. *Going forward, unless otherwise specified, we will assume that all means under discussion exist.*

2.5.2 LOTUS, Moments, and Variance

The next theorem often goes by the mysterious moniker **The Law of the Unconscious Statistician (LOTUS),** and generalizes our concept of expected value.

Theorem (LOTUS): The expected value of a function of X, say $h(X)$, is

$$\mathrm{E}[h(X)] = \begin{cases} \sum_x h(x) f(x) & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} h(x) f(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

Proof: See §2.8.3. □

$\mathrm{E}[h(X)]$ is a weighted function of $h(x)$, where the weights are the $f(x)$ values.

Remark: It looks like a definition, but it's really a theorem — they call it LOTUS because some statistics teachers aren't paying attention and present it as a definition!

Examples: LOTUS works on any reasonable function (assuming the resulting expected value is well-defined and finite). Here are several examples for a continuous random variable having pdf $f(x) = 3x^2$, $0 \leq x \leq 1$.

- $\mathrm{E}[X^2] = \int_{\mathbb{R}} x^2 f(x)\, dx = \int_0^1 3x^4\, dx = 3/5.$

- $\mathrm{E}[1/X] \;=\; \displaystyle\int_{\mathbb{R}} (1/x) f(x)\,dx \;=\; \int_0^1 3x\,dx \;=\; 3/2.$

- If we define the notation $y^+ \equiv \max\{y, 0\}$, then we have

$$\mathrm{E}\big[(X-0.5)^+\big] \;=\; \int_{\mathbb{R}} (x-0.5)^+ f(x)\,dx \;=\; \int_{0.5}^1 (x-0.5)\,3x^2\,dx \;=\; \frac{17}{64}.$$

- An application of integration by parts yields

$$\begin{aligned} \mathrm{E}\left[\frac{\sin(X)}{X}\right] \;&=\; \int_{\mathbb{R}} \frac{\sin(x)}{x} f(x)\,dx \;=\; \int_0^1 3x\sin(x)\,dx \\ &=\; 3\big(-x\cos(x)+\sin(x)\big)\big|_0^1 \;=\; 0.9035. \end{aligned}$$

- Finally, here's a cute and sometimes useful result that holds for any continuous RV X. Assuming that A is a well-behaved set, define the **indicator function**,

$$h(X) \;=\; 1_A(X) \;\equiv\; \begin{cases} 1 & \text{if } X \in A \text{ occurs} \\ 0 & \text{otherwise.} \end{cases}$$

  Then

$$\mathrm{E}\big[1_A(X)\big] \;=\; \int_{\mathbb{R}} 1_A(x) f(x)\,dx \;=\; \int_A f(x)\,dx \;=\; \mathrm{P}(A). \quad \square$$

Just a moment please... Now we'll discuss several special cases of LOTUS that are of particular importance.

Definition: The k^{th} **moment** of X is

$$\mathrm{E}[X^k] \;=\; \begin{cases} \sum_x x^k f(x) & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} x^k f(x)\,dx & \text{if } X \text{ is continuous.} \end{cases}$$

Example: Suppose $X \sim \mathrm{Bern}(p)$, so that $f(1) = p$ and $f(0) = q = 1-p$.

$$\mathrm{E}[X^k] \;=\; \sum_x x^k f(x) \;=\; (0^k \cdot q) + (1^k \cdot p) \;=\; p, \quad \text{for } \textit{all } k! \quad \text{Cool!} \quad \square$$

Example: $X \sim \text{Exp}(\lambda)$. $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$. Then

$$\begin{aligned}
\text{E}[X^k] &= \int_{\mathbb{R}} x^k f(x)\,dx \\
&= \int_0^\infty x^k \lambda e^{-\lambda x}\,dx \\
&= \int_0^\infty (y/\lambda)^k \lambda\, e^{-y} (1/\lambda)\,dy \quad \text{(substitute } y = \lambda x\text{)} \\
&= \frac{1}{\lambda^k} \int_0^\infty y^{(k+1)-1} e^{-y}\,dy \\
&= \frac{\Gamma(k+1)}{\lambda^k} \quad \text{(by definition of the gamma function)} \\
&= \frac{k!}{\lambda^k}. \quad \square
\end{aligned}$$

Definition: The k^{th} **central moment** of X is

$$\text{E}[(X-\mu)^k] = \begin{cases} \sum_x (x-\mu)^k f(x) & X \text{ is discrete} \\ \int_{\mathbb{R}} (x-\mu)^k f(x)\,dx & X \text{ is continuous.} \end{cases}$$

Definition: The **variance** of X is the second central moment, i.e., the expected value of the squared difference between the RV X and its mean μ. In other words,

$$\text{Var}(X) \equiv \text{E}\big[(X-\mu)^2\big].$$

Variance is a measure of *spread* or *dispersion* — a tight distribution has low variance, and a spread-out distribution has high variance.

Notation: $\sigma^2 \equiv \text{Var}(X)$.

Definition: The **standard deviation** of X is $\sigma \equiv +\sqrt{\text{Var}(X)}$.

Example: Suppose $X \sim \text{Bern}(p)$, so that $f(1) = p$ and $f(0) = q = 1 - p$.

Recall that $\mu = \text{E}[X] = p$. Then

$$\begin{aligned}
\text{Var}(X) &= \text{E}[(X-\mu)^2] \\
&= \sum_x (x-p)^2 f(x) \\
&= (0-p)^2 q + (1-p)^2 p \\
&= p^2 q + q^2 p \;=\; pq(p+q) \\
&= pq. \quad \square
\end{aligned}$$

The next results establish the fact that the expected value operator can pass through certain linear functions of X, and can then be used to obtain pleasant expressions for other expected values and variances.

Theorem: For any function of X, say $h(X)$, and constants a and b, we have[1]

$$\mathrm{E}[ah(X)+b] \;=\; a\mathrm{E}[h(X)]+b.$$

Proof (just do the continuous case): By LOTUS,

$$\begin{aligned}\mathrm{E}[ah(X)+b] &= \int_{\mathbb{R}}(ah(x)+b)f(x)\,dx\\ &= a\int_{\mathbb{R}}h(x)f(x)\,dx + b\int_{\mathbb{R}}f(x)\,dx\\ &= a\mathrm{E}[h(X)]+b. \quad \square\end{aligned}$$

Corollary: In particular,

$$\boxed{\mathrm{E}[aX+b] \;=\; a\mathrm{E}[X]+b.}$$

Similarly,

$$\mathrm{E}[g(X)+h(X)] \;=\; \mathrm{E}[g(X)]+\mathrm{E}[h(X)]. \tag{2.1}$$

Theorem: Here is what is sometimes an easier way to calculate variance:

$$\boxed{\mathrm{Var}(X) \;=\; \mathrm{E}[X^2]-(\mathrm{E}[X])^2.}$$

Proof: By Equation (2.1),

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{E}[(X-\mu)^2]\\ &= \mathrm{E}[X^2-2\mu X+\mu^2]\\ &= \mathrm{E}[X^2]-2\mu\mathrm{E}[X]+\mu^2\\ &= \mathrm{E}[X^2]-\mu^2. \quad \square\end{aligned}$$

Example: Suppose $X \sim \mathrm{Bern}(p)$. Recall the cool fact that $\mathrm{E}[X^k]=p$, for all $k=1,2,\ldots$. So,

$$\mathrm{Var}(X) \;=\; \mathrm{E}[X^2]-(\mathrm{E}[X])^2 \;=\; p-p^2 \;=\; pq. \quad \square$$

Example: $X \sim \mathrm{Unif}(a,b)$, so that $f(x)=1/(b-a)$, $a<x<b$. Then

$$\mathrm{E}[X] \;=\; \int_{\mathbb{R}} xf(x)\,dx \;=\; \int_a^b \frac{x}{b-a}\,dx \;=\; \frac{a+b}{2},$$

$$\mathrm{E}[X^2] \;=\; \int_{\mathbb{R}} x^2f(x)\,dx \;=\; \int_a^b \frac{x^2}{b-a}\,dx \;=\; \frac{a^2+ab+b^2}{3},$$

and

$$\mathrm{Var}(X) \;=\; \mathrm{E}[X^2]-(\mathrm{E}[X])^2 \;=\; \frac{(a-b)^2}{12} \quad \text{(after algebra)}. \quad \square$$

[1] "Shift" (i.e., b) happens

It's often the case that we'll need the variance of a linear function of X.

Theorem: For constants a and b,

$$\boxed{\text{Var}(aX+b) \;=\; a^2\text{Var}(X).}$$

Proof: By the definition of variance, and use of the previous theorem, we have

$$\begin{aligned}\text{Var}(aX+b) &= \text{E}\big[\big(aX+b-\text{E}[aX+b]\big)^2\big] \\ &= \text{E}\big[\big(aX+b-a\text{E}[X]-b\big)^2\big]] \\ &= a^2\text{E}\big[\big(X-\text{E}[X]\big)^2\big]. \quad \square\end{aligned}$$

Thus, multiplying a RV by a constant a results in a multiplicative bump of a^2 for the variance; but a constant additive shift b will not affect the variability of the RV at all.

Example: $X \sim \text{Bern}(0.3)$. Recall that

$$\text{E}[X] = p = 0.3 \quad \text{and} \quad \text{Var}(X) = pq = (0.3)(0.7) = 0.21.$$

Let $Y = 4X + 5$. Then

$$\text{E}[Y] \;=\; \text{E}[4X+5] \;=\; 4\text{E}[X]+5 \;=\; 6.2$$

and

$$\text{Var}(Y) \;=\; \text{Var}(4X+5) \;=\; 16\text{Var}(X) \;=\; 3.36. \quad \square$$

2.5.3 LOTUS via Taylor Series

Sometimes the function $h(X)$ is messy, and instead of using LOTUS directly, we may be able to evaluate $\text{E}[h(X)]$ via a Taylor series approach, as discussed in Meyer [5].

§2.5.3.1 Exact Taylor Series

Recall that the Taylor series for a real function $h(x)$ is given by

$$h(x) \;=\; \sum_{k=0}^{\infty} \frac{h^{(k)}(a)}{k!}(x-a)^k \quad \text{(for any constant } a\text{)},$$

where $h^{(k)}(a) = \frac{d^k}{dx^k}h(x)|_{x=a}$. Assuming that everything is well-defined and that we can use Equation (2.1) to move the expected value inside the (infinite) sum, we have

$$\text{E}[h(X)] \;=\; \sum_{k=0}^{\infty} \frac{h^{(k)}(a)}{k!}\text{E}\big[(X-a)^k\big]. \tag{2.2}$$

Example: Suppose that $h(X) = e^{tX}$. Then by Equation (2.2), with $a = 0$ and the fact that $h^{(k)}(x) = t^k e^{tx}$, we have

$$\mathrm{E}[e^{tX}] \;=\; \sum_{k=0}^{\infty} \frac{h^{(k)}(0)}{k!} \mathrm{E}[X^k] \;=\; \sum_{k=0}^{\infty} \frac{t^k}{k!}\, \mathrm{E}[X^k].$$

This is a pretty general result, but for now let's specifically suppose that $X \sim \mathrm{Exp}(\lambda = 1)$. Then by an example in §2.5.2, we know that $\mathrm{E}[X^k] = k!$, so

$$\mathrm{E}[e^{tX}] \;=\; \sum_{k=0}^{\infty} \frac{t^k}{k!}\, k! \;=\; \sum_{k=0}^{\infty} t^k \;=\; \frac{1}{1-t}, \quad \text{for } t < 1,$$

which is the same answer as the one we shall obtain via a slightly different method in §2.6.

And while we're at it, note that

$$\mathrm{Var}(e^{tX}) \;=\; \mathrm{E}[e^{2tX}] - \big(\mathrm{E}[e^{tX}]\big)^2 \;=\; \frac{1}{1-2t} - \frac{1}{(1-t)^2} \;=\; \frac{t^2}{(1-2t)(1-t)^2}. \quad \square$$

§2.5.3.2 Truncated Taylor Series Approximation

The success in calculating $\mathrm{E}[h(X)]$ in §2.5.3.1 is predicated on our willingness to calculate all of the moments $\mathrm{E}[(X-a)^k]$, $k = 1, 2, \ldots$. But if we want to be a little lazy, then it could be argued that merely the first three terms of the Taylor series with the mean $\mu = \mathrm{E}[X]$ substituted for a will do just fine as a crude approximation, i.e.,

$$h(X) \quad \approx \quad h(\mu) + h'(\mu)(X-\mu) + \frac{h''(\mu)}{2}(X-\mu)^2.$$

Then

$$\mathrm{E}[h(X)] \;\doteq\; h(\mu) + h'(\mu)\mathrm{E}[X-\mu] + \frac{h''(\mu)}{2}\mathrm{E}\big[(X-\mu)^2\big] \;=\; h(\mu) + \frac{h''(\mu)\,\sigma^2}{2},$$

where $\sigma^2 = \mathrm{Var}(X) = \mathrm{E}\big[(X-\mu)^2\big]$, and where we use "$\approx$" to denote random variables that have approximately the same distribution, and "$\doteq$" to indicate approximate equality of real numbers.

In addition, if we happen to be in the mood for an even cruder approximation, then we can use just the first *two* terms of the Taylor series to approximate the variance of $h(X)$,

$$\mathrm{Var}(h(X)) \;\doteq\; \mathrm{Var}\big(h(\mu) + h'(\mu)(X-\mu)\big) \;=\; (h'(\mu))^2\,\mathrm{Var}(X) \;=\; (h'(\mu))^2\,\sigma^2.$$

Example: Again suppose that $h(X) = e^{tX}$, with $X \sim \mathrm{Exp}(\lambda = 1)$. We'll now expand about $a = \mu = 1/\lambda = 1$ (instead of $a = 0$ as in the previous example) and note that $\sigma^2 = 1/\lambda^2 = 1$, to obtain the approximations

$$\mathrm{E}[e^{tX}] \;\doteq\; h(\mu) + \frac{h''(\mu)\sigma^2}{2} \;=\; h(1) + \frac{h''(1)}{2} \;=\; e^t + \frac{t^2 e^t}{2} \tag{2.3}$$

and

$$\mathrm{Var}(e^{tX}) \doteq (h'(\mu))^2\,\sigma^2 = (h'(1))^2 = t^2e^{2t}. \tag{2.4}$$

Recalling the exact results from §2.5.3.1 yielding $\mathrm{E}[e^{tX}] = \frac{1}{1-t}$ and $\mathrm{Var}(e^{tX}) = \frac{t^2}{(1-2t)(1-t)^2}$, we see that our approximation (2.3) for $\mathrm{E}[e^{tX}]$ matches up pretty well for $t < 0.4$; but the approximation (2.4) for $\mathrm{Var}(e^{tX})$ is acceptable only for very small t, before things start going haywire. □

Example: Suppose X has pdf $f(x) = 3x^2$, $0 \le x \le 1$, and we want to test out our approximations on the "complicated" random variable $Y = h(X) = X^{3/4}$. Well, it's not really that complicated, since we can calculate the *exact moments*, which we'll do now for the purpose of later comparison:

$$\begin{aligned}
\mathrm{E}[Y] &= \int_{\mathbb{R}} x^{3/4} f(x)\,dx = \int_0^1 3x^{11/4}\,dx = 4/5,\\
\mathrm{E}[Y^2] &= \int_{\mathbb{R}} x^{6/4} f(x)\,dx = \int_0^1 3x^{7/2}\,dx = 2/3, \quad \text{and}\\
\mathrm{Var}(Y) &= \mathrm{E}[Y^2] - (\mathrm{E}[Y])^2 = 2/75 = 0.0267.
\end{aligned}$$

Before we can do the approximations, note that

$$\begin{aligned}
\mu &= \mathrm{E}[X] = \int_{\mathbb{R}} x f(x)\,dx = \int_0^1 3x^3\,dx = 3/4,\\
\mathrm{E}[X^2] &= \int_{\mathbb{R}} x^2 f(x)\,dx = \int_0^1 3x^4\,dx = 3/5, \quad \text{and}\\
\sigma^2 &= \mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = 3/80 = 0.0375.
\end{aligned}$$

Now we can start the approximation portion of the exercise. We have

$$\begin{aligned}
h(\mu) &= \mu^{3/4} = (3/4)^{3/4} = 0.8059\\
h'(\mu) &= (3/4)\mu^{-1/4} = (3/4)(3/4)^{-1/4} = 0.8059\\
h''(\mu) &= -(3/16)\mu^{-5/4} = -0.2686.
\end{aligned}$$

Thus,

$$\mathrm{E}[Y] \doteq h(\mu) + \frac{h''(\mu)\sigma^2}{2} = 0.8059 - \frac{(0.2686)(0.0375)}{2} = 0.8009$$

and

$$\mathrm{Var}(Y) \doteq (h'(\mu))^2\sigma^2 = (0.8059)^2(0.0375) = 0.0243,$$

which are reasonably close to their true values, $\mathrm{E}[Y] = 0.8$ and $\mathrm{Var}(Y) = 0.0267$, respectively. □

2.6 Moment Generating Functions

We'll now discuss a tool that has a lot of very nice uses. Recall that $\mathrm{E}[X^k]$ is the k^{th} **moment** of X.

Definition: The **moment generating function (mgf)** of the random variable X is

$$M_X(t) \equiv \mathrm{E}\big[e^{tX}\big].$$

Remark: $M_X(t)$ is a function of t, *not* of X!

Example: $X \sim \mathrm{Bern}(p)$, so that $X = 1$ with probability p, and 0 with probability q. Then

$$M_X(t) = \mathrm{E}[e^{tX}] = \sum_x e^{tx} f(x) = e^{t\cdot 1}p + e^{t\cdot 0}q = pe^t + q. \quad \square$$

Example: $X \sim \mathrm{Exp}(\lambda)$. The pdf is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$. Then

$$\begin{aligned} M_X(t) &= \mathrm{E}[e^{tX}] \\ &= \int_{\mathbb{R}} e^{tx} f(x)\, dx \quad \text{(LOTUS)} \\ &= \lambda \int_0^\infty e^{(t-\lambda)x}\, dx \\ &= \frac{\lambda}{\lambda - t} \quad \text{if } \lambda > t. \quad \square \end{aligned}$$

Theorem (Why It's Called the Moment Generating Function): Under certain technical conditions (e.g., $M_X(t)$ must exist for all $t \in (-\epsilon, \epsilon)$, for some $\epsilon > 0$), we have

$$\mathrm{E}[X^k] = \frac{d^k}{dt^k} M_X(t)\Big|_{t=0}, \quad k = 1, 2, \ldots.$$

Thus, you can *generate* the moments of X from the mgf. (Sometimes, it's easier to get moments this way than directly.)

"Proof" (a little non-rigorous): By the Taylor series expression for the exponential function and some dainty swapping of expected values and sums, we have

$$M_X(t) = \mathrm{E}[e^{tX}] \text{ "=" } \mathrm{E}\left[\sum_{k=0}^\infty \frac{(tX)^k}{k!}\right] \text{ "=" } \sum_{k=0}^\infty \mathrm{E}\left[\frac{(tX)^k}{k!}\right] = \sum_{k=0}^\infty \frac{t^k}{k!}\mathrm{E}[X^k].$$

This (sort of) implies

$$\frac{d}{dt} M_X(t) \text{ "=" } \sum_{k=0}^\infty \frac{d}{dt}\frac{t^k}{k!}\mathrm{E}[X^k] = \sum_{k=1}^\infty \frac{t^{k-1}}{(k-1)!}\mathrm{E}[X^k] = \mathrm{E}[X] + t\,\mathrm{E}[X^2] + \cdots,$$

and so

$$\frac{d}{dt} M_X(t)\Big|_{t=0} = \mathrm{E}[X].$$

Same deal for higher-order moments. $\square$

Example: $X \sim \text{Bern}(p)$. Then $M_X(t) = pe^t + q$, and

$$\text{E}[X] \;=\; \frac{d}{dt} M_X(t)\Big|_{t=0} \;=\; \frac{d}{dt}(pe^t + q)\Big|_{t=0} \;=\; pe^t\big|_{t=0} \;=\; p.$$

In fact, it is easy to see that $\text{E}[X^k] = \frac{d^k}{dt^k} M_X(t)\big|_{t=0} = p$, for all k. □

Example: $X \sim \text{Exp}(\lambda)$. Then $M_X(t) = \lambda/(\lambda - t)$, for $\lambda > t$. So

$$\text{E}[X] \;=\; \frac{d}{dt} M_X(t)\Big|_{t=0} \;=\; \frac{\lambda}{(\lambda - t)^2}\Big|_{t=0} \;=\; 1/\lambda.$$

Further,

$$\text{E}[X^2] \;=\; \frac{d^2}{dt^2} M_X(t)\Big|_{t=0} \;=\; \frac{2\lambda}{(\lambda - t)^3}\Big|_{t=0} \;=\; 2/\lambda^2.$$

Thus,

$$\text{Var}(X) \;=\; \text{E}[X^2] - (\text{E}[X])^2 \;=\; \frac{2}{\lambda^2} - \frac{1}{\lambda^2} \;=\; 1/\lambda^2. \quad \square$$

You can do lots of other nice things with mgf's:

- Find the mgf of a linear function of X (next).
- Identify distributions (below).
- Probability inequality applications (§2.7).
- Find the mgf of the sum of independent random variables (§3.6).
- Convergence of random variables proofs (another course).

Theorem (mgf of a linear function of X): Suppose X has mgf $M_X(t)$, and let $Y = aX + b$. Then $M_Y(t) = e^{tb} M_X(at)$.

Proof: We have

$$M_Y(t) \;=\; \text{E}[e^{tY}] \;=\; \text{E}[e^{t(aX+b)}] \;=\; e^{tb}\text{E}[e^{(at)X}] \;=\; e^{tb} M_X(at). \quad \square$$

Example: Let $X \sim \text{Exp}(\lambda)$, and $Y = 3X + 2$. Then

$$M_Y(t) \;=\; e^{2t} M_X(3t) \;=\; e^{2t} \frac{\lambda}{\lambda - 3t}, \quad \text{if } \lambda > 3t. \quad \square$$

Theorem (identifying distributions): *In this text,* each distribution has a *unique mgf.*

Proof: Not here!

Example: Suppose that Y has mgf

$$M_Y(t) \;=\; e^{2t} \frac{\lambda}{\lambda - 3t}, \quad \text{for } \lambda > 3t.$$

Then by the previous example and the uniqueness of mgf's, it *must* be the case that $Y \sim 3X + 2$, where $X \sim \text{Exp}(\lambda)$. □

2.7 Some Probability Inequalities

Goal: Give results that provide general probability bounds. These are useful if we want rough estimates of probabilities, but can also be applied in many types of proofs.

Theorem (Markov's Inequality): If X is a nonnegative random variable and $c > 0$, then $\mathrm{P}(X \geq c) \leq \mathrm{E}[X]/c$. (This is a very crude upper bound.)

Proof: Because X is nonnegative, we have

$$\begin{aligned} \mathrm{E}[X] &= \int_{\mathbb{R}} x f(x)\, dx \\ &= \int_0^\infty x f(x)\, dx \\ &\geq \int_c^\infty x f(x)\, dx \\ &\geq c \int_c^\infty f(x)\, dx \\ &= c\,\mathrm{P}(X \geq c). \quad \square \end{aligned}$$

Theorem (Chebychev's Inequality)[2]**:** Suppose that $\mathrm{E}[X] = \mu$, and $\mathrm{Var}(X) = \sigma^2$. Then for any $c > 0$,

$$\boxed{\mathrm{P}\big(|X - \mu| \geq c\big) \;\leq\; \sigma^2/c^2.}$$

Proof: By Markov, with $|X - \mu|^2$ in place of X, and c^2 in place of c, we have

$$\mathrm{P}\big(|X - \mu| \geq c\big) \;=\; \mathrm{P}\big((X - \mu)^2 \geq c^2\big) \;\leq\; \frac{\mathrm{E}[(X - \mu)^2]}{c^2} \;=\; \sigma^2/c^2. \quad \square$$

Remarks:

- We can also write $\mathrm{P}\big(|X - \mu| < c\big) \geq 1 - \sigma^2/c^2$.
- Or, if $c = k\sigma$, then $\mathrm{P}\big(|X - \mu| \geq k\sigma\big) \leq 1/k^2$.
- Chebychev gives a bound on the probability that X deviates from the mean by more than a constant, in terms of the constant and the variance. You can always use Chebychev, but it's crude.

Example: Suppose $X \sim \mathrm{Unif}(0, 1)$, so that $f(x) = 1$, for $0 < x < 1$. Recall that for the $\mathrm{Unif}(a, b)$ distribution, we have $\mathrm{E}[X] = (a + b)/2 = 1/2$, and $\mathrm{Var}(X) = (b - a)^2/12 = 1/12$. Then Chebychev implies

$$\mathrm{P}\left(\left| X - \frac{1}{2} \right| \geq c \right) \;\leq\; \frac{1}{12c^2}.$$

[2]There are many, many ways to spell "Chebychev."

In particular, for $c = 1/3$,

$$\mathrm{P}\left(\left|X - \frac{1}{2}\right| \geq \frac{1}{3}\right) \;\leq\; \frac{3}{4} \quad \text{(Chebychev upper bound)}.$$

Let's compare the above bound to the *exact* answer:

$$\begin{aligned}
&\mathrm{P}\left(\left|X - \frac{1}{2}\right| \geq \frac{1}{3}\right) \\
&= \; 1 - \mathrm{P}\left(\left|X - \frac{1}{2}\right| < \frac{1}{3}\right) \\
&= \; 1 - \mathrm{P}\left(-\frac{1}{3} < X - \frac{1}{2} < \frac{1}{3}\right) \\
&= \; 1 - \mathrm{P}\left(\frac{1}{6} < X < \frac{5}{6}\right) \\
&= \; 1 - \int_{1/6}^{5/6} f(x)\,dx \\
&= \; 1 - \frac{2}{3} \;=\; 1/3.
\end{aligned}$$

So the Chebychev bound was pretty high in comparison. □

Bonus Theorem (Chernoff's Inequality):[3] For any c,

$$\mathrm{P}(X \geq c) \;\leq\; e^{-ct} M_X(t).$$

Proof: By Markov with e^{tX} in place of X and e^{tc} in place of c, we have

$$\mathrm{P}(X \geq c) \;=\; \mathrm{P}(e^{tX} \geq e^{tc}) \;\leq\; e^{-ct}\mathrm{E}[e^{tX}] \;=\; e^{-ct} M_X(t). \quad \square$$

Example: Suppose X has the standard normal distribution with pdf $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, for all $x \in \mathbb{R}$. It is easy to show (via a little calculus elbow grease involving completing a square) that the mgf of the standard normal is

$$M_X(t) \;=\; \mathrm{E}[e^{tX}] \;=\; \int_{\mathbb{R}} e^{tx}\phi(x)\,dx \;=\; e^{t^2/2}.$$

Then using Chernoff with $t = c$ immediately yields the "tail" probability,

$$\mathrm{P}(X \geq c) \;\leq\; e^{-c^2} M_X(c) \;=\; e^{-c^2/2}. \quad \square$$

2.8 Functions of a Random Variable

In §2.5.2, we learned about LOTUS, which gave us the expected value of a function of a random variable. Now we'll go for the *entire distribution* of that function.

[3] Seems like everyone's got their own inequality these days (just sayin'...)

For instance, if X is exponential, what is the distribution of X^2? This type of problem has tremendous applications all over the place, and it will often pop up as we proceed.

We'll begin the discussion with introductory material and basic examples in §2.8.1. Then §2.8.2 is concerned with what is known as the Inverse Transform Theorem, which is used extensively in computer simulation problems (among other applications) that require us to generate various random variables. §2.8.3 wraps up the discussion with a couple of "honors" results involving the nature of LOTUS.

2.8.1 Introduction and Baby Examples

Problem Statement:

- You have a random variable X, and you know its pmf/pdf $f(x)$.
- Define $Y \equiv h(X)$ (some function of X).
- Find $g(y)$, the pmf/pdf of Y.

We'll start with the case in which X is a discrete RV, and then we'll go to the continuous X case.

Discrete Case: X discrete implies Y discrete implies

$$g(y) \;=\; \mathrm{P}(Y = y) \;=\; \mathrm{P}(h(X) = y) \;=\; \mathrm{P}(\{x | h(x) = y\}) \;=\; \sum_{x|h(x)=y} f(x).$$

Example: Let X be the number of H's in two coin tosses. Suppose we want the pmf for $Y = h(X) = X^3 - X$.

x	0	1	2
$f(x) = \mathrm{P}(X = x)$	1/4	1/2	1/4
$y = x^3 - x$	0	0	6

This immediately gives us

$$g(0) = \mathrm{P}(Y = 0) = \mathrm{P}(X = 0 \text{ or } 1) = 3/4, \text{ and}$$
$$g(6) = \mathrm{P}(Y = 6) = \mathrm{P}(X = 2) = 1/4.$$

In other words,

$$g(y) \;=\; \begin{cases} 3/4 & \text{if } y = 0 \\ 1/4 & \text{if } y = 6. \end{cases} \quad \square$$

Example: Suppose X is discrete with

$$f(x) \;=\; \begin{cases} 1/8 & \text{if } x = -1 \\ 3/8 & \text{if } x = 0 \\ 1/3 & \text{if } x = 1 \\ 1/6 & \text{if } x = 2 \\ 0 & \text{otherwise.} \end{cases}$$

Let $Y = X^2$ (so Y can only equal 0, 1, or 4). Then

$$g(y) \;=\; \begin{cases} \mathrm{P}(Y=0) = f(0) = 3/8 \\ \mathrm{P}(Y=1) = f(-1) + f(1) = 11/24 \\ \mathrm{P}(Y=4) = f(2) = 1/6. \quad \square \end{cases}$$

Continuous Case: X continuous implies Y can be continuous *or* discrete.

Example: $Y = X^2$ (clearly continuous).

Example: $Y \;=\; \begin{cases} 0 & \text{if } X < 0 \\ 1 & \text{if } X \geq 0 \end{cases}$ is *not* continuous.

Method: If Y is continuous, compute its cdf and then differentiate to get its pdf.

- $G(y) \;\equiv\; \mathrm{P}(Y \leq y) \;=\; \mathrm{P}(h(X) \leq y) \;=\; \displaystyle\int_{x|h(x)\leq y} f(x)\,dx.$
- $g(y) \;=\; \dfrac{d}{dy} G(y).$

Example: Suppose that X has pdf $f(x) = |x|$, $-1 \leq x \leq 1$. Find the pdf of the random variable $Y = h(X) = X^2$. Let's first obtain the cdf of Y.

$$G(y) \;=\; \mathrm{P}(Y \leq y) \;=\; \mathrm{P}(X^2 \leq y) \;=\; \begin{cases} 0 & \text{if } y \leq 0 \\ 1 & \text{if } y \geq 1 \\ (\star) & \text{if } 0 < y < 1, \end{cases}$$

where

$$(\star) \;=\; \mathrm{P}(-\sqrt{y} \leq X \leq \sqrt{y}) \;=\; \int_{-\sqrt{y}}^{\sqrt{y}} f(x)\,dx \;=\; \int_{-\sqrt{y}}^{\sqrt{y}} |x|\,dx \;=\; y.$$

Thus,

$$G(y) \;=\; \begin{cases} 0 & \text{if } y \leq 0 \\ 1 & \text{if } y \geq 1 \\ y & \text{if } 0 < y < 1. \end{cases}$$

This implies

$$g(y) \;=\; G'(y) \;=\; \begin{cases} 0 & \text{if } y \leq 0 \text{ or } y \geq 1 \\ 1 & \text{if } 0 < y < 1. \end{cases}$$

This is the Unif(0,1) distribution! $\square$

2.8.2 Adolescent Inverse Transform Theorem Examples

In this subsection, we'll discuss a terrific result that has lots of applications. First, a motivating example.

Example: Suppose $U \sim \text{Unif}(0,1)$. Find the pdf of $Y = -\ell\text{n}(1-U)$. Using the usual recipe, we have

$$\begin{aligned} G(y) &= \text{P}(Y \le y) \\ &= \text{P}(-\ell\text{n}(1-U) \le y) \\ &= \text{P}(U \le 1 - e^{-y}) \\ &= \int_0^{1-e^{-y}} f(u)\,du \\ &= 1 - e^{-y} \quad (\text{since } f(u) = 1). \end{aligned}$$

Thus, $g(y) = G'(y) = e^{-y}$, $y > 0$, so that $Y \sim \text{Exp}(\lambda = 1)$. □

Wow! We plugged a uniform into a simple equation and obtained an exponential! In the following discussion, we'll generalize this result so that we can obtain almost *any* reasonable continuous random variable on demand merely by plugging a uniform into an appropriate equation.

Inverse Transform Theorem:[4] Suppose X is a continuous random variable having cdf $F(x)$. Then the *random variable* $F(X) \sim \text{Unif}(0,1)$. (Note that $F(X)$ is a RV, while $F(x)$ is a real function of x.)

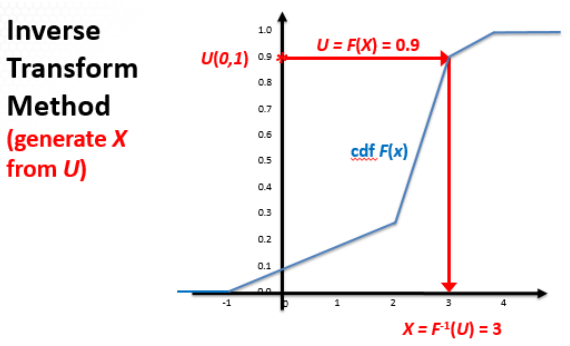
Proof: Let $Y = F(X)$. Then the cdf of Y is

$$\begin{aligned} G(y) &= \text{P}(Y \le y) \;=\; \text{P}(F(X) \le y) \\ &= \text{P}(F^{-1}(F(X)) \le F^{-1}(y)) \;\; (\text{the cdf is monotone increasing}) \\ &= \text{P}(X \le F^{-1}(y)) \;\; (F^{-1} \text{ and } F \text{ go poof!}) \\ &= F(F^{-1}(y)) \;\; (F(x) \text{ is the cdf of } X) \\ &= y. \quad \text{Uniform!} \quad \square \end{aligned}$$

Remark: What a nice theorem! It applies to any continuous random variable X!

Corollary: $X = F^{-1}(U)$, so you can plug a Unif(0,1) RV into the inverse cdf to generate a realization of a RV having X's distribution.

Method: Set $F(X) = U$ and solve for $X = F^{-1}(U)$ to generate X.



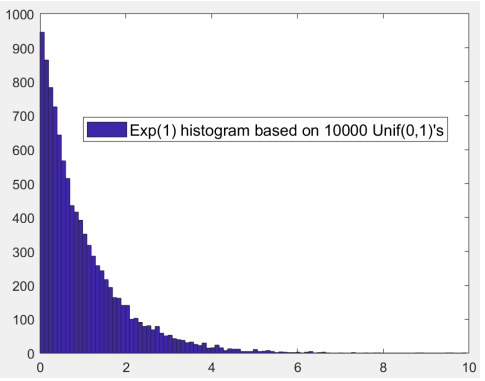
[4] Also known as the Probability Integral Transform.

Example: Suppose X is Exp(λ), so that it has cdf $F(x) = 1 - e^{-\lambda x}$. Similar to a previous example, set $F(X) = 1 - e^{-\lambda X} = U$, and generate an Exp($\lambda$) RV by solving for

$$X = F^{-1}(U) = -\frac{1}{\lambda}\ell\text{n}(1-U) \sim \text{Exp}(\lambda). \quad \square$$

Remark: So what does this result mean? If you want to obtain a beautiful Exp(λ) pdf, all you have to do is...

- Generate 10000 Unif(0,1)'s on a computer (e.g., use the `rand` function in Excel or `unifrnd` in Matlab),
- Plug the 10000 values into the equation for X above,
- Plot the histograms of the X's, and
- Admire your fine work.



Remark: This trick has tremendous applications in the field of computer simulation, where we often need to generate random variables such as customer interarrival times, resource service times, machine breakdown times, etc.

2.8.3 Grown-Up Honors Examples

Here's another, more-direct way to find the pdf of a function of a continuous random variable.

Honors Theorem: Suppose that $Y = h(X)$ is a monotonic function of a continuous RV X having pdf $f(x)$ and cdf $F(x)$. Then the pdf $g(y)$ of Y is given by

$$g(y) \;=\; f\big(h^{-1}(y)\big)\left|\frac{dh^{-1}(y)}{dy}\right|.$$

Proof: By definition, we have

$$\begin{aligned} g(y) &= \frac{d}{dy}G(y) \;=\; \frac{d}{dy}\mathrm{P}(Y \le y) \\ &= \frac{d}{dy}\mathrm{P}(h(X) \le y) \\ &= \frac{d}{dy}\mathrm{P}\big(X \le h^{-1}(y)\big) \quad (h(x) \text{ is monotone}) \\ &= \frac{d}{dy}F\big(h^{-1}(y)\big) \\ &= f\big(h^{-1}(y)\big)\left|\frac{dh^{-1}(y)}{dy}\right| \quad \text{(chain rule).} \quad \square \end{aligned}$$

Example: Suppose that $f(x) = 3x^2$, $0 < x < 1$. Let $Y = h(X) = X^{1/2}$, which is monotone increasing. Then the pdf of Y is

$$\begin{aligned} g(y) &= f\big(h^{-1}(y)\big)\left|\frac{dh^{-1}(y)}{dy}\right| \;=\; f(y^2)\left|\frac{d(y^2)}{dy}\right| \\ &= 3y^4(2y) \;=\; 6y^5, \quad 0 < y < 1. \quad \square \end{aligned}$$

Remark: Generalizations of this result are available for non-monotonic functions of a random variable, but we will save that discussion for another day.

Finally, here is why LOTUS works.

Honors Theorem: If $h(\cdot)$ is monotone increasing, then

$$\begin{aligned} \mathrm{E}[h(X)] &= \mathrm{E}[Y] \;=\; \int_{\mathbb{R}} y g(y)\,dy \\ &= \int_{\mathbb{R}} y f\big(h^{-1}(y)\big)\left|\frac{dh^{-1}(y)}{dy}\right| dy \\ &= \int_{\mathbb{R}} h(x) f(x)\left|\frac{dx}{dy}\right| dy \\ &= \int_{\mathbb{R}} h(x) f(x)\,dx. \quad \square \end{aligned}$$

Remark: Again, the monotonicity assumption isn't quite required, but it makes the proof and method easier.

2.9 Exercises

1. (§2.2) Suppose we toss a six-sided die five times. Let X denote the number of times that you see a 1, 2, or 3. Find $\mathrm{P}(X = 4)$.

2. (§2.2) Suppose $X \sim \mathrm{Pois}(2)$. Find $\mathrm{P}(X > 3)$.

3. (§2.2) My mom gives me candy from one of two boxes, A or B. Half of the candies in box A are yummy and half are icky. But the situation in box B is worse, because only 1/4 are yummy, while 3/4 are icky. The only good news is that when mom grabs a handful of candy, she's 90% likely to do it from box A.

   With this buildup in mind, suppose that mom brings me five candies from one of the boxes. If three of the candies are yummy (and two are icky), what's the chance that they came from box A?

   Hint: Try Bayes Theorem using binomial conditional probabilities.

4. (§2.3) Suppose that X is the lifetime of a lightbulb, and that $X \sim \mathrm{Exp}(2/\text{year})$.

   (a) Find the probability that the bulb will last at least one year, $\mathrm{P}(X > 1)$.

   (b) Suppose the bulb has already survived two years. What's the probability that it will survive another year, i.e., $\mathrm{P}(X > 3 \,|\, X > 2)$?

5. (§2.4) Suppose that the continuous random variable X has pdf $f(x) = x^2/9$ for $0 \le x \le 3$. What is its cdf?

6. (§2.5) I run a swanky car dealership. Today I'm going to sell no cars with probability 0.5, one car with probability 0.4, and two cars with probability 0.1. How many cars will I be expected to sell?

7. (§2.5) Suppose that the continuous random variable X has pdf $f(x) = x^2/9$ for $0 \le x \le 3$. What is its expected value?

8. (§2.5) Prove: If X is a *nonnegative* continuous random variable (always ≥ 0), then $\mathrm{E}[X] = \int_0^\infty \mathrm{P}(X > x)\, dx$.

9. (§2.5) Suppose X has the following discrete distribution.

x	-1	0	2	3
$\mathrm{P}(X = x)$	c	0.3	0.2	0.1

   (a) Find the value of c that will make the pmf sum to 1.

   (b) Calculate $\mathrm{P}(1 \le X \le 2)$.

   (c) Find the cdf $F(x)$ for all x.

   (d) Calculate $\mathrm{E}[X]$.

   (e) Calculate $\mathrm{Var}(X)$.

10. (§2.5) Suppose that X is continuous, with pdf $f(x) = cx^2$, $0 \le x \le 1$.

(a) Find the value of c that will make the pdf integrate to 1.

(b) Calculate $\mathrm{P}(0 \leq X \leq 1/2)$.

(c) Find the cdf $F(x)$ for all x.

(d) Calculate $\mathrm{E}[X]$.

(e) Calculate $\mathrm{Var}(X)$.

11. (§2.5) Let $\mathrm{E}[X] = 4$, $\mathrm{Var}(X) = 3$, and $Z = -4X + 7$. Find $\mathrm{E}[-3Z]$ and $\mathrm{Var}(-3Z)$.

12. (§2.5) Suppose that X is a discrete random variable having $X = -1$ with probability 0.2, and $X = 3$ with probability 0.8.

(a) Find $\mathrm{E}[X]$.

(b) Find $\mathrm{Var}(X)$.

(c) Find $\mathrm{E}[3 - \frac{1}{X}]$.

13. (§2.5) Suppose X is a continuous random variable with pdf $f(x) = 4x^3$, for $0 \leq x \leq 1$. Find $\mathrm{E}[1/X^2]$.

14. (§2.5) It's hot outside today, and my family wants to drink lemonade. The demand X for lemonade is known to have pdf $f(x) = 2x$, $0 \leq x \leq 1$, where x is in gallons.

(a) For some set $A \subseteq [0, 1]$, define the indicator function,

$$1_{X \in A} \equiv \begin{cases} 1 & \text{if } X \in A \\ 0 & \text{otherwise.} \end{cases}$$

Find $\mathrm{E}\big[1_{X \in [0.5,1]}\big]$.

(b) Unfortunately, I only have 1/2 gallon at the house today, so demand might not be met. The quantity $Y = (0.5 - X)^+ = \max\{0, 0.5 - X\}$ clearly denotes how much lemonade will be left at the end of the long, hot day. Find $\mathrm{E}[Y]$.

15. (§2.5) Suppose X is a continuous random variable, with mean $\mu = 2$ and variance $\sigma^2 = 10$. Consider the random function $Y = h(X) \equiv e^X$. Find an approximation for $\mathrm{E}[Y]$ using the Taylor series approach from the lesson, i.e., $\mathrm{E}[Y] \doteq h(\mu) + h''(\mu)\sigma^2/2$.

16. (§2.6) Suppose that $X = -2$ with probability 0.6, and $X = 4$ with probability 0.4. Find the moment generating function of X and use it to obtain $\mathrm{E}[X]$ and $\mathrm{Var}(X)$.

17. (§2.6) Suppose that $X \sim \mathrm{Pois}(\lambda)$. What is X's moment generating function? Use it to find $\mathrm{E}[X]$.

18. (§2.6) Suppose $X \sim \mathrm{Exp}(\lambda)$. Use the mgf of X to find $\mathrm{E}[X^k]$.

19. (§2.6) Suppose X is a discrete random variable whose only possible values are the nonnegative integers. We define the *probability generating function* (pgf) of X by

$$g_X(s) \;\equiv\; \mathrm{E}[s^X] \;=\; \sum_{k=0}^{\infty} s^k \mathrm{P}(X=k).$$

    (a) Assuming that the pgf exists, prove that $\mathrm{E}[X] = \frac{d}{ds}\, g_X(s)\,\big|_{s=1}$.
    (b) If $X \sim \mathrm{Pois}(\lambda)$, find $g_X(s)$. (Potentially helpful fact: $\sum_{k=0}^{\infty} y^k/k! = e^y$.)
    (c) Suppose that $X \sim \mathrm{Pois}(\lambda)$. Use (a) and (b) to find $\mathrm{E}[X]$.

20. (§2.6) If Y has mgf $M_Y(t) = 3e^{5t}/(3-4t)$, for $t < 3/4$, what is the distribution of Y?

21. (§2.7) Suppose that $X \sim \mathrm{Unif}(-1, 6)$. Compare the upper bound on the probability $\mathrm{P}\big(|X-\mu| \geq 1.5\sigma\big)$ obtained from Chebychev's inequality with the exact probability.

22. (§2.8) Suppose X is the result of a five-sided die toss having sides numbered $-2, -1, 0, 1, 2$. Find the probability mass function of $Y = X^2$.

23. (§2.8) Suppose $X \sim \mathrm{Unif}(1, 3)$. Find the pdf of $Z = e^{3X}$.

24. (§2.8) Suppose X has pdf $f(x) = 3x^2 e^{-x^3}$, $x \geq 0$. Find the pdf of $Z = X^2$.

25. (§2.8) Suppose X is a continuous random variable with pdf $f(x) = 2x$, for $0 < x < 1$. Find the pdf $g(y)$ of $Y = X^2$. (This may be easier than you think.)

26. (§2.8) Suppose X is a continuous random variable with pdf $f(x) = 2x$, for $0 < x < 1$. Find the pdf of $Y = \sqrt{X}$.

27. (§2.8) Suppose that X has the Pareto distribution with pdf $f(x) = \alpha x^{-(\alpha+1)}$, for $x \geq 1$ and parameter $\alpha > 1$. Prove that $Y = \ell\mathrm{n}(X) \sim \mathrm{Exp}(\alpha)$.

28. (§2.8) Computer Exercises — Random Variate Generation

    (a) Let's start out with something easy — the Unif(0,1) distribution. To generate a Unif(0,1) random variable in Excel, you simply use `rand()`. (We use Excel as an example in this exercise, but feel free to use the software of your choice.) Copy an entire column of 10,000 of these guys and make a histogram.
    (b) It's very easy to generate an Exp(1) random variable in Excel via the Inverse Transform technique. Just use

$$-\texttt{ln}(\texttt{rand}()) \quad \text{or} \quad -\texttt{ln}(1-\texttt{rand}()).$$

    In any case, generate 10,000 or so of these guys and make a nice histogram.
    (c) In Excel, you can generate a Normal(0,1) random variable using

$$\texttt{norminv}(\texttt{rand}(), 0, 1) \qquad \text{(inverse transform method)}.$$

    Generate a bunch of normals using this equation and make a histogram.

(d) As if you have nothing better to do, toss a die 10,000 times. How many times do each of the numbers come up? Approximately how many would you expect?

If you do this in Excel, try using

$$\texttt{int}(6 * \texttt{rand}()) + 1.$$

Why does this work?

(e) As if you still have nothing better to do, toss two dice 10,000 times. How many times do each of the possible sums come up? Approximately how many would you expect?

(f) Triangular distribution. Generate two columns of Unif(0,1)'s. In the third column, add up the respective entries from the previous two columns, e.g., `C1 = A1 + B1`, etc. Make a histogram of the third column. Guess what you get?

(g) Normal distribution from the Central Limit Theorem (which we'll learn about later). Generate *12* columns of Unif(0,1)'s. In the 13^{th} column, add up the respective entries from the previous 12 columns. Make a histogram of the 13^{th} column. Guess what you get this time?

29. (§2.8) Suppose that $f(x) = 4x^3$, for $0 < x < 1$. Let $Y = h(X) = X^2$, which is monotone increasing for $X > 0$. Find the pdf of Y using the "direct" equation,

$$g(y) \quad = \quad f(h^{-1}(y)) \left| \frac{dh^{-1}(y)}{dy} \right|.$$

Chapter 3

Bivariate Random Variables

In this chapter we explore what happens when we consider two random variables *simultaneously*. For example, suppose we choose a person at random and look at her height and weight (X, Y). Obviously, X and Y will be related somehow, and this relationship must be taken into account during any subsequent analysis. We'll look at various generalizations of the work from the previous chapter, as well as new concepts such as independence and correlation.

3.1 Introduction and Definitions

We consider discrete and continuous *bivariate* random variables. We'll see that things generalize naturally from our earlier discussion of univariate random variables in Chapter 2.

3.1.1 Discrete Case

Definition: If X and Y are discrete random variables, then (X, Y) is called a **jointly discrete bivariate random variable**. The **joint (or bivariate) pmf** is

$$f(x, y) \;=\; \mathrm{P}(X = x, Y = y), \quad \forall x, y.$$

Properties of $f(x, y)$:

- $0 \leq f(x, y) \leq 1, \ \forall x, y.$
- $\sum_x \sum_y f(x, y) = 1.$
- $A \subseteq \mathbb{R}^2 \Rightarrow \mathrm{P}((X, Y) \in A) = \sum \sum_{(x,y)\in A} f(x, y).$

Example: Consider three sox in a box (numbered 1,2,3). Suppose that we draw two sox at random without replacement.

Let X denote the number of the first sock, and Y the number of the second sock. The joint pmf $f(x, y)$ is depicted in the following table.

$f(x,y)$	$X=1$	$X=2$	$X=3$	$\mathrm{P}(Y=y)$
$Y=1$	0	1/6	1/6	1/3
$Y=2$	1/6	0	1/6	1/3
$Y=3$	1/6	1/6	0	1/3
$\mathrm{P}(X=x)$	1/3	1/3	1/3	1

We say (and will formally define later) that the bottom row $f_X(x) \equiv \mathrm{P}(X = x)$ defines the **marginal pmf of** X; and the right-most column $f_Y(y) \equiv \mathrm{P}(Y = y)$ defines the **marginal pmf of** Y. Of course, everything adds up to 1 (bottom-right entry), which we find satisfying and at the same time reassuring.

Continuing the example, note that by the Law of Total Probability,

$$\mathrm{P}(X = 1) \ = \ \sum_{y=1}^{3} \mathrm{P}(X = 1, Y = y) \ = \ 1/3.$$

In addition,

$$\begin{aligned} &\mathrm{P}(X \geq 2, Y \geq 2) \\ &= \ \sum_{x\geq 2} \sum_{y\geq 2} f(x, y) \\ &= \ f(2, 2) + f(2, 3) + f(3, 2) + f(3, 3) \\ &= \ 0 + 1/6 + 1/6 + 0 \ = \ 1/3. \quad \square \end{aligned}$$

3.1.2 Continuous Case

Definition: If X and Y are continuous random variables, then (X, Y) is a **jointly continuous bivariate random variable** if there exists a magic function $f(x, y)$, such that

- $f(x, y) \geq 0, \ \forall x, y.$
- $\int \int_{\mathbb{R}^2} f(x, y)\, dx\, dy = 1.$

- If $A \subseteq \mathbb{R}^2$, then $\mathrm{P}(A) = \mathrm{P}((X,Y) \in A) = \int\int_A f(x,y)\,dx\,dy$. $\mathrm{P}(A)$ is the volume between $f(x,y)$ and A.

In this case, $f(x,y)$ is called the **joint pdf of** (X,Y). Think of

$$f(x,y)\,dx\,dy \doteq \mathrm{P}(x < X < x+dx, y < Y < y+dy).$$

It is easy to see how all of this generalizes the one-dimensional pdf, $f(x)$.

Example: Choose a point (X,Y) at random in the interior of the circle inscribed in the unit square, i.e., $C \equiv (x-\frac{1}{2})^2 + (y-\frac{1}{2})^2 \leq \frac{1}{4}$. Since the area of the circle is $\pi/4$, and we are selecting a point uniformly in that area, the joint pdf of (X,Y) is

$$f(x,y) = \begin{cases} 4/\pi & \text{if } (x,y) \in C \\ 0 & \text{otherwise.} \end{cases}$$

We can use this formulation to calculate probabilities. For instance,

$$\begin{aligned} \mathrm{P}(X \geq 1/2, Y \geq 1/2) &= \int_{1/2}^{1}\int_{1/2}^{1} f(x,y)\,dx\,dy \\ &= \int\int_{x\geq 1/2,\, y\geq 1/2,\, (x,y)\in C} (4/\pi)\,dx\,dy \\ &= 1/4, \end{aligned}$$

where we just used a symmetry argument rather than attempt to formally integrate (which looks messier than it actually is). □

Application: Toss n darts randomly into the unit square. The probability that any individual dart will land in the circle is $\pi/4$. It stands to reason that the proportion of darts, $\hat{p}_n$, that land in the circle will be approximately $\pi/4$. So you can use $4\hat{p}_n$ to estimate π! For instance, if we toss 1000 darts and 752 land in the circle, then our estimate for π is $4(752/1000) = 3.08$. □

Example: Suppose that

$$f(x,y) = \begin{cases} 4xy & \text{if } 0 \leq x \leq 1,\ 0 \leq y \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

Let's find the probability (volume) of the region $0 \leq y \leq 1 - x^2$.

$$V = \int_0^1 \int_0^{1-x^2} 4xy\,dy\,dx = \int_0^1 \int_0^{\sqrt{1-y}} 4xy\,dx\,dy = 1/3.$$

Moral: If you are careful with limits, you can check your answer! □

3.1.3 Bivariate cdf's

Definition: The **joint (bivariate) cdf** of X and Y is $F(x,y) \equiv P(X \leq x, Y \leq y)$, for all x, y. In other words,

$$F(x,y) = \begin{cases} \sum\sum_{s\leq x,\, t\leq y} f(s,t) & \text{discrete} \\ \int_{-\infty}^{y}\int_{-\infty}^{x} f(s,t)\,ds\,dt & \text{continuous.} \end{cases}$$

Properties of cdf's:

- $F(x, y)$ is non-decreasing in both x and y.
- $\lim_{x\to-\infty} F(x, y) = \lim_{y\to-\infty} F(x, y) = 0$.
- $\lim_{x\to\infty} F(x, y) \equiv F_Y(y) = \mathrm{P}(Y \le y)$ **(marginal cdf of Y).**
- $\lim_{y\to\infty} F(x, y) \equiv F_X(x) = \mathrm{P}(X \le x)$ **(marginal cdf of X).**
- $\lim_{x\to\infty} \lim_{y\to\infty} F(x, y) = 1$.

In the continuous case, it is easy to go from cdf's to pdf's — just take the derivative(s):

- one-dimension: $f(x) = F'(x) = \frac{d}{dx} \int_{-\infty}^{x} f(t)\, dt$.
- two-dimensions: $f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y) = \frac{\partial^2}{\partial x \partial y} \int_{-\infty}^{x} \int_{-\infty}^{y} f(s, t)\, dt\, ds$.

Example: Suppose

$$F(x, y) = \begin{cases} 1 - e^{-x} - e^{-y} + e^{-(x+y)} & \text{if } x \ge 0,\ y \ge 0 \\ 0 & \text{if } x < 0 \text{ or } y < 0. \end{cases}$$

The marginal cdf of X is

$$F_X(x) = \lim_{y\to\infty} F(x, y) = \begin{cases} 1 - e^{-x} & \text{if } x \ge 0 \\ 0 & \text{if } x < 0. \end{cases}$$

The joint pdf is

$$\begin{aligned} f(x, y) &= \frac{\partial^2}{\partial x \partial y} F(x, y) \\ &= \frac{\partial}{\partial y}(e^{-x} - e^{-y} e^{-x}) \\ &= \begin{cases} e^{-(x+y)} & \text{if } x \ge 0,\ y \ge 0 \\ 0 & \text{if } x < 0 \text{ or } y < 0. \end{cases} \quad \square \end{aligned}$$

3.1.4 Marginal Distributions

We can use information from $f(x, y)$ and/or $F(x, y)$ to obtain the individual distributions of X and Y.

Definition: If X and Y are jointly *discrete*, then the **marginal pmf's** of X and Y are, respectively,

$$f_X(x) \;=\; \mathrm{P}(X = x) \;=\; \sum_y \mathrm{P}(X = x, Y = y) \;=\; \sum_y f(x, y)$$

and

$$f_Y(y) \;=\; \mathrm{P}(Y = y) \;=\; \sum_x \mathrm{P}(X = x, Y = y) \;=\; \sum_x f(x, y).$$

Remark: These definitions are just applications of the Law of Total Probability.

Example: Consider the following joint pmf $f(x, y) = \mathrm{P}(X = x, Y = y)$.

$f(x,y)$	$X=1$	$X=2$	$X=3$	$\mathrm{P}(Y=y)$
$Y=40$	0.01	0.07	0.12	0.2
$Y=60$	0.29	0.03	0.48	0.8
$\mathrm{P}(X=x)$	0.3	0.1	0.6	1

For example, by Total Probability,

$$\mathrm{P}(X = 1) = \mathrm{P}(X = 1, Y = \text{ any number}) = 0.3. \quad \square$$

Example: Consider another discrete pmf.

$f(x,y)$	$X=1$	$X=2$	$X=3$	$\mathrm{P}(Y=y)$
$Y=40$	0.06	0.02	0.12	0.2
$Y=60$	0.24	0.08	0.48	0.8
$\mathrm{P}(X=x)$	0.3	0.1	0.6	1

Remark: Hmmm.... Compared to the last example, this has the *same marginals* but a *different joint* distribution! That's because the joint distribution contains *much more information* than just the marginals, and therefore, many joint distributions can spawn the same marginals.

Definition: If X and Y are jointly *continuous*, then the **marginal pdf's** of X and Y are, respectively,

$$f_X(x) \;=\; \int_{\mathbb{R}} f(x, y)\, dy \quad \text{and} \quad f_Y(y) \;=\; \int_{\mathbb{R}} f(x, y)\, dx.$$

Remark: These definitions are clearly the continuous analogs of the discrete case (we now have integrals instead of sums). Note that, for instance, $f_X(x)$ is a function of x alone, since the y's have all been integrated away.

Example: Suppose that (X, Y) have bivariate pdf

$$f(x,y) \;=\; \begin{cases} e^{-(x+y)} & \text{if } x \geq 0,\, y \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

Then the marginal pdf of X is

$$f_X(x) \;=\; \int_{\mathbb{R}} f(x,y)\,dy \;=\; \int_0^\infty e^{-(x+y)}\,dy \;=\; \begin{cases} e^{-x} & \text{if } x \geq 0 \\ 0 & \text{otherwise,} \end{cases}$$

so that $X \sim \text{Exp}(1)$. It is also easy to show that $Y \sim \text{Exp}(1)$. □

Here is a trickier example that we will occasionally revisit.

Example: Consider the joint pdf

$$f(x,y) \;=\; \frac{21}{4}x^2 y, \quad \text{if } x^2 \leq y \leq 1,$$

and note the non-rectangular ***funny limits*** where the pdf is positive, i.e., $x^2 \leq y \leq 1$**.** After a bit of careful integration, we have

$$f_X(x) \;=\; \int_{\mathbb{R}} f(x,y)\,dy \;=\; \int_{x^2}^1 \frac{21}{4}x^2 y\,dy \;=\; \frac{21}{8}x^2(1-x^4), \quad \text{if } -1 \leq x \leq 1; \quad \text{and}$$

$$f_Y(y) \;=\; \int_{\mathbb{R}} f(x,y)\,dx \;=\; \int_{-\sqrt{y}}^{\sqrt{y}} \frac{21}{4}x^2 y\,dx \;=\; \frac{7}{2}y^{5/2}, \quad \text{if } 0 \leq y \leq 1. \quad \square$$

3.2 Conditional Distributions

Before we get into any new work, recall the old definition of conditional probability based on the events A and B: $\text{P}(A|B) = \text{P}(A \cap B)/\text{P}(B)$, if $\text{P}(B) > 0$. In this section, we'll apply the concept to probability distributions. To do so, let's establish the proper analogy using events involving random variables.

Suppose that X and Y are jointly discrete RV's. Then if $\text{P}(X = x) > 0$,

$$\text{P}(Y = y \,|\, X = x) \;=\; \frac{\text{P}(X = x \cap Y = y)}{\text{P}(X = x)} \;=\; \frac{f(x,y)}{f_X(x)}.$$

Thus, for instance, $\text{P}(Y = y \,|\, X = 2)$ defines the conditional probability distribution of Y, given the information that $X = 2$.

Now we are ready to proceed with the general definition.

Definition: If $f_X(x) > 0$, then the **conditional pmf/pdf of Y given $X = x$** is

$$f_{Y|X}(y \,|\, x) \;\equiv\; \frac{f(x,y)}{f_X(x)}, \quad \text{for all } y.$$

Remarks:

- To save valuable ink, we usually just write $f(y\,|\,x)$ instead of $f_{Y|X}(y\,|\,x)$.
- A conditional pmf/pdf is a legitimate pmf/pdf. Thus, if we regard x as fixed, we have $\sum_y f(y\,|\,x) = 1$ for the discrete case, and $\int_{\mathbb{R}} f(y\,|\,x)\,dy = 1$ for the continuous case.
- Of course, by notational symmetry, we have $f_{X|Y}(x\,|\,y) = f(x\,|\,y) = f(x,y)/f_Y(y)$.

Example: Consider the (discrete) joint pmf $f(x,y) = \mathrm{P}(X = x, Y = y)$ given below.

$f(x,y)$	$X=1$	$X=2$	$X=3$	$f_Y(y)$
$Y=40$	0.01	0.07	0.12	0.2
$Y=60$	**0.29**	**0.03**	**0.48**	**0.8**
$f_X(x)$	0.3	0.1	0.6	1

Then, for example,

$$f(x\,|\,y=60) \;=\; \frac{f(x,60)}{f_Y(60)} \;=\; \frac{f(x,60)}{0.8} \;=\; \begin{cases} \frac{29}{80} & \text{if } x=1 \\ \frac{3}{80} & \text{if } x=2 \\ \frac{48}{80} & \text{if } x=3, \end{cases}$$

which represents the updated pmf of X given that $Y = 60$. □

Old Continuous Example: Suppose that

$$f(x,y) \;=\; \frac{21}{4}x^2 y, \quad \text{if } x^2 \le y \le 1,$$

where we again note the funny limits and recall that

$$f_X(x) \;=\; \frac{21}{8}x^2(1-x^4), \quad \text{if } -1 \le x \le 1; \quad \text{and}$$

$$f_Y(y) \;=\; \frac{7}{2}y^{5/2}, \quad \text{if } 0 \le y \le 1.$$

Then the conditional pdf of Y given $X = x$ is

$$f(y\,|\,x) \;=\; \frac{f(x,y)}{f_X(x)} \;=\; \frac{\frac{21}{4}x^2 y}{\frac{21}{8}x^2(1-x^4)} \;=\; \frac{2y}{1-x^4}, \quad \text{if } x^2 \le y \le 1,$$

where we point out that the funny limits have been preserved (since x is still in play). Note that in the general expression for $f(y\,|\,x)$, the quantity $2/(1-x^4)$ is a *constant* with respect to y, and we can check to see that $f(y\,|\,x)$ is a legitimate conditional pdf:

$$\int_{\mathbb{R}} f(y\,|\,x)\,dy \;=\; \int_{x^2}^{1} \frac{2y}{1-x^4}\,dy \;=\; 1.$$

Continuing the example, let's see what happens when we have the specific conditional information that $X = 1/2$, i.e.,

$$f\big(y\,|\,1/2\big) \;=\; \frac{2y}{1-\frac{1}{16}} \;=\; \frac{32y}{15}, \quad \text{if } \tfrac{1}{4} \le y \le 1.$$

This conditional pdf allows us to calculate any relevant probabilities, e.g.,

$$\mathrm{P}\left(\frac{3}{4} \le Y \le 1 \,\middle|\, X = \frac{1}{2}\right) \;=\; \int_{3/4}^{1} f(y \,|\, 1/2)\, dy \;=\; \int_{3/4}^{1} \frac{32y}{15}\, dy \;=\; \frac{7}{15}. \quad \square$$

Generic Example: Given $f_X(x)$ and $f(y \,|\, x)$, find $\mathrm{E}[Y]$.

The game plan will be to find $f(x, y) = f_X(x) f(y \,|\, x)$, then $f_Y(y) = \int_{\mathbb{R}} f(x, y)\, dx$, and finally $\mathrm{E}[Y] = \int_{\mathbb{R}} y f_Y(y)\, dy$. For instance, ...

Example: Suppose that $f_X(x) = 2x$, for $0 < x < 1$. Given that $X = x$, suppose $Y \,|\, x \sim \mathrm{Unif}(0, x)$. Now find $\mathrm{E}[Y]$.

Solution: $Y \,|\, x \sim \mathrm{Unif}(0, x)$ implies that $f(y \,|\, x) = 1/x$, for $0 < y < x$ (note the funny limits). So

$$\begin{aligned} f(x, y) &= f_X(x) f(y \,|\, x) \\ &= 2x \cdot \frac{1}{x}, \quad \text{for } 0 < x < 1 \text{ and } 0 < y < x \\ &= 2, \quad \text{for } 0 < y < x < 1 \quad \text{(still have the funny limits)}. \end{aligned}$$

Thus,

$$f_Y(y) \;=\; \int_{\mathbb{R}} f(x, y)\, dx \;=\; \int_{y}^{1} 2\, dx \;=\; 2(1 - y),\; 0 < y < 1,$$

and then

$$\mathrm{E}[Y] \;=\; \int_{\mathbb{R}} y f_Y(y)\, dy \;=\; 2 \int_{0}^{1} (y - y^2)\, dy \;=\; 1/3. \quad \square$$

3.3 Independent Random Variables

Just as two events can be dependent, we can define independence with respect to random variables.

3.3.1 Definition and Basic Results

Recall that two events A and B are independent if $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$. Then

$$\mathrm{P}(A \,|\, B) \;=\; \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} \;=\; \frac{\mathrm{P}(A)\mathrm{P}(B)}{\mathrm{P}(B)} \;=\; \mathrm{P}(A).$$

And similarly, $\mathrm{P}(B \,|\, A) = \mathrm{P}(B)$.

Now we want to define independence for RV's, in which case the outcome of X doesn't influence the outcome of Y, and vice versa.

Definition: X and Y are **independent** random variables if

$$f(x, y) \;=\; f_X(x) f_Y(y) \quad \text{for all } x, y.$$

It turns out that the following are equivalent definitions:

$$F(x,y) \;=\; F_X(x)F_Y(y),\;\; \forall x,y; \quad \text{or}$$

$$\mathrm{P}(X \le x, Y \le y) \;=\; \mathrm{P}(X \le x)\mathrm{P}(Y \le y),\;\; \forall x,y.$$

If X and Y aren't independent, then they're **dependent**.

Intuitive Theorem: X and Y are independent iff $f(y \,|\, x) = f_Y(y)$, $\forall x,y$.

Proof: By assumption and then the definition of conditional, we have

$$f_Y(y) = f(y|x) = \frac{f(x,y)}{f_X(x)} \quad \Leftrightarrow \quad f(x,y) = f_X(x)f_Y(y) \;\Leftrightarrow\; \text{indep.} \quad \square$$

Similarly, X and Y independent implies $f(x \,|\, y) = f_X(x)$.

In other words, the outcomes of X and Y don't affect each other, e.g., the temperature on Mars and IBM's current stock price are unrelated.

Example: Consider the (discrete) bivariate pmf $f(x,y) = \mathrm{P}(X = x, Y = y)$ given by the following table.

$f(x,y)$	$X=1$	$X=2$	$f_Y(y)$
$Y=2$	0.12	0.28	0.4
$Y=3$	0.18	0.42	0.6
$f_X(x)$	0.3	0.7	1

Then X and Y are independent, since $f(x,y) = f_X(x)f_Y(y)$, $\forall x,y$. $\square$

Example: Sometimes, you have to be a little careful to make sure that *every* (x,y) pair behaves.

$f(x,y)$	$X=1$	$X=5$	$X=6$	$f_Y(y)$
$Y=2$	0.12	0.28	0	0.4
$Y=3$	0.18	0.32	0.10	0.6
$f_X(x)$	0.3	0.6	0.1	1

Even though $f(1,2)$ obediently factors, X and Y are *not* independent since, e.g., $f(5,2) \neq f_X(5)f_Y(2)$. ☹

Example: Consider the (continuous) bivariate pdf $f(x,y) = 6xy^2$, $0 \le x \le 1$, $0 \le y \le 1$. After some work (which can be avoided by the next theorem), we can derive

$$f_X(x) \;=\; \int_{\mathbb{R}} f(x,y)\,dy \;=\; 2x,\; 0 \le x \le 1; \quad \text{and}$$

$$f_Y(y) \;=\; \int_{\mathbb{R}} f(x,y)\,dx \;=\; 3y^2,\; 0 \le y \le 1.$$

Then X and Y are independent, since $f(x,y) = f_X(x)f_Y(y)$, $\forall x,y$. $\square$

Here is an easy way to tell if X and Y are independent (which we won't prove) that avoids the intermediate and possibly tedious calculation of $f_X(x)$ and $f_Y(y)$.

Theorem: X and Y are independent iff $f(x,y) = a(x)b(y)$, $\forall x, y$, for some functions $a(x)$ and $b(y)$ (not necessarily pdf's).

So if $f(x,y)$ factors into separate functions of x and y, then X and Y are independent.

But if there are non-rectangular *funny limits*, this makes factoring into marginals impossible. In that case, X and Y will be dependent — *watch out*!

In particular, for some (x,y), the funny limits might allow $f_X(x) > 0$ and $f_Y(y) > 0$, but $f(x,y) = 0$. This immediately messes up the factorization, because $f_X(x)f_Y(y) \neq f(x,y)$.

Example: $f(x,y) = 6xy^2$, for $0 \leq x \leq 1$, $0 \leq y \leq 1$.

Take $a(x) = 6x$, $0 \leq x \leq 1$, and $b(y) = y^2$, $0 \leq y \leq 1$. Thus, X and Y are independent (as above). ☺

Example: $f(x,y) = \frac{21}{4}x^2y$, for $x^2 \leq y \leq 1$ (recall that the calculations of $f_X(x)$ and $f_Y(y)$ were nasty).

Oops! Funny limits imply dependent! ☹

Example: $f(x,y) = \frac{c}{x+y}$, for $1 \leq x \leq 2$, $1 \leq y \leq 3$.

Unfortunately, you can't factor $f(x,y)$ into functions of x and y separately. Thus, X and Y are *not* independent. ☹

3.3.2 Consequences of Independence

Now that we can figure out if X and Y are independent, what can we do with that knowledge? Let's first introduce a useful tool.

Definition/Theorem (two-dimensional Law of the Unconscious Statistician): Let $h(X,Y)$ be a function of the random variables X and Y. Then

$$\mathrm{E}[h(X,Y)] = \begin{cases} \sum_x \sum_y h(x,y) f(x,y) & \text{discrete case} \\ \int_{\mathbb{R}} \int_{\mathbb{R}} h(x,y) f(x,y)\, dx\, dy & \text{continuous case.} \end{cases}$$

Examples: Just like its one-dimensional little brother, 2-D LOTUS can help us calculate all sorts of good stuff, some of which we will discuss in what follows. Here are some samples of its applicability in the continuous case.

- $\mathrm{E}[X+Y] = \int_{\mathbb{R}} \int_{\mathbb{R}} (x+y) f(x,y)\, dx\, dy$
- $\mathrm{E}[XY] = \int_{\mathbb{R}} \int_{\mathbb{R}} xy f(x,y)\, dx\, dy$
- $\mathrm{E}[X^2 \sin(Y)] = \int_{\mathbb{R}} \int_{\mathbb{R}} x^2 \sin(y) f(x,y)\, dx\, dy$

- Consider our old friend, $f(x,y) = \frac{21}{4}x^2 y$ if $x^2 \le y \le 1$. Then

$$\mathrm{E}\big[Y/X^2\big] \;=\; \int_{\mathbb{R}}\int_{\mathbb{R}} (y/x^2) f(x,y)\,dx\,dy \;=\; \int_{-1}^{1}\int_{x^2}^{1} \frac{21}{4} y^2\,dy\,dx \;=\; 3. \quad \square$$

One of the most-important and fundamental consequences of this two-dimensional LOTUS is the intuitive fact that expected values add up.

Theorem: *Whether or not* X and Y are independent, we have

$$\boxed{\mathrm{E}[X+Y] \;=\; \mathrm{E}[X] + \mathrm{E}[Y].}$$

Proof (continuous case):

$$\begin{aligned}
\mathrm{E}[X+Y] &= \int_{\mathbb{R}}\int_{\mathbb{R}} (x+y) f(x,y)\,dx\,dy \quad \text{(two-dimensional LOTUS)} \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} x f(x,y)\,dx\,dy + \int_{\mathbb{R}}\int_{\mathbb{R}} y f(x,y)\,dx\,dy \\
&= \int_{\mathbb{R}} x \int_{\mathbb{R}} f(x,y)\,dy\,dx + \int_{\mathbb{R}} y \int_{\mathbb{R}} f(x,y)\,dx\,dy \\
&= \int_{\mathbb{R}} x f_X(x)\,dx + \int_{\mathbb{R}} y f_Y(y)\,dy \\
&= \mathrm{E}[X] + \mathrm{E}[Y]. \quad \square
\end{aligned}$$

We can generalize this result to more than two RV's.

Corollary: If $X_1, X_2, \ldots, X_n$ are random variables (independent or not), then

$$\mathrm{E}\Big[\sum_{i=1}^{n} X_i\Big] \;=\; \sum_{i=1}^{n} \mathrm{E}[X_i].$$

Proof: Induction. $\square$

Sometimes, independence *does* matter...

Theorem: If X and Y are ***independent***, then $\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y]$.

Proof (continuous case):

$$\begin{aligned}
\mathrm{E}[XY] &= \int_{\mathbb{R}}\int_{\mathbb{R}} xy f(x,y)\,dx\,dy \quad \text{(two-dimensional LOTUS)} \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} xy f_X(x) f_Y(y)\,dx\,dy \quad (X \text{ and } Y \text{ are independent}) \\
&= \left(\int_{\mathbb{R}} x f_X(x)\,dx\right)\left(\int_{\mathbb{R}} y f_Y(y)\,dy\right) \\
&= \mathrm{E}[X]\mathrm{E}[Y]. \quad \square
\end{aligned}$$

Remark: The above theorem is *not* necessarily true if X and Y are *dependent*. See the discussion on covariance in §3.5 below.

The next theorem is a very important consequence of independence — it states that you can add up the variances if two RV's are independent.

Theorem: If X and Y are ***independent***, then

$$\boxed{\text{Var}(X+Y) \;=\; \text{Var}(X) + \text{Var}(Y).}$$

Proof:

$$\begin{aligned}
\text{Var}(X+Y) \;&=\; \text{E}[(X+Y)^2] - (\text{E}[X+Y])^2 \\
&= \text{E}[X^2 + 2XY + Y^2] - \big(\text{E}[X] + \text{E}[Y]\big)^2 \\
&= \text{E}[X^2] + 2\text{E}[XY] + \text{E}[Y^2] - \Big\{\big(\text{E}[X]\big)^2 + 2\text{E}[X]\text{E}[Y] + \big(\text{E}[Y]\big)^2\Big\} \\
&= \text{E}[X^2] + 2\text{E}[X]\text{E}[Y] + \text{E}[Y^2] - \big(\text{E}[X]\big)^2 - 2\text{E}[X]\text{E}[Y] - \big(\text{E}[Y]\big)^2 \\
&\quad \text{(since } X \text{ and } Y \text{ are independent)} \\
&= \text{E}[X^2] - \big(\text{E}[X]\big)^2 + \text{E}[Y^2] - \big(\text{E}[Y]\big)^2. \quad \square
\end{aligned}$$

Remark: The assumption of independence really is important here. If X and Y aren't independent, then the result might not hold! Again, see the upcoming discussion on covariance.

We can generalize the variance of the sum result...

Corollary: If $X_1, X_2, \ldots, X_n$ are *independent* random variables, then

$$\text{Var}\left(\sum_{i=1}^{n} a_i X_i + b\right) \;=\; \sum_{i=1}^{n} a_i^2 \,\text{Var}(X_i).$$

Proof: Induction. $\square$

Corollary (of Corollary): If $X_1, X_2, \ldots, X_n$ are *independent*, then

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) \;=\; \sum_{i=1}^{n} \text{Var}(X_i).$$

So far, all of the results in this subsection have looked at how independence affects (or doesn't affect) the expected value and variance of certain functions of random variables such as sums. We now present a useful result in which independence is required for more-challenging purposes.

Theorem: If X and Y are *independent* continuous RV's, then (using the obvious pdf and cdf notation)

$$\boxed{\text{P}(Y \le X) \;=\; \int_{\mathbb{R}} F_Y(x) f_X(x)\, dx.}$$

Proof: By definition of the event $\{Y < X\}$, we have

$$\begin{aligned} \mathrm{P}(Y \le X) &= \int\int_{y \le x} f(x,y)\,dx\,dy \\ &= \int_{\mathbb{R}} \int_{-\infty}^{x} f_Y(y) f_X(x)\,dy\,dx \quad (\text{since } X, Y \text{ are independent}) \\ &= \int_{\mathbb{R}} \mathrm{P}(Y \le x) f_X(x)\,dx. \quad \square \end{aligned}$$

Example: Suppose $X \sim \mathrm{Exp}(\alpha)$ is the time until the next male driver shows up at a parking lot (at rate α / hour), independent of that, $Y \sim \mathrm{Exp}(\beta)$ is the time for the next female driver (at rate β / hour). Then

$$\begin{aligned} \mathrm{P}(Y \le X) &= \int_{\mathbb{R}} F_Y(x) f_X(x)\,dx \\ &= \int_0^\infty (1 - e^{-\beta x}) \alpha e^{-\alpha x}\,dx \\ &= \frac{\beta}{\alpha + \beta}. \quad \square \end{aligned}$$

3.3.3 Random Samples

The concept of a random sample is easy to understand and extremely useful in practice.

Definition: The random variables $X_1, X_2, \ldots, X_n$ form a **random sample** if (i) the X_i's are all *independent*, and (ii) have the *same pmf/pdf*, say, $f(x)$.

Notation: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$ (**iid** = "independent and identically distributed").

Example/Theorem: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f(x)$, with $\mathrm{E}[X_i] = \mu$, and $\mathrm{Var}(X_i) = \sigma^2$. Define the **sample mean** as $\bar{X} \equiv \sum_{i=1}^n X_i/n$. Then

$$\mathrm{E}[\bar{X}] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^n X_i\right] = \frac{1}{n}\sum_{i=1}^n \mathrm{E}[X_i] = \frac{1}{n}\sum_{i=1}^n \mu = \frac{n\mu}{n} = \mu.$$

So the mean of $\bar{X}$ is the same as the mean of X_i. $\square$

Meanwhile, how about the *variance* of the sample mean?

$$\begin{aligned} \mathrm{Var}(\bar{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) \quad (X_i\text{'s are independent}) \\ &= \frac{1}{n^2}\sum_{i=1}^n \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \end{aligned}$$

So the mean of $\bar{X}$ is the same as the mean of X_i, but the *variance decreases*! This makes $\bar{X}$ a great ***estimator*** for μ (which is usually unknown in practice). The result is referred to as the **Law of Large Numbers**. Stay tuned for more about this beginning in Chapter 4. □

3.4 Extensions of Conditional Distributions

We've seen how conditional distributions reflect the changes in a pmf/pdf upon receipt of new information. We'll now study extensions of this important concept. Here's what's coming up in the following subsections. §3.4.1 defines conditional expectation and provides introductory examples. §3.4.2 presents the "double expectation" theorem that will allow us to indirectly calculate a surprising variety of expected values. §3.4.3 gives several honors problems involving conditional distributions.

3.4.1 Conditional Expectation

To begin, let's recall the usual definition of expectation. For instance, what's the expected weight of a male from some population (could be any height)?

$$\mathrm{E}[Y] \;=\; \begin{cases} \sum_y y f(y) & \text{discrete} \\ \int_{\mathbb{R}} y f(y)\, dy & \text{continuous} \end{cases}$$

Now, instead suppose we're interested in the *conditional* expected value, e.g., the mean weight of a male who is specifically 6' tall. One would surmise that tall people tend to have a higher expected weight than shorter people.

In any case, let $f(y \,|\, x)$ be the conditional pmf/pdf of Y given $X = x$.

Definition: The **conditional expectation** of Y given $X = x$ is

$$\mathrm{E}[Y|x] \;\equiv\; \mathrm{E}[Y|X = x] \;\equiv\; \begin{cases} \sum_y y f(y \,|\, x) & \text{discrete} \\ \int_{\mathbb{R}} y f(y \,|\, x)\, dy & \text{continuous.} \end{cases}$$

Note that $\mathrm{E}[Y|X = x]$ is a real function of x.

Discrete Example: Consider the following joint pmf.

$f(x, y)$	$X = 0$	$X = 3$	$f_Y(y)$
$Y = 2$	0.11	**0.34**	0.45
$Y = 5$	0.00	**0.05**	0.05
$Y = 10$	0.29	**0.21**	0.50
$f_X(x)$	0.40	**0.60**	1

The *unconditional* expectation is $\mathrm{E}[Y] = \sum_y y f_Y(y) = 6.15$. But conditional on,

e.g., $X = 3$, we have

$$f(y \,|\, x = 3) \;=\; \frac{f(3,y)}{f_X(3)} \;=\; \frac{f(3,y)}{0.60} \;=\; \begin{cases} \frac{34}{60} & \text{if } y = 2 \\ \frac{5}{60} & \text{if } y = 5 \\ \frac{21}{60} & \text{if } y = 10. \end{cases}$$

Then the expectation conditional on $X = 3$ is

$$\mathrm{E}[Y|X=3] \;=\; \sum_y y f(y \,|\, 3) \;=\; 2\Big(\frac{34}{60}\Big) + 5\Big(\frac{5}{60}\Big) + 10\Big(\frac{21}{60}\Big) \;=\; 5.05. \quad \square$$

This compares to the unconditional expectation $\mathrm{E}[Y] = 6.15$. So information that $X = 3$ pushes the conditional expected value of Y down to 5.05. $\square$

Old Continuous Example: Let's consider our good friend with the funny limits,

$$f(x,y) \;=\; \frac{21}{4}x^2 y, \quad \text{if } x^2 \le y \le 1.$$

Recall that

$$f_Y(y) \;=\; \frac{7}{2}y^{5/2},\ 0 \le y \le 1, \quad \text{so that} \quad \mathrm{E}[Y] \;=\; \int_{\mathbb{R}} y f_Y(y)\, dy \;=\; \frac{7}{9}.$$

Further recall that

$$f(y \,|\, x) \;=\; \frac{2y}{1-x^4} \quad \text{if } x^2 \le y \le 1.$$

Thus,

$$\mathrm{E}[Y|x] \;=\; \int_{\mathbb{R}} y f(y \,|\, x)\, dy \;=\; \frac{2}{1-x^4}\int_{x^2}^1 y^2\, dy \;=\; \frac{2}{3}\cdot\frac{1-x^6}{1-x^4}, \quad -1 \le x \le 1.$$

So, e.g., $\mathrm{E}[Y|X = 0.5] = \frac{2}{3}\cdot\frac{63}{64}/\frac{15}{16} = 0.70$ (a bit lower than $\mathrm{E}[Y] = 7/9$). $\square$

3.4.2 Double Expectation

And now we have arrived at the most-interesting theorem of the chapter...

Theorem (Double Expectation):

$$\mathrm{E}\big[\mathrm{E}(Y|X)\big] \;=\; \mathrm{E}[Y].$$

Remarks: This funny-looking result has several ramifications.

- In plain English: The expected value (averaged over all X's) of the conditional expected value (of $Y|X$) is the plain old expected value (of Y).
- Think of the outside expected value as the expected value of $h(X) = \mathrm{E}(Y|X)$. Then LOTUS miraculously gives us $\mathrm{E}[Y]$.

- Believe it or not, sometimes it's easier to calculate $\mathrm{E}[Y]$ indirectly by using our double expectation trick.

Proof (continuous case): By the Unconscious Statistician,

$$\begin{aligned}
\mathrm{E}\big[\mathrm{E}(Y|X)\big] &= \int_{\mathbb{R}} \mathrm{E}(Y|x) f_X(x)\,dx \\
&= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} y f(y\,|\,x)\,dy \right) f_X(x)\,dx \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} y f(y\,|\,x) f_X(x)\,dx\,dy \\
&= \int_{\mathbb{R}} y \int_{\mathbb{R}} f(x,y)\,dx\,dy \\
&= \int_{\mathbb{R}} y f_Y(y)\,dy \\
&= \mathrm{E}[Y]. \quad \square
\end{aligned}$$

Old Example: Suppose $f(x,y) = \frac{21}{4}x^2 y$, if $x^2 \le y \le 1$. We will find $\mathrm{E}[Y]$ in *two ways*. To do so, we note that, by previous examples,

$$\begin{aligned}
f_X(x) &= \frac{21}{8}x^2(1-x^4), \quad \text{for } -1 \le x \le 1, \\
f_Y(y) &= \frac{7}{2}y^{5/2}, \quad \text{for } 0 \le y \le 1, \\
\mathrm{E}[Y|x] &= \frac{2}{3} \cdot \frac{1-x^6}{1-x^4}, \quad \text{for } -1 \le x \le 1.
\end{aligned}$$

Solution #1 (old, boring way):

$$\mathrm{E}[Y] = \int_{\mathbb{R}} y f_Y(y)\,dy = \int_0^1 \frac{7}{2} y^{7/2}\,dy = \frac{7}{9}.$$

Solution #2 (new, exciting way):

$$\begin{aligned}
\mathrm{E}[Y] &= \mathrm{E}\big[\mathrm{E}(Y|X)\big] \\
&= \int_{\mathbb{R}} \mathrm{E}(Y|x) f_X(x)\,dx \\
&= \int_{-1}^{1} \left(\frac{2}{3} \cdot \frac{1-x^6}{1-x^4} \right) \left(\frac{21}{8}x^2(1-x^4) \right) dx = \frac{7}{9}.
\end{aligned}$$

Notice that both answers are the same (whew!). ☺

3.4.3 Honors Applications

This subsection discusses a number of interesting applications of the concepts presented so far. §3.4.3.1 is concerned with what we call the "standard conditioning argument" — a powerful tool that can be used to derive various useful results. §3.4.3.2 presents applications of double expectation, including examples involving so-called first-step analysis and random sums of random variables.

§3.4.3.1 Standard Conditioning Argument Applications

We give a couple of examples invoking the general method known as the **standard conditioning argument** for computing all sorts of interesting probabilities. The method follows directly from the Law of Total Probability (see §1.8) adapted for use with bivariate random variables.

Example/Theorem: Suppose X and Y are independent continuous RV's, with pdf $f_X(\cdot)$ and cdf $F_Y(\cdot)$, respectively. Then the sum $Z = X + Y$ has cdf

$$\boxed{\mathrm{P}(Z \le z) \;=\; \int_{\mathbb{R}} F_Y(z-x) f_X(x)\, dx.}$$

An expression such as the above for $\mathrm{P}(Z \le z)$ is often called a **convolution**.

Proof: Denote the pdf of Z by $f_Z(z)$, the joint pdf of (X, Z) as $f_{X,Z}(x, z)$, and the conditional pdf of $Z|X = x$ by $f_{Z|X}(z|x)$. Then we have

$$\begin{aligned}
\mathrm{P}(Z \le z) &= \int_{-\infty}^{z} f_Z(t)\, dt \\
&= \int_{-\infty}^{z} \int_{\mathbb{R}} f_{X,Z}(x,t)\, dx\, dt \\
&= \int_{\mathbb{R}} \int_{-\infty}^{z} f_{Z|X}(t|x) f_X(x)\, dt\, dx \quad \text{(flip integrals)} \\
&= \int_{\mathbb{R}} \mathrm{P}(Z \le z \,|\, X = x) f_X(x)\, dx \\
&= \int_{\mathbb{R}} \mathrm{P}(X + Y \le z \,|\, X = x) f_X(x)\, dx \\
&= \int_{\mathbb{R}} \mathrm{P}(Y \le z - x \,|\, X = x) f_X(x)\, dx \\
&= \int_{\mathbb{R}} \mathrm{P}(Y \le z - x) f_X(x)\, dx \quad (X, Y \text{ are independent}). \quad \square
\end{aligned}$$

Example: Suppose $X, Y \overset{\text{iid}}{\sim} \mathrm{Exp}(\lambda)$, and let $Z = X + Y$. Then

$$\begin{aligned}
\mathrm{P}(Z \le z) &= \int_{\mathbb{R}} F_Y(z-x) f_X(x)\, dx \\
&= \int_{0}^{z} F_Y(z-x) f_X(x)\, dx \quad (z - x \ge 0,\, x \ge 0 \;\Rightarrow\; 0 \le x \le z) \\
&= \int_{0}^{z} (1 - e^{-\lambda(z-x)}) \lambda e^{-\lambda x}\, dx \\
&= 1 - e^{-\lambda z} - \lambda z e^{-\lambda z}.
\end{aligned}$$

Thus, the pdf of Z is

$$\frac{d}{dz}\mathrm{P}(Z \le z) \;=\; \lambda^2 z e^{-\lambda z}, \quad z \ge 0.$$

This turns out to mean that $X + Y \sim \mathbf{Gamma}(2, \lambda)$, aka $\mathbf{Erlang}_2(\lambda)$. $\square$

One can do similar kinds of convolutions with discrete RV's. We state the following result without proof (which is, in any case, straightforward).

Example/Theorem: Suppose X and Y are two independent integer-valued RV's, with pmf's $f_X(x)$ and $f_Y(y)$. Then the pmf of $Z = X + Y$ is

$$\boxed{f_Z(z) \;=\; \mathrm{P}(Z = z) \;=\; \sum_{x=-\infty}^{\infty} f_X(x) f_Y(z - x).}$$

Example: Suppose X and Y are iid Bern(p). In addition, recall that, for any set A and element z, the indicator function $1_A(z) = 1$ if $z \in A$ and $1_A(z) = 0$ otherwise. Then the pmf of $Z = X + Y$ is

$$\begin{aligned}
f_Z(z) &= \sum_{x=-\infty}^{\infty} f_X(x) f_Y(z - x) \\
&= f_X(0) f_Y(z) + f_X(1) f_Y(z - 1) \quad (X \text{ can only} = 0 \text{ or } 1) \\
&= f_X(0) f_Y(z) 1_{\{0,1\}}(z) + f_X(1) f_Y(z - 1) 1_{\{1,2\}}(z) \\
&\qquad (1_{\{\cdot\}}(z) \text{ functions indicate nonzero } f_Y(\cdot)\text{'s}) \\
&= p^0 q^{1-0} p^z q^{1-z} 1_{\{0,1\}}(z) + p^1 q^{1-1} p^{z-1} q^{2-z} 1_{\{1,2\}}(z) \\
&= p^z q^{2-z} \big[1_{\{0,1\}}(z) + 1_{\{1,2\}}(z)\big] \\
&= \binom{2}{z} p^z q^{2-z}, \quad z = 0, 1, 2.
\end{aligned}$$

Thus, $Z \sim \mathrm{Bin}(2, p)$, a fond blast from the past! □

§3.4.3.2 Double Expectation Applications

Double expectation can be used in clever ways to obtain interesting findings.

Example: We'll use a **"first-step" method** to calculate the mean of $Y \sim \mathrm{Geom}(p)$. We motivate the example by regarding Y as the number of coin flips until H appears for the first time, where $\mathrm{P}(\mathrm{H}) = p$.

Let's start out by looking at the *first* flip, just by itself. Let $X = 1$ if the first flip is H and 0 if T. Of course, if $X = 1$, then we are done immediately, in which case $Y = 1$. If, however, $X = 0$, then it's just like we've wasted the first toss and have to start the exercise over. In other words, $\mathrm{E}[Y|X = 1] = 1$, but $\mathrm{E}[Y|X = 0] = 1 + \mathrm{E}[Y]$. Thus, based on the result X of the first step, we have

$$\begin{aligned}
\mathrm{E}[Y] &= \mathrm{E}\big[\mathrm{E}(Y|X)\big] \\
&= \sum_x \mathrm{E}[Y|x]\, f_X(x) \\
&= \mathrm{E}[Y|X = 0]\mathrm{P}(X = 0) + \mathrm{E}[Y|X = 1]\mathrm{P}(X = 1) \\
&= (1 + \mathrm{E}[Y])(1 - p) + (1)(p).
\end{aligned}$$

Solving, we get $\mathrm{E}[Y] = 1/p$ (which is indeed the correct answer)! $\square$

Example: Consider a sequence of coin flips. What is the expected number of flips Y until HT appears for the first time? Clearly, $Y = A + B$, where A is the number of flips until the first H appears, and B is the number of subsequent flips until T appears for the first time after the sequence of H's begins. For instance, the sequence TTTHHT corresponds to $Y = A + B = 4 + 2 = 6$.

In any case, it's obvious that A and B are iid Geom$(p = 1/2)$, so by the previous example, $\mathrm{E}[Y] = \mathrm{E}[A] + \mathrm{E}[B] = (1/p) + (1/p) = 4$. $\square$

The previous example didn't involve first-step analysis (besides using the expected value of a geometric RV). But the next related example will...

Example: Again consider a sequence of coin flips. What is the expected number of flips Y until HH appears for the first time? For instance, the sequence TTHTTHH corresponds to $Y = 7$ tries. Using an enhanced first-step analysis, we see that

$$\begin{aligned}
\mathrm{E}[Y] &= \mathrm{E}[Y|\mathrm{T}]\,\mathrm{P}(\mathrm{T}) + \mathrm{E}[Y|\mathrm{H}]\,\mathrm{P}(\mathrm{H}) \\
&= \mathrm{E}[Y|\mathrm{T}]\,\mathrm{P}(\mathrm{T}) \\
&\quad + \big\{\mathrm{E}[Y|\mathrm{HH}]\,\mathrm{P}(\mathrm{HH}|\mathrm{H}) + \mathrm{E}[Y|\mathrm{HT}]\,\mathrm{P}(\mathrm{HT}|\mathrm{H})\big\}\mathrm{P}(\mathrm{H}) \\
&= \big(1 + \mathrm{E}[Y]\big)(0.5) + \big\{(2)(0.5) + \big(2 + \mathrm{E}[Y]\big)(0.5)\big\}(0.5) \\
&\quad \text{(since we have to start over once we see a T)} \\
&= 1.5 + 0.75\,\mathrm{E}[Y].
\end{aligned}$$

Solving, we obtain $\mathrm{E}[Y] = 6$, which is perhaps surprising given the result from the previous example. $\square$

Bonus Theorem (the expectation of the sum of a random number of RV's): Suppose that $X_1, X_2, \ldots$ are independent RV's, all with the same mean. Also suppose that N is a nonnegative, integer-valued RV that's independent of the X_i's. Then

$$\mathrm{E}\Big[\sum_{i=1}^{N} X_i\Big] \;=\; \mathrm{E}[N]\,\mathrm{E}[X_1].$$

Remark: Careful! In particular, note that $\mathrm{E}\big[\sum_{i=1}^{N} X_i\big] \neq N\mathrm{E}[X_1]$, since the left-hand side is a number while the right-hand side is a random variable.

Proof (cf. Ross [7]): By double expectation, and the fact that N is independent of

the X_i's, we have

$$\begin{aligned}
\mathrm{E}\Big[\sum_{i=1}^{N} X_i\Big] &= \mathrm{E}\Big[\mathrm{E}\Big(\sum_{i=1}^{N} X_i \Big| N\Big)\Big] \\
&= \sum_{n=1}^{\infty} \mathrm{E}\Big(\sum_{i=1}^{N} X_i \Big| N = n\Big)\mathrm{P}(N=n) \\
&= \sum_{n=1}^{\infty} \mathrm{E}\Big(\sum_{i=1}^{n} X_i \Big| N = n\Big)\mathrm{P}(N=n) \\
&= \sum_{n=1}^{\infty} \mathrm{E}\Big(\sum_{i=1}^{n} X_i\Big)\mathrm{P}(N=n) \quad (N \text{ and } X_i\text{'s independent}) \\
&= \sum_{n=1}^{\infty} n\mathrm{E}[X_1]\mathrm{P}(N=n) \\
&= \mathrm{E}[X_1]\sum_{n=1}^{\infty} n\mathrm{P}(N=n). \quad \square
\end{aligned}$$

Example: Suppose the number of times we roll a die is $N \sim \text{Pois}(10)$. If X_i denotes the value of the i^{th} toss, then the expected total of all of the rolls is

$$\mathrm{E}\Big[\sum_{i=1}^{N} X_i\Big] \;=\; \mathrm{E}[N]\mathrm{E}[X_1] \;=\; 10(3.5) \;=\; 35. \quad \square$$

Theorem: Under the same conditions as before,

$$\mathrm{Var}\Big(\sum_{i=1}^{N} X_i\Big) \;=\; \mathrm{E}[N]\mathrm{Var}(X_1) + (\mathrm{E}[X_1])^2\mathrm{Var}(N).$$

Proof: See, for instance, Ross [7]. $\square$

3.5 Covariance and Correlation

Covariance and Correlation are measures used to define the degree of *association* between X and Y if they don't happen to be independent. We'll begin our discussion in §3.5.1 with the basics of covariance and correlation. §3.5.2 discusses the relationship between correlation and causation — does the fact that two items are correlated imply that one causes the other? (The answer is: *not necessarily.*) We work some numerical examples in detail in §3.5.3. And finally, §3.5.4 spells out a number of theorems that will be useful in the sequel.

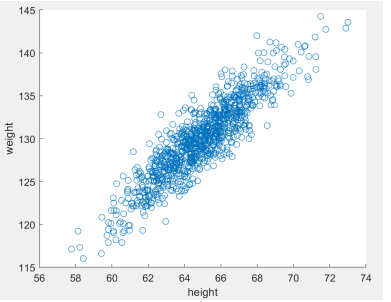
3.5.1 Basics

Definition: The **covariance** between X and Y is

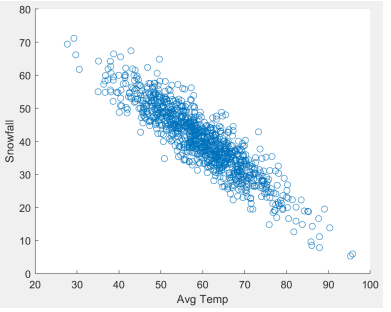
$$\mathrm{Cov}(X,Y) \;\equiv\; \sigma_{XY} \;\equiv\; \mathrm{E}\big[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\big].$$

Remarks:

- $\text{Cov}(X, X) = \text{E}\big[(X - \text{E}[X])^2\big] = \text{Var}(X)$.
- If X and Y have positive covariance, then X and Y move "in the same direction." Think height and weight.

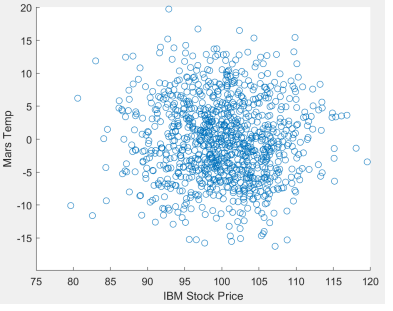


- If X and Y have negative covariance, then X and Y move "in opposite directions." Think snowfall and temperature.



- If X and Y are *independent*, then of course they have no association with each other. In fact, we'll prove below that independence implies that the covariance is 0 (but not the other way around).

  Example: IBM stock price vs. temperature on Mars are independent. (At least that's what they want you to believe!)



Theorem (easier way to calculate covariance):

$$\text{Cov}(X,Y) \;=\; \text{E}[XY] - \text{E}[X]\text{E}[Y].$$

Proof:

$$\begin{aligned}\text{Cov}(X,Y) &= \text{E}\big[(X-\text{E}[X])(Y-\text{E}[Y])\big] \\ &= \text{E}\big[XY - X\text{E}[Y] - Y\text{E}[X] + \text{E}[X]\text{E}[Y]\big] \\ &= \text{E}[XY] - \text{E}[X]\text{E}[Y] - \text{E}[Y]\text{E}[X] + \text{E}[X]\text{E}[Y]. \quad \square\end{aligned}$$

Theorem: X and Y independent implies that $\text{Cov}(X,Y) = 0$.

Proof: By a theorem from §3.3.2, we recall that X and Y independent implies that $\text{E}[XY] = \text{E}[X]\text{E}[Y]$. Then

$$\text{Cov}(X,Y) \;=\; \text{E}[XY] - \text{E}[X]\text{E}[Y] \;=\; \text{E}[X]\text{E}[Y] - \text{E}[X]\text{E}[Y]. \quad \square$$

Danger Will Robinson! $\text{Cov}(X,Y) = 0$ **does not imply** that X and Y are independent!

Example: Suppose $X \sim \text{Unif}(-1,1)$, and $Y = X^2$ (so that X and Y are clearly *dependent*). But

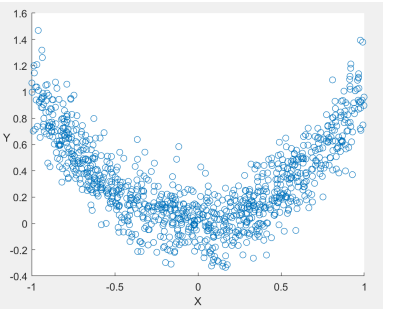
$$\text{E}[X] \;=\; \int_{-1}^{1} x \cdot \frac{1}{2}\, dx \;=\; 0 \text{ and}$$

$$\text{E}[XY] \;=\; \text{E}[X^3] \;=\; \int_{-1}^{1} x^3 \cdot \frac{1}{2}\, dx \;=\; 0.$$

So

$$\text{Cov}(X,Y) \;=\; \text{E}[XY] - \text{E}[X]\text{E}[Y] \;=\; 0. \quad ※$$

Here's a graphical illustration of this zero-covariance dependence phenomenon, where we've actually added some normal noise to Y to make it look prettier.



Now, on to covariance's sibling.

Definition: The **correlation** between X and Y is

$$\rho \;\equiv\; \text{Corr}(X,Y) \;\equiv\; \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \;=\; \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.$$

Remark: Covariance has "square" units, while correlation is unitless. In fact, correlation can be regarded as a "standardized" version of covariance.

Theorem: It can be shown that $-1 \leq \rho \leq 1$.

Corollary (of a previous theorem): X and Y independent implies $\rho = 0$.

Remark: $\rho \doteq 1$ is regarded as "high" correlation; $\rho \doteq 0$ is "low"; and $\rho \doteq -1$ is "high" negative correlation.

Example: Height is *highly* correlated with weight. Temperature on Mars has *low* correlation with IBM stock price.

3.5.2 Correlation and Causation

NOTE! Correlation does not necessarily imply causality! This is a very common pitfall in many areas of data analysis and public discourse. We present some examples that illustrate the relevant issues.

Example in which correlation does imply causality: As explained earlier, height and weight are positively correlated. In this case, it appears that larger height does indeed tend to cause greater weight. □

Example in which correlation does not imply causality: Temperature and lemonade sales have positive correlation, and it is fair to say that temperature has some causal influence on lemonade sales. Similarly, temperature and overheating cars are positively correlated with a causal relationship. It is also likely that lemonade sales and overheating cars are positively correlated, but there's obviously no causal relationship there. □

Example of a zero correlation relationship with causality! We saw earlier that it is possible for two dependent random variables to have zero correlation. □

To prove that X causes Y, one must establish that:

- X occurred before Y (in some sense).
- The relationship between X and Y is not completely due to random chance.
- Nothing else accounts for the relationship (which is violated in the lemonade sales/overheating cars example above).

These items can be often be established via mathematical analysis, statistical analysis of appropriate data, or consultation with appropriate experts.

The three examples above seem to give conflicting guidance with respect to the relationship between correlation and causality. How can we interpret these findings in a meaningful way? Here are the takeaways:

- If the correlation between X and Y is (significantly) nonzero, there is some type of relationship between the two items, which may or may not be causal, but this should raise our curiosity.
- If the correlation between X and Y is 0, we are not quite out of the woods with respect to dependence and causality. In order to definitively rule out a relationship between X and Y, it is always highly recommended protocol to, at the very least,
  - Plot data from X and Y against each other to see if there is a nonlinear relationship, as in the zero correlation yet dependent example.
  - Consult with appropriate experts.

3.5.3 A Couple of Worked Numerical Examples

Discrete Example: Suppose X is the GPA of a university student, and Y is the average hours per week that the student is on social media. Here's the joint pmf.

$f(x,y)$	$X=2$	$X=3$	$X=4$	$f_Y(y)$
$Y=40$	0.0	0.2	0.2	0.4
$Y=50$	0.1	0.1	0.0	0.2
$Y=60$	0.4	0.0	0.0	0.4
$f_X(x)$	0.5	0.3	0.2	1

We can easily calculate

$$\begin{aligned} \mathrm{E}[X] &= \sum_x x f_X(x) = 2.7, \\ \mathrm{E}[X^2] &= \sum_x x^2 f_X(x) = 7.9, \quad \text{and} \\ \mathrm{Var}(X) &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = 0.61. \end{aligned}$$

Similarly, $\mathrm{E}[Y] = 50$, $\mathrm{E}[Y^2] = 2580$, and $\mathrm{Var}(Y) = 80$. Finally,

$$\begin{aligned} \mathrm{E}[XY] &= \sum_x \sum_y xy f(x,y) \\ &= 2(40)(0.0) + 3(40)(0.2) + \cdots + 4(60)(0.0) = 129, \\ \mathrm{Cov}(X,Y) &= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] = -6.0, \quad \text{and} \\ \rho &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = -0.859. \end{aligned}$$

Hmm.... In this example, more social media seems to correlate with lower grades (though we have not actually proven cause and effect). □

Continuous Example: Suppose $f(x,y) = 10x^2y$, for $0 \le y \le x \le 1$.

Note that funny limits often result in nonzero correlation. We have

$$\begin{aligned} f_X(x) &= \int_0^x 10x^2 y\,dy \;=\; 5x^4, \;\; 0 \le x \le 1, \\ \mathrm{E}[X] &= \int_0^1 5x^5\,dx \;=\; 5/6, \\ \mathrm{E}[X^2] &= \int_0^1 5x^6\,dx \;=\; 5/7, \quad \text{and} \\ \mathrm{Var}(X) &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \;=\; 0.01984. \end{aligned}$$

Similarly,

$$\begin{aligned} f_Y(y) &= \int_y^1 10x^2 y\,dx \;=\; \frac{10}{3} y(1-y^3), \;\; 0 \le y \le 1, \\ \mathrm{E}[Y] &= 5/9, \quad \text{and} \quad \mathrm{Var}(Y) \;=\; 0.04850. \end{aligned}$$

And finally,

$$\begin{aligned} \mathrm{E}[XY] &= \int_{\mathbb{R}}\int_{\mathbb{R}} xy f(x,y)\,dx\,dy \;=\; \int_0^1\int_0^x 10x^3y^2\,dy\,dx \;=\; 10/21, \\ \mathrm{Cov}(X,Y) &= \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] \;=\; 0.01323, \quad \text{and} \\ \rho &= \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \;=\; 0.4265. \quad \square \end{aligned}$$

3.5.4 Additional Useful Theorems Involving Covariance

In this subsection, we'll discuss several theorems that will be of great use later on.

Theorem: *Whether or not* X and Y are independent,

$$\mathrm{Var}(X+Y) \;=\; \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y).$$

Remark: If X and Y are independent, then the covariance term goes away. But if X and Y are dependent, don't forget the covariance term!

Proof: By the work we did on a previous proof,

$$\begin{aligned} \mathrm{Var}(X+Y) &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 + \mathrm{E}[Y^2] - (\mathrm{E}[Y])^2 + 2\big(\mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]\big) \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y). \quad \square \end{aligned}$$

This result can be generalized...

Theorem:

$$\mathrm{Var}\Big(\sum_{i=1}^n X_i\Big) \;=\; \sum_{i=1}^n \mathrm{Var}(X_i) + 2\sum\sum_{i<j} \mathrm{Cov}(X_i, X_j).$$

Proof: Induction.

Corollary (which happens to be an old friend from §3.3.2): If all of the X_i's are *independent*, then

$$\mathrm{Var}\Big(\sum_{i=1}^{n} X_i\Big) \;=\; \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Now we have a result on the covariance of linear functionals of X and Y.

Theorem: $\mathrm{Cov}(aX, bY + c) = ab\,\mathrm{Cov}(X, Y)$.

Proof:

$$\begin{aligned}
\mathrm{Cov}(aX, bY + c) &= \mathrm{E}[aX \cdot (bY + c)] - \mathrm{E}[aX]\,\mathrm{E}[bY + c] \\
&= \mathrm{E}[abXY] + \mathrm{E}[acX] - \mathrm{E}[aX]\,\mathrm{E}[bY] - \mathrm{E}[aX]\,\mathrm{E}[c] \\
&= ab\,\mathrm{E}[XY] - ab\,\mathrm{E}[X]\,\mathrm{E}[Y] + ac\mathrm{E}[X] - ac\mathrm{E}[X] \\
&= ab\,\mathrm{Cov}(X, Y). \quad \square
\end{aligned}$$

Putting the above two theorems together, we get a really general result.

Theorem:

$$\mathrm{Var}\Big(\sum_{i=1}^{n} a_i X_i + c\Big) \;=\; \sum_{i=1}^{n} a_i^2\,\mathrm{Var}(X_i) + 2\sum\sum_{i<j} a_i a_j \mathrm{Cov}(X_i, X_j).$$

Example: Here is a useful expression for $\mathrm{Var}(X - Y)$ that comes up a lot in statistics and computer simulation applications.

$$\begin{aligned}
\mathrm{Var}(X - Y) &= (1)^2\mathrm{Var}(X) + (-1)^2\mathrm{Var}(Y) + 2(1)(-1)\mathrm{Cov}(X, Y) \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y). \quad \square
\end{aligned}$$

Example: If $\mathrm{Var}(X) = \mathrm{Var}(Y) = \mathrm{Var}(Z) = 10$, $\mathrm{Cov}(X, Y) = 3$, $\mathrm{Cov}(X, Z) = -2$, and $\mathrm{Cov}(Y, Z) = 0$, then

$$\begin{aligned}
&\mathrm{Var}(X - 2Y + 3Z) \\
&\quad = (1)^2\mathrm{Var}(X) + (-2)^2\mathrm{Var}(Y) + (3)^2\mathrm{Var}(Z) \\
&\qquad +2(1)(-2)\mathrm{Cov}(X, Y) + 2(1)(3)\mathrm{Cov}(X, Z) + 2(-2)(3)\mathrm{Cov}(Y, Z) \\
&\quad = 10 + 4(10) + 9(10) - 4(3) + 6(-2) - 12(0) \;=\; 116. \quad \square
\end{aligned}$$

3.6 Moment Generating Functions, Revisited

There is still some mileage left in our moment generating function (mgf) car. Now that we know a little bit about joint distributions, we'll take another drive.

Old Definition: $M_X(t) \equiv \mathrm{E}[e^{tX}]$ is the **mgf** of the RV X.

Old Example: If $X \sim \mathrm{Bern}(p)$, then

$$M_X(t) \;=\; \mathrm{E}[e^{tX}] \;=\; \sum_x e^{tx} f(x) \;=\; e^{t\cdot 1}p + e^{t\cdot 0}q \;=\; pe^t + q. \quad \square$$

Old Example: If $X \sim \text{Exp}(\lambda)$, then

$$M_X(t) \;=\; \text{E}[e^{tX}] \;=\; \int_{\mathbb{R}} e^{tx} f(x)\,dx \;=\; \frac{\lambda}{\lambda - t} \quad \text{if } \lambda > t. \quad \square$$

Old Theorem (why it's called the mgf): Under certain technical conditions,

$$\text{E}[X^k] \;=\; \frac{d^k}{dt^k} M_X(t)\Big|_{t=0}, \quad k = 1, 2, \ldots.$$

New Theorem: Suppose $X_1, \ldots, X_n$ are *independent*. Let $Y = \sum_{i=1}^n X_i$. Then the mgf of the sum is the product of the mgf's,

$$M_Y(t) \;=\; \prod_{i=1}^n M_{X_i}(t).$$

Proof: We have

$$M_Y(t) \;=\; \text{E}[e^{tY}] \;=\; \text{E}\Big[e^{t\sum_{i=1}^n X_i}\Big] \;=\; \text{E}\Big[\prod_{i=1}^n e^{tX_i}\Big] \;=\; \prod_{i=1}^n \text{E}[e^{tX_i}],$$

where the last step follows because the X_I's are independent. $\square$

Corollary: If $X_1, \ldots, X_n$ are iid, and $Y = \sum_{i=1}^n X_i$, then $M_Y(t) = [M_{X_1}(t)]^n$.

Example: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$. Then, by a previous example,

$$M_Y(t) \;=\; [M_{X_1}(t)]^n \;=\; (pe^t + q)^n. \quad \square$$

We can use results like this with our old friend...

Old Theorem (identifying distributions): *In this book*, each distribution has a unique mgf.

Proof: Not here.

Example/Theorem: The sum Y of n iid $\text{Bern}(p)$ random variables has the same distribution as a $\text{Bin}(n, p)$ random variable.

Proof: By the previous example and uniqueness, all we need to show is that the mgf of $Z \sim \text{Bin}(n, p)$ matches $M_Y(t) = (pe^t + q)^n$. To this end, we have

$$\begin{aligned}
M_Z(t) &= \text{E}[e^{tZ}] \\
&= \sum_z e^{tz} \text{P}(Z = z) \\
&= \sum_{z=0}^n e^{tz} \binom{n}{z} p^z q^{n-z} \\
&= \sum_{z=0}^n \binom{n}{z} (pe^t)^z q^{n-z} \\
&= (pe^t + q)^n \quad \text{(by the Binomial Theorem).} \quad \square
\end{aligned}$$

Example: You can identify a distribution by its mgf. For instance, $M_X(t) = \left(\frac{3}{4}e^t + \frac{1}{4}\right)^{15}$ implies that $X \sim \text{Bin}(15, 0.75)$. □

Old Theorem (mgf of a linear function of X): Suppose X has mgf $M_X(t)$, and let $Y = aX + b$. Then $M_Y(t) = e^{tb}M_X(at)$.

Example: Suppose that a RV Y has the following mgf,

$$M_Y(t) \;=\; e^{-2t}\left(\frac{3}{4}e^{3t} + \frac{1}{4}\right)^{15} \;=\; e^{bt}(pe^{at} + q)^n,$$

where $a = 3$, $b = -2$, $p = 3/4$, $q = 1/4$, and $n = 15$. By the previous example and theorem, we immediately have that

$$M_Y(t) \;=\; e^{bt}M_X(at),$$

where $X \sim \text{Bin}(n, p) \sim \text{Bin}(15, 0.75)$. Thus, $Y = aX + b = 3X - 2$, so that Y is a "shifted" binomial RV. □

Theorem (additive property of binomials): If $X_1, \ldots, X_k$ are independent, with $X_i \sim \text{Bin}(n_i, p)$ (where p is the same for all X_i's), then

$$Y \;\equiv\; \sum_{i=1}^{k} X_i \;\sim\; \text{Bin}\Big(\sum_{i=1}^{k} n_i, p\Big).$$

Proof: We have

$$\begin{aligned} M_Y(t) &= \prod_{i=1}^{k} M_{X_i}(t) \quad \text{(mgf of independent sum)} \\ &= \prod_{i=1}^{k} (pe^t + q)^{n_i} \quad (\text{Bin}(n_i, p) \text{ mgf}) \\ &= (pe^t + q)^{\sum_{i=1}^{k} n_i}. \end{aligned}$$

This is the mgf of the $\text{Bin}\big(\sum_{i=1}^{k} n_i, p\big)$, so we're done. □

Remark: You can use the mgf technique to add up lots of things. Here are some examples (stated without proofs or relevant parameters).

- The sum of iid Geom(p) random variables is a negative binomial random variable (see §4.1.3.2).
- The sum of independent Poissons is another Poisson (see Exercise 3.8.22).
- The sum of iid Exp(λ)'s is an Erlang — a type of gamma distribution (see §4.2.2.2).
- The sum of independent normals is another normal (see §4.3.1).
- And many more....

3.7 Bivariate Functions of Random Variables

3.7.1 Introduction and Basic Theory

So far, we've looked at a variety of properties of functions of one or more random variables:

- Functions of a single variable, e.g., what is the expected value of $h(X)$? (This is LOTUS, from §§2.5.2 and 2.8.3.)
- What is the distribution of $h(X)$? (See the discussion on functions of RV's in §2.8.)
- And sometimes even functions of two (or more) variables. For example, if the X_i's are independent, what is the $\text{Var}(\sum_{i=1}^n X_i)$? (Take a look at §3.3.3.)
- Use a standard conditioning argument to get the distribution of $X+Y$. (See §3.4.3.)

Goal: Now let's give a *general result* on the distribution of functions of *two* random variables, the proof of which is beyond the scope of the text.

Honors Theorem: Suppose X and Y are continuous random variables with joint pdf $f(x,y)$, and $V = h_1(X,Y)$ and $W = h_2(X,Y)$ are functions of X and Y, where

$$X \;=\; k_1(V,W) \quad \text{and} \quad Y \;=\; k_2(V,W),$$

for suitably chosen inverse functions k_1 and k_2. Then the joint pdf of V and W is

$$g(v,w) \;=\; f\big(k_1(v,w),\, k_2(v,w)\big)\,|J(v,w)|,$$

where $|J|$ is the absolute value of the *Jacobian* (determinant) of the transformation, i.e.,

$$J(v,w) \;=\; \begin{vmatrix} \frac{\partial x}{\partial v} & \frac{\partial x}{\partial w} \\ \frac{\partial y}{\partial v} & \frac{\partial y}{\partial w} \end{vmatrix} \;=\; \frac{\partial x}{\partial v}\,\frac{\partial y}{\partial w} - \frac{\partial y}{\partial v}\,\frac{\partial x}{\partial w}.$$

Wow, that was a mouthful!

Corollary: If X and Y are *independent*, then the joint pdf of V and W is

$$g(v,w) \;=\; f_X\big(k_1(v,w)\big) f_Y\big(k_2(v,w)\big)\,|J(v,w)|.$$

Remark: These results generalize the one-dimensional method from §2.8. In fact, you can use this method to find all sorts of cool stuff, e.g., the distribution of $X+Y$, X/Y, etc., as well as the joint pdf of any functions of X and Y.

Remark: Although the notation is nasty, the application isn't really so bad.

3.7.2 Examples

Example: Suppose X and Y are iid Exp(λ). Find the pdf of $X+Y$.

Trick: We'll set $V = X+Y$, along with the "dummy" random variable $W = X$. This yields

$$X \;=\; W \;=\; k_1(V,W), \quad \text{and} \quad Y \;=\; V-W \;=\; k_2(V,W).$$

To get the Jacobian term, we calculate $\frac{\partial x}{\partial v} = 0$, $\frac{\partial x}{\partial w} = 1$, $\frac{\partial y}{\partial v} = 1$ and $\frac{\partial y}{\partial w} = -1$, so that

$$|J(v,w)| \;=\; \left|\frac{\partial x}{\partial v}\frac{\partial y}{\partial w} - \frac{\partial y}{\partial v}\frac{\partial x}{\partial w}\right| \;=\; |0(-1) - 1(1)| \;=\; 1.$$

This implies that the joint pdf of V and W is

$$\begin{aligned} g(v,w) &= f\big(k_1(v,w),\, k_2(v,w)\big)\,|J(v,w)| \\ &= f(w, v-w)\cdot 1 \\ &= f_X(w) f_Y(v-w) \quad (X \text{ and } Y \text{ independent}) \\ &= \lambda e^{-\lambda w}\cdot \lambda e^{-\lambda(v-w)}, \quad \text{for } w>0 \text{ and } v-w>0 \\ &= \lambda^2 e^{-\lambda v}, \quad \text{for } 0<w<v. \end{aligned}$$

And, finally, we obtain the desired pdf of the sum V (after carefully noting the region of integration),

$$g_V(v) \;=\; \int_{\mathbb{R}} g(v,w)\,dw \;=\; \int_0^v \lambda^2 e^{-\lambda v}\,dw \;=\; \lambda^2 v e^{-\lambda v}, \quad \text{for } v>0.$$

This is the Gamma$(2,\lambda)$ pdf, which matches our answer in §3.4.3. □

Honors Example

Suppose X and Y are iid Unif(0,1). Find the joint pdf of $V = X+Y$ and $W = X/Y$. After some algebra, we obtain

$$X \;=\; \frac{VW}{W+1} \;=\; k_1(V,W), \quad \text{and} \quad Y \;=\; \frac{V}{W+1} \;=\; k_2(V,W).$$

After more algebra, we calculate

$$\frac{\partial x}{\partial v} = \frac{w}{w+1}, \quad \frac{\partial x}{\partial w} = \frac{v}{(w+1)^2}, \quad \text{and} \quad \frac{\partial y}{\partial v} = \frac{1}{w+1}, \quad \frac{\partial y}{\partial w} = \frac{-v}{(w+1)^2},$$

so that after still more algebra,

$$|J| \;=\; \left|\frac{\partial x}{\partial v}\frac{\partial y}{\partial w} - \frac{\partial y}{\partial v}\frac{\partial x}{\partial w}\right| \;=\; \frac{v}{(w+1)^2}.$$

This implies that the joint pdf of V and W is

$$\begin{aligned}
g(v,w) &= f\big(k_1(v,w),\, k_2(v,w)\big)\,|J(v,w)| \\
&= f\left(\frac{vw}{w+1},\, \frac{v}{w+1}\right)\cdot\frac{v}{(w+1)^2} \\
&= f_X\left(\frac{vw}{w+1}\right) f_Y\left(\frac{v}{w+1}\right)\frac{v}{(w+1)^2} \quad (X \text{ and } Y \text{ independent}) \\
&= 1\cdot 1\cdot\frac{v}{(w+1)^2},\ \text{for } 0<x,y<1 \quad (\text{since } X,Y\sim\text{Unif}(0,1)) \\
&= \frac{v}{(w+1)^2},\ \text{for } 0<x=\tfrac{vw}{w+1}<1,\ \text{and } 0<y=\tfrac{v}{w+1}<1 \\
&= \frac{v}{(w+1)^2},\ \text{for } 0<v<1+\min\{\tfrac{1}{w},w\},\ \text{and } w>0 \ \ (\text{algebra!}).
\end{aligned}$$

Note that you have to be careful about the limits of v and w, but this thing really does double integrate to 1! □

We can also get the marginal pdf's. First of all, for the ratio of the uniforms, we have

$$\begin{aligned}
g_W(w) &= \int_{\mathbb{R}} g(v,w)\,dv \\
&= \int_0^{1+\min\{1/w,w\}} \frac{v}{(w+1)^2}\,dv \\
&= \frac{\big(1+\min\{1/w,w\}\big)^2}{2(w+1)^2} \\
&= \begin{cases} \frac{1}{2}, & \text{if } w\le 1 \\ \frac{1}{2w^2}, & \text{if } w>1, \end{cases}
\end{aligned}$$

which is a little weird-looking and unexpected to me (it's flat for $w \le 1$, and then decreases to 0 pretty quickly for $w > 1$). □

For the pdf of the sum of the uniforms, we have to calculate $g_V(v) = \int_{\mathbb{R}} g(v,w)\,dw$. But first we need to deal with some inequality constraints so that we can integrate over the proper region, namely,

$$0 \le v \le 1+\min\{1/w, w\}, \quad 0\le v\le 2, \quad \text{and} \quad w\ge 0.$$

With a little thought, we see that if $0 \le v \le 1$, then there is no constraint on w except for it being positive. On the other hand, if $1 < v \le 2$, then you can show (it takes a little work) that $v-1 \le w \le \frac{1}{v-1}$. Thus, we have

$$\begin{aligned}
g_V(v) &= \begin{cases} \int_0^{\infty} g(v,w)\,dw, & \text{if } 0\le v\le 1 \\ \int_{v-1}^{1/(v-1)} g(v,w)\,dw, & \text{if } 1<v\le 2 \end{cases} \\
&= \begin{cases} v, & \text{if } 0\le v\le 1 \\ 2-v, & \text{if } 1<v\le 2 \end{cases} \quad (\text{after algebra}).
\end{aligned}$$

This is a **Triangular(0,1,2)** pdf. Can you see why? Is there an intuitive explanation for this pdf? □

And Now a Word From Our Sponsor...

We are finally done with what is perhaps the most-difficult material of the text! Congratulations and Felicitations! Things will get easier from now on! Happy days are here again! ☺☺

3.8 Exercises

1. (§3.1) Suppose that $f(x, y) = 6x$, for $0 \le x \le y \le 1$.
   (a) Find $\mathrm{P}(X < 1/2 \text{ and } Y < 1/2)$.
   (b) Find the marginal pdf $f_X(x)$ of X.

2. (§3.1) Suppose X and Y are discrete random variables with the following joint pmf, where any letters denote probabilities that you'll need to figure out.

$f(x,y)$	$X=-3$	$X=0$	$X=5$	$\mathrm{P}(Y=y)$
$Y=8$	0.2	a	b	0.3
$Y=27$	c	0.3	0.3	d
$\mathrm{P}(X=x)$	e	0.3	g	h

   (a) Complete the table with the correct values for the letters.
   (b) Find $\mathrm{P}(X \le 1)$.
   (c) Find $\mathrm{P}(X = 5 \,|\, Y = 27)$.

3. (§3.2) Suppose that $f(x, y) = cxy^2$, for $0 \le x \le y^2 \le 1$ and $0 \le y \le 1$.
   (a) Find c.
   (b) Find the marginal pdf of X, $f_X(x)$.
   (c) Find the marginal pdf of Y, $f_Y(y)$.
   (d) Find $\mathrm{E}[X]$.
   (e) Find $\mathrm{E}[Y]$.
   (f) Find the conditional pdf of X, given $Y = y$, i.e., $f(x \,|\, y)$.

4. (§3.2) The following table gives the joint pmf $f(x, y)$ of two random variables: X (the GPA of a university student), and Y (the average hours a week the student spends on social media).

$f(x,y)$	$X=2$	$X=3$	$X=4$
$Y=40$	0.4	0.1	0.1
$Y=50$	0.1	0.2	0.1

   (a) What's the probability that a random student spends 50 hours on social media?

(b) What's the conditional probability that a random student spends 50 hours a week on social media given that his GPA = 2?

5. (§3.3) Consider the random variables X and Y having the by-now-familiar joint pmf

$f(x,y)$	$X=2$	$X=3$	$X=4$	$f_Y(y)$
$Y=40$	0.4	0.1	0.1	
$Y=50$	0.1	0.2	0.1	
$f_X(x)$				1

Are X and Y independent?

6. (§3.3) Suppose that $f(x,y) = c(x+y)$ for $0 < x < y < 1$, for appropriate constant c. Are X and Y independent?

7. (§3.3) If $\mathrm{E}[X] = 7$, $\mathrm{E}[Y] = -3$, $\mathrm{Var}(X) = 1$, $\mathrm{Var}(Y) = 9$, and X and Y are *independent*, find $\mathrm{Var}(X+Y)$.

8. (§3.3) We have two brands of lightbulbs. Brand X has an $\mathrm{Exp}(\mu = 1)$ lifetime with a mean of 1 year. Brand Y has an $\mathrm{Exp}(\lambda = 1/2)$ lifetime with a mean of 2 years. Assuming that X and Y are independent, what's the chance that X will last longer than Y?

9. (§3.3) Suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Poisson}(5)$ with pmf $\mathrm{P}(X_i = x) = e^{-5}5^x/x!$, $x = 0, 1, 2, \ldots$. Let $\bar{X}_n = \sum_{i=1}^n X_i/n$ denote the sample mean of the X_i's. What are $\mathrm{E}[\bar{X}]$, $\mathrm{Var}(\bar{X})$, $\lim_{n\to\infty} \mathrm{E}[\bar{X}]$, and $\lim_{n\to\infty} \mathrm{Var}(\bar{X})$?

10. (§3.4) Which are legitimate expressions for $\mathrm{E}[Y|x]$?

(a) $\mathrm{E}[Y]$
(b) $\int_{\mathbb{R}} y f(y \,|\, x)\, dy$
(c) $\int_{\mathbb{R}} y f(y \,|\, x)\, dx$
(d) $\frac{1}{f_X(x)} \int_{\mathbb{R}} y f(x,y)\, dy$
(e) Both (b) and (d)

11. (§3.4) Suppose that

$$f(x,y) \;=\; 6x \quad \text{for } 0 \le x \le y \le 1.$$

You may already have seen someplace that the marginal pdf of X turns out to be

$$f_X(x) \;=\; 6x(1-x) \quad \text{for } 0 \le x \le 1.$$

(a) Find the conditional pdf of Y given that $X = x$.
(b) Find $\mathrm{E}[Y|X = x]$.
(c) Find $\mathrm{E}\big[\mathrm{E}[Y|X]\big]$.

12. (§3.4) Back to Question 3.
    (a) Find the conditional expectation, $\mathrm{E}[X|y]$.
    (b) Find the double conditional expectation, $\mathrm{E}\big[\mathrm{E}[X|Y]\big]$.

13. (§3.4) Consider the set-up from Question 4 involving the joint pmf $f(x, y)$ of two random variables: X (the GPA of a university student) and Y (the average hours per week spent on social media), along with the resulting marginals.

$f(x,y)$	$X=2$	$X=3$	$X=4$	$f_Y(y)$
$Y=40$	0.4	0.1	0.1	0.6
$Y=50$	0.1	0.2	0.1	0.4
$f_X(x)$	0.5	0.3	0.2	1

    What is $\mathrm{E}\big[\mathrm{E}[Y|X]\big]$?

14. (§3.4) **[NEW HONORS!]** Suppose that X and Y are iid Unif$(0, 1)$. Find the conditional pdf of X given that $XY = t$.

15. (§3.4) Consider a sequence of flips of a *biased* coin — the probability of H is 1/3, and T is 2/3. What is the expected value of the number of flips Y until "HT" appears for the first time? (For instance, the sequence TTHHT corresponds to $Y = 5$.)

16. (§3.4) I'm a good gambler. When I play a game, I either lose \$10 with probability 0.4, or I make \$20 w.p. 0.6. Let's suppose that I play a Poisson$(\lambda = 9.5)$ random number of games before I get tired and go home. Furthermore, assume that everything is independent (i.e., all of the games as well as the number of games). What are my expected winnings by the time I go home?

17. (§3.5) Suppose that the correlation between December snowfall and temperature in Siberacuse, NY is -0.5. Further suppose that $\mathrm{Var}(S) = 100$ in^2 and $\mathrm{Var}(T) = 25$ (degrees F)^2. Find $\mathrm{Cov}(S, T)$ (in units of degree inches, whatever those are).

18. (§3.5) TRUE or FALSE? If X and Y are positively correlated, and Y and Z are positively correlated, then it must be the case that either X causes Z or vice versa.

19. (§3.5) Consider the following joint pmf for two discrete RVs X and Y.

$f(x,y)$	$X=-1$	$X=1$	$f_Y(y)$
$Y=-1$	0.4	0.1	0.5
$Y=1$	0.1	0.4	0.5
$f_X(x)$	0.5	0.5	1

    Find the correlation between X and Y.

20. (§3.5) If X and Y both have mean -7 and variance 4, and $\mathrm{Cov}(X, Y) = 1$, find $\mathrm{Var}(3X - Y)$.

21. (§3.5) Let $\text{Var}(X) = \text{Var}(Y) = 10$, $\text{Var}(Z) = 20$, $\text{Cov}(X, Y) = 2$, $\text{Cov}(X, Z) = -3$, and $\text{Cov}(Y, Z) = -4$. Find both $\text{Corr}(X, Y)$ and $\text{Var}(X - 2Y + 5Z)$.

22. (§3.6) Suppose that $X_1, X_2, \ldots, X_n$ are independent Poisson RV's with $X_i \sim \text{Pois}(\lambda_i)$. Use mgf's to show that $Y = \sum_{i=1}^{n} X_i \sim \text{Pois}\big(\sum_{i=1}^{n} \lambda_i\big)$.

23. (§3.6) Identify random variable Y's probability distribution if it has mgf $M_Y(t) = (0.3e^t + 0.7)^4$.

24. (§3.7) Suppose $X \sim \text{Exp}(\lambda)$, $Y \sim \text{Exp}(\mu)$, and X and Y are independent. Find the pdf of $X + Y$.

Chapter 4

Distributions

This chapter will discuss lots of interesting probability distributions. We'll provide a compendium of results, some of which we'll prove, and some of which we've already seen. Special emphasis will be placed on the **normal distribution**, because it's so important and has so many implications, including the **Central Limit Theorem**.

4.1 Discrete Distributions

4.1.1 Bernoulli and Binomial Distributions

We'll begin with the most-basic distributions.

Definition: The **Bernoulli(p) distribution** is given by

$$X = \begin{cases} 1 & \text{w.p. } p \quad (\text{“success”}) \\ 0 & \text{w.p. } q = 1-p \quad (\text{“failure”}). \end{cases}$$

Recall: $\mathrm{E}[X] = p$, $\mathrm{Var}(X) = pq$, and the moment generating function (mgf) $M_X(t) = pe^t + q$.

Definition: The **binomial distribution** with parameters n and p is given by the probability mass function (pmf)

$$\mathrm{P}(Y = k) = \binom{n}{k} p^k q^{n-k}, \quad k = 0, 1, \ldots, n.$$

Theorem: $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p) \Rightarrow Y \equiv \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$.

Proof: This kind of result can easily be proved by a moment generating function uniqueness argument, as we mentioned in §3.6. □

Think of the $\text{Bin}(n, p)$ as the number of successes from n $\text{Bern}(p)$ trials.

Example: Toss two dice and take the sum; repeat the experiment five times. Let Y be the number of 7's you see. Each experiment can be regarded as a $\text{Bern}(p)$ trial with $p = \text{P}(\text{Sum} = 7) = 1/6$, so that $Y \sim \text{Bin}(5, 1/6)$. Then, e.g.,

$$\text{P}(Y = 4) \;=\; \binom{5}{4}\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^{5-4}. \quad \square$$

Theorem: $Y \sim \text{Bin}(n, p)$ implies

$$\text{E}[Y] \;=\; \text{E}\left[\sum_{i=1}^n X_i\right] \;=\; \sum_{i=1}^n \text{E}[X_i] \;=\; np,$$

and similarly,

$$\text{Var}(Y) \;=\; npq.$$

We saw in §3.6 that $M_Y(t) = (pe^t + q)^n$.

Theorem: Certain binomials add up: If $Y_1, \ldots, Y_k$ are *independent* and $Y_i \sim \text{Bin}(n_i, p)$, then

$$\sum_{i=1}^k Y_i \;\sim\; \text{Bin}\Big(\sum_{i=1}^k n_i, p\Big).$$

4.1.2 Hypergeometric Distribution

Definition: Suppose that you have a objects of Type 1, and b objects of Type 2. Select n objects without replacement from the $a + b$. Let X be the number of Type 1's selected. Then X has the **hypergeometric distribution** with pmf

$$\text{P}(X = k) \;=\; \frac{\binom{a}{k}\binom{b}{n-k}}{\binom{a+b}{n}}, \quad k = \max\{0, n-b\}, \ldots, \min\{a, n\}.$$

The support of the pmf looks a little complicated, but if $n \le \min\{a, b\}$, then $k = 0, 1, \ldots, n$.

Example: Suppose there are 25 socks in a box — 15 red and 10 blue. Pick seven without replacement.

$$\text{P}(\text{exactly three reds are picked}) \;=\; \frac{\binom{15}{3}\binom{10}{4}}{\binom{25}{7}}. \quad \square$$

Theorem: After some algebra, it turns out that

$$\text{E}[X] \;=\; n\left(\frac{a}{a+b}\right), \quad \text{and} \quad \text{Var}(X) \;=\; n\left(\frac{a}{a+b}\right)\left(1 - \frac{a}{a+b}\right)\left(\frac{a+b-n}{a+b-1}\right).$$

Remark: Here, $\frac{a}{a+b}$ plays the role of p in the binomial distribution. And then the corresponding $Y \sim \text{Bin}(n, p)$ results would be

$$\text{E}[Y] \;=\; n\left(\frac{a}{a+b}\right), \quad \text{and} \quad \text{Var}(Y) \;=\; n\left(\frac{a}{a+b}\right)\left(1-\frac{a}{a+b}\right).$$

So the binomial has the same mean as the hypergeometric, but slightly larger variance.

4.1.3 Geometric and Negative Binomial Distributions

This subsection discusses two distributions that are closely related to the Bernoulli and the binomial.

§4.1.3.1 Geometric Distribution

Definition: Suppose we consider an infinite sequence of independent Bern(p) trials. Let Z equal the number of trials *until the first success* is obtained. The event $Z = k$ corresponds to $k-1$ failures, followed by a success. For example, FFFFS yields $Z = 5$. Thus,

$$\text{P}(Z = k) \;=\; q^{k-1}p, \quad k = 1, 2, \ldots$$

and we say that Z has the **geometric distribution** with parameter p.

Notation: $X \sim \text{Geom}(p)$.

We'll get the mean and variance of the geometric via the mgf...

Theorem: The mgf of the Geom(p) is

$$M_Z(t) \;=\; \frac{pe^t}{1-qe^t}, \quad \text{for } t < \ell\text{n}(1/q).$$

Proof: We have

$$\begin{aligned} M_Z(t) &= \text{E}[e^{tZ}] \\ &= \sum_{k=1}^{\infty} e^{tk} q^{k-1} p \\ &= pe^t \sum_{k=0}^{\infty} (qe^t)^k \\ &= \frac{pe^t}{1-qe^t}, \text{ for } qe^t < 1. \quad \square \end{aligned}$$

Corollary: $\text{E}[Z] = 1/p$.

Proof: Using the usual mgf theorem for the expected value,

$$\begin{aligned} \mathrm{E}[Z] &= \frac{d}{dt} M_Z(t)\Big|_{t=0} \\ &= \frac{(1-qe^t)(pe^t) - (-qe^t)(pe^t)}{(1-qe^t)^2}\Big|_{t=0} \\ &= \frac{pe^t}{(1-qe^t)^2}\Big|_{t=0} \\ &= \frac{p}{(1-q)^2} \;=\; 1/p. \quad \square \end{aligned}$$

Remark: We could also have proven this directly from the definition of expected value, as in §2.5.1.

Similarly, after a lot of algebra,

$$\mathrm{E}[Z^2] \;=\; \frac{d^2}{dt^2} M_Z(t)\Big|_{t=0} \;=\; \frac{2-p}{p^2},$$

so that

$$\mathrm{Var}(Z) \;=\; \mathrm{E}[Z^2] - (\mathrm{E}[Z])^2 \;=\; \frac{2-p}{p^2} - \frac{1}{p^2} \;=\; \frac{q}{p^2}. \quad \square$$

Example: Toss a die repeatedly. What's the probability that we observe a 3 for the first time on the eighth toss?

Answer: The number of tosses we need is $Z \sim \mathrm{Geom}(1/6)$. Thus, $\mathrm{P}(Z = 8) = (5/6)^7(1/6)$. $\square$

How many tosses would we expect to take?

Answer: $\mathrm{E}[Z] = 1/p = 6$ tosses. $\square$

Here's a really important property of the geometric.

Theorem (Memoryless Property of the Geometric): Suppose that $Z \sim \mathrm{Geom}(p)$. Then for positive integers s, t, we have

$$\mathrm{P}(Z > s + t \,|\, Z > s) \;=\; \mathrm{P}(Z > t).$$

Why is it called the Memoryless Property? If an event hasn't occurred by time s, then the probability that it will occur after an additional t time units is the same as the (unconditional) probability that it would have originally occurred after time t — it forgot that it made it past time s!

Proof: First of all, for any $t = 0, 1, 2, \ldots$, the tail probability is

$$\mathrm{P}(Z > t) \;=\; \mathrm{P}(t \text{ Bern}(p) \text{ failures in a row}) \;=\; q^t.$$

This immediately implies that

$$\mathrm{P}(Z > s + t \,|\, Z > s) = \frac{\mathrm{P}(Z > s + t \cap Z > s)}{\mathrm{P}(Z > s)} = \frac{\mathrm{P}(Z > s + t)}{\mathrm{P}(Z > s)} = \frac{q^{s+t}}{q^s} = q^t.$$

In other words, $\mathrm{P}(Z > s+t \,|\, Z > s) = \mathrm{P}(Z > t)$. □

Example: Let's toss a die until a 5 appears for the first time. Suppose that we've already made four tosses without success. What's the probability that we'll need more than two additional tosses before we observe a 5?

Let Z be the number of tosses required. By the Memoryless Property (with $s = 4$ and $t = 2$), we want

$$\mathrm{P}(Z > 6 \,|\, Z > 4) \;=\; \mathrm{P}(Z > 2) \;=\; (5/6)^2. \quad \square$$

Fun Fact: The Geom(p) is the only discrete distribution with the memoryless property.

Not-as-Fun Fact: Some books define the Geom(p) as the number of Bern(p) *failures* until you observe a success; i.e., the number of failures = the number of trials − 1. You should be aware of this inconsistency, but don't worry about it for now.

§4.1.3.2 Negative Binomial Distribution

Definition: Suppose we consider an infinite sequence of independent Bern(p) trials. Let W denote the number of trials *until the* r^{th} *success* is obtained, so that $W = r, r+1, \ldots$. The event $W = k$ corresponds to exactly $r-1$ successes by time $k-1$, and then the r^{th} success at time k. We say that W has the **negative binomial distribution** (aka the Pascal distribution), with parameters r and p.

Notation: $W \sim \mathrm{NegBin}(r, p)$.

Example: FFFFSFS corresponds to $W = 7$ trials until the $r = 2^{\text{nd}}$ success.

Remark: As with the Geom(p), the exact definition of the NegBin depends on which book you're reading.

Let's obtain the pmf of W. Note that $W = k$ iff you get exactly $r-1$ successes (and $k-r$ failures) by time $k-1$, and then the r^{th} success at time k. So,

$$\mathrm{P}(W = k) \;=\; \left[\binom{k-1}{r-1} p^{r-1} q^{k-r}\right] p \;=\; \binom{k-1}{r-1} p^r q^{k-r}, \quad k = r, r+1, \ldots$$

Example: Toss a die until a 5 appears for the third time. What's the probability that we'll need exactly seven tosses?

Let W be the number of tosses required. Clearly, $W \sim \mathrm{NegBin}(3, 1/6)$. Then

$$\mathrm{P}(W = 7) \;=\; \binom{7-1}{3-1} (1/6)^3 (5/6)^{7-3}. \quad \square$$

Theorem: If $Z_1, \ldots, Z_r \overset{\text{iid}}{\sim} \mathrm{Geom}(p)$, then $W = \sum_{i=1}^{r} Z_i \sim \mathrm{NegBin}(r, p)$. In other words, Geom($p$)'s add up to a NegBin.

Proof: We won't do it here, but you can use the mgf technique. (Or, you can even argue intuitively.) □

The theorem makes sense if you think of Z_i as the number of trials after the $(i-1)^{\text{st}}$ success, up to and including the i^{th} success.

Properties (of the NegBin): Since the Z_i's are iid, the above theorem gives:

$$\begin{aligned} \mathrm{E}[W] &= r\mathrm{E}[Z_i] = r/p, \\ \mathrm{Var}(W) &= r\mathrm{Var}(Z_i) = rq/p^2, \\ M_W(t) &= [M_{Z_i}(t)]^r = \left(\frac{pe^t}{1-qe^t}\right)^r. \end{aligned}$$

How are the Binomial and NegBin Related?

If $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$, then

$$Y \equiv \sum_{i=1}^{n} X_i \sim \text{Bin}(n,p), \text{ with } \mathrm{E}[Y] = np \text{ and } \mathrm{Var}(Y) = npq.$$

If $Z_1, \ldots, Z_r \overset{\text{iid}}{\sim} \text{Geom}(p)$, then

$$W \equiv \sum_{i=1}^{r} Z_i \sim \text{NegBin}(r,p), \text{ with } \mathrm{E}[W] = r/p \text{ and } \mathrm{Var}(W) = rq/p^2.$$

4.1.4 Poisson Processes and the Poisson Distribution

We'll first talk about Poisson processes and then how they lead to the Poisson distribution, which we have already met in passing.

4.1.4.1 Poisson Processes

Let $N(t)$ be a **counting process**. That is, $N(t)$ is the number of occurrences (or arrivals, or events) of some process over the time interval $[0,t]$. $N(t)$ looks like a step function.

Examples: $N(t)$ could be any of the following:

(a) Cars entering a shopping center (by time t).

(b) Defects on a wire (of length t).

(c) Raisins in cookie dough (of volume t).

Let $\lambda > 0$ be the average number of occurrences per unit time (or length or volume). In the above examples, we might have: (a) $\lambda = 10$ / min, (b) $\lambda = 0.5$ / ft, (c) $\lambda = 4$ / in^3. □

Before proceeding, let's enjoy some useful notation.

Notation: $o(h)$ ("little-oh") is a generic function such that $o(h)/h \to 0$ as $h \to 0$, i.e., $o(h)$ goes to zero "faster" than h goes to zero. For example, $f(h) = h^2 = o(h)$, because $f(h)/h = h \to 0$; but $g(h) = 2h$ is *not* $o(h)$, because in that case $g(h)/h = 2$, which doesn't go to 0.

A Poisson process is a specific counting process.

Definitions: For $a < b$, the **increment** $N(b) - N(a)$ is a random variable representing the number of arrivals in $(a, b]$. A **Poisson process (PP)** is a counting process that satisfies the following conditions on increments:

(i) There is a short enough interval of time h, such that, for all t,

$$\begin{aligned} \mathrm{P}\big(N(t+h) - N(t) = 0\big) &= 1 - \lambda h + o(h) \\ \mathrm{P}\big(N(t+h) - N(t) = 1\big) &= \lambda h + o(h) \\ \mathrm{P}\big(N(t+h) - N(t) \geq 2\big) &= o(h). \end{aligned}$$

(ii) The distribution of the increment $N(t+h) - N(t)$ only depends on the length h.

(iii) If $a < b < c < d$, then the two increments $N(d) - N(c)$ and $N(b) - N(a)$ are *independent* random variables.

English translations of the Poisson process assumptions:

(i) Arrivals pretty much occur one-at-a-time, and then at rate λ/unit time.

(ii) The arrival pattern is **stationary** — it doesn't change over time; in particular, the rate λ is constant.

(iii) The numbers of arrivals in two disjoint time intervals are independent.

Poisson Process Example: Neutrinos hit a detector. Occurrences are rare enough that they really do happen one-at-a-time — you never get arrivals of groups of neutrinos.[1] Furthermore, the rate doesn't vary over time, and all arrivals are independent of each other. □

Anti-Example: Customers arrive at a restaurant. They tend to show up in groups, rather than one-at-a-time. The rate varies over the day (e.g., more at dinner time). Moreover, arrivals may not be independent. This ain't a Poisson process. □

4.1.4.2 Poisson Distribution

Definition: Let X be the number of occurrences in a Poisson(λ) process in a *unit interval* of time. Then X has the **Poisson distribution** with parameter λ.

Notation: $X \sim \mathrm{Pois}(\lambda)$.

[1] Neutrinos are solitary creatures.

Theorem/Definition: $X \sim \text{Pois}(\lambda) \Rightarrow \text{P}(X = k) = e^{-\lambda}\lambda^k/k!$, $k = 0, 1, 2, \ldots$.

Proof: Follows from the Poisson process assumptions and involves some simple differential equations.

To begin with, let's define $P_x(t) \equiv \text{P}(N(t) = x)$, i.e., the probability of exactly x arrivals by time t. We note that the probability that there haven't been any arrivals by time $t+h$ can be written in terms of the probability that there haven't been any arrivals by time t.

$$
\begin{aligned}
& P_0(t+h) \\
&\quad = \text{P}(N(t+h) = 0) \\
&\quad = \text{P}(\text{no arrivals by time } t \text{ and then no arrivals by time } t+h) \\
&\quad = \text{P}\big(\{N(t) = 0\} \cap \{N(t+h) - N(t) = 0\}\big) \\
&\quad = \text{P}(N(t) = 0)\text{P}(N(t+h) - N(t) = 0) \quad \text{(by independent increments (iii))} \\
&\quad = \text{P}(N(t) = 0)\text{P}(N(h) = 0) \quad \text{(by stationary increments (ii))} \\
&\quad \doteq P_0(t)(1 - \lambda h) \quad \text{(by definition and (i))}.
\end{aligned}
$$

Thus,

$$\frac{P_0(t+h) - P_0(t)}{h} \doteq -\lambda P_0(t).$$

Taking the limit as $h \to 0$, we have

$$P_0'(t) = -\lambda P_0(t). \tag{4.1}$$

Similarly, for $x > 0$, we have

$$
\begin{aligned}
P_x(t+h) &= \text{P}(N(t+h) = x) \\
&= \text{P}\big(N(t+h) = x \text{ and no arrivals during } [t, t+h]\big) \\
&\quad + \text{P}\big(N(t+h) = x \text{ and} \geq 1 \text{ arrival during } [t, t+h]\big) \\
&\quad \text{(Law of Total Probability)} \\
&\doteq \text{P}\big(\{N(t) = x\} \cap \{N(t+h) - N(t) = 0\}\big) \\
&\quad + \text{P}\big(\{N(t) = x-1\} \cap \{N(t+h) - N(t) = 1\}\big) \\
&\quad \text{(by (i), only consider case of one arrival in } [t, t+h]) \\
&= \text{P}(N(t) = x)\text{P}(N(t+h) - N(t) = 0) \\
&\quad + \text{P}(N(t) = x-1)\text{P}(N(t+h) - N(t) = 1) \\
&\quad \text{(by independent increments (iii))} \\
&\doteq P_x(t)(1 - \lambda h) + P_{x-1}(t)\lambda h.
\end{aligned}
$$

Taking the limits as $h \to 0$, we obtain

$$P_x'(t) = \lambda\big[P_{x-1}(t) - P_x(t)\big], \quad x = 1, 2, \ldots. \tag{4.2}$$

The solution of differential equations (4.1) and (4.2) is easily shown to be

$$P_x(t) = \frac{(\lambda t)^x e^{-\lambda t}}{x!}, \quad x = 0, 1, 2, \ldots.$$

(If you don't believe me, just plug in and see for yourself!)
Noting that $t = 1$ for the Pois(λ) case finally completes the proof. □

Remark: The value of λ can be changed simply by changing the units of time. (It's up to you to keep track of the units that you're working with.)

Examples:

$$\begin{aligned} X &= \text{number of phone calls in one minute} \sim \text{Pois}(\lambda = 3\,/\,\text{min}) \\ Y &= \text{number of phone calls in five minutes} \sim \text{Pois}(\lambda = 15\,/\,5\text{ min}) \\ Z &= \text{number of phone calls in 10 seconds} \sim \text{Pois}(\lambda = 0.5\,/\,10\text{ sec}). \quad \square \end{aligned}$$

Theorem: $X \sim \text{Pois}(\lambda) \Rightarrow$ mgf is $M_X(t) = e^{\lambda(e^t - 1)}$.
Proof:

$$M_X(t) \;=\; \mathrm{E}[e^{tX}] \;=\; \sum_{k=0}^{\infty} e^{tk} \left(\frac{e^{-\lambda}\lambda^k}{k!} \right) \;=\; e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda e^t)^k}{k!} \;=\; e^{-\lambda} e^{\lambda e^t}. \quad \square$$

Theorem: $X \sim \text{Pois}(\lambda) \Rightarrow \mathrm{E}[X] = \mathrm{Var}(X) = \lambda$.
Proof (using mgf):

$$\mathrm{E}[X] \;=\; \frac{d}{dt} M_X(t)\Big|_{t=0} \;=\; \frac{d}{dt} e^{\lambda(e^t - 1)}\Big|_{t=0} \;=\; \lambda e^t M_X(t)\Big|_{t=0} \;=\; \lambda.$$

Similarly,

$$\begin{aligned} \mathrm{E}[X^2] &= \frac{d^2}{dt^2} M_X(t)\Big|_{t=0} \\ &= \frac{d}{dt}\Big(\frac{d}{dt} M_X(t) \Big)\Big|_{t=0} \\ &= \lambda \frac{d}{dt}\big(e^t M_X(t) \big)\Big|_{t=0} \\ &= \lambda \Big[e^t M_X(t) + e^t \frac{d}{dt} M_X(t) \Big]\Big|_{t=0} \\ &= \lambda e^t \Big[M_X(t) + \lambda e^t M_X(t) \Big]\Big|_{t=0} \\ &= \lambda(1 + \lambda). \end{aligned}$$

Thus, $\mathrm{Var}(X) = \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \lambda(1 + \lambda) - \lambda^2 = \lambda$. Done. □

Example: Calls to a center arrive as a Poisson process with rate three per minute.
Let $X =$ number of calls in one minute.
So $X \sim \text{Pois}(3)$, $\mathrm{E}[X] = \mathrm{Var}(X) = 3$, $\mathrm{P}(X \le 4) = \sum_{k=0}^{4} e^{-3} 3^k / k!$.
Let $Y =$ number of calls in 40 seconds.

So $Y \sim \text{Pois}(2)$, $\text{E}[Y] = \text{Var}(Y) = 2$, $\text{P}(Y \le 4) = \sum_{k=0}^{4} e^{-2}2^k/k!$. □

Theorem (Additive Property of Poissons): Suppose $X_1, \ldots, X_n$ are *independent* with $X_i \sim \text{Pois}(\lambda_i)$, $i = 1, \ldots, n$. Then

$$Y \equiv \sum_{i=1}^{n} X_i \sim \text{Pois}\Big(\sum_{i=1}^{n} \lambda_i\Big).$$

Proof: Since the X_i's are independent, a theorem from §3.6 implies that

$$M_Y(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} e^{\lambda_i(e^t-1)} = e^{\left(\sum_{i=1}^{n}\lambda_i\right)(e^t-1)},$$

which is the mgf of the $\text{Pois}\big(\sum_{i=1}^{n}\lambda_i\big)$ distribution. □

Example: Cars driven by males [females] arrive at a parking lot according to a Poisson process with a rate of $\lambda_1 = 3$ / hr [$\lambda_2 = 5$ / hr]. All arrivals are independent. What's the probability of exactly two arrivals in the next 30 minutes?

Answer: The total number of arrivals is $\text{Pois}(\lambda_1 + \lambda_2 = 8/\text{hr})$, and so the total in the next 30 minutes is $X \sim \text{Pois}(4)$. So $\text{P}(X = 2) = e^{-4}4^2/2!$. □

4.2 Continuous Distributions

And now, we continue on to some of our continuous friends...

4.2.1 Uniform Distribution

Definition: The random variable X has the **uniform distribution** with bounds a and b if it has pdf and cdf:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad F(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & x \ge b. \end{cases}$$

Notation: $X \sim \text{Unif}(a, b)$.

Previous work showed that

$$\text{E}[X] = \frac{a+b}{2} \quad \text{and} \quad \text{Var}(X) = \frac{(a-b)^2}{12}.$$

We can also derive the mgf,

$$M_X(t) = \text{E}[e^{tX}] = \int_a^b e^{tx}\frac{1}{b-a}\,dx = \frac{e^{tb}-e^{ta}}{t(b-a)}.$$

4.2.2 Exponential, Erlang, and Gamma Distributions

The exponential distribution has a robust family tree, some branches of which we'll discuss here.

§4.2.2.1 Exponential Distribution

Definition: The **exponential distribution** with rate parameter λ has pdf

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Notation: $X \sim \text{Exp}(\lambda)$.

Previous work showed that the cdf $F(x) = 1 - e^{-\lambda x}$, $\text{E}[X] = 1/\lambda$, and $\text{Var}(X) = 1/\lambda^2$. We also derived the mgf,

$$M_X(t) = \text{E}[e^{tX}] = \int_0^\infty e^{tx} f(x)\, dx = \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$

The exponential distribution has some really cool properties.

Theorem (Memoryless Property of the Exponential): Suppose that $X \sim \text{Exp}(\lambda)$. Then for positive s, t, we have

$$\text{P}(X > s + t \,|\, X > s) = \text{P}(X > t).$$

Similar to the discrete geometric distribution, the probability that X will survive an additional t time units is the (unconditional) probability that it will survive at least t — i.e., it forgot that it already made it past time s. It's always "like new!"

Proof: We have

$$\begin{aligned} \text{P}(X > s + t \,|\, X > s) &= \frac{\text{P}(X > s + t \cap X > s)}{\text{P}(X > s)} = \frac{\text{P}(X > s + t)}{\text{P}(X > s)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = \text{P}(X > t). \quad \square \end{aligned}$$

Example: Suppose that the life of a lightbulb is exponential with a mean of 10 months. If the light survives 20 months, what's the probability that it'll survive another 10?

$$\text{P}(X > 30 \,|\, X > 20) = \text{P}(X > 10) = e^{-\lambda x} = e^{-(1/10)(10)} = e^{-1} \quad \square$$

Example: If the time to the next bus is exponentially distributed with a mean of 10 minutes, and you've already been waiting 20 minutes, you can expect to wait 10 more. $\square$

Remark: The exponential is the *only* continuous distribution with the Memoryless Property.

Remark: Notice how similar $\text{E}[X]$ and $\text{Var}(X)$ are for the geometric and exponential distributions. (Not a coincidence!)

Definition: If X is a continuous random variable with pdf $f(x)$ and cdf $F(x)$, then its **failure rate function** is

$$S(t) \equiv \frac{f(t)}{\mathrm{P}(X > t)} = \frac{f(t)}{1 - F(t)},$$

which can loosely be regarded as X's instantaneous rate of death, given that it has so far survived to time t.

Example: If $X \sim \mathrm{Exp}(\lambda)$, then $S(t) = \lambda e^{-\lambda t}/e^{-\lambda t} = \lambda$. So if X is the exponential lifetime of a lightbulb, then its instantaneous burn-out rate is always λ — always good as new! This is clearly a result of the Memoryless Property. □

The exponential (which is continuous) is also related to the Poisson (which is discrete)!

Theorem: Let X be the amount of time until the *first arrival* of a Poisson process with rate λ. Then $X \sim \mathrm{Exp}(\lambda)$.

Proof: The cdf of $F(x)$ is

$$\begin{aligned} F(x) &= \mathrm{P}(X \le x) \\ &= 1 - \mathrm{P}(\text{no arrivals in } [0, x]) \\ &= 1 - \frac{e^{-\lambda x}(\lambda x)^0}{0!} \quad (\text{the number of arrivals in } [0, x] \text{ is } \mathrm{Pois}(\lambda x)) \\ &= 1 - e^{-\lambda x}. \quad \square \end{aligned}$$

Theorem: Amazingly, it can be shown (after a lot of work) that the interarrival times of a Poisson process are *all* iid $\mathrm{Exp}(\lambda)$! See for yourself when you take an advanced stochastic processes course.

Example: Suppose that arrivals to a shopping center are from a Poisson process with rate $\lambda = 20\,/\,\mathrm{hr}$. What's the probability that the time between the 13^{th} and 14^{th} customers will be at least four minutes?

Answer: Let the time between customers 13 and 14 be X. Since we have a Poisson process, the interarrivals are iid $\mathrm{Exp}(\lambda = 20\,/\,\mathrm{hr})$, so

$$\mathrm{P}(X > 4 \text{ min}) = \mathrm{P}(X > 1/15 \text{ hr}) = e^{-\lambda t} = e^{-20/15}. \quad \square$$

§4.2.2.2 Erlang and Gamma Distributions

There are several distributions that generalize the exponential distribution.

Definition: Suppose that $X_1, \ldots, X_k \overset{\text{iid}}{\sim} \mathrm{Exp}(\lambda)$, and let $Y = \sum_{i=1}^{k} X_i$. Then Y has the **Erlang distribution** with parameters k and λ, denoted $Y \sim \mathrm{Erlang}_k(\lambda)$.

The Erlang is simply the sum of iid exponentials. For instance, if you buy a package of lightbulbs, then the Erlang may be a good way to model the entire lifetime of all of the bulbs. Of course, it's obvious that $\mathrm{Erlang}_1(\lambda) \sim \mathrm{Exp}(\lambda)$.

Theorem: The pdf and cdf of the Erlang are

$$f(y) = \frac{\lambda^k e^{-\lambda y} y^{k-1}}{(k-1)!}, \quad y \geq 0, \quad \text{and}$$

$$F(y) = 1 - \sum_{i=0}^{k-1} \frac{e^{-\lambda y}(\lambda y)^i}{i!}, \quad y \geq 0.$$

Notice that the cdf is the sum of a bunch of Poisson probabilities. (Although we won't do it here, this observation helps in the derivation of the cdf.)

Expected value, variance, and mgf of the Erlang:

$$\begin{aligned} \mathrm{E}[Y] &= \mathrm{E}\Big[\sum_{i=1}^{k} X_i\Big] = \sum_{i=1}^{k} \mathrm{E}[X_i] = k/\lambda \\ \mathrm{Var}(Y) &= k/\lambda^2 \\ M_Y(t) &= \left(\frac{\lambda}{\lambda - t}\right)^k, \quad t < \lambda. \end{aligned}$$

Example: Suppose X_1 and X_2 are iid Exp(3). Find $\mathrm{P}(X_1 + X_2 < 1)$. Since $X_1 + X_2 \sim \mathrm{Erlang}_{k=2}(\lambda = 3)$, we have

$$\mathrm{P}(X_1 + X_2 < 1) = 1 - \sum_{i=0}^{k-1} \frac{e^{-\lambda y}(\lambda y)^i}{i!} = 1 - \sum_{i=0}^{2-1} \frac{e^{-(3\cdot 1)}(3\cdot 1)^i}{i!} = 0.801. \quad \square$$

We're not quite done with our generalization festival...

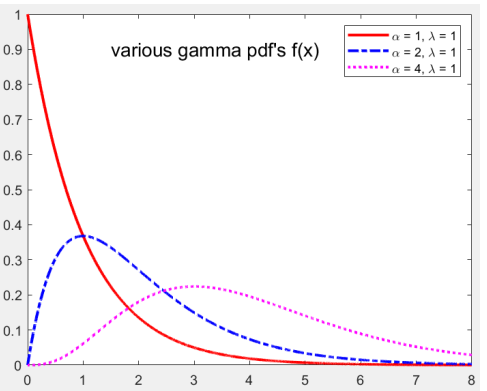
Definition: X has the **gamma distribution** with parameters $\alpha > 0$ and $\lambda > 0$, if it has pdf

$$f(x) = \frac{\lambda^\alpha e^{-\lambda x} x^{\alpha-1}}{\Gamma(\alpha)}, \quad x \geq 0,$$

where

$$\Gamma(\alpha) \equiv \int_0^\infty s^{\alpha-1} e^{-s}\, ds$$

is the *gamma function*. We write $X \sim \mathrm{Gam}(\alpha, \lambda)$.



Remark: The gamma distribution generalizes the Erlang distribution (where α has to be a positive integer). It has the same expected value, variance, and mgf as the Erlang, with α in place of k. (See Exercise 4.6.13.)

Remark: It is easy to show that $\Gamma(1) = 1$. Moreover, if $\alpha > 0$, then $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$. In particular, if α is a positive integer, then $\Gamma(\alpha + 1) = \alpha!$. And here's a tidbit that will make you popular at parties: $\Gamma(1/2) = \sqrt{\pi}$. (See Exercise 4.6.14.)

4.2.3 Other Continuous Distributions

We'll briefly mention a few other continuous distributions that are good to know about, though we'll save the normal distribution for its very own, special section.

Triangular(a, b, c) **distribution** — good for modeling RV's on the basis of limited data (minimum, mode, maximum). The pdf and moments are:

$$f(x) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)}, & a < x \le b \\ \frac{2(c-x)}{(c-b)(c-a)}, & b < x < c \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathrm{E}[X] = \frac{a+b+c}{3} \quad \text{and} \quad \mathrm{Var}(X) = \text{mess.}$$

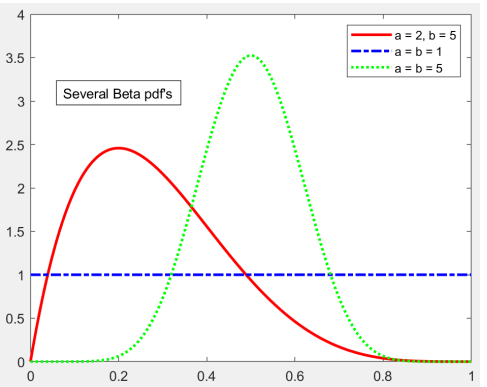
Beta(a, b) **distribution** — good for modeling RV's that are restricted to an interval. The pdf and moments are:

$$f(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \ 0 < x < 1,$$

$$\mathrm{E}[X] = \frac{a}{a+b} \quad \text{and} \quad \mathrm{Var}(X) = \frac{ab}{(a+b)^2(a+b+1)} \quad \text{(see Exercise 4.6.15).}$$

This distribution gets its name from the *beta function*, which is defined as

$$\beta(a, b) \equiv \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} = \int_0^1 x^{a-1}(1-x)^{b-1}\, dx.$$

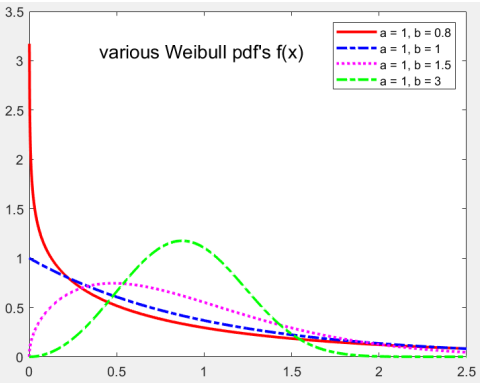


Remark: Certain versions of the uniform ($a = b = 1$) and triangular distributions are special cases of the beta.

Weibull(a, b) **distribution** — good for modeling reliability models. The constant a is known as the "scale" parameter, and b is the "shape" parameter. The pdf, cdf, expected value, and variance are (see Exercise 4.6.16)

$$f(x) = ab(ax)^{b-1}e^{-(ax)^b} \quad \text{and} \quad F(x) = 1 - \exp[-(ax)^b], \; x > 0,$$

$$\mathrm{E}[X] = \frac{1}{a}\Gamma\left(1 + \frac{1}{b}\right) \quad \text{and} \quad \mathrm{Var}(X) = \frac{1}{a^2}\left[\Gamma\left(1 + \frac{2}{b}\right) - \Gamma^2\left(1 + \frac{1}{b}\right)\right].$$



Remark: The exponential is also a special case of the Weibull.

Example: The time-to-failure T for a transmitter has a Weibull distribution with parameters $a = 1/(200 \text{ hours})$ and $b = 1/3$. Then

$$\mathrm{E}[T] = 200\,\Gamma(1 + 3) = 1200 \text{ hours}.$$

The probability that it fails before 2000 hours is

$$F(2000) = 1 - \exp[-(2000/200)^{1/3}] = 0.884. \quad \square$$

Cauchy distribution — A "fat-tailed" distribution good for disproving things!

$$f(x) = \frac{1}{\pi(1 + x^2)} \quad \text{and} \quad F(x) = \frac{1}{2} + \frac{\arctan(x)}{\pi}, \quad x \in \mathbb{R}.$$

Theorem: The Cauchy distribution has an undefined mean and infinite variance!

Weird Fact: $X_1, \ldots, X_n \overset{\text{iid}}{\sim}$ Cauchy $\Rightarrow \sum_{i=1}^n X_i/n \sim$ Cauchy. Even if you take the average of a bunch of Cauchys, you're right back where you started!

Alphabet Soup of Other Distributions

- χ^2 distribution — coming up when we talk Statistics.

- t distribution — coming up.
- F distribution — coming up.
- Distributions named after old statisticians: Pareto, LaPlace, Rayleigh, Gumbel, Johnson.
- Etc. . . .

4.3 The Normal Distribution and the Central Limit Theorem

The normal distribution is so important that we're giving it an entire section!

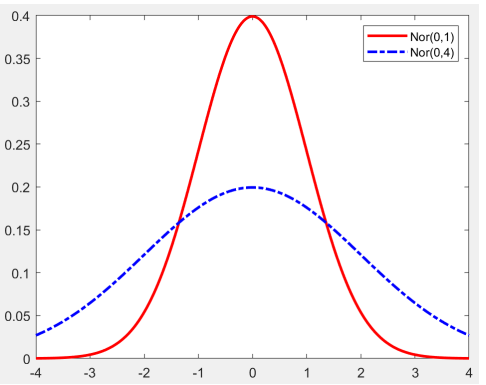
4.3.1 Basics

Definition: The random variable X has the **normal distribution with parameters μ and σ^2** if it has pdf

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right], \quad \forall x \in \mathbb{R}.$$

$f(x)$ is "bell-shaped" and symmetric around $x = \mu$, with tails falling off quickly as you move away from μ.

A normal with a small σ^2 corresponds to a "tall, skinny" bell curve; a large σ^2 is associated with a "short, fat" bell curve.



Remark: The normal distribution is also called the **Gaussian distribution.**

Examples: Heights, weights, SAT scores, crop yields, and averages of things tend to be normal.

Notation: $X \sim \text{Nor}(\mu, \sigma^2)$.

Fun Fact (1): $\int_{\mathbb{R}} f(x)\,dx = 1$.

Proof: Let $z = (x-\mu)/\sigma$ and note that

$$\int_{\mathbb{R}} f(x)\,dx \;=\; \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] dx \;=\; \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\,dz.$$

The result now follows from the polar coordinates example in §1.2.2.6. □

Fun Fact (2): The cdf is

$$F(x) \;=\; \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y-\mu)^2}{2\sigma^2}\right] dy \;= ??.$$

Remark: There is no closed-form expression for this quantity, though very precise approximations obtained via numerical methods are widely available. Stay tuned.

Fun Fact (3): If X is *any* normal RV, then

$$\begin{aligned} \mathrm{P}(\mu-\sigma < X < \mu+\sigma) &\doteq 0.6827 \\ \mathrm{P}(\mu-2\sigma < X < \mu+2\sigma) &\doteq 0.9545 \\ \mathrm{P}(\mu-3\sigma < X < \mu+3\sigma) &\doteq 0.9973. \end{aligned}$$

So almost all of the probability is contained within three standard deviations of the mean. (This is sort of what Toyota is referring to when it brags about "six-sigma" quality.)

Fun Fact (4): The mgf is $M_X(t) = \mathrm{E}[e^{tX}] = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$.

Proof: By LOTUS, we have

$$\begin{aligned} M_X(t) &= \int_{\mathbb{R}} e^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \\ &= \int_{\mathbb{R}} e^{t(\mu+\sigma z)} \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-z^2/2}\,\sigma\,dz \quad (\text{letting } z = (x-\mu)/\sigma) \\ &= \exp\left[\mu t + \frac{\sigma^2 t^2}{2}\right] \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(z-\sigma t)^2}{2}\right] dz, \end{aligned}$$

and we are done since the integrand is simply the pdf of a $\mathrm{Nor}(\sigma t, 1)$ distribution and thus integrates to 1. □

Fun Facts (5) and (6): $\mathrm{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

Proof: By Fun Fact (4), the chain rule, and the obvious result that $M_X(0) = 1$, we have

$$\mathrm{E}[X] \;=\; \frac{d}{dt} M_X(t)\big|_{t=0} \;=\; (\mu + \sigma^2 t) M_X(t)\big|_{t=0} \;=\; \mu M_X(0) \;=\; \mu. \quad \square$$

Similarly, but with the added use of the product rule, we have

$$\begin{aligned}
\mathrm{E}[X^2] &= \frac{d^2}{dt^2} M_X(t)\big|_{t=0} \\
&= \frac{d}{dt}\big[(\mu+\sigma^2 t) M_X(t)\big]\big|_{t=0} \\
&= \big[\sigma^2 M_X(t) + (\mu+\sigma^2 t) M_X'(t)\big]\big|_{t=0} \\
&= \big[\sigma^2 M_X(t) + (\mu+\sigma^2 t)^2 M_X(t)\big]\big|_{t=0} \\
&= \sigma^2 + \mu^2.
\end{aligned}$$

Thus, $\mathrm{Var}(X) = \mathrm{E}[X^2] - \big(\mathrm{E}[X]\big)^2 = (\sigma^2+\mu^2) - \mu^2 = \sigma^2$. □

The following very general result says that we can add up independent normals and still get a normal.

Theorem (Additive Property of Normals): If $X_1, \ldots, X_n$ are *independent* with $X_i \sim \mathrm{Nor}(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$, then

$$Y \equiv \sum_{i=1}^n a_i X_i + b \sim \mathrm{Nor}\Big(\sum_{i=1}^n a_i\mu_i + b, \sum_{i=1}^n a_i^2\sigma_i^2\Big).$$

So a linear combination of independent normals is itself normal.

Proof: Since Y is a linear function,

$$\begin{aligned}
M_Y(t) &= M_{\sum_i a_i X_i + b}(t) = e^{tb} M_{\sum_i a_i X_i}(t) \\
&= e^{tb} \prod_{i=1}^n M_{a_i X_i}(t) \quad (X_i\text{'s independent}) \\
&= e^{tb} \prod_{i=1}^n M_{X_i}(a_i t) \quad (\text{mgf of linear function}) \\
&= e^{tb} \prod_{i=1}^n \exp\Big[\mu_i(a_i t) + \frac{1}{2}\sigma_i^2 (a_i t)^2\Big] \quad (\text{normal mgf}) \\
&= \exp\Big[\Big(\sum_{i=1}^n \mu_i a_i + b\Big)t + \frac{1}{2}\Big(\sum_{i=1}^n a_i^2\sigma_i^2\Big)t^2\Big],
\end{aligned}$$

and we are done by mgf uniqueness. □

Remark: A normal distribution is *completely characterized* by its mean and variance.

By the above, we know that a linear combination of independent normals is still normal. Therefore, when we add up independent normals, all we have to do is figure out the mean and variance — the normality of the sum comes for free.

Example: Suppose that $X \sim \mathrm{Nor}(3,4)$, $Y \sim \mathrm{Nor}(4,6)$, and X and Y are independent. Find the distribution of $2X - 3Y$.

Solution: This is *normal* with

$$\mathrm{E}[2X - 3Y] = 2\mathrm{E}[X] - 3\mathrm{E}[Y] = 2(3) - 3(4) = -6$$

and

$$\text{Var}(2X - 3Y) \;=\; 4\text{Var}(X) + 9\text{Var}(Y) \;=\; 4(4) + 9(6) \;=\; 70.$$

Thus, $2X - 3Y \sim \text{Nor}(-6, 70)$. □

Corollary (of Additive Property):

$$X \;\sim\; \text{Nor}(\mu, \sigma^2) \;\Rightarrow\; aX + b \;\sim\; \text{Nor}(a\mu + b, a^2\sigma^2).$$

Proof: Immediate from the Additive Property, after noting that $\text{E}[aX+b] = a\mu+b$ and $\text{Var}(aX + b) = a^2\sigma^2$. □

Here is an extremely important corollary of the previous corollary.

Corollary (Standardizing a Normal):

$$X \;\sim\; \text{Nor}(\mu, \sigma^2) \;\;\Rightarrow\;\; Z \;\equiv\; \frac{X - \mu}{\sigma} \;\sim\; \text{Nor}(0, 1).$$

Proof: Use the above corollary with $a = 1/\sigma$ and $b = -\mu/\sigma$. □

The $\text{Nor}(0, 1)$ distribution is special, as we shall now see...

4.3.2 The Standard Normal Distribution

Definition: The $\text{Nor}(0, 1)$ distribution is called the **standard normal distribution**, and is often denoted by Z. The pdf and cdf of the Nor(0,1) are

$$\phi(z) \;\equiv\; \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad \text{and} \quad \Phi(z) \;\equiv\; \int_{-\infty}^{z} \phi(y)\, dy, \quad z \in \mathbb{R},$$

respectively, where the pdf / cdf get their own Greek letter ϕ / Φ to use instead of boring ol' f / F!

Remarks: The $\text{Nor}(0, 1)$ is nice because there are tables available for its cdf. For instance, see Table B.1 in the Appendix. You can standardize any normal random variable X into a standard normal by applying the transformation $Z = (X - \mu)/\sigma$. Then you can use the cdf tables.

More Remarks: The following results are easy to derive via symmetry arguments and should be committed to memory.

- $\text{P}(Z \le a) \;=\; \Phi(a)$
- $\text{P}(Z \ge b) \;=\; 1 - \Phi(b)$
- $\text{P}(a \le Z \le b) \;=\; \Phi(b) - \Phi(a)$
- $\Phi(0) \;=\; \text{P}(Z \le 0) \;=\; 1/2$
- $\Phi(-b) \;=\; \text{P}(Z \le -b) \;=\; \text{P}(Z \ge b) \;=\; 1 - \Phi(b)$

- $\mathrm{P}(-b \le Z \le b) \;=\; \Phi(b) - \Phi(-b) \;=\; 2\Phi(b) - 1$
- And more generally,

$$\mathrm{P}(\mu - k\sigma \le X \le \mu + k\sigma) \;=\; \mathrm{P}(-k \le Z \le k) \;=\; 2\Phi(k) - 1.$$

So the probability that *any* normal RV is within k standard deviations of its mean doesn't depend on the mean or variance.

Famous Nor$(0,1)$ table values. A lot of people end up memorizing[2] the values in the following table. ☺ In addition, Table B.1 anxiously awaits your perusal. Or you can use convenient software calls, like `normdist` in Excel (which calculates the cdf for *any* normal distribution).

z	$\Phi(z) = \mathrm{P}(Z \le z)$
0.00	0.5000
1.00	0.8413
1.28	$0.8997 \doteq 0.90$
1.645	0.9500
1.96	0.9750
2.33	$0.9901 \doteq 0.99$
3.00	0.9987
4.00	$\doteq 1.0000$

By the earlier "Fun Facts" and the subsequent discussion, the probability that *any* normal RV is within k standard deviations of its mean is given in the following table for various k values.

k	$\mathrm{P}(-k \le Z \le k) \;=\; 2\Phi(k) - 1$
1	0.6827
2	0.9545
3	0.9973
4	0.99994
5	0.9999994
6	1.0000000

Famous Inverse Nor$(0,1)$ table values. The p^{th} **quantile** of Z is the value of z such that $\Phi(z) = \mathrm{P}(Z \le z) = p$; it is denoted by $z_{1-p} \equiv \Phi^{-1}(p)$.

Some of these values $\Phi^{-1}(p)$ are easy to memorize. But in general, you can always use Table B.1 or software such as Excel's `norminv` function, which actually calculates inverses for *any* normal distribution, not just the standard normal.

[2] Maybe "a lot of people" need to get out more.

p	$z_{1-p} = \Phi^{-1}(p)$
0.90	1.28
0.95	1.645
0.975	1.96
0.99	2.33
0.995	2.58

Example: Suppose $X \sim \text{Nor}(21, 4)$. Find $\text{P}(19 < X < 22.5)$.

Answer: **Standardizing,** we get

$$\begin{aligned}
\text{P}(19 < X < 22.5) &= \text{P}\left(\frac{19-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{22.5-\mu}{\sigma}\right) \\
&= \text{P}\left(\frac{19-21}{2} < Z < \frac{22.5-21}{2}\right) \\
&= \text{P}(-1 < Z < 0.75) \\
&= \Phi(0.75) - \Phi(-1) \\
&= \Phi(0.75) - [1 - \Phi(1)] \\
&= 0.7734 - [1 - 0.8413] \quad \text{(from Table B.1)} \\
&= 0.6147. \quad \square
\end{aligned}$$

Example: Suppose that heights of men are $M \sim \text{Nor}(68, 4)$, and heights of women are $W \sim \text{Nor}(65, 1)$. Find the probability that a random woman is taller than a random man.

Answer: Note that

$$\begin{aligned}
W - M &\sim \text{Nor}\big(\text{E}[W-M], \text{Var}(W-M)\big) \\
&\sim \text{Nor}(65 - 68, 1 + 4) \sim \text{Nor}(-3, 5).
\end{aligned}$$

Then

$$\begin{aligned}
\text{P}(W > M) &= \text{P}(W - M > 0) \\
&= \text{P}\left(Z > \frac{0+3}{\sqrt{5}}\right) \\
&= 1 - \Phi(3/\sqrt{5}) \\
&\doteq 1 - 0.910 = 0.090. \quad \square
\end{aligned}$$

4.3.3 The Sample Mean of Normal Observations

Recall that the **sample mean** of the random variables $X_1, \ldots, X_n$ is $\bar{X} \equiv \sum_{i=1}^{n} X_i/n$.

Corollary (of old Theorem): $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2) \Rightarrow \bar{X} \sim \text{Nor}(\mu, \sigma^2/n)$.

Proof: By previous work, as long as $X_1, \ldots, X_n$ are iid something, we have $\text{E}[\bar{X}] = \mu$ and $\text{Var}(\bar{X}) = \sigma^2/n$. Since $\bar{X}$ is a linear combination of independent normals, it's also normal. Done. $\square$

Remark: This result is *very significant!* As the number of observations increases, $\mathrm{Var}(\bar{X})$ gets *smaller* while $\mathrm{E}[\bar{X}] = \mathrm{E}[X_i] = \mu$ remains constant. In fact, in the next example and theorem, we'll see some consequences of the corollary.

First of all, here's a nice little application in which we determine the sample size necessary to ensure that $\bar{X}$ will have a good chance of being close to μ.

Theorem: Suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Nor}(\mu, \sigma^2)$. Then the sample size n that guarantees the requirement

$$\mathrm{P}\big(|\bar{X} - \mu| \le c\big) \;\ge\; \gamma$$

is

$$n \;\ge\; \frac{\sigma^2}{c^2}\left[\Phi^{-1}\left(\frac{1+\gamma}{2}\right)\right]^2.$$

Proof: Note that $\bar{X} \sim \mathrm{Nor}(\mu, \sigma^2/n)$. Then

$$\begin{aligned}
\mathrm{P}\big(|\bar{X} - \mu| \le c\big) &= \mathrm{P}(-c \le \bar{X} - \mu \le c) \\
&= \mathrm{P}\left(\frac{-c}{\sigma/\sqrt{n}} \le \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \le \frac{c}{\sigma/\sqrt{n}}\right) \\
&= \mathrm{P}\left(\frac{-c\sqrt{n}}{\sigma} \le Z \le \frac{c\sqrt{n}}{\sigma}\right) \\
&= 2\Phi(c\sqrt{n}/\sigma) - 1.
\end{aligned}$$

Now we have to find n such that this probability is at least γ. To this end,

$$\begin{aligned}
2\Phi(c\sqrt{n}/\sigma) - 1 \;\ge\; \gamma \quad &\Leftrightarrow \quad \Phi(c\sqrt{n}/\sigma) \;\ge\; \frac{1+\gamma}{2} \\
&\Leftrightarrow \quad \frac{c\sqrt{n}}{\sigma} \;\ge\; \Phi^{-1}\left(\frac{1+\gamma}{2}\right),
\end{aligned}$$

and we are done after squaring. □

Example: Suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Nor}(\mu, 16)$. Find the sample size n such that

$$\mathrm{P}\big(|\bar{X} - \mu| \le 1\big) \;\ge\; 0.95.$$

Solution: By the above result, we have

$$n \;\ge\; \frac{4^2}{1^2}\left[\Phi^{-1}\left(\frac{1+0.95}{2}\right)\right]^2 \;=\; 16\left[\Phi^{-1}(0.975)\right]^2 \;=\; 16(1.96)^2 \;=\; 61.47.$$

So if you take the average of 62 observations, then $\bar{X}$ has a 95% chance of being within 1 of the true (but unknown) mean μ. □

Thus, the example shows that as the number of normal observations n increases, the sample mean $\bar{X}$ tends to hang around the neighborhood of μ.

At this point, we state (and simultaneously prove) the **(Weak) Law of Large Numbers (LLN)**, which makes the same conclusion *without having to assume that the underlying observations are normal!*

Law of Large Numbers: Suppose $X_1, X_2, \ldots, X_n$ are iid ***anything*** with finite mean μ and finite variance σ^2. Then, for any fixed $c > 0$, the sample mean $\bar{X}$ *converges* to μ in the following sense.

$$\mathrm{P}\big(\big|\bar{X} - \mu\big| > c\big) \;\leq\; \frac{\mathrm{Var}(\bar{X})}{c^2} \;=\; \frac{\sigma^2}{nc^2} \;\to\; 0, \quad \text{as } n \to \infty.$$

Proof: Already done in the statement of the result, thanks to Chebychev's Inequality! □

The LLN shows that the probability that the sample mean deviates from the true mean by more than any fixed amount goes to zero as the sample size $n \to \infty$! Or in plain English,

The sample mean $\bar{X}$ converges to the true mean μ as $n \to \infty$, i.e, $\bar{X} \to \mu$.

In the upcoming statistics portion of the course (see, e.g., Chapter 5), we'll learn that this makes the sample mean $\bar{X}$ an excellent **estimator** for the mean $\mathrm{E}[X_i] = \mu$, which is typically unknown in practice. Can you just imagine what the *Strong* LLN can do? ☺

4.3.4 The Central Limit Theorem

Before we get to the good stuff, let's start off with a useful definition regarding the convergence of a sequence of cdf's to a limiting cdf.

Definition: The sequence of random variables $Y_1, Y_2, \ldots$ with respective cdf's $F_{Y_1}(y), F_{Y_2}(y), \ldots$ **converges in distribution** to the random variable Y having cdf $F_Y(y)$ if $\lim_{n\to\infty} F_{Y_n}(y) = F_Y(y)$ for all y belonging to the continuity set of Y, i.e., the set of y-values for which $F_Y(y)$ is continuous.

Notation: $Y_n \xrightarrow{d} Y$.

Idea: If $Y_n \xrightarrow{d} Y$ and n is large, then you ought to be able to approximate the distribution of Y_n by the limit distribution of Y.

Example: Suppose $Y_n \sim \mathrm{Exp}(\lambda + \frac{1}{n})$, for $n = 1, 2, \ldots$, and $Y \sim \mathrm{Exp}(\lambda)$. Then the cdf of Y_n is $F_{Y_n}(y) = 1 - \exp[-(\lambda + \frac{1}{n})y]$, $y > 0$, and Y's is $F_Y(y) = 1 - e^{-\lambda y}$, $y > 0$. It is clear that $F_{Y_n}(y) \to F_Y(y)$ for all $y > 0$, so that $Y_n \xrightarrow{d} Y$. □

We now offer you the most-important theorem in probability and statistics, which shows that sample means and sums of *any iid observations* tend to become normally distributed as the sample size increases.

Central Limit Theorem (CLT): Suppose $X_1, \ldots, X_n$ are iid ***anything*** with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Then as $n \to \infty$,

$$Z_n \equiv \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\bar{X} - \mathrm{E}[\bar{X}]}{\sqrt{\mathrm{Var}(\bar{X})}} \xrightarrow{d} \mathrm{Nor}(0,1).$$

Remarks: We have a lot to say about such an important theorem.

- The main take-away: If n is large, then

$$\bar{X} \approx \mathrm{Nor}\big(\mathrm{E}[\bar{X}], \mathrm{Var}(\bar{X})\big) \sim \mathrm{Nor}(\mu, \sigma^2/n) \quad \text{and} \quad \sum_{i=1}^n X_i \approx \mathrm{Nor}(n\mu, n\sigma^2).$$

- The X_i's **don't have to be normal** for the CLT to work! It even works on discrete distributions!
- You usually need $n \geq 30$ observations for the approximation to work well. (Need fewer observations if the X_i's come from a symmetric distribution.)
- You can almost always use the CLT if the observations are iid — you just have to have finite μ and σ^2.
- In fact, the CLT is actually a lot more general than the theorem presented here! In some cases (not discussed here), it works for RV's that are modestly correlated and/or not from the same distribution!

Honors Proof of CLT

Suppose that the mgf $M_X(t)$ of the X_i's exists and satisfies certain technical conditions that you don't need to know about. Moreover, without loss of generality (since we're standardizing anyway) and for notational convenience, we'll assume that $\mu = 0$ and $\sigma^2 = 1$. We will be done if we can show that the mgf of Z_n converges to the mgf of $Z \sim \mathrm{Nor}(0,1)$. That is, we need to show that $M_{Z_n}(t) \to e^{t^2/2}$ as $n \to \infty$.

To get things going, the mgf of Z_n is

$$\begin{aligned} M_{Z_n}(t) &= M_{\sum_{i=1}^n X_i/\sqrt{n}}(t) \\ &= M_{\sum_{i=1}^n X_i}(t/\sqrt{n}) \quad \text{(mgf of a linear function of a RV)} \\ &= \big[M_X(t/\sqrt{n})\big]^n \quad (X_i\text{'s are iid}). \end{aligned}$$

Thus, taking logs, our goal is to show that

$$\lim_{n\to\infty} \ell\mathrm{n}\big(M_{Z_n}(t)\big) = \lim_{n\to\infty} n\,\ell\mathrm{n}\big(M_X(t/\sqrt{n})\big) = t^2/2.$$

If we let $y = 1/\sqrt{n}$, our revised goal is to show that

$$\lim_{y\to 0} \frac{\ell\mathrm{n}\big(M_X(ty)\big)}{y^2} = t^2/2.$$

Before proceeding further, note that

$$\lim_{y\to 0} \ell\mathrm{n}\big(M_X(ty)\big) \;=\; \ell\mathrm{n}\big(M_X(0)\big) \;=\; \ell\mathrm{n}(1) \;=\; 0 \tag{4.3}$$

and

$$\lim_{y\to 0} M_X'(ty) \;=\; M_X'(0) \;=\; \mathrm{E}[X] \;=\; \mu \;=\; 0, \tag{4.4}$$

where the last equality is from our standardization assumption.

So after all of this build-up, we have

$$\begin{aligned}
&\lim_{y\to 0} \frac{\ell\mathrm{n}\big(M_X(ty)\big)}{y^2} \\
&\quad= \lim_{y\to 0} \frac{t\,M_X'(ty)}{2y\,M_X(ty)} \quad \text{(by (4.3) et L'Hôpital to deal with } 0\,/\,0) \\
&\quad= \lim_{y\to 0} \frac{t^2\,M_X''(ty)}{2\,M_X(ty) + 2yt\,M_X'(ty)} \quad \text{(by (4.4) et L'Hôpital encore)} \\
&\quad= \frac{t^2\,M_X''(0)}{2\,M_X(0)+0} \;=\; \frac{t^2\,\mathrm{E}[X^2]}{2} \;=\; \frac{t^2}{2} \quad (\mathrm{E}[X^2] = \sigma^2 + \mu^2 = 1). \quad \text{☺}
\end{aligned}$$

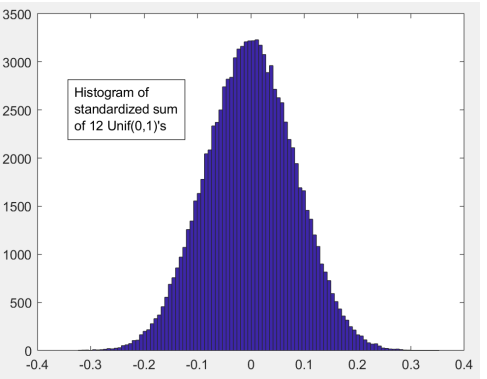
4.3.5 CLT Examples

Just to show that the CLT really works, we'll give a few illustrative examples.

Example: Let's add up some iid Unif(0,1)'s, $U_1, \ldots, U_n$. Let $S_n = \sum_{i=1}^n U_i$. Note that $\mathrm{E}[S_n] = n\mathrm{E}[U_i] = n/2$, and $\mathrm{Var}(S_n) = n\mathrm{Var}(U_i) = n/12$. Therefore, the CLT says that, for large n,

$$Z_n \;\equiv\; \frac{S_n - n/2}{\sqrt{n/12}} \;\approx\; \mathrm{Nor}(0,1).$$

The figure below is a histogram compiled using 100,000 simulation realizations of $Z_{12} = S_{12} - 6$ for $n = 12$. Even with this seemingly small value of $n = 12$, the resulting histogram looks beautifully normal — almost certainly due to the symmetry of the underlying Unif(0,1) pdf of the constituents of S_n. □



Example: Suppose $X_1, \ldots, X_{100} \overset{\text{iid}}{\sim} \text{Exp}(1/1000)$. Find $\text{P}(950 \le \bar{X} \le 1050)$.

Solution: Recall that if $X_i \sim \text{Exp}(\lambda)$, then $\text{E}[X_i] = 1/\lambda$ and $\text{Var}(X_i) = 1/\lambda^2$. Further, since $\bar{X}$ is the sample mean based on n observations, then

$$\text{E}[\bar{X}] \;=\; \text{E}[X_i] \;=\; 1/\lambda \quad \text{and}$$

$$\text{Var}(\bar{X}) \;=\; \text{Var}(X_i)/n \;=\; 1/(n\lambda^2).$$

For our problem, $\lambda = 1/1000$ and $n = 100$, so that $\text{E}[\bar{X}] = 1000$ and $\text{Var}(\bar{X}) = 10000$. These results and the CLT immediately imply that

$$\bar{X} \;\approx\; \text{Nor}\big(\text{E}[\bar{X}],\, \text{Var}(\bar{X})\big) \;\sim\; \text{Nor}\big(1000,\, 10000\big).$$

Thus,

$$\begin{aligned}
&\text{P}(950 \le \bar{X} \le 1050) \\
&\quad = \quad \text{P}\Big(\frac{950 - \text{E}[\bar{X}]}{\sqrt{\text{Var}(\bar{X})}} \le \frac{\bar{X} - \text{E}[\bar{X}]}{\sqrt{\text{Var}(\bar{X})}} \le \frac{1050 - \text{E}[\bar{X}]}{\sqrt{\text{Var}(\bar{X})}}\Big) \\
&\quad \doteq \quad \text{P}\Big(\frac{950 - 1000}{100} \le Z \le \frac{1050 - 1000}{100}\Big) \quad (\text{where } Z \sim \text{Nor}(0,1)) \\
&\quad = \quad \text{P}(-0.5 \le Z \le 0.5) \;=\; 2\Phi(1/2) - 1 \;=\; 0.383. \quad \square
\end{aligned}$$

Remark: Since the X_i's are iid $\text{Exp}(1/1000)$, we know that

$$100\,\bar{X} \;=\; \sum_{i=1}^{100} X_i \;\sim\; \text{Erlang}_{100}(1/1000) \;\sim\; \text{Gam}(100, 1/1000).$$

This indicates that this problem can be solved *exactly* — for instance, the Excel gamma cdf function, $\texttt{gammadist}(x, n, 1/\lambda, 1)$, yields

$$\begin{aligned}
\text{P}(950 \le \bar{X} \le 1050) \;&=\; \text{P}\left(95000 \le \sum_{i=1}^{100} X_i \le 105000\right) \\
&= \quad \texttt{gammadist}(105000, 100, 1000, 1) - \texttt{gammadist}(95000, 100, 1000, 1).
\end{aligned}$$

And what do you know, you end up with precisely the same answer of 0.383, so the CLT normal approximation is really good! $\square$

Example: Suppose $X_1, \ldots, X_{100}$ are iid from some distribution with mean 1000 and standard deviation 1000. Find $\text{P}(950 \le \bar{X} \le 1050)$.

Solution: By exactly the same manipulations as in the previous example, the answer $\doteq 0.383$. Notice that we don't care whether or not the data comes from an exponential distribution. We just need the mean and variance. $\square$

Normal Approximation to the Binomial: Suppose $Y \sim \text{Bin}(n, p)$, where n is very large. In such cases, we usually approximate the binomial via an appropriate normal distribution. The CLT applies since $Y = \sum_{i=1}^{n} X_i$, where the X_i's are iid $\text{Bern}(p)$. Then

$$\frac{Y - \text{E}[Y]}{\sqrt{\text{Var}(Y)}} \;=\; \frac{Y - np}{\sqrt{npq}} \;\approx\; \text{Nor}(0,1),$$

in which case $Y \approx \text{Nor}(np,\, npq)$.

The usual rule of thumb for the normal approximation to the binomial is that it works pretty well as long as $np \geq 5$ and $nq \geq 5$.

Why do we need such an approximation?

Example: Suppose $Y \sim \text{Bin}(100,\, 0.8)$, and we want to calculate

$$\text{P}(Y \geq 84) \;=\; \sum_{i=84}^{100} \binom{100}{i} (0.8)^i (0.2)^{100-i}.$$

Good luck with the binomial coefficients (they're too big) and the number of terms to sum up (it's going to get tedious). I'll come back to visit you in an hour. □

The next example shows how to use the approximation. Note that it incorporates a **"continuity correction"** to account for the fact that the binomial is *discrete* while the normal is *continuous*. (But if you don't want to use it, don't worry too much.)

Example: The Atlanta Braves play 100 independent baseball games, each of which they have probability 0.8 of winning. What's the probability that they win at least 84 games?

Answer: $Y \sim \text{Bin}(100, 0.8)$, and we want $\text{P}(Y \geq 84)$ (as in the last example). So,

$$\begin{aligned}
\text{P}(Y \geq 84) &= \text{P}(Y \geq 83.5) \quad (\text{"continuity correction," since } 84 \doteq [83.5,\, 84.5]) \\
&\doteq \text{P}\left(Z \geq \frac{83.5 - np}{\sqrt{npq}} \right) \quad (\text{CLT, with } Z \sim \text{Nor}(0,1)) \\
&= \text{P}\left(Z \geq \frac{83.5 - 80}{\sqrt{16}} \right) \\
&= \text{P}(Z \geq 0.875) \;=\; 0.1908.
\end{aligned}$$

The actual answer (using the true $\text{Bin}(100, 0.8)$ distribution) turns out to be 0.1923 — pretty close! □

4.4 Extensions of the Normal Distribution

The normal distribution has an enormous number of extensions/applications. We present some of these here.

4.4.1 Bivariate Normal Distribution

Definition: (X, Y) has the **bivariate normal distribution** if it has joint pdf

$$f(x,y) \;=\; C \exp\left\{ \frac{-\left[z_X^2(x) - 2\rho z_X(x) z_Y(y) + z_Y^2(y) \right]}{2(1-\rho^2)} \right\},$$

where μ_X, μ_Y, σ_X^2, and σ_Y^2 are the obvious marginal moments of X and Y,

$$\rho \equiv \text{Corr}(X,Y), \quad C \equiv \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}},$$

$$z_X(x) \equiv \frac{x-\mu_X}{\sigma_X} \quad \text{and} \quad z_Y(y) \equiv \frac{y-\mu_Y}{\sigma_Y}.$$

Pretty nasty joint pdf, eh?

Fun Fact: The marginals $X \sim \text{Nor}(\mu_X, \sigma_X^2)$ and $Y \sim \text{Nor}(\mu_Y, \sigma_Y^2)$.

Example: (X,Y) could be a person's (height, weight). The two quantities are marginally normal, but positively correlated. □

If you want to calculate bivariate normal probabilities, you'll need to evaluate quantities like

$$\text{P}(a < X < b,\, c < Y < d) = \int_c^d \int_a^b f(x,y)\,dx\,dy,$$

which will probably require numerical integration techniques. There are also easy-to-use function calls available in statistics packages such as R.

Fun Fact (which comes up in the study of Regression): The conditional distribution of Y given that $X = x$, is also normal. In particular,

$$Y|X = x \sim \text{Nor}\big(\mu_Y + \rho(\sigma_Y/\sigma_X)(x-\mu_X),\, \sigma_Y^2(1-\rho^2)\big).$$

Thus, information about X helps to update the distribution of Y.

Example: Consider students at a tough university. Let X be their combined SAT scores (Math and Verbal), and Y their freshman GPA (out of 4). Suppose a study reveals that

$$\mu_X = 1300, \quad \mu_Y = 2.3, \quad \sigma_X^2 = 6400, \quad \sigma_Y^2 = 0.25, \quad \text{and} \quad \rho = 0.6.$$

Find $\text{P}(Y \geq 2 \,|\, X = 900)$.

Solution: First of all,

$$\begin{aligned} \text{E}[Y|X=900] &= \mu_Y + \rho(\sigma_Y/\sigma_X)(x-\mu_X) \\ &= 2.3 + (0.6)(\sqrt{0.25/6400}\,)(900-1300) \\ &= 0.8, \end{aligned}$$

indicating that the expected GPA of a kid with 900 SAT's will be 0.8.

Secondly,

$$\text{Var}(Y|X=900) = \sigma_Y^2(1-\rho^2) = 0.16.$$

Thus,

$$Y|X = 900 \sim \text{Nor}(0.8, 0.16).$$

Now we can calculate

$$\text{P}(Y \geq 2 \,|\, X = 900) = \text{P}\left(Z \geq \frac{2-0.8}{\sqrt{0.16}}\right) = 1 - \Phi(3) = 0.0013.$$

This guy doesn't have much chance of having a good GPA. □

The bivariate normal distribution is easily generalized to the multivariate case.

Honors Definition: The random vector $\boldsymbol{X} = (X_1, \ldots, X_k)^{\mathrm{T}}$ has the **multivariate normal distribution**, with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)^{\mathrm{T}}$ and $k \times k$ **covariance matrix** $\boldsymbol{\Sigma} = (\sigma_{ij})$, if it has multivariate pdf

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}\right\}, \quad \boldsymbol{x} \in \mathbb{R}^k,$$

where $|\boldsymbol{\Sigma}|$ and $\boldsymbol{\Sigma}^{-1}$ are the determinant and inverse of $\boldsymbol{\Sigma}$, respectively.

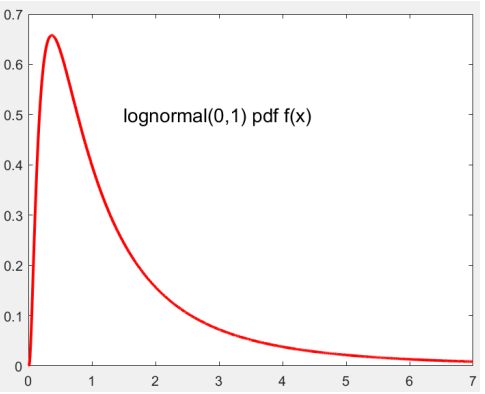
It turns out that for $i, j = 1, 2, \ldots, k$,

$$\mathrm{E}[X_i] = \mu_i, \quad \mathrm{Var}(X_i) = \sigma_{ii}, \quad \text{and} \quad \mathrm{Cov}(X_i, X_j) = \sigma_{ij}.$$

Notation: $\boldsymbol{X} \sim \mathrm{Nor}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

4.4.2 Lognormal Distribution

Definition: If $Y \sim \mathrm{Nor}(\nu, \tau^2)$, then $X \equiv e^Y$ has the **lognormal distribution** with parameters (ν, τ^2). This distribution has tremendous uses, e.g., in the pricing of certain stock options.



Fun Facts: Let $\phi(x)$ and $\Phi(x)$ respectively denote our old friends, the pdf and cdf of the standard normal distribution. Then the pdf, cdf, and moments of the lognormal are

$$\begin{aligned}
f(x) &= \frac{1}{\tau x}\,\phi\left(\frac{\ell\mathrm{n}(x)-\nu}{\tau}\right) = \frac{1}{x\tau\sqrt{2\pi}}\exp\left\{-\frac{[\ell\mathrm{n}(x)-\nu]^2}{2\tau^2}\right\}, \; x > 0, \\
F(x) &= \Phi\left(\frac{\ell\mathrm{n}(x)-\nu}{\tau}\right), \; x > 0, \quad \text{and} \\
\mathrm{E}[X^k] &= \exp\left\{k\nu + \frac{k^2\tau^2}{2}\right\}, \quad k = 1, 2, \ldots.
\end{aligned}$$

In particular,

$$\mathrm{E}[X] \;=\; e^{\nu+\frac{\tau^2}{2}} \quad \text{and} \quad \mathrm{Var}(X) \;=\; e^{2\nu+\tau^2}\big(e^{\tau^2}-1\big).$$

Example: Suppose $Y \sim \mathrm{Nor}(10, 4)$, and let $X = e^Y$. Then

$$\mathrm{P}(X \le 1000) \;=\; \mathrm{P}\left(Z \le \frac{\ell\mathrm{n}(1000)-10}{2}\right) \;=\; \Phi(-1.55) \;=\; 0.061,$$

and

$$\mathrm{E}[X] \;=\; e^{10+\frac{4}{2}} \;=\; e^{12},$$

which certainly gives rise to high expectations!

Crazy Fact: Although all of the moments of the lognormal exist, the mgf *doesn't!* This is because $\mathrm{E}[e^{tX}]$ is infinite for $t > 0$. ☹

Honors Example: How to Win a Nobel Prize

It is well-known that stock prices are closely related to the lognormal distribution. In fact, it's common to use the following model for a stock price at a fixed time $t > 0$,

$$S(t) \;=\; S(0)\exp\left\{\left(\mu-\frac{\sigma^2}{2}\right)t+\sigma\sqrt{t}Z\right\},$$

where μ is related to the "drift" of the stock price (i.e., the natural rate of increase), σ is its "volatility" (how much the stock bounces around), $S(0)$ is the initial price, and Z is a standard normal RV.

An active area of finance is to estimate **option prices**. For example, a so-called **European call option** C permits its owner, who pays an up-front fee for the privilege, to purchase the stock at a pre-agreed **strike price** k, at a predetermined **expiry date** T.

For instance, suppose IBM is currently selling for \$100 a share. If I think that the stock will go up in value, I may want to pay \$3/share now for the right to buy IBM at \$105 three months from now.

- If IBM is worth \$120 three months from now, I'll be able to buy it for only \$105, and will have made a profit of \$120 − \$105 − \$3 = \$12.
- If IBM is selling for \$107 three months hence, I can still buy it for \$105, but will lose \$1 (recouping \$2 from my original option purchase).
- If IBM sells for \$95, then I won't exercise my option, and will walk away with my tail between my legs having lost my original \$3.

So what's the option worth (and what should I pay for it)? Its expected value is

$$\mathrm{E}[C] \;=\; e^{-rT}\mathrm{E}\big[(S(T)-k)^+\big],$$

where $x^+ \equiv \max\{0, x\}$, and

- r is the "risk-free" interest rate, e.g., what you can get from a U.S. Treasury bond. This is used instead of the drift μ.
- The term e^{-rT} denotes the time-value of money, i.e., a depreciation term corresponding to the interest I could have made had I used my money to buy a Treasury note.

Using the standard conditioning argument, we can calculate

$$\begin{aligned}
\mathrm{E}[C] &= e^{-rT}\mathrm{E}\left[S(0)\exp\left\{\left(r-\frac{\sigma^2}{2}\right)T+\sigma\sqrt{T}Z\right\}-k\right]^+ \\
&= e^{-rT}\int_{-\infty}^{\infty}\left[S(0)\exp\left\{\left(r-\frac{\sigma^2}{2}\right)T+\sigma\sqrt{T}z\right\}-k\right]^+\phi(z)\,dz \\
&= S(0)\Phi(b+\sigma\sqrt{T})-ke^{-rT}\Phi(b) \quad \text{(after lots of algebra)},
\end{aligned}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the Nor(0,1) pdf and cdf (as usual), and

$$b \equiv \frac{rT-\frac{\sigma^2 T}{2}-\ell\mathrm{n}(k/S(0))}{\sigma\sqrt{T}}.$$

Black and Scholes did the original calculation in 1968 (albeit in a different way). In 1997, Scholes and Merton won a Nobel for this work (poor Black had already reached his own expiry date ☹). There are many, many generalizations of this work that are used in practical finance applications, but this is a great starting point. Meanwhile, get your tickets to Norway or Sweden or wherever they give out the Nobel Prize in Economics! ☺

4.5 Computer Considerations

4.5.1 Evaluating pmf's / pdf's and cdf's

We can use various computer packages (Excel, Minitab, R, SAS, etc.) to calculate values of pmf's/pdf's and cdf's for a variety of common distributions. For instance, in Excel, we find the functions:

- `binomdist` = Binomial distribution
- `expondist` = Exponential
- `negbinomdist` = Negative Binomial
- `normdist` and `normsdist` = Normal and Standard Normal
- `poisson` = Poisson.

Excel functions such as `normsinv`, `tinv`, etc., can calculate the inverses of the standard normal, t, and other distributions. This functionality is useful for obtaining distribution quantiles. It can also be used for purposes of *simulating* various RV's.

4.5.2 Simulating Random Variables

How would one simulate RV's? This is an important problem because simulations are used to evaluate a variety of real-world processes that contain inherent randomness (queueing systems, inventory systems, manufacturing systems, etc.). Although this is not a simulation text, we'll quickly mention a few basics with respect to RV generation.

To start at the very beginning,[3] you can use the Excel function `rand` to simulate a Unif(0,1) RV, say U. Then, building on the Inverse Transform Theorem (§2.8.2), you can simulate any other continuous RV X having cdf $F(x)$, simply by plugging into $X = F^{-1}(U)$.

Of course, you have to have the inverse function on hand in order to be able to do this calculation, and sometimes that's difficult because the inverse doesn't always exist in closed form. The standard statistics packages partially take care of this issue for you. For instance, you can use the Excel standard normal inverse function to generate a standard normal RV by using `normsinv(rand())`. Inverse transform methods for discrete RV's are similar, but you have to be a little careful because the inverse is not always unique for the discrete case.

There are numerous methods for RV generation in addition to Inverse Transform. If you are intrigued, check out Exercise 28 from §2.9 for additional insight.

We end this chapter with a remarkable RV generation example that uses one of the alternative methods...

Theorem (Box and Muller): If $U_1, U_2 \overset{\text{iid}}{\sim} \text{Unif}(0,1)$, then

$$Z_1 = \sqrt{-2\ell\text{n}(U_1)}\cos(2\pi U_2) \quad \text{and} \quad Z_2 = \sqrt{-2\ell\text{n}(U_1)}\sin(2\pi U_2)$$

are iid Nor(0,1).

Example: Suppose that $U_1 = 0.3$ and $U_2 = 0.8$ are realizations of two iid Unif$(0,1)$'s. Use the Box–Muller method to generate two iid standard normals.

Solution: We have

$$\begin{aligned} Z_1 &= \sqrt{-2\ell\text{n}(U_1)}\cos(2\pi U_2) = 0.480 \\ Z_2 &= \sqrt{-2\ell\text{n}(U_1)}\sin(2\pi U_2) = -1.476. \quad \square \end{aligned}$$

Remarks:

- There are lots of ways to generate Nor$(0,1)$'s, but this may be the easiest.
- The cosine and sine terms are calculated in *radians*, not degrees.
- To get $X \sim \text{Nor}(\mu, \sigma^2)$ from $Z \sim \text{Nor}(0,1)$, just take $X = \mu + \sigma Z$.
- Amazingly, it's "Muller," not "Müller!"

[3] The very beginning is a very good place to start.

Honors Proof of Box–Muller

We follow the method given by the main corollary of §3.7.1; namely, if we can express $U_1 = k_1(Z_1, Z_2)$ and $U_2 = k_2(Z_1, Z_2)$ for some functions $k_1(\cdot,\cdot)$ and $k_2(\cdot,\cdot)$, then the joint pdf of (Z_1, Z_2) is given by

$$\begin{aligned} g(z_1, z_2) &= f_{U_1}\big(k_1(z_1, z_2)\big) f_{U_2}\big(k_2(z_1, z_2)\big) \left| \frac{\partial u_1}{\partial z_1}\frac{\partial u_2}{\partial z_2} - \frac{\partial u_2}{\partial z_1}\frac{\partial u_1}{\partial z_2} \right| \\ &= \left| \frac{\partial u_1}{\partial z_1}\frac{\partial u_2}{\partial z_2} - \frac{\partial u_2}{\partial z_1}\frac{\partial u_1}{\partial z_2} \right| \quad \big(U_1 \text{ and } U_2 \text{ are iid Unif}(0,1)\big). \end{aligned}$$

In order to obtain the functions $k_1(Z_1, Z_2)$ and $k_2(Z_1, Z_2)$, note that

$$Z_1^2 + Z_2^2 = -2\,\ell\mathrm{n}(U_1)\big[\cos^2(2\pi U_2) + \sin^2(2\pi U_2)\big] = -2\,\ell\mathrm{n}(U_1),$$

so that

$$U_1 = e^{-(Z_1^2+Z_2^2)/2}.$$

This immediately implies that

$$\begin{aligned} Z_1^2 &= -2\,\ell\mathrm{n}(U_1)\,\cos^2(2\pi U_2) \\ &= -2\,\ell\mathrm{n}\big(e^{-(Z_1^2+Z_2^2)/2}\big)\,\cos^2(2\pi U_2) \\ &= (Z_1^2 + Z_2^2)\,\cos^2(2\pi U_2), \end{aligned}$$

so that

$$U_2 = \frac{1}{2\pi}\arccos\left(\pm\sqrt{\frac{Z_1^2}{Z_1^2+Z_2^2}}\right) = \frac{1}{2\pi}\arccos\left(\sqrt{\frac{Z_1^2}{Z_1^2+Z_2^2}}\right),$$

where we are (non-rigorously) getting rid of the "$\pm$" to balance off the fact that the range of $y = \arccos(x)$ is only regarded to be $0 \le y \le \pi$ (not $0 \le y \le 2\pi$).

Now some derivative fun. Let's start off with the easy one first.

$$\frac{\partial u_1}{\partial z_i} = \frac{\partial}{\partial z_i} e^{-(z_1^2+z_2^2)/2} = -z_i\, e^{-(z_1^2+z_2^2)/2},\ i = 1, 2.$$

Then the more-challenging fellow.

$$\begin{aligned}
\frac{\partial u_2}{\partial z_1} &= \frac{\partial}{\partial z_1}\,\frac{1}{2\pi}\arccos\left(\sqrt{\frac{z_1^2}{z_1^2+z_2^2}}\right) \\
&= \frac{1}{2\pi}\,\frac{-1}{\sqrt{1-\frac{z_1^2}{z_1^2+z_2^2}}}\,\frac{\partial}{\partial z_1}\sqrt{\frac{z_1^2}{z_1^2+z_2^2}} \quad \text{(chain rule)} \\
&= \frac{1}{2\pi}\,\frac{-1}{\sqrt{\frac{z_2^2}{z_1^2+z_2^2}}}\,\frac{1}{2}\left(\frac{z_1^2}{z_1^2+z_2^2}\right)^{-1/2}\frac{\partial}{\partial z_1}\,\frac{z_1^2}{z_1^2+z_2^2} \quad \text{(chain rule again)} \\
&= \frac{-(z_1^2+z_2^2)}{4\pi z_1 z_2}\,\frac{2z_1 z_2^2}{(z_1^2+z_2^2)^2} \\
&= \frac{-z_2}{2\pi(z_1^2+z_2^2)}, \quad \text{and} \\
\frac{\partial u_2}{\partial z_2} &= \frac{z_1}{2\pi(z_1^2+z_2^2)} \quad \text{(after similar algebra).}
\end{aligned}$$

Then we finally have

$$\begin{aligned}
g(z_1,z_2) &= \left|\frac{\partial u_1}{\partial z_1}\,\frac{\partial u_2}{\partial z_2} - \frac{\partial u_2}{\partial z_1}\,\frac{\partial u_1}{\partial z_2}\right| \\
&= \left|-z_1\,e^{-(z_1^2+z_2^2)/2}\,\frac{z_1}{2\pi(z_1^2+z_2^2)} - \frac{z_2}{2\pi(z_1^2+z_2^2)}\,z_2\,e^{-(z_1^2+z_2^2)/2}\right| \\
&= \frac{1}{2\pi}\,e^{-(z_1^2+z_2^2)/2} \;=\; \left(\frac{1}{\sqrt{2\pi}}\,e^{-z_1^2/2}\right)\left(\frac{1}{\sqrt{2\pi}}\,e^{-z_2^2/2}\right),
\end{aligned}$$

which is the product of two iid Nor(0,1) pdf's, so we are done! ☺

4.6 Exercises

1. (§4.1.1) My consulting business is undertaking six independent projects, each of which has an estimated success probability of 0.9. What's the probability that at least five will be successful?

2. (§4.1.1) Find the mean and variance of the Bin(n,p) distribution using the moment generating function.

3. (§4.1.2) Suppose we have a box of sox — four red sox and five blue sox. Let's sample three sox *without replacement*. Find the expected number of red sox that you get.

4. (§4.1.3) I can make 80% of my basketball free throws. Assuming independent shots, what's the probability that my first missed shot occurs with the fourth toss?

5. (§4.1.3) If the number of orders at a production center this month is a Geom(0.7) random variable, find the probability that we'll have at most three orders.

6. (§4.1.3) Suppose that the questions on a test are iid in the sense that you will be able to answer any question correctly with probability 0.9. What is the expected number of questions that you will complete until you make your second incorrect answer?

7. (§4.1.4) Customers arrive at a bakery according to a Pois(10 / hour) process. The bakery can handle up to 15 customers in an hour without becoming overloaded. What's the probability of an overload in the next hour?

8. (§4.2.1) Suppose that $U_1, U_2, \ldots$ is an iid sample of Unif$(0,1)$ random variables. I'm carrying out an experiment in which I sequentially sample the U_i's until one of them is greater than 0.75, at which point the experiment stops. Let's denote that stopping time by the random variable N. So, for instance, if I observe $U_1 = 0.62$, $U_2 = 0.17$, $U_3 = 0.84$, then I stop at sample $N = 3$. Find the expected stopping time, $\mathrm{E}[N]$.

9. (§4.2.1) Recall that the mgf of the Unif(a,b) distribution is

$$M_X(t) \;=\; \mathrm{E}[e^{tX}] \;=\; \int_a^b e^{tx} \frac{1}{b-a}\, dx \;=\; \frac{e^{tb} - e^{ta}}{t(b-a)}.$$

   Use this fact to find the expected value of X.

10. (§4.2.2) The time to failure of an air conditioner is exponential with a mean of four years.

    (a) What's the probability that the air conditioner will fail before the four-year mark?

    (b) Suppose the air conditioner has already lasted two years. What is the probability that it will fail before the six-year mark?

11. (§4.2.2) Children arrive at a house to do Halloween trick-or-treating according to a Poisson process at the unlucky rate of 13/hour. What is the probability that the time between the 15^{th} and 16^{th} arrivals will be more than 4 minutes? (Hint: Think exponential.)

12. (§4.2.2) Suppose that X and Y are iid exponential random variables with rate $\lambda = 1/3$. Find $\mathrm{P}(1 \le X + Y \le 2)$.

13. (§4.2.2) If $X \sim \mathrm{Gam}(\alpha, \lambda)$, find $M_X(t)$, $\mathrm{E}[X]$, and $\mathrm{Var}(X)$.

14. (§4.2.2) (a) Prove that if $\alpha > 0$, then $\Gamma(\alpha + 1) = (\alpha)\Gamma(\alpha)$. (b) While you're at it, prove that $\Gamma(1/2) = \sqrt{\pi}$.

15. (§4.2.3) If $X \sim \mathrm{Beta}(a,b)$, derive $\mathrm{E}[X]$ and $\mathrm{Var}(X)$.

16. (§4.2.3) If $X \sim \mathrm{Weibull}(a,b)$, derive $\mathrm{E}[X]$ and $\mathrm{Var}(X)$.

17. (§4.2.3) A random variable X is said to have the *Pareto distribution*, with parameters $\lambda > 0$ and $\beta > 0$, if it has cdf

$$F(x) \;=\; 1 - (\lambda/x)^{\beta}, \quad \text{for } x \geq \lambda.$$

The Pareto is a heavy-tailed distribution that has a variety of uses in statistical modeling. Find $\mathrm{E}[X]$ (but be careful when $\beta \leq 1$, because that's when the heavy tails cause problems).

18. (§4.3.1) Suppose that $X \sim \mathrm{Nor}(3, 10)$, $Y \sim \mathrm{Nor}(-4, 3)$, and X and Y are independent. Find the distribution of $W = -2X + Y$.

19. (§4.3.2) Suppose Z is standard normal. Find

    (a) $\mathrm{P}(-2 < Z < 0)$.
    (b) $\mathrm{P}(-1 < Z < 1)$.
    (c) $\mathrm{P}(Z > 1.65)$.
    (d) $\mathrm{P}(Z > -1.96)$.
    (e) $\mathrm{P}(|Z| > 1.2)$.

20. (§4.3.2) Find z such that $\Phi(z) = 0.92$.

21. (§4.3.2) If $\Phi(x)$ is the standard normal cdf, use any method in your arsenal to find x such that $\Phi(x) = 2x$. I'd like you to find an answer that's correct to at least two decimal places.

22. (§4.3.2) If X has a normal distribution with mean 2 and variance 9, find the probability that $X \leq 5$.

23. (§4.3.2) Suppose the SAT math score of an ABC University student can be approximated by a normal distribution with mean 700 and variance 225. Find the probability that the student will score at least a 715.

24. (§4.3.2) If W, X, Y, Z are iid Nor(0,1) RV's, find $\mathrm{P}(W + X + Y + Z \leq 2)$.

25. (§4.3.2) If $Z \sim \mathrm{Nor}(0, 1)$, find $\mathrm{E}[Z^{2k}]$ for $k = 1, 2, 3, 4$.

26. (§4.3.3) Suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Nor}(\mu, 9)$, where μ is unknown. Denote the sample mean by $\bar{X} = \sum_{i=1}^{n} X_i/n$. Find the sample size n such that

$$\mathrm{P}\big(|\bar{X} - \mu| \leq 1\big) \;\geq\; 0.95.$$

In other words, how many observations should you take so that $\bar{X}$ will have a 95% chance of being close to μ?

27. (§4.3.4) What is the most important theorem in the universe?

    (a) Eastern Limit Theorem
    (b) Central Limit Theorem
    (c) Central Limit Serum
    (d) Central Simit Theorem (simit is a tasty Turkish bagel)

28. (§4.3.5) Suppose that $X_1, \ldots, X_{600}$ are iid with values 1, 0 and -1, each with probability 1/3. (This is what is known as a **random walk**.) Find the approximate probability that the sum $\sum_{i=1}^{600} X_i$ will be at most 40.

29. (§4.3.5) 100 big marbles are packed in a box. Each weighs an average of 2 ounces, with a standard deviation of 0.2 ounces. Find the approximate probability that a box weighs less than 202 ounces.

30. (§4.3.5) If $X_1, \ldots, X_{400}$ are iid from some distribution with mean 1 and variance 400, find the approximate probability that the sample mean $\bar{X}$ is greater than 2.

31. (§4.3.5) A production process produces items, of which 6% are defective. A random sample of 200 items is selected every day and the number of defective items X is counted. Using the normal approximation to the binomial, find $\mathrm{P}(X \leq 10)$.

32. (§4.4.1) Denote the pair (height, weight) of a random male from a certain population by (X, Y). Suppose that (X, Y) is bivariate normal with $\mu_X = 70$ inches, $\sigma_X^2 = 100$ in^2, $\mu_Y = 150$ pounds, $\sigma_Y^2 = 225$ lb^2, and $\rho = \mathrm{Corr}(X, Y) = 0.8$.

    Find $\mathrm{P}(Y \geq 165 \,|\, X = 75)$.

33. (§4.4.2) As we discussed in the text, you can use properties of the normal / lognormal distributions to estimate option prices for stocks. I'm not going to have you do that rigorously or via simulation, but I'm going to give you a quick look-up assignment. As I write this on July 10, 2020, IBM is currently selling at \$118.35 per share. Suppose I'm interested in guaranteeing that I can buy a share of IBM for at most \$130 on Oct. 16, 2020. Look up (maybe using something like FaceTube on the internets) various IBM option prices.

34. (§4.4.2) Suppose $Y \sim \mathrm{Nor}(1, 4)$, so that $X = e^Y$ is lognormal. Find $\mathrm{P}(X > e)$.

35. (§4.4.2) Suppose that the random variable Y has a $\mathrm{Nor}(50, 25)$ distribution. Find the mean and variance of $X = e^Y$, and then $\mathrm{P}\big(X \leq \mathrm{E}[X]\big)$.

36. (§4.5) Suppose we have two iid Unif(0,1)'s, $U_1 = 0.6$ and $U_2 = 0.9$.

    (a) Use the Box–Muller method to generate two iid Nor(0,1) random variables, Z_1 and Z_2.

    (b) What is the distribution of $Z_1^2 + Z_2^2$. (Hint: Use the forms of Z_1 and Z_2 from Box–Muller, and see if you come up with anything interesting.)

    (c) Use the results from the two previous parts to generate an Exp(1/2) random variable.

    (d) We're all disappointed that "Box–Muller" doesn't have an umlaut. Can you at least find some consumer products or heavy metal bands that have umlauts?

37. Name That Distribution!

    (a) If $Z \sim \mathrm{Nor}(0, 1)$, what's the distribution of $3Z - 2$?

(b) If $Z \sim \text{Nor}(0,1)$ and $\Phi(\cdot)$ is the standard normal cdf, what's the distribution of the nasty random variable $\Phi(Z)$?

(c) If $U \sim \text{Unif}(0,1)$, what's the distribution of $-10\,\ell\text{n}(\sqrt{U}\,)$?

(d) If U_1, U_2, U_3 are iid Unif(0,1), what's the distribution of

$$-3\ell\text{n}\big(U_1(1-U_2)(1-U_3)\big)?$$

(e) If U and V are iid Unif(0,1), what's the distribution of

$$-2 + \sqrt{-\ell\text{n}(U)}\cos(2\pi V) + \sqrt{-\ell\text{n}(U)}\sin(2\pi V)?$$

(Hint: Think Box–Muller from §4.5.2.)

38. Mathemusical Bonus: Suppose that a, k, and e are all nonzero. Use Beatles lyrics to prove that $m = t$.

39. Half of a famous Normal Inequality — Add this problem somewhere

Let Z be a standard normal r.v., and let

$$\zeta(x) \;\equiv\; \text{P}(Z > x) - \frac{x}{x^2+1}\phi(x) \quad \text{for } x \geq 0.$$

Note that $\zeta(0) = 1/2$, $\lim_{x\to\infty}\zeta(x) = 0$, and (after a little algebra),

$$\frac{d}{dx}\zeta(x) \;=\; \frac{-2\phi(x)}{(x^2+1)^2} \;<\; 0 \quad \text{for } x \geq 0.$$

This implies that $\zeta(x) \geq 0$ for all $x \geq 0$; and so

$$\Phi(-x) \;=\; \text{P}(Z > x) \;\geq\; \frac{x\phi(x)}{x^2+1} \quad \text{for } x \geq 0,$$

which can be rewritten as

$$\frac{1}{\Phi(-x)} \;=\; \frac{1}{\text{P}(Z > x)} \;\leq\; \frac{x^2+1}{x\phi(x)} \quad \text{for } x \geq 0. \tag{4.5}$$

Chapter 5

Descriptive Statistics

Now on to the second major portion of the text — Statistics! Up to this point, we have worked with several probability distributions that are appropriate for modeling real-world phenomena, such as the number of customer arrivals to a call center, electronic component lifetimes, weights of humans, etc. But how do we know that the specific models we use are actually reasonable? And how can we make appropriate decisions given these models?

Statistics uses observational sample data to draw general conclusions about the population from which a sample was taken. We'll begin this chapter in §5.1 by going over some basic *descriptive statistics* (histograms, performance measures such as sample means, etc.), and then reviewing situations that are appropriate for the use of certain distributions. For instance, when might you want to use the Poisson distribution? We'll then discuss in §5.2 a variety of techniques to *estimate* any unknown parameters associated with these distributions. For example, if you think that a part repair time might be exponential, then you still have to estimate the rate λ. The classes of techniques we'll learn about include unbiased, maximum likelihood, and method of moments parameter estimation.

Finally, we'll make a brief aside in §5.3 to learn about several *sampling distributions* that are especially important in statistics. In particular, we'll make the acquaintance of some new friends — the χ^2, Student t, and F distributions. This material will prepare us for Chapter 6, where we'll derive a variety of *confidence intervals* that can be used to provide certain lower and upper bounds on the values of any unknown distributional parameters; and then Chapter 7, where we'll describe various *hypothesis tests* to determine if our final choice of distribution (including its parameters) adequately fits the data set that we observe.

5.1 Introduction to Statistics

We'll ease into the subject with a high-level discussion on data, along with some simple data analysis techniques that can be used as a first pass.

5.1.1 What is Statistics?

Statistics forms a rational basis for decision-making using observed or experimental **data**. We make these decisions in the face of uncertainty, and statistics helps us answer questions concerning:

- The analysis of one population (or system).
- The comparison of two or more populations. Which is the best system?

Examples: Statistics is everywhere!

- Election polling.
- Coke vs. Pepsi.
- The effect of cigarette smoking on the probability of getting cancer.
- The effect of a new vaccine on the probability of contracting hepatitis.
- What's the most popular TV show during a certain time period?
- The effect of various heat-treating methods on steel tensile strength.
- Which fertilizers improve crop yield?

Idea (Election polling example): We can't poll every single voter. Thus, we take a (small, efficient, representative) **sample** of data from the (large) **population** of voters, and we try to make reasonable conclusions based on the sample.

Game Plan: Statistics tells us how to conduct the sampling (i.e., how many observations to take, how to take them, etc.), and then how to draw conclusions from the sampled data. In general, a statistical study is carried out in roughly the following way.

1. Collect and summarize data for analysis (discussed in this section).
2. Determine / estimate the underlying distribution (along with associated parameters), e.g., Nor(30,8) (discussed in §5.2 and Chapter 6).
3. Conduct a statistical test to see if your distribution or some associated hypothesis is "approximately" correct (Chapter 7).
4. Loop among these steps as necessary, as we become curious about additional questions and/or encounter the need for more data.

Types of Data:

The way we sample and analyze data sometimes depends on the type of data that is available.

- **Discrete variables:** Such data can only take on specific values, e.g., the number of accidents at a factory this week, or the possible rolls of a pair of dice.
- **Continuous variables:** Data that can take on any real value in a certain interval. For example, the lifetime of a light bulb, or the weight of a newborn child.
- **Categorical variables:** This type of data isn't numerical. For instance, what's your favorite (broadcast) TV show during a certain time slot?

5.1.2 Descriptive Statistics

In this subsection, we discuss how to go about summarizing data.

A picture is worth 1000 words. Thus, one should always plot out data before doing anything else, if only to identify obvious issues such as nonstandard distributions, missing data points, outliers, etc.

Histograms provide a quick, succinct look at what you are dealing with. It can be shown (by what is known as the Glivenko–Cantelli Theorem), that if you take enough observations, the histogram will eventually converge to the true distribution, thus showing that histograms are reliable at portraying data. One thorny issue is determining the proper number of cells in which to partition the data (Figure 5.1). Luckily, many software packages (e.g., Excel, R, Matlab) can be coaxed into doing this chore automatically.

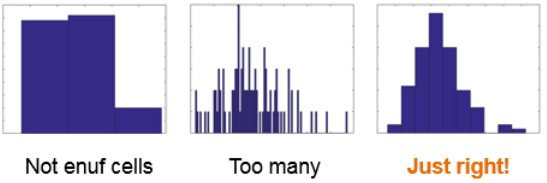


Figure 5.1: Different histograms depicting the same data.

Example: Grades on a test (i.e., raw data):

23	62	91	83	82	64	73	94	94	52
67	11	87	99	37	62	40	33	80	83
99	90	18	73	68	75	75	90	36	55

Stem-and-Leaf Diagram of grades. This old friend is an easy way to write down all of the data. It saves some space, and looks like a sideways histogram.

9	9944100
8	73320
7	5533
6	87422
5	52
4	0
3	763
2	3
1	81

Grouped Data. Sometimes it's informative to group the data into interval buckets, along with information about the frequencies and cumulative frequencies of items in the buckets.

Range	Frequency	Cumulative Frequency	Proportion of observations so far
0–20	2	2	2/30
21–40	5	7	7/30
41–60	2	9	9/30
61–80	10	19	19/30
81–100	11	30	1

It's nice to have lots of data. But sometimes it's too much of a good thing! So we often need to succinctly summarize data sets numerically.

Summary Statistics paint an extremely succinct picture of the data by giving us information about data characteristics such as the sample size, sample mean, sample median, and sample standard deviation.

In the running data set, we have...

- $n = 30$ observations.
- If X_i is the i^{th} score, then the **sample mean** is

$$\bar{X} \equiv \sum_{i=1}^{n} X_i/n = 66.5.$$

  The sample mean is a measure of *central tendency*, and it is an estimator of the true but typically unknown mean $\mathrm{E}[X_i]$.

- The **sample median** is another measure of central tendency, and is simply the "middle" observation when the X_i's are arranged numerically.

  Examples: The data set 16, 7, 83 gives a sample median of 16. When the sample size is even, a "reasonable" approach is to just use the average of the two middle values, e.g., the set 16, 7, 83, 20 gives a "reasonable" sample median of $\frac{16+20}{2} = 18$. The sample median for our running data set is 73.

Remark: The sample median is less susceptible to "outlier" data than the sample mean. One bad number can spoil the sample mean's entire day.

Example: 7, 7, 7, 672, 7 has a sample mean of 140 and a sample median of 7.

- The **sample variance** (an estimator of the true but unknown variance $\mathrm{Var}(X_i)$) is
$$S^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = 630.6.$$
The variance is a measure of spread or dispersion.
- Other such measures of spread are the **sample standard deviation**, $S = +\sqrt{S^2}$, and the **sample range**, $\max_i X_i - \min_i X_i$.

Remark: Before you take any observations, $\bar{X}$ and S^2 must be regarded as *random variables* and are said to be **estimators** of the unknown mean and variance, respectively. Once the data is taken, then $\bar{X}$ and S^2 become *constants* and are then **estimates** of the unknown mean and variance.

Remark: Suppose the data takes p different values $X_1, \ldots, X_p$, with frequencies $f_1, \ldots, f_p$, respectively. To calculate $\bar{X}$ and S^2 quickly and efficiently, simply take
$$\bar{X} = \frac{1}{n}\sum_{j=1}^{p} f_j X_j \quad \text{and} \quad S^2 = \frac{\sum_{j=1}^{p} f_j X_j^2 - n\bar{X}^2}{n-1}.$$

Example: Suppose we roll a die 10 times.

X_j	1	2	3	4	5	6
f_j	2	1	1	3	0	3

Then $\bar{X} = (2 \cdot 1 + 1 \cdot 2 + \cdots + 3 \cdot 6)/10 = 3.7$, and
$$S^2 = \frac{(2 \cdot 1^2 + 1 \cdot 2^2 + \cdots + 3 \cdot 6^2) - 10(3.7)^2}{9} = 3.789. \quad \square$$

Remark: If the individual observations can't be precisely categorized in frequency distributions, you might just break up the observations[1] into c intervals and simply approximate $\bar{X}$ and S^2.

Example: Consider the following illustration with $c = 3$, where we denote the midpoint of the j^{th} interval by m_j, $j = 1, 2, 3$, and the total sample size $n = \sum_{j=1}^{c} f_j = 30$.

X_j interval	m_j	f_j
100–150	125	10
150–200	175	15
200–300	250	5

[1] Breaking up is hard to do.

Then we have the approximations

$$\begin{aligned}
\bar{X} &\doteq \frac{\sum_{j=1}^{c} f_j m_j}{n} = \frac{10(125) + 15(175) + 5(250)}{30} = 170.833, \\
S^2 &\doteq \frac{\sum_{j=1}^{c} f_j m_j^2 - n\bar{X}^2}{n-1} \\
&= \frac{\left[10(125)^2 + 15(175)^2 + 5(250)^2\right] - 30(170.833)^2}{29} \\
&= 1814. \quad \square
\end{aligned}$$

Note that, in general, it's a good idea to be a bit careful to maintain proper numerical precision in the above calculations for S^2 — mainly because we might be subtracting big numbers from other big numbers, in which case it would be easy to lose significant digits.

5.1.3 Candidate Distributions

Now that we have used descriptive statistics to obtain a rough idea about the nature of the data, perhaps it's time to make an informed guess about the type of probability distribution we're dealing with. We'll look at more-formal methodology for fitting distributions in Chapter 7.

We may get lucky in the sense that a "standard" distribution (such as the geometric, exponential or normal) results in a perfectly fine, intuitive data fit; or we may have to punt and use a nonstandard solution. But how do we even start to decide? Let's begin by thinking about the following questions.

- Is the data from a discrete, continuous, or mixed distribution?
- Univariate/multivariate?
- How much data is available?
- Are experts around to ask about nature of the data?
- What if we do not have much/any data — can we at least guess at a good distribution?

If the distribution is a discrete random variable, then we have a number of familiar choices to select from.

- Bernoulli(p) (success with probability p)
- Binomial(n, p) (number of successes in n Bern(p) trials)
- Geometric(p) (number of Bern(p) trials until first success)
- Negative Binomial (number of Bern(p) trials until multiple successes)
- Poisson(λ) (counts the number of arrivals over time)
- Empirical (the all-purpose "sample" distribution based on the histogram)

And if the data suggests a continuous distribution...

- Uniform (not much is known from the data, except perhaps the minimum and maximum possible values)
- Triangular (at least we have an idea regarding the minimum, maximum, and "most likely" values)
- Exponential(λ) (e.g., interarrival times from a Poisson process)
- Normal (a good model for heights, weights, IQ's, sample means, etc.)
- Beta (good for specifying bounded data)
- Gamma, Weibull, Gumbel, lognormal (reliability data)
- Empirical (our all-purpose friend)

Our game plan for now is to: Choose a "reasonable" distribution and estimate the relevant parameters in §5.2; do that aside we promised on sampling distributions in §5.3; undertake additional confidence-interval analysis in Chapter 6; and then eventually put everything together by formally proposing and evaluating our distributional choice in Chapter 7.

5.2 Point Estimation

Suppose that, based on histograms from the collected data as well as past experience, we believe that certain interarrival times follow an Exp(λ). Before we can conduct formal hypothesis tests in Chapter 7, we need to estimate the exponential distribution's unknown parameter value λ. In this section, we'll introduce the concept of estimation, and then present a variety of estimation techniques.

5.2.1 Introduction to Estimation

Definition: A **statistic** is a function of the observations $X_1, \ldots, X_n$, and not explicitly dependent on any unknown **parameters**.

Examples (of statistics): The *sample mean* and *sample variance*, commonly denoted as $\bar{X} \equiv \sum_{i=1}^{n} X_i/n$ and $S^2 \equiv \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$, respectively.

Examples (of parameters): The true but unknown expected value of the population, μ, and the population variance, σ^2. Thus, $(\bar{X} - \mu)/\sigma$ is not a statistic, because that quantity contains unknown parameters.

Remark: Statistics are *random variables*. If we take two different samples, we'd expect to get two different values of a statistic. But note that after the X_i's are observed (at which point they are no longer random), you can actually calculate the statistics — there are no unknown parameters involved.

A statistic is usually used to estimate some unknown parameter from the underlying probability distribution of the X_i's.

Let $X_1, \ldots, X_n$ be iid RV's and let the function $T(\mathbf{X}) \equiv T(X_1, \ldots, X_n)$ be a statistic based on the X_i's. Suppose we use $T(\mathbf{X})$ to estimate some unknown parameter θ. Then $T(\mathbf{X})$ is called a **point estimator** for θ.

Examples: The sample mean $\bar{X}$ is usually a point estimator for the true mean $\mu = \mathrm{E}[X_i]$; and the sample variance S^2 is often a point estimator for the true variance $\sigma^2 = \mathrm{Var}(X_i)$.

It would be nice if $T(\mathbf{X})$ had certain desirable properties (which we shall discuss in detail in a minute).

- Its expected value should equal the parameter it is trying to estimate, i.e., $\mathrm{E}[T(\mathbf{X})] = \theta$, or at least it should be close.
- It should have low variance.

5.2.2 Unbiased Estimation

A good property for an estimator to have is that it is correct "on average."

Definition: $T(\mathbf{X})$ is **unbiased** for θ if $\mathrm{E}[T(\mathbf{X})] = \theta$.

First, we look at the unbiasedness of $\bar{X}$ as an estimator of μ.

Theorem: Suppose $X_1, \ldots, X_n$ are iid anything with mean μ. Then

$$\mathrm{E}[\bar{X}] \;=\; \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] \;=\; \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[X_i] \;=\; \mathrm{E}[X_1] \;=\; \mu.$$

So $\bar{X}$ is always unbiased for μ. That's why $\bar{X}$ is called the **sample mean**. □

Baby Example: In particular, suppose $X_1, \ldots, X_n$ are iid Exp(λ). Then $\bar{X}$ is unbiased for $\mu = \mathrm{E}[X_i] = 1/\lambda$.

Remark: But be careful! It turns out (see Exercise 5.9) that $1/\bar{X}$ is *biased* for λ in this exponential case, i.e., $\mathrm{E}[1/\bar{X}] \neq 1/\mathrm{E}[\bar{X}] = \lambda$. □

Now we establish the unbiasedness of S^2 as an estimator of σ^2.

Theorem: If $X_1, \ldots, X_n$ are iid anything with mean μ and variance σ^2, we have

$$\mathrm{E}[S^2] \;=\; \frac{1}{n-1}\mathrm{E}\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] \;=\; \mathrm{Var}(X_i) \;=\; \sigma^2.$$

Thus, in the iid case, S^2 is always unbiased for σ^2. This is why S^2 is called the **sample variance** (and it's why we divide by $(n-1)$ instead of by n). □

Proof: First, some standard algebra gives

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\bar{X})^2 &= \sum_{i=1}^{n}(X_i^2-2\bar{X}X_i+\bar{X}^2)\\ &= \sum_{i=1}^{n}X_i^2-2\bar{X}\sum_{i=1}^{n}X_i+n\bar{X}^2\\ &= \sum_{i=1}^{n}X_i^2-2n\bar{X}^2+n\bar{X}^2\\ &= \sum_{i=1}^{n}X_i^2-n\bar{X}^2. \qquad (5.1)\end{aligned}$$

Then

$$\begin{aligned}\mathrm{E}[S^2] &= \frac{1}{n-1}\mathrm{E}\Big[\sum_{i=1}^{n}X_i^2-n\bar{X}^2\Big] \quad \text{(by definition of } S^2 \text{ and Equation (5.1))}\\ &= \frac{1}{n-1}\Big(\sum_{i=1}^{n}\mathrm{E}[X_i^2]-n\mathrm{E}[\bar{X}^2]\Big)\\ &= \frac{n}{n-1}\big(\mathrm{E}[X_1^2]-\mathrm{E}[\bar{X}^2]\big) \quad \text{(since the } X_i\text{'s are iid)}\\ &= \frac{n}{n-1}\big(\mathrm{Var}(X_1)+(\mathrm{E}[X_1])^2-\mathrm{Var}(\bar{X})-(\mathrm{E}[\bar{X}])^2\big)\\ &= \frac{n}{n-1}(\sigma^2-\sigma^2/n) \quad \text{(since } \mathrm{E}[X_1]=\mathrm{E}[\bar{X}] \text{ and } \mathrm{Var}(\bar{X})=\sigma^2/n)\\ &= \sigma^2. \quad \text{Done.} \quad \square\end{aligned}$$

Baby Example: Suppose $X_1,\ldots,X_n$ are iid $\mathrm{Exp}(\lambda)$. Then S^2 is unbiased for $\mathrm{Var}(X_i)=1/\lambda^2$. $\square$

Remark: Careful! It can be shown (see Exercise 5.10) that the sample standard deviation S is *not* unbiased for the standard deviation σ.

The individual unbiased estimators $\bar{X}$ and S^2 for the respective unknown parameters μ and σ^2 make good intuitive sense. Nevertheless, it is sometimes possible to have *multiple* reasonable unbiased estimators for an unknown parameter, in which case we will have to take certain measures to determine which estimator is somehow "the best."

Big Example: Suppose that $X_1,X_2,\ldots,X_n \overset{\text{iid}}{\sim} \mathrm{Unif}(0,\theta)$, so that the pdf is $f(x)=1/\theta$, $0<x<\theta$. Here's the motivation: I give you a bunch of random numbers between 0 and θ, and you have to guess what θ is.

We'll look at *three* unbiased estimators for θ — the good, the better, and the ugly (we'll explain our choice of names in a little while):

$$Y_1 = 2\bar{X}, \quad Y_2 = \frac{n+1}{n}\max_{1\le i\le n}X_i, \quad \text{and} \quad Y_3 = \begin{cases}12\bar{X} & \text{w.p. } 1/2\\ -8\bar{X} & \text{w.p. } 1/2.\end{cases}$$

We'll start out by proving that all three estimators are indeed unbiased. First consider the "good" estimator, $Y_1 = 2\bar{X}$.

Proof (that Y_1 is unbiased): $\mathrm{E}[Y_1] = 2\mathrm{E}[\bar{X}] = 2\mathrm{E}[X_i] = 2(\theta/2) = \theta$. □

Second, let's study the "better" estimator, $Y_2 = \frac{n+1}{n}M$, where $M \equiv \max_{1 \le i \le n} X_i$. Why might this estimator for θ make intuitive sense? Answer: Because it takes the largest observation so far, and gives it a little boost (the factor $(n+1)/n$) in an attempt to nudge it closer to θ.

Proof (that Y_2 is unbiased): Instead of proving that $\mathrm{E}[Y_2] = \frac{n+1}{n}\mathrm{E}[M] = \theta$, we'll equivalently prove that $\mathrm{E}[M] = n\theta/(n+1)$. To do so, let's begin by obtaining the cdf of M:

$$\begin{aligned}
\mathrm{P}(M \le y) &= \mathrm{P}(X_1 \le y \text{ and } X_2 \le y \text{ and } \cdots \text{ and } X_n \le y) \\
&= \mathrm{P}(X_1 \le y)\mathrm{P}(X_2 \le y) \cdots \mathrm{P}(X_n \le y) \quad (X_i\text{'s independent}) \\
&= [\mathrm{P}(X_1 \le y)]^n \quad (X_i\text{'s identically distributed}) \\
&= \left[\int_{-\infty}^{y} f_{X_1}(x)\, dx \right]^n \\
&= \left[\int_{0}^{y} (1/\theta)\, dx \right]^n \\
&= (y/\theta)^n.
\end{aligned}$$

This implies that the pdf of M is

$$f_M(y) \equiv \frac{d}{dy}(y/\theta)^n = \frac{n y^{n-1}}{\theta^n}, \quad 0 < y < \theta,$$

so that

$$\mathrm{E}[M] = \int_{\mathbb{R}} y f_M(y)\, dy = \int_0^\theta \frac{n y^n}{\theta^n}\, dy = \frac{n\theta}{n+1}.$$

Whew! This finally shows that $Y_2 = \frac{n+1}{n}M$ is an unbiased estimator for θ! □

Lastly, let's look at the "ugly" estimator for θ,

$$Y_3 = \begin{cases} 12\bar{X} & \text{w.p. } 1/2 \\ -8\bar{X} & \text{w.p. } 1/2. \end{cases}$$

Note that it's possible to get a *negative* estimate for θ, which is really strange since $\theta > 0$! Nevertheless...

Proof (that Y_3 is unbiased):

$$\mathrm{E}[Y_3] = 12\mathrm{E}[\bar{X}] \cdot \frac{1}{2} - 8\mathrm{E}[\bar{X}] \cdot \frac{1}{2} = 2\mathrm{E}[\bar{X}] = \theta. \quad \square$$

If multiple estimators are all unbiased for θ, which one's the best? Of course, it's *good* for an estimator to be unbiased, but the "ugly" estimator Y_3 shows that unbiased estimators can sometimes be goofy. We can break the unbiasedness tie by looking at other properties that an estimator can have. For instance, consider the *variance* of the estimators.

Big Example (cont'd): Again suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$. Recall that the "good" $Y_1 = 2\bar{X}$, the "better" $Y_2 = \frac{n+1}{n} M$, and the "ugly" Y_3 are all unbiased for θ.

Let's find their variances. First of all,

$$\text{Var}(Y_1) \;=\; 4\text{Var}(\bar{X}) \;=\; \frac{4}{n} \cdot \text{Var}(X_i) \;=\; \frac{4}{n} \cdot \frac{\theta^2}{12} \;=\; \frac{\theta^2}{3n},$$

which isn't too shabby, since the variance goes to 0 as $n \to \infty$. ☺

Meanwhile,

$$\begin{aligned}
\text{Var}(Y_2) &= \text{E}[Y_2^2] - \big(\text{E}[Y_2]\big)^2 \\
&= \left(\frac{n+1}{n}\right)^2 \text{E}[M^2] - \theta^2 \quad \text{(since } Y_2 \text{ is unbiased)} \\
&= \left(\frac{n+1}{n}\right)^2 \int_0^\theta \frac{n y^{n+1}}{\theta^n}\, dy - \theta^2 \\
&= \theta^2 \cdot \frac{(n+1)^2}{n(n+2)} - \theta^2 \;=\; \frac{\theta^2}{n(n+2)},
\end{aligned}$$

which goes to 0 *very* quickly — better than Y_1's rate! 😀

Finally,

$$\begin{aligned}
\text{Var}(Y_3) &= \text{E}[Y_3^2] - \big(\text{E}[Y_3]\big)^2 \\
&= \frac{1}{2}\big(144\,\text{E}[\bar{X}^2] + 64\,\text{E}[\bar{X}^2]\big) - \theta^2 \quad \text{(since } Y_3 \text{ is unbiased)} \\
&= 104\,\text{E}[\bar{X}^2] - \theta^2 \\
&= 104\left[\text{Var}(\bar{X}) + \big(\text{E}[\bar{X}]\big)^2\right] - \theta^2 \\
&= 104\left[\frac{\theta^2}{12n} + \frac{\theta^2}{4}\right] - \theta^2 \quad \text{(by previous work)} \\
&= \left(\frac{26}{3n} + 25\right)\theta^2,
\end{aligned}$$

and this mess doesn't even go to 0 as $n \to \infty$! ☹

Thus, Y_1, Y_2, and Y_3 are all unbiased, but Y_2 has *much lower variance* than Y_1, while Y_3 has crazy high variance. Now you can see why they're called the "good," the "better," and the "ugly." In any case, we can break the "unbiasedness tie" by choosing Y_2, which has the lowest variance. □

5.2.3 Mean Squared Error

We saw in §5.2.2 that unbiasedness is a good thing, but that it doesn't tell the entire story; variance must also enter into the conversation. *Mean squared error*

(MSE) takes both bias and variance into consideration when evaluating estimator performance.

Definition: The **mean squared error** of an estimator $T(\mathbf{X})$ of θ is

$$\text{MSE}(T(\mathbf{X})) \equiv \text{E}\big[(T(\mathbf{X}) - \theta)^2\big].$$

Before giving an easier interpretation of MSE, we define the **bias** of an estimator for the parameter θ,

$$\text{Bias}(T(\mathbf{X})) \equiv \text{E}[T(\mathbf{X})] - \theta.$$

Theorem: An easier interpretation: $\text{MSE} = \text{Bias}^2 + \text{Var}$.

Proof: By definition, we have

$$\begin{aligned}
\text{MSE}(T(\mathbf{X})) &= \text{E}\big[(T(\mathbf{X}) - \theta)^2\big] \\
&= \text{E}[T^2] - 2\theta\text{E}[T] + \theta^2 \\
&= \text{E}[T^2] - \big(\text{E}[T]\big)^2 + \big(\text{E}[T]\big)^2 - 2\theta\text{E}[T] + \theta^2 \\
&= \text{Var}(T) + \big(\underbrace{\text{E}[T] - \theta}_{\text{Bias}}\big)^2. \quad \square
\end{aligned}$$

So the MSE combines the bias and variance of an estimator. The lower the MSE the better. If $T_1(\mathbf{X})$ and $T_2(\mathbf{X})$ are two estimators of θ, we'd usually prefer the one with the lower MSE — even if it happens to have higher bias.

Definition: The **relative efficiency** of $T_2(\mathbf{X})$ to $T_1(\mathbf{X})$ is the ratio $\text{MSE}(T_1)/\text{MSE}(T_2)$. If this quantity is < 1, then we would prefer T_1.

Example: Suppose that estimator A has bias $= 3$ and variance $= 10$, while estimator B has bias $= -2$ and variance $= 14$. Which estimator (A or B) has the lower mean squared error?

Solution: $\text{MSE} = \text{Bias}^2 + \text{Var}$, so

$$\text{MSE(A)} = 9 + 10 = 19 \qquad \text{and} \qquad \text{MSE(B)} = 4 + 14 = 18.$$

Thus, B has lower MSE. $\square$

Example: $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$. Recall that we examined two pretty decent *unbiased* estimators for θ, namely, $Y_1 = 2\bar{X}$, and $Y_2 = \frac{n+1}{n} \max_i X_i$. We found that $\text{Var}(Y_1) = \theta^2/(3n)$, and $\text{Var}(Y_2) = \theta^2/(n^2 + 2n)$, so that

$$\frac{\text{MSE}(Y_1)}{\text{MSE}(Y_2)} = \frac{\theta^2/(3n)}{\theta^2/(n^2 + 2n)} = \frac{n+2}{3} > 1, \quad \text{for } n > 1.$$

This indicates that Y_2 is the better estimator. $\square$

5.2.4 Fisher Information

This section discusses Fisher information and some of its consequences.

- The definition of Fisher information, followed by examples (right now).
- The Cramér–Rao Lower Bound for the variance of an unbiased estimator (right after).
- Asymptotic normality of maximum likelihood estimators (later on in §5.2.5.4).

Definition: Consider an iid sample $X_1, X_2, \ldots, X_n$ from pmf/pdf $f(x)$ having an underlying unknown parameter θ. The **Fisher information** of these n observations is

$$I_n(\theta) \;\equiv\; n\,\mathrm{E}\left\{\left[\frac{\partial}{\partial\theta}\,\ell\mathrm{n}\big(f(X)\big)\right]^2\right\}. \tag{5.2}$$

Remarks:

- Don't panic — it looks awful, but it's actually easier than LOTUS! ☺
- Fisher information is related to the sensitivity of changes in the pmf/pdf $f(x)$ to changes in θ — informally, the larger $I_n(\theta)$ is (i.e., the more Fisher information you have), the easier it is to distinguish changes in $f(x)$.
- Fisher information is related to the variance of unbiased and maximum likelihood estimators.
- *Many of the results on Fisher information will be stated without proof and are subject to several assumptions that we won't talk about*, except to say that, for instance,
  - The domain of X can't depend on the parameter θ, e.g., as one would encounter with a Unif$(0,\theta)$ random variable;
  - All of the expected values have to exist; and
  - You should be able to legally swap certain expected values and partial derivatives, e.g., $\frac{\partial}{\partial\theta}\mathrm{E}\big[\cdots\big] = \mathrm{E}\big[\frac{\partial}{\partial\theta}\cdots\big]$.
- Here is an alternative (sometimes easier) way to calculate the Fisher information.

$$I_n(\theta) \;=\; -n\,\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\,\ell\mathrm{n}\big(f(X)\big)\right]. \tag{5.3}$$

Example: Suppose that $X \sim \mathrm{Bern}(p)$, and recall from long, long ago that the pmf is $f(x) = p^x(1-p)^{1-x}$ for $x = 0, 1$, and $\mathrm{E}[X^k] = p$ for all k. Then

$$\ell\mathrm{n}(f(x)) \;=\; x\,\ell\mathrm{n}(p) + (1-x)\,\ell\mathrm{n}(1-p) \quad \text{and}$$

$$\frac{\partial}{\partial p}\,\ell\mathrm{n}(f(x)) \;=\; \frac{x}{p} - \frac{1-x}{1-p} \;=\; \frac{x-p}{p(1-p)}.$$

So, by Equation (5.2), the Fisher information is

$$\begin{aligned} I_n(p) &= n\,\mathrm{E}\left\{\left[\frac{\partial}{\partial p}\,\ell\mathrm{n}(f(X))\right]^2\right\} = n\,\mathrm{E}\left[\left(\frac{X-p}{p(1-p)}\right)^2\right] \\ &= \frac{n}{p^2(1-p)^2}\mathrm{E}\big[X^2 - 2pX + p^2\big] \\ &= \frac{n(p-2p^2+p^2)}{p^2(1-p)^2} = \frac{n}{p(1-p)}. \quad \square \end{aligned} \tag{5.4}$$

We can also use the alternative Equation (5.3) to calculate $I_n(p)$. First, note that

$$\frac{\partial^2}{\partial p^2}\,\ell\mathrm{n}(f(x)) = -\frac{x}{p^2} - \frac{1-x}{(1-p)^2},$$

so

$$\begin{aligned} I_n(p) &= -n\,\mathrm{E}\left[\frac{\partial^2}{\partial p^2}\,\ell\mathrm{n}\big(f(X)\big)\right] \\ &= n\,\mathrm{E}\left[\frac{X}{p^2} + \frac{1-X}{(1-p)^2}\right] \\ &= \frac{np}{p^2} + \frac{n(1-p)}{(1-p)^2} \\ &= \frac{n}{p(1-p)}, \end{aligned}$$

which matches the answer from the definition of $I_n(p)$! $\square$

Example: For $X \sim \mathrm{Exp}(\lambda)$, recall that $f(x) = \lambda e^{-\lambda x}$, $x \geq 0$, and $\mathrm{E}[X^k] = \lambda^k/k!$. Then

$$\ell\mathrm{n}(f(x)) = \ell\mathrm{n}(\lambda) - \lambda x, \quad \frac{\partial}{\partial \lambda}\,\ell\mathrm{n}(f(x)) = \frac{1}{\lambda} - x, \quad \text{and} \quad \frac{\partial^2}{\partial \lambda^2}\,\ell\mathrm{n}(f(x)) = -\frac{1}{\lambda^2},$$

so that

$$\begin{aligned} I_n(\lambda) &= n\,\mathrm{E}\left\{\left[\frac{\partial}{\partial \lambda}\,\ell\mathrm{n}(f(X))\right]^2\right\} = n\,\mathrm{E}\left[\left(\frac{1}{\lambda} - X\right)^2\right] \\ &= n\,\mathrm{E}\left[\frac{1}{\lambda^2} - \frac{2X}{\lambda} + X^2\right] = \frac{n}{\lambda^2} - \frac{2n}{\lambda^2} + \frac{2n}{\lambda^2} = \frac{n}{\lambda^2}. \quad \square \end{aligned}$$

Alternatively, and with less work,

$$I_n(\lambda) = -n\,\mathrm{E}\left[\frac{\partial^2}{\partial \lambda^2}\,\ell\mathrm{n}\big(f(X)\big)\right] = -n\,\mathrm{E}\left[-\frac{1}{\lambda^2}\right] = \frac{n}{\lambda^2}. \quad \square$$

Fisher information turns up in all sorts of surprising places. For instance, we can provide a lower bound on the variance of an unbiased estimator.

Cramér–Rao Lower Bound (CRLB): If $T(X_1, X_2, \ldots, X_n)$ is an *unbiased* estimator for a parameter θ, then $\mathrm{Var}(T) \geq 1/I_n(\theta)$.

The CRLB is the best an unbiased estimator can do (assuming that all of the mysterious assumptions that we haven't talked about hold).

Example: If $X_1, X_2, \ldots, X_n$ are iid Bern(p), recall that the sample mean $\bar{X}$ is unbiased for p. Moreover, by Equation (5.4), the Fisher information is $I_n(p) = n/[p(1-p)]$. Fortuitously, $\bar{X}$ achieves the CRLB exactly, since

$$\mathrm{Var}(\bar{X}) \;=\; \mathrm{Var}(X_i)/n \;=\; p(1-p)/n \;=\; 1/I_n(p). \quad \square$$

Example: If $X_1, X_2, \ldots, X_n$ are iid anything, recall that the sample variance S^2 is unbiased for $\sigma^2 = \mathrm{Var}(X_i)$. Let's see how $\mathrm{Var}(S^2)$ compares to the CRLB in the special case that the X_i's are Nor(μ, σ^2). By an upcoming result from Equation (6.1) in the future, it turns out that $S^2 \sim \sigma^2 \chi^2(n-1)/(n-1)$; and then Exercise 5.29 shows that $\mathrm{Var}(S^2) = 2\sigma^4/(n-1)$.

Meanwhile, we'll obtain the Fisher information for σ^2 by starting with the normal pdf

$$f(x) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\},$$

in which case

$$\ell\mathrm{n}(f(x)) \;=\; -\frac{1}{2}\ell\mathrm{n}(2\pi) - \frac{1}{2}\ell\mathrm{n}(\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2},$$

$$\frac{\partial}{\partial\sigma^2}\, \ell\mathrm{n}(f(x)) \;=\; -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2(\sigma^2)^2}, \quad \text{and}$$

$$\frac{\partial^2}{\partial(\sigma^2)^2}\, \ell\mathrm{n}(f(x)) \;=\; \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6}.$$

This implies that

$$\begin{aligned} I_n(\sigma^2) \;&=\; -n\,\mathrm{E}\left[\frac{\partial^2}{\partial(\sigma^2)^2}\, \ell\mathrm{n}\big(f(X)\big) \right] \;=\; \mathrm{E}\left[\frac{1}{2\sigma^4} - \frac{(X-\mu)^2}{\sigma^6} \right] \\ &=\; -n\left[\frac{1}{2\sigma^4} - \frac{\mathrm{Var}(X)}{\sigma^6} \right] \;=\; -n\left[\frac{1}{2\sigma^4} - \frac{\sigma^2}{\sigma^6} \right] \;=\; \frac{n}{2\sigma^4}, \end{aligned}$$

where the middle step follows by the definition of variance. Thus, after all of the algebraic fireworks,

$$\mathrm{Var}(S^2) \;=\; \frac{2\sigma^4}{n-1} \;>\; \frac{2\sigma^4}{n} \;=\; 1/I_n(\sigma^2) \;=\; \mathrm{CRLB},$$

though it was pretty close! $\square$

5.2.5 Maximum Likelihood Estimation

Maximum likelihood estimators (MLEs) serve as an alternative approach to unbiased estimation (though MLEs are sometimes unbiased themselves). MLEs have numerous desirable properties and are important components of the goodness-of-fit tests that will be discussed in Chapter 7.

§5.2.5.1 Introduction to MLEs and Baby Examples

Definition: Consider an iid random sample $X_1, X_2, \ldots, X_n$, where each X_i has pmf/pdf $f(x)$. Further, suppose that θ is some unknown parameter from X_i. The **likelihood function** is $L(\theta) \equiv \prod_{i=1}^{n} f(x_i)$.

Definition: The **maximum likelihood estimator** of θ is the value of θ that maximizes $L(\theta)$. The MLE is a function of the X_i's and is a random variable.

Remark: We can *very informally* regard the MLE as the "most likely" estimate of θ.

Example: Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda)$. Let's find the MLE for λ. We start with the likelihood function,

$$L(\lambda) \;=\; \prod_{i=1}^{n} f(x_i) \;=\; \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \;=\; \lambda^n \exp\Big(-\lambda \sum_{i=1}^{n} x_i\Big).$$

Now the task is to maximize $L(\lambda)$ with respect to λ. We could take the derivative and plow through all of the horrible algebra in the usual way, but this is typically too tedious. A useful trick that often applies is simply to perform the maximization on the natural logarithm of the likelihood function. Since the natural log function is one-to-one, it is easy to see that the λ that maximizes $L(\lambda)$ also maximizes $\ell\text{n}(L(\lambda))$. For this purpose, we have

$$\ell\text{n}(L(\lambda)) \;=\; \ell\text{n}\bigg(\lambda^n \exp\Big(-\lambda \sum_{i=1}^{n} x_i\Big)\bigg) \;=\; n\ell\text{n}(\lambda) - \lambda \sum_{i=1}^{n} x_i,$$

so that

$$\frac{d}{d\lambda}\ell\text{n}(L(\lambda)) \;=\; \frac{d}{d\lambda}\Big(n\ell\text{n}(\lambda) - \lambda \sum_{i=1}^{n} x_i\Big) \;=\; \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \;\equiv\; 0.$$

Solving for the critical point (and doing a second-derivative test not illustrated here), we find that the MLE is $\hat{\lambda} = 1/\bar{X}$. □

Remarks:

- The MLE $\hat{\lambda} = 1/\bar{X}$ makes sense, since $\text{E}[X_1] = 1/\lambda$ for the exponential distribution.
- Perhaps surprisingly, even though we showed in §5.2.2 that $\bar{X} = 1/\hat{\lambda}$ is unbiased for $\text{E}[X_1] = 1/\lambda$, it turns out that the MLE $\hat{\lambda}$ is *slightly biased* for λ (see Exercise 5.9). This is actually a minor drawback of unbiased estimation.
- At the end of our toils, we put a little $\widehat{\text{hat}}$ over λ to indicate that this is the MLE. It's like a party hat!
- And finally, we make all of the little x_i's into big X_i's to indicate that this is a random variable. Our MLE is all grown up!

- You should always perform a second-derivative test, but maybe we won't blame you if you don't.

MLEs work well for discrete distributions too.

Example: Suppose $X_1, X_2 \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$. We will find the MLE for p. Since $\text{Bern}(p) \sim \text{Bin}(1, p)$, we can write the pmf as

$$f(x) \;=\; p^x (1-p)^{1-x}, \quad x = 0, 1.$$

Thus, the likelihood function is

$$L(p) \;=\; \prod_{i=1}^{n} f(x_i) \;=\; \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \;=\; p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i},$$

so that

$$\ell\text{n}(L(p)) \;=\; \sum_{i=1}^{n} x_i \, \ell\text{n}(p) + \left(n - \sum_{i=1}^{n} x_i\right) \ell\text{n}(1-p).$$

We set the derivative with respect to p to zero,

$$\frac{d}{dp}\ell\text{n}(L(p)) \;=\; \frac{\sum_i x_i}{p} - \frac{n - \sum_i x_i}{1-p} \;\equiv\; 0,$$

and finally solve for the MLE, $\hat{p} = \bar{X}$. This is intuitively nice since $\text{E}[X] = p$. □

For instance, suppose we observe the following 30 $\text{Bern}(p)$ trial outcomes (1 is a success, 0 a failure):

0	1	1	0	0	1	1	1	1	1	0	1	1	0	1
1	1	1	1	0	0	1	1	1	1	1	1	0	1	0

Then $\hat{p} = \bar{X} = 21/30 = 0.7$. □

§5.2.5.2 Adolescent MLE Examples

MLEs can also be used to attack more-substantial problems. First, we offer a normal distribution example in which we find MLEs for two parameters at once.

Example: Suppose $X_1, X_2 \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$. We will find the *simultaneous* MLEs for μ and σ^2. The likelihood function is

$$\begin{aligned} L(\mu, \sigma^2) \;&=\; \prod_{i=1}^{n} f(x_i) \\ &=\; \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right\} \\ &=\; \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}. \end{aligned}$$

This implies that

$$\ell\mathrm{n}(L(\mu,\sigma^2)) \;=\; -\frac{n}{2}\ell\mathrm{n}(2\pi) - \frac{n}{2}\ell\mathrm{n}(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Thus, taking the partial derivative with respect to μ, we obtain

$$\frac{\partial}{\partial\mu}\ell\mathrm{n}(L(\mu,\sigma^2)) \;=\; \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) \;\equiv\; 0,$$

and so the MLE is $\hat{\mu} = \bar{X}$ (which again makes sense).
Now do the same thing for σ^2 by taking the partial with respect to σ^2 (*not* σ),

$$\frac{\partial}{\partial\sigma^2}\ell\mathrm{n}(L(\mu,\sigma^2)) \;=\; -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 \;\equiv\; 0$$

(where we snuck in $\hat{\mu}$ for μ), and eventually get

$$\widehat{\sigma^2} \;=\; \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2. \quad \square$$

Remark: Notice how similar $\widehat{\sigma^2}$ is to the (unbiased) sample variance,

$$S^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \;=\; \frac{n}{n-1}\,\widehat{\sigma^2}.$$

The estimator $\widehat{\sigma^2}$ is a little bit biased, but it has slightly less variance than S^2. In any case, as n gets big, S^2 and $\widehat{\sigma^2}$ become the same.

The MLEs for the normal distribution's two parameters μ and σ^2 are more-or-less uncoupled in such a way that the respective calculations were easy. That is not quite the case for the gamma distribution, which also has two parameters.

Example: The pdf of the gamma distribution with parameters r and λ is

$$f(x) \;=\; \frac{\lambda^r}{\Gamma(r)}x^{r-1}e^{-\lambda x}, \quad x>0.$$

Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Gam}(r,\lambda)$. We will find the MLEs for r and λ. First of all, the likelihood function is

$$L(r,\lambda) \;=\; \prod_{i=1}^{n} f(x_i) \;=\; \frac{\lambda^{nr}}{\big[\Gamma(r)\big]^n}\Big(\prod_{i=1}^{n}x_i\Big)^{r-1}e^{-\lambda\sum_{i=1}^{n}x_i},$$

so that

$$\ell\mathrm{n}(L) \;=\; nr\,\ell\mathrm{n}(\lambda) - n\,\ell\mathrm{n}\big(\Gamma(r)\big) + (r-1)\,\ell\mathrm{n}\Big(\prod_{i=1}^{n}x_i\Big) - \lambda\sum_{i=1}^{n}x_i.$$

Take the partial derivative with respect to λ and set this to 0 to obtain

$$\frac{\partial}{\partial\lambda}\ell\mathrm{n}(L) \;=\; \frac{nr}{\lambda} - \sum_{i=1}^{n} x_i \;\equiv\; 0,$$

so that the MLE for λ is $\hat{\lambda} = \hat{r}/\bar{X}$.

The trouble in River City (A pre-*Hamilton* rap) is that we need to find $\hat{r}$, which will not be easy — but we will try our best. Similar to the above work, we get

$$\frac{\partial}{\partial r}\ell\mathrm{n}(L) \;=\; n\,\ell\mathrm{n}(\lambda) - \frac{n}{\Gamma(r)}\,\frac{d}{dr}\Gamma(r) + \ell\mathrm{n}\Big(\prod_{i=1}^{n} x_i\Big) \;\equiv\; 0, \tag{5.5}$$

where $\psi(r) \equiv \Gamma'(r)/\Gamma(r)$ is known as the **digamma function.**

At this point, we substitute $\hat{\lambda} = \hat{r}/\bar{X}$ into Equation (5.5), and use a *computer algorithm* (e.g., bisection, Newton's method, etc.) to search for the value of r that solves

$$n\,\ell\mathrm{n}(r/\bar{x}) - n\psi(r) + \ell\mathrm{n}\Big(\prod_{i=1}^{n} x_i\Big) \;\equiv\; 0.$$

The gamma function is readily available in any reasonable software package; but if the *di*gamma function happens to be unavailable in your town, you can take advantage of the approximation

$$\Gamma'(r) \;\doteq\; \frac{\Gamma(r+h) - \Gamma(r)}{h} \quad \text{(for any small } h \text{ of your choosing).}$$

For instance, choosing $h = 0.01$, we find that

$$\Gamma'(1.5) \;\doteq\; \frac{\Gamma(1.51) - \Gamma(1.5)}{0.01} \;=\; \frac{0.8865917 - 0.8862269}{0.01} \;=\; 0.03648,$$

which compares nicely to the actual value (obtained via Matlab) rounded to 0.03649.

See §7.5.3 where we carry out an extended numerical example involving the gamma distribution, including the search component. □

Finally, an interesting example where calculus is not applied directly.

Example: Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$. We will find the MLE for θ. First of all, recall that the pdf is $f(x) = 1/\theta$, $0 < x < \theta$, and you need to beware of the funny limits. The likelihood function is

$$L(\theta) \;=\; \prod_{i=1}^{n} f(x_i) \;=\; \begin{cases} 1/\theta^n & \text{if } 0 \le x_i \le \theta,\ \forall i \\ 0 & \text{otherwise.} \end{cases}$$

In order to have $L(\theta) > 0$, we must have $0 \le x_i \le \theta$, *for all* i. In other words, we must have $\theta \ge \max_i x_i$. Subject to this constraint, $L(\theta) = 1/\theta^n$ is maximized at the smallest possible θ value, namely, $\hat{\theta} = \max_i X_i$. This makes sense in light of the similar (unbiased) estimator, $Y_2 = \frac{n+1}{n}\max_i X_i$, that we looked at previously in §5.2.2. □

Remark: We used very little calculus in this example!

§5.2.5.3 Invariance Property of MLEs

All of our work so far on MLEs leads up to the *Invariance Property*, which allows us to use MLEs in extremely general ways, notably, on the goodness-of-fit tests that we will discuss in Chapter 7.

Theorem (Invariance Property of MLEs): If $\hat{\theta}$ is the MLE of some parameter θ, and $h(\cdot)$ is any reasonable function, then $h(\hat{\theta})$ is the MLE of $h(\theta)$.

Remark: We noted before that such a property does *not* hold for unbiasedness. For instance, although $\mathrm{E}[S^2] = \sigma^2$, it is usually the case that $\mathrm{E}\left[\sqrt{S^2}\right] \neq \sigma$.

Remark: The proof of the Invariance Property is "easy" when $h(\cdot)$ is a one-to-one function. It's not so easy — but still usually true — when $h(\cdot)$ is nastier. In any case, we will not carry out the proof here.

Example: Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Nor}(\mu, \sigma^2)$.

Recall that the MLE for σ^2 is $\widehat{\sigma^2} = \sum_{i=1}^n (X_i - \bar{X})^2/n$. If we consider the function $h(y) = +\sqrt{y}$, then the Invariance Property says that the MLE of σ is

$$\hat{\sigma} \;=\; \sqrt{\widehat{\sigma^2}} \;=\; \left[\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2\right]^{1/2}. \quad \square$$

Example: Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Bern}(p)$.

We saw that the MLE for p is $\hat{p} = \hat{X}$. Then Invariance says that $\hat{p}(1-\hat{p}) = \bar{X}(1-\bar{X})$ is the MLE for $\mathrm{Var}(X_i) = p(1-p)$. $\square$

And, finally, here is an example that has tremendous applications in survival analysis and the actuarial sciences, where we are interested in whether or not an item (or human) will survive beyond a certain time.

Example: Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Exp}(\lambda)$.

Recall that the MLE for λ is $\hat{\lambda} = 1/\bar{X}$. Meanwhile, we define the **survival function** as

$$\bar{F}(x) \;\equiv\; \mathrm{P}(X_1 > x) \;=\; 1 - F(x) \;=\; e^{-\lambda x}.$$

Then Invariance says that the MLE of $\bar{F}(x)$ is

$$\widehat{\bar{F}(x)} \;=\; e^{-\hat{\lambda}x} \;=\; e^{-x/\bar{X}},$$

where that thing over $\bar{F}(x)$ is more a roof than a hat! $\square$

Consider the following 30 battery lifetimes (in years) that I've generated from an $\mathrm{Exp}(\lambda)$ distribution.

4.082	1.575	4.031	2.008	4.491	4.003	3.041	0.898	0.401	2.005
0.084	0.396	3.761	2.407	3.776	5.296	3.672	0.620	2.528	5.255
3.930	2.750	3.931	1.502	1.741	0.218	1.727	2.923	0.714	0.290

An easy calculation reveals that the sample mean $\bar{X} = 2.469$. Thus, the MLE for λ is $\hat{\lambda} = 1/\bar{X} = 0.405$. Therefore, the MLE of the probability that a particular part will last more than 3 years is

$$\widehat{\bar{F}(x)} \;=\; e^{-x/\bar{X}} \;=\; e^{-3(0.405)} \;=\; 0.297. \quad \square$$

§5.2.5.4 Asymptotic Normality of MLEs

MLEs are (usually) asymptotically normal! This fact will prove useful when we derive a variety of *confidence intervals* for different parameters θ in Chapter 6.

Theorem (Asymptotic Normality of MLEs): Suppose that $\hat{\theta}_n$ is the MLE of some parameter θ, and $I_n(\theta)$ is the associated Fisher information from Equations (5.2) or (5.3). Assuming that all of the mysterious assumptions hinted at in §5.2.4 hold, then as the sample size $n \to \infty$, we have

$$\sqrt{I_n(\theta)}\,\big(\hat{\theta}_n - \theta\big) \;\xrightarrow{d}\; \text{Nor}(0,1).$$

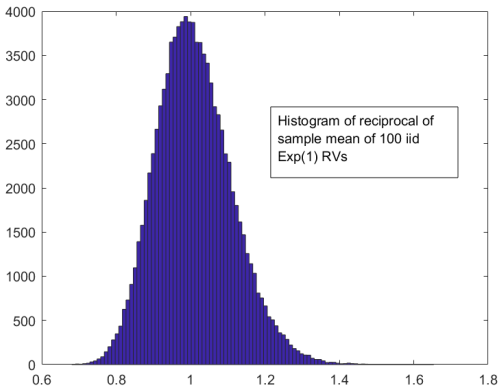
Example: If $X_1, X_2, \ldots, X_n$ are iid Exp(λ), then

$$\sqrt{I_n(\lambda)}\big(\hat{\lambda}_n - \lambda\big) \;=\; \frac{\sqrt{n}}{\lambda}\left(\frac{1}{\bar{X}} - \lambda\right) \;\xrightarrow{d}\; \text{Nor}(0,1).$$

In other words,

$$\frac{1}{\bar{X}} \;\approx\; \text{Nor}\left(\lambda,\, \frac{\lambda^2}{n}\right),$$

which is more-or-less borne out by the histogram below based on 100,000 simulated data points for the case $\lambda = 1$ and $n = 100$, for which Exercise 5.9 establishes $\text{E}[1/\bar{X}] = n\lambda/(n-1) = 1.010$.



This result is interesting in light of the fact that the usual CLT implies that $\bar{X}$ is also approximately normal, i.e.,

$$\bar{X} \;\approx\; \text{Nor}\left(\frac{1}{\lambda},\, \frac{1}{n\lambda^2}\right). \quad \square$$

5.2.6 Method of Moments

Another rich class of estimators arises from the *method of moments* (MoM). We have seen that MLEs sometimes take a great deal of work to come up with (don't even ask about the Weibull distribution until we get to §7.5.3!). MoM is an alternative to MLEs that is often easier to apply.

To begin, recall that the k^{th} **moment** of a random variable X is

$$\mu_k \;\equiv\; \text{E}[X^k] \;=\; \begin{cases} \sum_x x^k f(x) & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} x^k f(x)\, dx & \text{if } X \text{ is continuous.} \end{cases}$$

Definition: Suppose $X_1, X_2, \ldots, X_n$ are iid random variables. Then the **method of moments estimator** for $\mu_k = \text{E}[X^k]$ is the **sample moment** $m_k \equiv \sum_{i=1}^n X_i^k/n$, for $k = 1, 2, \ldots$.

Remark: As $n \to \infty$, the Law of Large Numbers implies that $\sum_{i=1}^n X_i^k/n \to \text{E}[X^k]$, i.e., $m_k \to \mu_k$ (so this is a good estimator).

Remark: You should always love your MoM!

Examples:

- The MoM estimator for true mean $\mu_1 = \mu = \text{E}[X_i]$ is the sample mean $m_1 = \bar{X} = \sum_{i=1}^n X_i/n$.
- The MoM estimator for $\mu_2 = \text{E}[X_i^2]$ is $m_2 = \sum_{i=1}^n X_i^2/n$.
- The MoM estimator for $\text{Var}(X_i) = \text{E}[X_i^2] - (\text{E}[X_i])^2 = \mu_2 - \mu_1^2$ is

$$m_2 - m_1^2 \;=\; \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2 \;=\; \frac{n-1}{n} S^2.$$

  (For large n, it's also okay to use S^2.) $\square$

General Game Plan: Express the parameter of interest in terms of the true moments $\mu_k = \text{E}[X^k]$. Then substitute in the sample moments $m_k = \sum_{i=1}^n X_i^k/n$.

Example: Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Pois}(\lambda)$. Recall that $\lambda = \text{E}[X_i]$, so a simple MoM estimator for λ is $\bar{X}$.

But also note that $\lambda = \text{Var}(X_i)$, so *another* MoM estimator for λ is $\frac{n-1}{n} S^2$ (or plain old S^2). $\square$

Remark: Use the easier-looking MoM estimator if you have a choice.

Example: Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$. MoM estimators for μ and σ^2 are $\bar{X}$ and $\frac{n-1}{n}S^2$ (or S^2), respectively. So for this example, these estimators are the same as the MLEs. □

Let's finish up with a less-trivial example.

Example: Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Beta}(a, b)$. The pdf is

$$f(x) \;=\; \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1.$$

It turns out (after lots of algebra) that

$$\text{E}[X] \;=\; \frac{a}{a+b} \quad \text{and} \quad \text{Var}(X) \;=\; \frac{ab}{(a+b)^2(a+b+1)}.$$

Let's estimate a and b via MoM. To do so, note that

$$\text{E}[X] \;=\; \frac{a}{a+b},$$

which implies

$$a \;=\; \frac{b\text{E}[X]}{1-\text{E}[X]} \;\doteq\; \frac{b\bar{X}}{1-\bar{X}} \tag{5.6}$$

and

$$\text{Var}(X) \;=\; \frac{ab}{(a+b)^2(a+b+1)} \;=\; \frac{\text{E}[X]\, b}{(a+b)(a+b+1)}. \tag{5.7}$$

Get the MoM for b, plug the following into Equation (5.7): $\bar{X}$ for $\text{E}[X]$, S^2 for $\text{Var}(X)$, and $b\bar{X}/(1-\bar{X})$ for a. Here's what we get after a corvée of algebra:

$$b \;\doteq\; \frac{(1-\bar{X})^2\bar{X}}{S^2} - 1 + \bar{X}.$$

To finish up, plug back into Equation (5.6) to get the MoM estimator for a. □

Example: Consider the following data set consisting of $n = 10$ observations that we have obtained from a beta distribution.

0.86 0.77 0.84 0.38 0.83 0.54 0.77 0.94 0.37 0.40

We immediately have $\bar{X} = 0.67$ and $S^2 = 0.04971$. The MoM estimators are

$$b \;\doteq\; \frac{(1-\bar{X})^2\bar{X}}{S^2} - 1 + \bar{X} \;=\; 1.1377,$$

and then

$$a \;\doteq\; \frac{b\bar{X}}{1-\bar{X}} \;=\; 2.310. \quad \square$$

5.3 Sampling Distributions

Goal: We'll discuss a number of distributions that will be needed later when we learn about confidence intervals (Chapter 6) and hypothesis tests (Chapter 7). In particular, we'll give brief synopses of the normal distribution (which, of course, we already know a great deal about), along with some new friends — the χ^2, t, and F distributions.

Definition: Recall that a **statistic** is just a function of the observations $X_1, X_2, \ldots, X_n$ from a random sample. The function does not depend explicitly on any unknown parameters.

Examples: $\bar{X}$, S^2, and $\max\{X_1, X_2, \ldots, X_n\}$ are statistics, but $(\bar{X} - \mu)/\sigma$ is not.

Since statistics are random variables, it's useful to figure out their distributions. The distribution of a statistic is called a **sampling distribution**. We'll look at several sampling distributions: the normal, χ^2, Student t, and F in §§5.3.1–5.3.4, respectively.

5.3.1 Normal Distribution

Main Take-Away: We have seen many times that $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$ implies that the sample mean $\bar{X} \sim \text{Nor}(\mu, \sigma^2/n)$.

This result is often used to get confidence intervals for μ and to conduct hypothesis tests. Stay tuned!

We'll now introduce some other important sampling distributions...

5.3.2 χ^2 Distribution

Definition/Theorem: If $Z_1, \ldots, Z_k \overset{\text{iid}}{\sim} \text{Nor}(0, 1)$, then $Y \equiv \sum_{i=1}^{k} Z_i^2$ has the **chi-squared distribution with k degrees of freedom** (df), and we write $Y \sim \chi^2(k)$.

Remarks: The term "df" informally corresponds to the number of "independent pieces of information" that you have. For example, suppose that $\sum_{i=1}^{n} X_i = c$, where c is a known constant after you actually observe the X_i's. Then you might have $n-1$ df, since knowledge of any $n-1$ of the X_i's gives you the remaining X_i.

We also informally "lose" a degree of freedom every time we have to estimate a parameter. For instance, If we have access to n observations, but have to estimate two parameters μ and σ^2 (say, for the normal distribution), then we might only end up with $n-2$ df.

In reality, df corresponds to the number of dimensions of a certain mathematical space (not covered in this course)!

Fun $\chi^2(k)$ **Facts:** The pdf is

$$f_Y(y) \;=\; \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} y^{\frac{k}{2}-1} e^{-y/2}, \;\; y > 0.$$

We can also show that $\mathrm{E}[Y] = k$ (see Exercise 5.29), and $\mathrm{Var}(Y) = 2k$.

You can easily compare pdf's to see that the exponential distribution is a special case, i.e., $\chi^2(2) \sim \mathrm{Exp}(1/2)$. More generally, if the df is even, we get a little surprise, namely, $\chi^2(2k) \sim \mathrm{Erlang}_k(1/2)$. Finally, for large k, the $\chi^2(k)$ distribution is approximately normal (by the CLT).

Definition: The $(1-\alpha)$ **quantile** of a (continuous) random variable X with cdf $F(x)$ is the value x_α such that $F(x_\alpha) = P(X \le x_\alpha) = 1-\alpha$. Note that $x_\alpha = F^{-1}(1-\alpha)$, where $F^{-1}(\cdot)$ is the **inverse cdf** of X.

Notation: If $Y \sim \chi^2(k)$, then we denote the $(1-\alpha)$ quantile with the special symbol $\chi^2_{\alpha,k}$ (instead of x_α). In other words, the so-called "tail" probability $\mathrm{P}(Y > \chi^2_{\alpha,k}) = \alpha$. You can look $\chi^2_{\alpha,k}$ up, e.g., in Table B.3 at the back of the book or via the Excel function `chisq.inv`$(1-\alpha, k)$.

Example: If $Y \sim \chi^2(10)$, then $\mathrm{P}(Y > \chi^2_{0.05,10}) = 0.05$, where we can look up $\chi^2_{0.05,10} = 18.31$. □

Theorem: χ^2's add up. If $Y_1, Y_2, \ldots, Y_n$ are *independent* with $Y_i \sim \chi^2(d_i)$, for all i, then $\sum_{i=1}^n Y_i \sim \chi^2\left(\sum_{i=1}^n d_i\right)$.

Proof: Just use a standard mgf argument, though no details here. □

So where does the χ^2 distribution come up in statistics? It usually arises when we try to estimate σ^2. See Chapters 6 and 7.

Example: If $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \mathrm{Nor}(\mu, \sigma^2)$, then we'll show in Chapter 6 that

$$S^2 \;=\; \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 \;\sim\; \frac{\sigma^2 \chi^2(n-1)}{n-1}. \quad \square$$

5.3.3 Student t Distribution

Definition/Theorem: Suppose that $Z \sim \mathrm{Nor}(0,1)$ and $Y \sim \chi^2(k)$, and that Z and Y are *independent*. Then $T \equiv Z/\sqrt{Y/k}$ has the **Student** t **distribution with** k **degrees of freedom**, and we write $T \sim t(k)$.

Fun Facts: The pdf is (see Exercise 5.30)

$$f_T(x) \;=\; \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)} \left(\frac{x^2}{k} + 1\right)^{-(k+1)/2}, \quad x \in \mathbb{R}.$$

The $t(k)$ looks like the Nor(0,1), except the t has fatter tails. The $k = 1$ case gives the **Cauchy** distribution, which has *really* fat tails (as we've seen before).

As the df k becomes large, $t(k) \to \text{Nor}(0,1)$ (see Exercise 5.31).

It is easy to show that $\text{E}[T] = 0$ for $k > 1$, and $\text{Var}(T) = \frac{k}{k-2}$ for $k > 2$.

Notation: If $T \sim t(k)$, then we denote the $(1-\alpha)$ quantile by $t_{\alpha,k}$. In other words, $\text{P}(T > t_{\alpha,k}) = \alpha$. See Table B.2 in the Appendix.

Example: If $T \sim t(10)$, then $\text{P}(T > t_{0.05,10}) = 0.05$, where we find $t_{0.05,10} = 1.812$ in the back of the book or via the Excel function `t.inv(`$1-\alpha, k$`)`. □

So what do we use the t distribution for in statistics? It comes about when we find confidence intervals and conduct hypothesis tests for the mean μ, especially in the case where the variance is unknown. Again, stay tuned for Chapters 6 and 7.

The Story Behind the Name: By the way, why did we originally call it the **Student** t distribution? "Student" is the pseudonym of William Gossett,[2] who first derived the distribution. Gossett was a statistician at the Guinness Brewery in Ireland and had to work on his research anonymously so that Guinness wouldn't risk its trade secrets.

5.3.4 F Distribution

Definition/Theorem: Suppose that $X \sim \chi^2(n)$ and $Y \sim \chi^2(m)$, and that X and Y are *independent*. Then $F \equiv \frac{X/n}{Y/m} = mX/(nY)$ has the F **distribution**, with n and m df (yes, you have to specify *two* df values). It's denoted by $F \sim F(n,m)$.

Fun Facts: The nasty-looking pdf is

$$f_F(x) \;=\; \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \; \frac{\left(\frac{n}{m}\right)^{\frac{n}{2}} x^{\frac{n}{2}-1}}{\left(\frac{n}{m}x+1\right)^{\frac{n+m}{2}}}, \quad x > 0.$$

The $F(n,m)$ is usually a bit skewed to the right.

It can be shown that $\text{E}[F] = m/(m-2)$ for $m > 2$, and $\text{Var}(F) = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$ for $m > 4$.

The t distribution is a special case — can you figure out which?

Notation: If $F \sim F(n,m)$, then we denote the $(1-\alpha)$ quantile by $F_{\alpha,n,m}$. That is, $\text{P}(F > F_{\alpha,n,m}) = \alpha$. Tables B.4–B.7 can be found in back of the book for various α, n, m, or you can use the Excel function `f.inv(`$1-\alpha, n, m$`)`.

Example: If $F \sim F(5,10)$, then $\text{P}(F > F_{0.05,5,10}) = 0.05$, where we find $F_{0.05,5,10} = 3.326$. □

Remark: The F distribution has the interesting property that $F_{1-\alpha,m,n} = 1/F_{\alpha,n,m}$. Use this fact if you have to find something like $F_{0.95,\mathbf{10},\mathbf{5}} = 1/F_{0.05,\mathbf{5},\mathbf{10}} = 1/3.326$, where we note that the df's are swapped. This property mitigates the need to provide quantiles below 0.5 in the tables.

[2] Not to be confused with Academy Award winning actor, Louis Gossett, Jr.

So what do we use the F distribution for in statistics? It's used when we find confidence intervals and conduct hypothesis tests for the ratio of variances from two different populations. For instance, are the variances from the two populations about the same, or are they "significantly" different? Details coming up later in Chapters 6 and 7.

5.4 Exercises

1. (§5.1.1) I'm conducting a poll today in which I ask respondents who their favorite political candidate is — Smith, Jones, or Thomas. What type of data am I collecting?

   (a) Continuous

   (b) Discrete

   (c) Categorical

   (d) Time Series

   (e) Ordinal

2. (§5.1.2) Suppose we collect the following observations: 7, −2, 1, 6. What are the sample mean and the sample variance?

3. (§5.1.2) Here are the IQ's for the students in my Probability and Statistics class at Georgia Tech. My students are awesome!

146	149	154	161
141	152	156	165
136	148	150	182
145	146	154	149
154	147	140	147
140	142	149	153
145	131	167	149
170	168	146	160
150	140	137	151
169	137	163	153

   Construct a histogram and calculate the sample mean and sample variance from this data.

4. (§5.1.2) Here's a stem-and-leaf diagram of scores from a recent test I gave. Find the sample mean, sample median, and sample standard deviation.

4	4
5	8 3
6	9 6 6 2
7	7 7 5 4 3 1 1 0
8	9 7 7 6 5 3 2 1 1
9	6 4 3 1
10	0 0

5. (§5.1.2) Let's generate 1000 Unif(0,1) random variables in Excel via the `rand()` function (or use whatever your favorite programming language is). Find the sample mean and variance of your observations.

6. (§5.1.2) What value of c minimizes the weighted sum of squares $\sum_{i=1}^{n} w_i(X_i - c)^2$, where $w_1, \ldots, w_n$ constitute a given set of weights?

7. (§5.1.3) We are interested in modeling a simple single-server queueing system, for example, a one-man barber shop. Customers arrive at the barber shop randomly, but always one-at-a-time. Moreover, the overall arrival rate is pretty much the same all day, and the numbers of arrivals in disjoint time periods are known to be independent. If, upon his arrival, a customer sees a line of a least one other person waiting (the "barber queue" ☺), then there is a 25% chance that he gets annoyed and leaves. Otherwise, he joins the line, which is handled in a first-in-first-out manner. Eventually, the customer gets his turn to be served by the barber. Now, this is a great barber, and he does a million little things comprising the overall service time, each of which takes iid time — hair wash, trim, head massage, talk about sports, mani, pedi, pay up, etc., etc. After all of this stuff, the customer leaves. How might you model the various components of this system?

   (a) Poisson times between arrivals, Bernoulli decision process about whether to enter the line, normal overall service times

   (b) Exponential times between arrivals, Bernoulli decision process about whether to enter the line, normal overall service times

   (c) Poisson times between arrivals, Bernoulli decision process about whether to enter the line, uniform overall service times

   (d) Exponential times between arrivals, geometric decision process about whether to enter the line, exponential overall service times

8. (§5.2.1) Suppose that $X_1, X_2, \ldots, X_n$ are iid with unknown mean μ and unknown σ^2. Which of the following could be considered a statistic?

   (a) The sample mean, $\bar{X}$

   (b) The sample standard deviation, S

   (c) The standardized sample mean, $(\bar{X} - \mu)/\sqrt{\sigma^2/n}$

   (d) Both (a) and (b)

   (e) None of the above

9. (§5.2.2) Suppose that $X_1, X_2, \ldots, X_n$ are iid Exp(λ). We know that the sample mean $\bar{X}$ is unbiased for the mean $1/\lambda$. But prove that $1/\bar{X}$ is a little bit *biased* for λ.

10. (§5.2.2) Suppose that the two observations X_1 and X_2 are iid.

    (a) Show that for this special case of just two observations, we have the algebraic identity $S^2 = (X_1 - X_2)^2/2$.

    (b) Suppose, specifically, that the two observations X_1 and X_2 are iid Bern(p). We know that the sample variance S^2 is unbiased for the variance $\sigma^2 = pq$, where $q = 1 - p$. But prove that $S = |X_1 - X_2|/\sqrt{2}$ is *biased* for $\sigma = \sqrt{pq}$.

11. (§5.2.2) Suppose that $X_1, X_2, \ldots, X_n$ are iid with $\mathrm{P}(X_i = 1) = 1 - p$, and $\mathrm{P}(X_i = 2) = p$ (essentially a Bern(p) + 1 distribution).

    (a) Prove that the estimator $\hat{p} = \bar{X} - 1$ is unbiased for p.

    (b) For finite sample size n, is the estimator $\hat{p}^2 = (\bar{X} - 1)^2$ unbiased for p^2?

12. (§5.2.3) Suppose $X_1, X_2, \ldots$ are iid with mean μ and variance σ^2. Denote the sample mean based on n observations by $\bar{X}_n = \sum_{i=1}^{n} X_i/n$ for $n \geq 1$. Which of $\bar{X}_{10}$ and $\bar{X}_{20}$ is the better estimator of μ? Explain your choice.

13. (§5.2.3) Consider two estimators, T_1 and T_2, for an unknown parameter θ. Suppose that $\mathrm{Bias}(T_1) = 0$, $\mathrm{Bias}(T_2) = \theta$, $\mathrm{Var}(T_1) = 4\theta^2$, and $\mathrm{Var}(T_2) = \theta^2$. Which estimator might you decide to use and why?

14. (§5.2.3) Suppose that $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$ are estimators of θ, and we know that

$$\mathrm{E}[\hat{\theta}_1] = \mathrm{E}[\hat{\theta}_2] = \theta, \quad \text{and} \quad \mathrm{E}[\hat{\theta}_3] = \theta + 3,$$

    and

$$\mathrm{Var}(\hat{\theta}_1) = 12, \quad \mathrm{Var}(\hat{\theta}_2) = 10, \quad \text{and} \quad \mathrm{Var}(\hat{\theta}_3) = 4.$$

    Which of these estimators has the lowest MSE?

15. (§5.2.4) Suppose that $X_1, X_2, \ldots, X_n$ are iid Pois(λ).

    (a) Based on this sample, find $I_n(\lambda)$, the Fisher information for λ.

    (b) What is the Cramér–Rao Lower Bound?

    (c) Recall that the sample mean $\bar{X}$ is unbiased for $\lambda = \mathrm{E}[X_i]$. Does $\mathrm{Var}(\bar{X})$ achieve the CRLB?

16. (§5.2.4) Suppose that $X_1, X_2, \ldots, X_n$ are iid Nor(μ, σ^2).

    (a) Based on this sample, find $I_n(\mu)$, the Fisher information for μ.

    (b) What is the Cramér–Rao Lower Bound?

    (c) Recall that the sample mean $\bar{X}$ is unbiased for $\mu = \mathrm{E}[X_i]$. Does $\mathrm{Var}(\bar{X})$ achieve the CRLB?

17. (§5.2.5) Suppose that we have observed three iid Exp(λ) customer service times, namely, 2, 4, and 9 minutes. What is the maximum likelihood estimate of λ?

18. (§5.2.5) Suppose that $X_1, X_2, \ldots, X_n$ are iid Pois(λ). Find the MLE of λ.

19. (§5.2.5) Consider an iid sample of size $n = 4$ from a Unif$(0, \theta)$ distribution, where $\theta > 0$ is unknown. If $U_1 = 3.7$, $U_2 = 16.3$, $U_3 = 1.6$, and $U_4 = 7.9$, find the MLE $\hat{\theta}$.

20. (§5.2.5) If $X_1, X_2, \ldots, X_n$ are iid Nor(μ, σ^2), where μ and σ^2 are unknown, what is the expected value of the MLE for σ^2?

21. (§5.2.5) By hook or by crook, evaluate the digamma function at the point 2.5, that is, compute $\psi(2.5) = \Gamma'(2.5)/\Gamma(2.5)$.

22. (§5.2.5) Suppose $X_1, X_2, \ldots, X_n$ are iid Exp(λ), and we observe the sample mean $\bar{X} = 2$. What is the maximum likelihood estimate of $\mathrm{P}(X_1 > 2)$?

23. (§5.2.5) [Bonus Extension of Problems 9 and 22!] Suppose that $X_1, X_2, \ldots, X_n$ are iid Exp(λ). Recall that the MLE for the survival function is $\widehat{\mathrm{P}}(X_1 > t) = e^{-t/\bar{X}}$. Let's show that the MLE is a tad biased for $\mathrm{P}(X_1 > t) = e^{-\lambda t}$.

    (a) Establish that

$$\mathrm{E}\big[\widehat{\mathrm{P}}(X_1 > t)\big] \;=\; \frac{\lambda^n}{(n-1)!} \int_0^\infty y^{n-1} \exp\left[-\frac{nt}{y} - \lambda y\right] dy.$$

    (b) Unfortunately, we can't really simplify the integral into an easy closed-form expression. That being said, you can certainly use numerical integration (e.g., via Matlab or Mathematica) to graph $\mathrm{E}\big[\widehat{\mathrm{P}}(X_1 > t)\big]$ for $\lambda = 1/2$ and $t = 2$ as n increases. And what do you think the expected value converges to?

    Fun Fact: Don't worry about this, but the integral in (a) is actually equivalent to $\frac{2(n\lambda t)^{n/2}}{(n-1)!} K_n\big(2\sqrt{n\lambda t}\,\big)$, where $K_n(\cdot)$ is a *Bessel function* (see Gradshteyn and Ryzhik [3], §3.471.9 and Chapter 8, for details).

24. (§5.2.5) Suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Geom}(p)$.

    (a) Find the maximum likelihood estimator of p.

    (b) Suppose you have $n = 5$ observations: 1, 5, 1, 2, 3. What is the resulting MLE value?

    (c) What is the MLE of $2p^3$?

25. (§5.2.5) Suppose that $X_1, X_2, \ldots, X_n$ are iid from the Pareto distribution, i.e., X_i has pdf $f(x) = \alpha x^{-(\alpha+1)}$, for $x \geq 1$ and parameter $\alpha > 1$.

    (a) Based on this sample, find $I_n(\alpha)$, the Fisher information for α.

    (b) What is the Cramér–Rao Lower Bound?

    (c) Find the MLE $\hat{\alpha}_n$ of α.

(d) What is the asymptotic distribution of the MLE?

(e) Find the mean and variance of the MLE. Hint: See Exercises 2.27 and 5.9, and roll up your sleeves.

(f) Find c_n so that $c_n\hat{\alpha}_n$ is unbiased for α. How does $\text{Var}(c_n\hat{\alpha}_n)$ compare to the CRLB?

26. (§5.2.6) Let $X_1, \ldots, X_n$ be an iid sample of Bernoulli random variables with parameter p. Find a MOM estimator for p. Are there multiple MOM estimators?

27. (§5.3) A population of male college students has heights that are normally distributed with a mean of 68 inches and a standard deviation of 3 inches. A random sample of 10 students is selected. Specify the sampling distribution of the sample mean, $\bar{X}$.

28. (§5.3) It's Quantile Time!

(a) Suppose $Y \sim \chi^2(8)$. Find $\text{P}(Y \leq 2.73)$.

(b) Find $\chi^2_{0.025,6}$, i.e., the 97.5^{th} quantile of the $\chi^2(6)$ distribution such that $\text{P}\big(\chi^2(6) \leq \chi^2_{0.025,6}\big) = 0.975$.

(c) Find $t_{0.25,9}$, i.e., the 75^{th} quantile of the $t(9)$ distribution such that $\text{P}\big(t(9) \leq t_{0.25,9}\big) = 0.75$.

(d) Suppose $T \sim t(1000)$. What's $\text{P}(T > 1.96)$?

(e) Find $F_{0.05,4,9}$, i.e., the quantile such that $\text{P}\big(F(4,9) \leq F_{0.05,4,9}\big) = 0.95$.

(f) Find $F_{0.975,4,5}$, i.e., the quantile such that $\text{P}\big(F(4,5) \leq F_{0.975,4,5}\big) = 0.025$.

29. (§5.3) Find the expected value of the $\chi^2(k)$ distribution.

30. (§5.3) Recall that if $Z \sim \text{Nor}(0,1)$, $Y \sim \chi^2(k)$, and Z and Y are *independent*, then $T \equiv Z/\sqrt{Y/k} \sim t(k)$. Use the corollary of the Honors Theorem of §3.7.1 to obtain the pdf of T.

31. (§5.3) *Stirling's approximation* says that

$$\frac{\Gamma(n+1)}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n} \;\to\; 1 \quad \text{as } n \to \infty.$$

Use this result to show that the pdf of the t distribution with k df converges to that of the standard normal as k becomes large.

32. (§5.3) Suppose that $X \sim \text{Nor}(0,1)$ and $Y \sim \chi^2(3)$, and X and Y are independent. Name the distribution of $3X^2/Y$. (Hint: Recall that $[\text{Nor}(0,1)]^2 \sim \chi^2(1)$.)

Chapter 6

Confidence Intervals

In this chapter we finally begin our exploration into the exciting topic of *confidence intervals* (CIs)! We've all heard about confidence intervals, which provide an efficient way to report the uncertainty around a statistic. For instance, suppose that according to a new poll, President Smith has a popularity of 56%, plus or minus 3%. This can be loosely interpreted as a statement that we are 95% sure that Smith's true popularity is somewhere in the range $[0.53, 0.59]$.

In this chapter, we'll formalize such concepts in order to tackle a wide range of problems. §6.1 starts things off with a succinct introduction to the topic. The remaining sections present a potpourri of different types of CIs. §6.2 looks at the special case of a CI for the mean of a normal distribution when we happen to know the underlying variance (which may or may not be a big assumption). §6.3 considers two normal distributions with known variances, and gives a CI for the difference in their means. §§6.4 and 6.5 do the same as their two predecessors, but in the more-realistic scenario in which the underlying variances are *unknown*. §§6.6–6.8 give CIs for the normal distribution's variance, the ratio of variances from two normal distributions, and the Bernoulli distribution's success probability, respectively. §6.9 takes advantage of the asymptotic normality of many maximum likelihood estimators to obtain a general class of CIs for a variety of parameters.

6.1 Introduction to Confidence Intervals

Idea: Instead of estimating a parameter by a **point estimator** alone, find a (random) **interval** that contains the unknown parameter with a certain probability. This gives us a way to express our confidence in the point estimator.

Example: The sample mean $\bar{X}$ is a point estimator for the unknown true mean μ. A 95% **confidence interval** for μ might look something like

$$\mu \in \left[\bar{X} - z_{0.025}\sqrt{\sigma^2/n},\ \ \bar{X} + z_{0.025}\sqrt{\sigma^2/n}\right],$$

where $z_{0.025} = 1.96$ is the 0.975 quantile of the Nor$(0, 1)$ distribution (see Table B.1 in the Appendix). The interpretation is that we have confidence of 0.95 (i.e., $1.0 - 2(0.025)$) that μ lies in the interval.

Definition: A $100(1-\alpha)\%$ **confidence interval** for an unknown parameter θ is given by two random variables L and U satisfying $\mathrm{P}(L \le \theta \le U) = 1 - \alpha$.

The random variables L and U are the **lower** and **upper** confidence limits.

The quantity $(1-\alpha)$ is the **confidence coefficient**, specified in advance. There is a $(1-\alpha)$ chance that θ actually lies between L and U.

Since $L \le \theta \le U$, we call $[L, U]$ a **two-sided** CI for θ.

If L is such that $\mathrm{P}(L \le \theta) = 1 - \alpha$, then $[L, \infty)$ is a $100(1-\alpha)\%$ **one-sided lower** CI for θ.

Similarly, if U is such that $\mathrm{P}(\theta \le U) = 1 - \alpha$, then $(-\infty, U]$ is a $100(1-\alpha)\%$ **one-sided upper** CI for θ.

Example: We're 95% sure that President Smith's popularity is $56\% \pm 3\%$.

Example: Table 6.1 illustrates realizations of confidence intervals arising from ten independent samples, each consisting of 100 different observations. From each sample, we use the 100 observations to recalculate L and U, in order to obtain a 95% confidence interval for θ. In each case, $[L, U]$ either *covers* (contains) or doesn't cover the unknown true value of θ (which, let's say for purposes of this discussion, is 2).

Some salient points from the table:

- The parameter θ is a fixed unknown constant (the true value of $\theta = 2$, but we wouldn't know this in practice), so it does not change from sample to sample.
- Some of the CIs are skinny (e.g., sample 6), and some are a bit more rotund (e.g., sample 7).
- Sometimes CIs miss θ by being too low, i.e., $\theta > U$ (sample 10).
- Sometimes CIs miss θ by being too high, i.e., $\theta < L$ (sample 3).
- We see that only eight out of the ten CIs actually cover θ. This is completely fine, even for a 95% CI, because...

Sample #	L	U	θ	CI covers θ?
1	1.86	2.23	2	Yes
2	1.90	2.31	2	Yes
3	3.21	3.86	2	No ☹
4	1.75	2.10	2	Yes
5	1.72	2.03	2	Yes
6	1.92	2.01	2	Yes
7	0.63	3.52	2	Yes
8	1.51	2.62	2	Yes
9	1.79	2.07	2	Yes
10	1.62	1.98	2	No ☹

Table 6.1: Ten confidence intervals from ten samples.

- As the number of samples gets large, the proportion of CIs that cover the unknown θ will approach $1 - \alpha = 0.95$. So, happily, it all works out in the end (assuming we are using mathematically valid CIs)!
- However, note that in the real world, you usually only get one sample and thus one shot at a CI — but you at least have probability $1 - \alpha$ of getting it right!

Confidence intervals are everywhere! We'll look at lots of CIs for the mean and variance of normal distributions, as well as CIs for the success probability of Bernoulli distributions. We'll also extend these results to compare competing normal distributions (e.g., which of two normals has the larger mean?), as well as competing Bernoulli distributions.

6.2 Confidence Interval for Normal Mean (Variance Known)

We'll start off with what is pretty much the easiest possible case.

Goal: Sample from a normal distribution with unknown mean μ and *known variance* σ^2. Use these observations to obtain a CI for μ.

Remark: Admittedly, this is an unrealistic case, since if we didn't know μ in real life, then we probably wouldn't know σ^2 either. But it's a good place to start the discussion.

Setup: Suppose that $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$, where σ^2 is *known*. Use $\bar{X} = \sum_{i=1}^{n} X_i/n$ as our point estimator for μ. (It's unbiased! It's the MLE! It's our

MoM!) Recall that

$$\bar{X} \sim \text{Nor}(\mu, \sigma^2/n) \Rightarrow Z \equiv \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim \text{Nor}(0, 1).$$

The quantity Z is called a **pivot**. It's a "starting point" for us. The definition of Z implies that

$$\begin{aligned}
1 - \alpha &= \text{P}\big(-z_{\alpha/2} \le Z \le z_{\alpha/2}\big) \\
&= \text{P}\left(-z_{\alpha/2} \le \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \le z_{\alpha/2}\right) \\
&= \text{P}\Big(-z_{\alpha/2}\sqrt{\sigma^2/n} \le \bar{X} - \mu \le z_{\alpha/2}\sqrt{\sigma^2/n}\Big) \\
&= \text{P}\Big(\underbrace{\bar{X} - z_{\alpha/2}\sqrt{\sigma^2/n}}_{L} \le \mu \le \underbrace{\bar{X} + z_{\alpha/2}\sqrt{\sigma^2/n}}_{U}\Big) \\
&= \text{P}(L \le \mu \le U).
\end{aligned}$$

Thus, we have

$100(1-\alpha)\%$ **two-sided CI for μ:**

$$\bar{X} - z_{\alpha/2}\sqrt{\sigma^2/n} \le \mu \le \bar{X} + z_{\alpha/2}\sqrt{\sigma^2/n}.$$

Remarks:

- Notice how we used the pivot and algebra to "isolate" μ all by itself to the middle of the inequalities.
- After you observe $X_1, \ldots, X_n$, you can calculate L and U. At that point, nothing is unknown, since L and U don't involve μ.
- Sometimes we'll write the CI in the intuitive form $\mu \in \bar{X} \pm H$, where the **half-width** (aka **half-length**) is

$$H \equiv z_{\alpha/2}\sqrt{\sigma^2/n}.$$

Example: Suppose we observe the weights of $n = 25$ interior linemen on a college football team. Assume that these are iid observations from a $\text{Nor}(\mu, \sigma^2)$ distribution, where we somehow *know* that the variance $\sigma^2 = 324$.

262.4	274.6	245.0	307.7	281.6
320.5	261.7	283.5	273.5	308.5
257.7	296.2	294.8	279.1	242.1
299.3	254.5	281.1	297.4	281.8
315.8	284.1	286.2	251.8	258.6

This data yields a sample mean of $\bar{X} = 279.98$. Since $z_{\alpha/2} = z_{0.025} = 1.96$, we obtain the following two-sided $100(1-\alpha) = 95\%$ CI for μ.

$$\mu \in \bar{X} \pm z_{\alpha/2}\sqrt{\sigma^2/n} = 280.0 \pm z_{0.025}\sqrt{324/25} = 280.0 \pm 7.1.$$

So a 95% CI for μ is $272.9 \le \mu \le 287.1$. □

Sample-Size Calculation: If we had taken more observations, then the CI would have gotten shorter, since $H = z_{\alpha/2}\sqrt{\sigma^2/n}$. In fact, how many observations should be taken to make the half-length (or "error") $\le \epsilon$? We easily see that

$$z_{\alpha/2}\sqrt{\sigma^2/n} \le \epsilon \quad \text{iff} \quad n \ge \sigma^2 z_{\alpha/2}^2/\epsilon^2. \quad \square$$

Example: Suppose, in the previous example, that we want the half-length to be ≤ 3, i.e., $\mu \in \bar{X} \pm 3$. What should n be?

$$n \ge \sigma^2 z_{\alpha/2}^2/\epsilon^2 = 324\,(1.96)^2/9 = 138.3.$$

Just to make n an integer, round up to $n = 139$. □

One-Sided Confidence Intervals: We can similarly obtain one-sided CIs for μ (if we're just interested in a bound in one direction): Consider the same pivot $Z = (\bar{X} - \mu)/\sqrt{\sigma^2/n} \sim \text{Nor}(0,1)$ that we used before. Then, by definition of the $(1-\alpha)$ quantile z_α, we have

$$\begin{aligned} 1-\alpha &= \mathrm{P}\big(Z \le z_\alpha\big) \\ &= \mathrm{P}\left(\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \le z_\alpha\right) \\ &= \mathrm{P}\Big(\bar{X} - \mu \le z_\alpha\sqrt{\sigma^2/n}\Big) \\ &= \mathrm{P}\Big(\bar{X} - z_\alpha\sqrt{\sigma^2/n} \le \mu\Big). \end{aligned}$$

This immediately gives us a $100(1-\alpha)\%$ *lower CI* for μ,

$$\mu \ge \bar{X} - z_\alpha\sqrt{\sigma^2/n}.$$

And by symmetry, a $100(1-\alpha)\%$ *upper CI* for μ is

$$\mu \le \bar{X} + z_\alpha\sqrt{\sigma^2/n}.$$

Note that we use the $1-\alpha$ quantile z_α for the *one-sided* CIs — *not* the $(1-\alpha/2)$ quantile, $z_{\alpha/2}$. This "reallocation" of α moves the one-sided bound closer to $\bar{X}$, and is therefore a bit more informative.

Example: Continuing yet again with our football data, let's get a 95% lower bound on the mean weight of those big boys. Since $z_{0.05} = 1.645$, we have

$$\mu \ge \bar{X} - z_\alpha\sqrt{\sigma^2/n} = 280.0 - z_{0.05}\sqrt{324/25} = 280.0 - 5.9 = 274.1. \quad \square$$

6.3 Confidence Interval for Difference of Normal Means (Variances Known)

Goal: Sample from two normal distributions with unknown means and *known variances*. Use these observations to obtain a CI for the *difference* of the two means.

Remark: This will give us information about which distribution is "better" than the other, at least in terms of their means. We'll do the more-realistic unknown variance case a little later in §6.5.

Example: Give a 95% confidence interval for the mean difference in the revenues produced by inventory policies X and Y.

Setup: Suppose we have samples of sizes n and m from the two competing populations:

$$\begin{aligned} X_1, X_2, \ldots, X_n &\overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma_x^2) \quad \text{(population 1)} \quad \text{and} \\ Y_1, Y_2, \ldots, Y_m &\overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma_y^2) \quad \text{(population 2)}, \end{aligned}$$

where the means μ_x and μ_y are *unknown*, while σ_x^2 and σ_y^2 are somehow *known*.

Also assume that the X_i's are *independent* of the Y_i's.

Let's find a CI for the difference in means, $\mu_x - \mu_y$. To do so, define the sample means from populations 1 and 2,

$$\bar{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \bar{Y} \equiv \frac{1}{m}\sum_{i=1}^{m} Y_i.$$

Obviously,

$$\bar{X} \sim \text{Nor}(\mu_x, \sigma_x^2/n) \quad \text{and} \quad \bar{Y} \sim \text{Nor}(\mu_y, \sigma_y^2/m),$$

so that

$$\bar{X} - \bar{Y} \sim \text{Nor}\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right).$$

This gives us the pivot

$$Z \equiv \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim \text{Nor}(0, 1),$$

so that

$$\text{P}(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha.$$

Using the same manipulations as in the single-population case, we immediately obtain

$100(1-\alpha)\%$ **two-sided CI for** $\mu_x - \mu_y$**:**

$$\mu_x - \mu_y \in \bar{X} - \bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}.$$

One-sided CIs: Similarly, we have a one-sided upper CI,

$$\mu_x - \mu_y \;\leq\; \bar{X} - \bar{Y} + z_\alpha \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}},$$

and a one-sided lower CI,

$$\mu_x - \mu_y \;\geq\; \bar{X} - \bar{Y} - z_\alpha \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}.$$

Example: A traveling professor gives the same test to XXX University students (X) and YYY University (Y) students. She assumes that the test scores are normally distributed with known standard deviations of $\sigma_x = 20$ points and $\sigma_y = 12$ points, respectively. She takes random samples of 40 XXX scores and 24 YYY tests and observes sample means of $\bar{X} = 95$ points and $\bar{Y} = 60$, respectively.

Let's find the 90% two-sided CI for $\mu_x - \mu_y$. Plugging into the two-sided version with $z_{\alpha/2} = z_{0.05}$, we get

$$\begin{aligned} \mu_x - \mu_y \quad &\in \quad \bar{X} - \bar{Y} \pm z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}} \\ &= \quad 35 \pm 1.645 \sqrt{\frac{400}{40} + \frac{144}{24}} \\ &= \quad 35 \pm 6.58, \end{aligned}$$

implying that $28.42 \leq \mu_x - \mu_y \leq 41.58$. In other words, we're 90% sure that $\mu_x - \mu_y$ lies in this interval. This means that we can informally conclude that, on average, XXX students score significantly higher than YYY kids. □

Remark: Let's assume that both sample sizes equal n. To obtain a half-length $\leq \epsilon$, we use reasoning analogous to that in §6.2 to find that we require

$$n \;\geq\; \frac{z_{\alpha/2}^2 (\sigma_x^2 + \sigma_y^2)}{\epsilon^2}.$$

The next section will examine the more-realistic case in which the variance of the underlying observations is unknown.

6.4 Confidence Interval for Normal Mean (Variance Unknown)

In this section, we'll look at the more-realistic case in which the variance of the underlying normal random variables is *unknown*. This takes a little more work, but has many more applications.

Setup: $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$, with σ^2 *unknown*.

Facts: Here are three results that we'll need in the upcoming discussion.

$$\text{(a)} \qquad \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \;\sim\; \text{Nor}(0,1).$$

$$\text{(b)} \qquad \bar{X} \text{ and } S^2 \text{ are } \textit{independent}.$$

$$\text{(c)} \qquad S^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \;\sim\; \frac{\sigma^2\chi^2(n-1)}{n-1}. \tag{6.1}$$

Sketch of Proof: We already proved Fact (a) back in §4.3.3. The proof of Fact (b) uses what is known as Cochran's Theorem, which is a bit beyond our scope. In order to derive Fact (c), we first apply some easy algebra to establish the identity

$$\begin{aligned}
\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}\big[(X_i-\bar{X})+(\bar{X}-\mu)\big]^2 \\
&= \sum_{i=1}^{n}(X_i-\bar{X})^2 + 2\sum_{i=1}^{n}(X_i-\bar{X})(\bar{X}-\mu) + \sum_{i=1}^{n}(\bar{X}-\mu)^2 \\
&= \sum_{i=1}^{n}(X_i-\bar{X})^2 + 2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X}) + n(\bar{X}-\mu)^2 \\
&= \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2 \quad \big(\text{since } \textstyle\sum_i(X_i-\bar{X})=0\big),
\end{aligned}$$

which can be rewritten as

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\mu)^2 \;=\; S^2 + \frac{n}{n-1}(\bar{X}-\mu)^2. \tag{6.2}$$

Note that $X_i \sim \text{Nor}(\mu,\sigma^2)$ together with the fact that $\big[\text{Nor}(0,1)\big]^2 \sim \chi^2(1)$ imply that $(X_i-\mu)^2 \sim \sigma^2\chi^2(1)$, for all i. Thus, the additive property of independent χ^2's (from §5.3.2) implies that

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\mu)^2 \;\sim\; \frac{\sigma^2\chi^2(n)}{n-1}. \tag{6.3}$$

Similarly, Fact (a) implies that

$$\frac{n}{n-1}(\bar{X}-\mu)^2 \;\sim\; \frac{\sigma^2\chi^2(1)}{n-1}. \tag{6.4}$$

Since $\bar{X}$ and S^2 are independent (by Fact (b)), we can substitute the results from Equations (6.3) and (6.4) into Equation (6.2) and then invoke χ^2 additivity again to obtain

$$\frac{\sigma^2\chi^2(n)}{n-1} \;\sim\; S^2 + \frac{\sigma^2\chi^2(1)}{n-1} \quad\Rightarrow\quad S^2 \;\sim\; \frac{\sigma^2\chi^2(n-1)}{n-1}. \quad \square$$

(Now it's nearly time for $t\ldots$)

Using Facts (a)–(c), we have, by the definition of the t distribution,

$$\frac{(\bar{X}-\mu)/\sqrt{\sigma^2/n}}{\sqrt{S^2/\sigma^2}} \sim \frac{\text{Nor}(0,1)}{\sqrt{\chi^2(n-1)/(n-1)}} \sim t(n-1).$$

In other words,

$$\frac{\bar{X}-\mu}{\sqrt{S^2/n}} \sim t(n-1).$$

Remark: This expression doesn't contain the unknown σ^2. It's been replaced by S^2, which we can calculate. This process is called **"standardizing and Studentizing."**

Now, by the *same* manipulations as in the known-variance case, we can obtain

$100(1-\alpha)\%$ **two-sided CI for μ:**

$$\mu \in \bar{X} \pm t_{\alpha/2,n-1}\sqrt{S^2/n},$$

where the t quantile can be found in Table B.2 in the Appendix or via software.

One-Sided CIs: The corresponding $100(1-\alpha)\%$ lower and upper CIs are, respectively,

$$\mu \geq \bar{X} - t_{\alpha,n-1}\sqrt{S^2/n} \quad \text{and} \quad \mu \leq \bar{X} + t_{\alpha,n-1}\sqrt{S^2/n}.$$

Remark: Here we use t-distribution quantiles (instead of normal quantiles as in the known-variance case from §6.2). The t quantile tends to be larger than the corresponding $\text{Nor}(0,1)$ quantile, so these unknown-variance CIs tend to be a little longer than the known-variance CIs. The longer CIs are the result of the fact that we lack precise information about the variance.

Example: Consider our data set from §6.2, which listed the weights of $n = 25$ college football players. Recall that the sample mean of the observations was $\bar{X} = 279.98$. In addition, this time around, let's assume that we don't know the variance σ^2, and instead we estimate it using the usual sample variance, which turns out to be $S^2 = 479.65$.

In order to obtain a two-sided 95% confidence interval for μ, we can use the Excel function `t.inv(0.975,24)` to get $t_{\alpha/2,n-1} = t_{0.025,24} = 2.064$. Then the half-length of the CI is

$$H = t_{\alpha/2,n-1}\sqrt{S^2/n} = 2.064\sqrt{479.65/25} = 9.04.$$

Thus, the CI is $\mu \in \bar{X} \pm H = 279.98 \pm 9.04$, or $270.94 \leq \mu \leq 289.02$, which happens to be a little wider than the analogous known-variance CI in §6.2. □

Here is R code for the above example:

```
x <- data.frame(x=c(
262.4, 274.6, 245.0, 307.7, 281.6,
320.5, 261.7, 283.5, 273.5, 308.5,
```

```
257.7, 296.2, 294.8, 279.1, 242.1,
299.3, 254.5, 281.1, 297.4, 281.8,
315.8, 284.1, 286.2, 251.8, 258.6))
print(confint(lm(x~1,x)))
  2.5 \%   97.5 \%
(Intercept) 270.94  289.02
```

6.5 Confidence Intervals for Difference of Normal Means (Variances Unknown)

This section studies CIs for the difference between the means of two competing normal populations, and is analogous the discussion started in §6.3. Here, however, we assume that the underlying population variances are *unknown*, and so our CIs will involve t-distribution quantiles instead of normal quantiles.

Setup: Suppose we have samples of sizes n and m from the two populations,

$$\begin{aligned} X_1, X_2, \ldots, X_n &\overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma_x^2) \quad \text{(population 1), and} \\ Y_1, Y_2, \ldots, Y_m &\overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma_y^2) \quad \text{(population 2).} \end{aligned}$$

We assume that the means μ_x and μ_y are *unknown*, and the variances σ_x^2 and σ_y^2 are also *unknown*.

Goal: Find a CI for the difference in means, $\mu_x - \mu_y$.

Our discussion will be broken up into three cases in the subsequent subsections, each involving different assumptions.

- (§6.5.1) The random samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ are themselves independent of each other, and the unknown variances are *equal*, i.e., $\sigma_x^2 = \sigma_y^2 = \sigma^2$. This is a slightly unrealistic case, but can be useful when there is evidence that the variances are at least approximately equal.
- (§6.5.2) The random samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_m$ are themselves independent of each other, but the unknown variances are arbitrary (and so may not be equal), i.e., $\sigma_x^2 \neq \sigma_y^2$. This is perhaps the most-realistic scenario, and the most popular for the purpose of comparing means.
- (§6.5.3) The random samples $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are of the same size, and are not necessarily independent of each other. In particular, the two components within any particular *pair* (X_i, Y_i) may be correlated, i.e., $\text{Corr}(X_i, Y_i) \neq 0$. Further, the unknown variances are arbitrary. This setup can be used to great advantage — at least in scenarios with significant positive correlation within each pair.

In each case, we'll make use of the usual sample means,

$$\bar{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \bar{Y} \equiv \frac{1}{m}\sum_{i=1}^{m} Y_i.$$

Now, on to the three cases...

6.5.1 Variances Unknown but Equal

Setup: We assume that the two samples $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma^2)$ and $Y_1, Y_2, \ldots, Y_m \overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma^2)$ are independent of each other, but have the *same* unknown variance σ^2.

Example: We compare the means of two industrial weight scales, both of which have the same variation, but which may be centered at different points. □

Having already defined the sample means, let's remind ourselves of the sample variances,

$$\begin{aligned} S_x^2 &\equiv \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \frac{\sigma^2\chi^2(n-1)}{n-1} \quad \text{and} \\ S_y^2 &\equiv \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2 \sim \frac{\sigma^2\chi^2(m-1)}{m-1}, \end{aligned}$$

where the distributional results follow by Equation (6.1). Both S_x^2 and S_y^2 are estimators of the *common variance* σ^2. A better estimator is the **pooled** estimator of σ^2, which linearly interpolates the information from *both* S_x^2 and S_y^2,

$$S_p^2 \equiv \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}.$$

Theorem:

$$S_p^2 \sim \frac{\sigma^2\chi^2(n+m-2)}{n+m-2}.$$

Proof: The result follows immediately from the fact that independent χ^2 random variables add up. □

After some of the usual algebra (and the sneaky fact that $\bar{X} - \bar{Y}$ is independent of S_p^2), we see that the pivot

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}}}{\sqrt{S_p^2/\sigma^2}} \sim \frac{\text{Nor}(0,1)}{\sqrt{\frac{\chi^2(n+m-2)}{n+m-2}}} \sim t(n+m-2),$$

which has more degrees of freedom than a pivot from either $\bar{X}$ or $\bar{Y}$ separately — more df is good!

So, after the usual additional algebra, we get

$100(1-\alpha)\%$ **two-sided pooled CI for** $\mu_x - \mu_y$**:**

$$\mu_x - \mu_y \in \bar{X} - \bar{Y} \pm t_{\alpha/2,n+m-2}\, S_p\sqrt{\frac{1}{n} + \frac{1}{m}},$$

where $t_{\alpha/2,n+m-2}$ is the appropriate t-distribution quantile.

One-Sided CIs: If we let $A \equiv t_{\alpha,n+m-2}\, S_p \sqrt{\frac{1}{n}+\frac{1}{m}}$, then the one-sided lower CI is $\mu_x - \mu_y \geq \bar{X} - \bar{Y} - A$, and the one-sided upper CI is $\mu_x - \mu_y \leq \bar{X} - \bar{Y} + A$.

Example: Let's consider IQ's of students at Tech School of Technology and Clever University, where we'll assume a common but unknown variance σ^2.

TST students: $X_1, \ldots, X_{25} \overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma^2)$.

CU students: $Y_1, \ldots, Y_{36} \overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma^2)$.

Suppose it turns out that

$$\bar{X} = 120, \quad \bar{Y} = 80, \quad S_x^2 = 100, \quad \text{and} \quad S_y^2 = 95.$$

The two sample variances are pretty close, so we'll go ahead and feel good about our common σ^2 assumption, and we'll use the pooled variance estimator,

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2} = \frac{(24)(100)+(35)(95)}{59} = 97.03.$$

Thus, since $t_{0.025,59} = 2.00$, we see that our two-sided 95% CI for $\mu_x - \mu_y$ is

$$\begin{aligned} \mu_x - \mu_y &\in \bar{X} - \bar{Y} \pm t_{\alpha/2,n+m-2}\, S_p \sqrt{\frac{1}{n}+\frac{1}{m}} \\ &= 120 - 80 \pm 2.00\sqrt{97.03}\sqrt{0.0678} \\ &= 40 \pm 5.13, \end{aligned}$$

which can be rewritten as $34.87 \leq \mu_x - \mu_y \leq 45.13$.

The above CI doesn't contain 0 (not even close), making it all the more obvious that TST students are somewhat more astute than CU students. □

6.5.2 Variances Unknown and Unequal

We'll again compare the means from two competing normal populations, but now they might have *unequal* variances — which seems to be the most-realistic case.

Setup: Here we assume that the samples $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma_x^2)$ and $Y_1, Y_2, \ldots, Y_m \overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma_y^2)$ are independent of each other, but that they might have *different* unknown variances $\sigma_x^2 \neq \sigma_y^2$.

As before, start by calculating sample means and variances, $\bar{X}$, $\bar{Y}$, S_x^2, and S_y^2. Since the variances are possibly unequal, we can't use the pooled estimator S_p^2. Instead, we use what is known as Welch's approximation method,

$$t^\star \equiv \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}} \approx t(\nu),$$

where the *approximate* degrees of freedom is given by the formidable expression

$$\nu \;\equiv\; \frac{\left(\frac{S_x^2}{n} + \frac{S_y^2}{m}\right)^2}{\frac{(S_x^2/n)^2}{n-1} + \frac{(S_y^2/m)^2}{m-1}}.$$

After the usual pivot-related algebra, we obtain the following CI that has very wide application in practice.

Approximate $100(1-\alpha)\%$ **two-sided Welch CI for** $\mu_x - \mu_y$**:**

$$\mu_x - \mu_y \;\in\; \bar{X} - \bar{Y} \pm t_{\alpha/2,\nu} \sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}.$$

One-sided CIs: Define the quantity $B \equiv t_{\alpha,\nu} \sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}$. Then a one-sided lower CI is $\mu_x - \mu_y \geq \bar{X} - \bar{Y} - B$, and a one-sided upper CI is $\mu_x - \mu_y \leq \bar{X} - \bar{Y} + B$.

Example: Here are some times (in minutes) for people to solve a certain type of test problem. The X_i's denote folks who attempted to solve the problem *without training*, and the Y_i's denote people who *received training* and then did the problem. The X_i's and Y_i's assumed to be normal with respective unknown means μ_x and μ_y, and are from completely different sets of people (so a total of $5+5 = 10$ independent participants have been used in the study).

Untrained person X_i	Trained person Y_i
20	21
40	6
5	31
35	1
20	16

We easily calculate the following sample means and sample variances.

$$\bar{X} \;=\; 24, \quad \bar{Y} \;=\; 15, \quad S_x^2 \;=\; 192.5, \quad S_y^2 \;=\; 142.5.$$

Then we have

$$\nu \;=\; \frac{\left(\frac{S_x^2}{n} + \frac{S_y^2}{n}\right)^2}{\frac{(S_x^2/n)^2}{n-1} + \frac{(S_y^2/n)^2}{n-1}} \;=\; \frac{4\big(192.5 + 142.5\big)^2}{(192.5)^2 + (142.5)^2} \;=\; 7.83 \;\doteq\; 8,$$

where we've rounded up in this case because it was close (sometimes we round down to be conservative). This gives us the following 90% approximate Welch CI for $\mu_x - \mu_y$,

$$\mu_x - \mu_y \;\in\; \bar{X} - \bar{Y} \pm t_{0.05,8} \sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{n}} \;=\; 9 \pm 1.860\sqrt{67} \;=\; 9 \pm 15.22.$$

So even though the trained people take 9 minutes less (on average), the confidence interval is so wide (± 15.22) that it contains 0, which informally means that we can't really be sure that training actually helps. ☹

I wonder if we can do any better...?

6.5.3 Paired Observations

...Yes, sometimes we can do better than Welch!

Setup: Again consider two competing normal populations. Suppose now we collect observations from the two populations in *pairs*. The random variables between *different* pairs are *independent*. But the two observations within the *same* pair may *not* be independent.

$$\text{independent} \left\{ \begin{array}{ll} \text{Pair } 1: & (X_1, Y_1) \\ \text{Pair } 2: & (X_2, Y_2) \\ \vdots & \vdots \\ \text{Pair } n: & \underbrace{(X_n, Y_n)}_{\text{might not be indep.}} \end{array} \right.$$

In particular, pair (X_i, Y_i) is independent of pair (X_j, Y_j) for all $i \neq j$, but within any pair i, X_i and Y_i might not be independent.

Example: Think of sets of twins. One twin takes a new drug, the other takes a placebo. The reactions of twins Sal and Sally will likely have some correlation, but they will be independent of Dennis and Denise's reactions.

Idea: By setting up such experiments, we hope to be able to capture the difference between the two normal populations more precisely, since we're using the pairs to eliminate extraneous noise.

More Details: Take n pairs of observations, (X_i, Y_i), $i = 1, 2, \ldots, n$, such that

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma_x^2) \quad \text{and} \quad Y_1, Y_2, \ldots, Y_n \overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma_y^2).$$

We assume that the means μ_x and μ_y are *unknown*, and the variances σ_x^2 and σ_y^2 are also *unknown* and possibly *unequal*.

Further, pair i is independent of pair j (*between* pairs), but X_i may not be independent of Y_i (*within* a pair). (An additional technical requirement is that all (X_i, Y_i) pairs are bivariate normal.)

Goal: Find a CI for the difference in means, $\mu_x - \mu_y$.

To do so, define the pair-wise differences,

$$D_i \equiv X_i - Y_i, \quad i = 1, 2, \ldots, n,$$

so that

$$D_1, D_2, \ldots, D_n \overset{\text{iid}}{\sim} \text{Nor}(\mu_d, \sigma_d^2),$$

where $\mu_d \equiv \mu_x - \mu_y$ (which is what we want the CI for), and

$$\sigma_d^2 \equiv \sigma_x^2 + \sigma_y^2 - 2\text{Cov}(X_i, Y_i).$$

Now the problem reduces to the old $\text{Nor}(\mu, \sigma^2)$ case with unknown μ and σ^2. So let's calculate the sample mean and variance as before:

$$\bar{D} \equiv \frac{1}{n}\sum_{i=1}^{n} D_i \sim \text{Nor}(\mu_d, \sigma_d^2/n) \quad \text{and}$$

$$S_d^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2 \sim \frac{\sigma_d^2 \chi^2(n-1)}{n-1}.$$

And just like before, we get

$$\frac{\bar{D} - \mu_d}{\sqrt{S_d^2/n}} \sim t(n-1).$$

Then after the usual algebra, we obtain

$100(1-\alpha)\%$ **two-sided paired CI for $\mu_x - \mu_y$:**

$$\mu_d = \mu_x - \mu_y \in \bar{D} \pm t_{\alpha/2, n-1}\sqrt{S_d^2/n}.$$

One-sided CIs: Define the quantity $C \equiv t_{\alpha, n-1}\sqrt{S_d^2/n}$. Then a one-sided lower CI is $\mu_d \geq \bar{D} - C$, and a one-sided upper CI is $\mu_d \leq \bar{D} + C$.

The purpose of the following example is to compare the paired-t method against the "usual" Welch CI from §6.5.2, when both are used as intended.

Example: Let's revisit the example from §6.5.2 in which we are interested in problem-solving times for people who have not received training vs. those who have. Now we will have an untrained person work the problem, then train that person, and then have that same person do a similar problem after having received the training. So instead of conducting the experiment with two completely different sets of five people (for a total of ten people), we use the *same* five folks on both the "before" and "after" components of the experiment. Here are the results of that sampling scheme.

Person	Before Training X_i	After Training Y_i	Difference D_i
1	20	10	10
2	40	25	15
3	5	5	0
4	35	20	15
5	20	15	5

Though the people are independent, the before and after times for the same individual to solve the test problem are probably correlated. (For instance, a sharp person is likely to be relatively fast before and after the training, whereas a dull person..., well, you know.)

Let's assume that all times are normal. We want a 90% two-sided CI for $\mu_d = \mu_b - \mu_a$. We see that $n = 5$, $\bar{D} = 9$, and $S_d^2 = 42.5$.

(Note that we "rigged" the numbers here so that we have the same sample mean of the differences as in the analogous example from §6.5.2, namely, $\bar{X} - \bar{Y} = 9$ in both cases. This will allow us to make a fair, apples-to-apples comparison of the Welch vs. paired-t methods.)

In any case, we find that the 90% two-sided paired-t CI is

$$\mu_d \in \bar{D} \pm t_{0.05,4} \sqrt{S_d^2/n} = 9 \pm 2.13 \sqrt{42.5/5} = 9 \pm 6.21. \quad \square$$

Remarks:

- The approximate df ν from the Welch method is larger than the df $n-1$ from the paired-t method. Larger df tends to make the CI shorter. This works in favor of Welch.
- The paired-t CI tends to remove a certain amount of natural noise by intentionally inducing positive correlation between X_i and Y_i. This works in favor of paired-t.
- In the current example, the paired-t confidence interval's width (± 6.21) is *much shorter* than that of the Welch CI from §6.5.2 (± 15.22). Evidently, in this example, the reduction of natural noise more than balances out the difference in the two df's.
- In fact, the paired-t CI does not contain 0, which informally indicates that the training actually has a significant effect on problem-solving time — a conclusion that could not have been reached via the Welch CI results.

Moral: Use paired-t when you can.

6.6 Confidence Interval for Normal Variance

To address how much variability we can expect from some system, we'll now obtain a confidence interval for the variance σ^2 of a normal distribution (instead of the mean).

Setup: $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$.

Recall that the distribution of the sample variance is

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \frac{\sigma^2 \chi^2(n-1)}{n-1}.$$

This implies that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

so that some algebra and appropriate χ^2 quantiles (see Table B.3) yield

$$\begin{aligned} 1-\alpha &= \mathrm{P}\left(\chi^2_{1-\frac{\alpha}{2},n-1} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{\frac{\alpha}{2},n-1}\right) \\ &= \mathrm{P}\left(\frac{1}{\chi^2_{1-\frac{\alpha}{2},n-1}} \ge \frac{\sigma^2}{(n-1)S^2} \ge \frac{1}{\chi^2_{\frac{\alpha}{2},n-1}}\right) \\ &= \mathrm{P}\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right). \end{aligned}$$

Thus, we have

$100(1-\alpha)\%$ **two-sided CI for σ^2:**

$$\sigma^2 \in \left[\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right].$$

Remark: The CI for σ^2 is directly proportional to the sample variance S^2.

Remark: This CI contains no reference to the unknown μ!

One-Sided CIs: $100(1-\alpha)\%$ lower and upper CIs for σ^2 are, respectively,

$$\sigma^2 \ge \frac{(n-1)S^2}{\chi^2_{\alpha,n-1}} \quad \text{and} \quad \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}}.$$

Example: Once again, consider our data set from §6.2, listing the weights of $n = 25$ football players. Recall that the sample variance was $S^2 = 479.65$. Here's a 95% upper CI for the variance σ^2.

$$\sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}} = \frac{(24)(479.65)}{\chi^2_{0.95,24}} = \frac{11512}{13.85} = 831.2. \quad \square$$

6.7 Confidence Interval for Ratio of Normal Variances

We investigate which of two normal distributions is less / more variable, e.g., two types of industrial scales.

Setup: $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma_x^2)$ and $Y_1, Y_2, \ldots, Y_m \overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma_y^2)$, where all X's and Y's are independent with unknown means and variances.

In order to compare the variability of the two distributions, we'll get a CI for the *ratio* σ_x^2/σ_y^2.

Recall the distributions of the two sample variances:

$$S_x^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \;\sim\; \frac{\sigma_x^2\,\chi^2(n-1)}{n-1}, \quad \text{and}$$

$$S_y^2 \;=\; \frac{1}{m-1}\sum_{i=1}^{m}(Y_i-\bar{Y})^2 \;\sim\; \frac{\sigma_y^2\,\chi^2(m-1)}{m-1}.$$

Thus,

$$\frac{S_y^2/\sigma_y^2}{S_x^2/\sigma_x^2} \;\sim\; \frac{\chi^2(m-1)/(m-1)}{\chi^2(n-1)/(n-1)} \;\sim\; F(m-1,n-1).$$

Using the little trick that turns left-tail F quantiles into right-tail quantiles, and then some algebra, we get

$$\begin{aligned} 1-\alpha &= \mathrm{P}\left(F_{1-\frac{\alpha}{2},m-1,n-1} \le \frac{S_y^2/\sigma_y^2}{S_x^2/\sigma_x^2} \le F_{\frac{\alpha}{2},m-1,n-1}\right) \\ &= \mathrm{P}\left(\frac{S_x^2}{S_y^2}\,\frac{1}{F_{\frac{\alpha}{2},n-1,m-1}} \le \frac{\sigma_x^2}{\sigma_y^2} \le \frac{S_x^2}{S_y^2}\,F_{\frac{\alpha}{2},m-1,n-1}\right). \end{aligned}$$

So we have

$100(1-\alpha)\%$ **two-sided CI for σ_x^2/σ_y^2:**

$$\frac{\sigma_x^2}{\sigma_y^2} \;\in\; \left[\frac{S_x^2}{S_y^2}\,\frac{1}{F_{\frac{\alpha}{2},n-1,m-1}},\;\; \frac{S_x^2}{S_y^2}\,F_{\frac{\alpha}{2},m-1,n-1}\right].$$

Remark: The CI for σ_x^2/σ_y^2 is proportional to the ratio of the sample variances, S_x^2/S_y^2, and contains no reference to μ_x or μ_y.

One-Sided CIs: $100(1-\alpha)\%$ lower and upper CIs for σ_x^2/σ_y^2 are, respectively,

$$\frac{\sigma_x^2}{\sigma_y^2} \;\ge\; \frac{S_x^2}{S_y^2}\,\frac{1}{F_{\alpha,n-1,m-1}} \quad \text{and} \quad \frac{\sigma_x^2}{\sigma_y^2} \;\le\; \frac{S_x^2}{S_y^2}\,F_{\alpha,m-1,n-1}.$$

Remark: If you want CIs for σ_y^2/σ_x^2, just flip all of the X's and Y's in the various CIs discussed above.

Example: Suppose 25 people take test A, and 16 people take test B. Assume all scores are normal and independent.

If $S_{\mathrm{A}}^2 = 100$ and $S_{\mathrm{B}}^2 = 70$, let's find a 95% upper CI for the ratio $\sigma_{\mathrm{A}}^2/\sigma_{\mathrm{B}}^2$. Looking up the $F_{\alpha,n_{\mathrm{B}}-1,n_{\mathrm{A}}-1} = F_{0.05,15,24}$ quantile (see Tables B.4–B.7), we obtain

$$\frac{\sigma_{\mathrm{A}}^2}{\sigma_{\mathrm{B}}^2} \;\le\; \frac{S_{\mathrm{A}}^2}{S_{\mathrm{B}}^2}\,F_{\alpha,n_{\mathrm{B}}-1,n_{\mathrm{A}}-1} \;=\; \frac{100}{70}\,(2.11) \;=\; 3.01. \quad \square$$

6.8 Confidence Interval for Bernoulli Success Probability

Goal: Obtain a confidence interval for the Bernoulli success probability p.

Examples: There are plenty of practical problems to motivate this material.

- We are 95% sure that the President's popularity is somewhere in the range of $53\% \pm 2.5\%$.
- We have 90% confidence that the proportion of defective items produced by a manufacturer is $2.3\% \pm 0.2\%$.

Setup: Suppose that $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$.

Since $\sum_{i=1}^n X_i \sim \text{Bin}(n,p)$, we know that $\bar{X} \sim \frac{1}{n}\text{Bin}(n,p)$. Let's assume that n is "large," so that we'll be able to use the Central Limit Theorem to approximate the binomial distribution by the normal — and don't worry about the associated "continuity correction." By "large," we mean that $np \geq 5$ and $n(1-p) \geq 5$, whatever we think p might happen to be — maybe make a preliminary informed guess about p. (If n isn't large, then we'll have to use nasty binomial tables, which we don't want to deal with here!)

Note that

$$\text{E}[\bar{X}] \;=\; \text{E}[X_i] \;=\; p \quad \text{and} \quad \text{Var}(\bar{X}) \;=\; \text{Var}(X_i)/n \;=\; pq/n.$$

Then for large n, the CLT implies

$$\frac{\bar{X} - \text{E}[\bar{X}]}{\sqrt{\text{Var}(\bar{X})}} \;=\; \frac{\bar{X} - p}{\sqrt{pq/n}} \;\approx\; \text{Nor}(0,1).$$

Now let's do something crazy and estimate pq by its maximum likelihood estimator, $\bar{X}(1-\bar{X})$. This gives

$$\frac{\bar{X} - p}{\sqrt{\bar{X}(1-\bar{X})/n}} \;\approx\; \text{Nor}(0,1).$$

Then the usual algebra implies that

$$\begin{aligned} 1-\alpha \;&\doteq\; \text{P}\left(-z_{\alpha/2} \leq \frac{\bar{X} - p}{\sqrt{\bar{X}(1-\bar{X})/n}} \leq z_{\alpha/2}\right) \\ &=\; \text{P}\left(\bar{X} - z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leq p \leq \bar{X} + z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right). \end{aligned}$$

This results in

Approximate $100(1-\alpha)\%$ two-sided CI for p:

$$p \in \bar{X} \pm z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}.$$

One-Sided CIs: Approximate $100(1-\alpha)\%$ lower and upper CIs for p are, respectively,

$$p \geq \bar{X} - z_\alpha\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \quad \text{and} \quad p \leq \bar{X} + z_\alpha\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}.$$

Example: The probability that a student correctly answers a certain test question is p. Suppose that a random sample of 200 students yields 160 correct answers to the question. A 95% two-sided CI for p is given by

$$p \in \bar{X} \pm z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} = 0.8 \pm 1.96\sqrt{\frac{(0.8)(0.2)}{200}} = 0.8 \pm 0.055. \quad \square$$

Sample-Size Calculation: The half-width of the two-sided CI is

$$z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}.$$

How many observations should we take so that the half-length is $\leq \epsilon$?

$$z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \leq \epsilon \iff n \geq (z_{\alpha/2}/\epsilon)^2\bar{X}(1-\bar{X}).$$

Of course, $\bar{X}$ is unknown before taking observations. A *conservative* choice for n arises by maximizing $\bar{X}(1-\bar{X}) = 1/4$. Then we have

$$n \geq z^2_{\alpha/2}/(4\epsilon^2).$$

On the other hand, if we can somehow make a *preliminary estimate* $\hat{p}$ of p (based on a preliminary sample mean from a pilot study, for instance), we could use

$$n \geq z^2_{\alpha/2}\,\hat{p}(1-\hat{p})/\epsilon^2.$$

Example: Suppose we want to take enough samples so that a 95% two-sided CI for p will be no bigger than $\pm 3\%$. How many observations will we need to take?

$$n \geq z^2_{\alpha/2}/(4\epsilon^2) = \frac{(1.96)^2}{4\,(0.03)^2} = 1067.1.$$

So take $n = 1067$ samples.

Now let's instead suppose that a pilot study gives an estimate of $\hat{p} = 0.65$. How should n be revised?

$$n \geq z^2_{\alpha/2}\,\hat{p}(1-\hat{p})/\epsilon^2 = \frac{(1.96)^2}{(0.03)^2}(0.65)(0.35) = 971.1.$$

So now we only have to take $n = 971$. □

Remark: We can also derive approximate CIs for the difference between two competing Bernoulli parameters. For instance, suppose that $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p_x)$ and $Y_1, Y_2, \ldots, Y_m \overset{\text{iid}}{\sim} \text{Bern}(p_y)$, where the two samples are independent of each other. Then, using the techniques already developed in this section, we can obtain in a straightforward manner the following CI for the difference of the success proportions.

Approximate $100(1-\alpha)\%$ **two-sided CI for** $p_x - p_y$**:**

$$p_x - p_y \;\in\; \bar{X} - \bar{Y} \pm z_{\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n} + \frac{\bar{Y}(1-\bar{Y})}{m}}.$$

Example: The probabilities that Tech School of Technology and Universal College University students earn at least a 700 on the Math portion of the SAT test are $p_{\text{☺}}$ and $p_{\text{☹}}$, respectively. A random sample of 200 TST students yielded 160 students who scored ≥ 700 on the test. But, sadly, a sample of 400 UCU students revealed only 50 who succeeded. Then a 95% CI for the difference in success probabilities is

$$\begin{aligned} p_{\text{☺}} - p_{\text{☹}} \;&\in\; \bar{X} - \bar{Y} \pm z_{\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n} + \frac{\bar{Y}(1-\bar{Y})}{m}} \\ &=\; 0.8 - 0.125 \pm 1.96 \sqrt{\frac{(0.8)(0.2)}{200} + \frac{(0.125)(0.875)}{400}} \\ &=\; 0.675 \pm 0.064. \end{aligned}$$

Not looking real good for UCU. □

6.9 Confidence Intervals Based on Maximum Likelihood Estimators

Recall from §5.2.5.4 that the MLE $\hat{\theta}_n$ for an unknown parameter θ is apt to be asymptotically normal as the sample size n becomes large. This result provides a basis for us to obtain approximate confidence intervals for a rich class of parameters. To proceed, let $I_n(\theta)$ denote the associated Fisher information from Equations (5.2) and (5.3). Under the various assumptions alluded to in §5.2.4, we have

$$\sqrt{I_n(\theta)}\,\big(\hat{\theta}_n - \theta\big) \;\xrightarrow{d}\; \text{Nor}(0,1) \quad \text{as } n \to \infty.$$

This implies that $\hat{\theta}_n \approx \text{Nor}(\theta, 1/I_n(\theta))$, and so by the usual algebra,

$$\text{P}\left(\theta \;\in\; \hat{\theta}_n \pm \frac{z_{\alpha/2}}{\sqrt{I_n(\theta)}}\right) \;\doteq\; 1 - \alpha.$$

Since θ is unknown, the Fisher information $I_n(\theta)$ will itself be unknown, in which case we substitute $\hat{\theta}_n$ in place of θ. To summarize,

MLE-based approximate $100(1-\alpha)\%$ two-sided CI for θ:

$$\theta \in \hat{\theta}_n \pm \frac{z_{\alpha/2}}{\sqrt{I_n(\hat{\theta}_n)}}.$$

Example: If $X_1, X_2, \ldots, X_n$ are iid Exp(λ), we recall that the MLE is $\hat{\lambda}_n = 1/\bar{X}$ and the Fisher information is $I_n(\lambda) = n/\lambda^2$, so that $I(\hat{\lambda}_n) = n\bar{X}^2$. Thus, the MLE-based approximate two-sided CI for λ is given by

$$\lambda \in \hat{\lambda}_n \pm \frac{z_{\alpha/2}}{\sqrt{I_n(\hat{\lambda}_n)}} = \frac{1}{\bar{X}}\Big(1 \pm \frac{z_{\alpha/2}}{\sqrt{n}}\Big),$$

i.e.,

$$\mathrm{P}\bigg[\frac{1}{\bar{X}}\Big(1 - \frac{z_{\alpha/2}}{\sqrt{n}}\Big) < \lambda < \frac{1}{\bar{X}}\Big(1 + \frac{z_{\alpha/2}}{\sqrt{n}}\Big)\bigg] \doteq 1 - \alpha.$$

We can also derive CIs for more-complicated quantities such as the survival function, $\mathrm{P}(X > t) = e^{-\lambda t}$. In that case, we simply exponentiate and make sure to be a little careful when flipping the limits because of the negative power,

$$\mathrm{P}\bigg[\exp\Big\{-\frac{t}{\bar{X}}\Big(1 + \frac{z_{\alpha/2}}{\sqrt{n}}\Big)\Big\} < e^{-\lambda t} < \exp\Big\{-\frac{t}{\bar{X}}\Big(1 - \frac{z_{\alpha/2}}{\sqrt{n}}\Big)\Big\}\bigg] \doteq 1 - \alpha. \quad (6.5)$$

For example, if we have on hand a sample of size $n = 100$ yielding a sample mean of $\bar{X} = 25$, then we obtain the following 95% approximate CIs.

$$\lambda \in \bigg[\frac{1}{25}\Big(1 - \frac{1.96}{\sqrt{100}}\Big), \frac{1}{25}\Big(1 + \frac{1.96}{\sqrt{100}}\Big)\bigg] = [0.0322, 0.0478],$$

and

$$\begin{aligned}\mathrm{P}(X > 50) &\in \bigg[\exp\Big\{-\frac{50}{25}\Big(1 + \frac{1.96}{\sqrt{100}}\Big)\Big\}, \exp\Big\{-\frac{50}{25}\Big(1 - \frac{1.96}{\sqrt{100}}\Big)\Big\}\bigg] \\ &= [0.0914, 0.2003]. \quad \square\end{aligned}$$

6.10 Exercises

1. (§6.1) Consider the following 95% confidence interval for the unknown mean of a certain population:

$$\mu \in \bar{X} \pm z_{0.025}\sqrt{\sigma^2/n} = -5 \pm 15.$$

Which of the following statements are true?

(a) We are 95% sure that μ is somewhere between -20 and 10.

(b) If we instead wanted a 99% confidence interval for μ, it would be wider.

(c) The corresponding 95% confidence interval for 2μ is -10 ± 30.

(d) If I collect 100 sets of data and calculate the corresponding 100 confidence intervals, then exactly 95 of them will contain μ.

2. (§6.2) The amount of time that it takes to drive from Atlanta to San Francisco is approximately normally distributed, with a standard deviation of $\sigma = 25$ hours.

   (a) Suppose that 20 people each make an iid trip, and the sample mean of the times it takes them is $\bar{X} = 100$ hours. Construct a 95% two-sided confidence interval for the mean trip time.

   (b) Suppose we wanted to be 95% confident that the error in estimating the mean trip time is less than 2 hours. What sample size should be used?

3. (§6.3) I love corn on the cob! Let's study the lengths of ears of two types of corn — white and yellow.[1] Suppose we (somehow) know that the lengths of both types have approximately the same standard deviation, $\sigma_w = \sigma_y = 4$ cm. Two random samples of sizes $n_w = 30$ and $n_y = 20$, respectively, are tested, and the sample mean lengths are $\bar{W} = 22$ and $\bar{Y} = 26$ cm. Construct a 95% confidence interval on the mean difference in lengths.

4. (§6.4) Suppose we collect the following observations: 7, −2, 1, 6. Let's assume that these guys are iid from a normal distribution with *unknown* variance σ^2. Give me a two-sided 95% confidence interval for the mean μ.

5. (§6.4) The weights of 16 Atlanta Falcons football players are given below (in pounds, not kilos).

225	301	281	263
216	237	249	204
250	238	300	217
318	255	275	295

   Construct a 95% two-sided confidence interval for the mean weight.

6. (§6.4) Consider the following data set of $n = 15$ observations, conveniently presented in numerical order.

−123	−112	−100	−84	−83
−61	−7	17	20	21
26	33	33	43	80

   Construct a one-sided 95% *upper* confidence interval for the mean.

7. (§6.4) We love to vacation at our palatial cabin in the Georgia mountains. We've made 25 iid trips over the years. On average, it's taken 6.10 gallons of gas, with a sample standard deviation of 0.16 gallons. Find a 90% lower confidence interval for the mean number of gallons per trip.

[1] How many kernels are there on a typical ear of corn? 800! That makes a lot of popcorn!

8. (§6.4) Suppose that $\mu \in [-30, 90]$ is a 90% confidence interval for the mean yearly cost incurred by a certain inventory policy. Further suppose that this interval was based on four iid observations of the underlying inventory system, which we are assuming to be normal, but with unknown variance. Unfortunately, the boss has decided that she wants a 95% confidence interval. Can you supply it?

9. (§6.4) Consider the usual two-sided confidence interval for the normal mean when the variance is unknown,

$$\mu \in \bar{X} \pm t_{\alpha/2,n-1}\sqrt{S^2/n}.$$

Let $H = t_{\alpha/2,n-1}\sqrt{S^2/n}$ denote the *half-width* of this interval. It can be shown that

$$H \sim \sigma\, t_{\alpha/2,n-1}\frac{\chi(n-1)}{\sqrt{n(n-1)}} \quad \text{(the } \textbf{chi distribution!}\text{)}.$$

Then this, in turn, leads to the result

$$\mathrm{E}[H] = \sigma\, t_{\alpha/2,n-1}\sqrt{\frac{2}{n(n-1)}}\,\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})},$$

where $\Gamma(\cdot)$ is the gamma function and has the well-known property that $\Gamma(a) = (a-1)\Gamma(a-1)$ for $a > 1$. In particular, after some algebra (including the fact that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$), it turns out that if $n \geq 4$ is *even*, then the expected half-length is proportional to

$$\zeta(n) \equiv \frac{t_{\alpha/2,n-1}}{\sqrt{n(n-1)}}\,\frac{(\frac{n-2}{2})!}{(\frac{n-3}{2})\,(\frac{n-5}{2})\cdots\frac{1}{2}}.$$

Using the above equation with $\alpha = 0.05$, determine which of $n = 4, 6, 8$ gives the smallest expected width.

10. (§6.5.1) Random samples of size 20 were taken from two independent normal populations. The sample means and standard deviations were $\bar{X} = 23.10$, $S_x^2 = 12.96$, $\bar{Y} = 20.00$, and $S_y^2 = 12.00$. Assuming that $\sigma_x^2 = \sigma_y^2$, find a 95% two-sided confidence interval for $\mu_x - \mu_y$.

11. (§6.5.2) Consider two normal populations for which it turns out that

$$\begin{array}{llll} n = 25 & \bar{X} = 100 & S_x^2 = 400 \\ m = 16 & \bar{Y} = 80 & S_y^2 = 100. \end{array}$$

You can tell from S_x^2 and S_y^2 that there's no way the two variances are equal. Obtain a two-sided 95% CI for $\mu_x - \mu_y$.

12. (§6.5.2) We are studying the times required to solve two elementary math problems. Suppose we ask four students to attempt Problem A, and four other students to attempt Problem B. Assume everything is independent and normal. The results are presented below (in seconds).

Problem A	Problem B
20	35
30	20
15	50
40	40

Assuming that the *variances are unequal*, find a two-sided 95% confidence interval for the difference in the means between Problems A and B.

13. (§6.5.3) We are again studying the times required to solve two elementary math problems. Suppose we ask four students to attempt *both* Problem A and Problem B. Assume the students are independent and all results are normally distributed, but note that a particular student's times on the two questions are likely to be positively correlated. These results are presented below (in seconds).

Student	Problem A	Problem B
1	20	35
2	30	40
3	15	20
4	40	50

Again find a two-sided 95% CI for the difference in the means of A and B.

14. (§6.5.3) Suppose that we want to see if a new drug lowers cholesterol levels in people. We administer the new drug to four people, and we obtain the following "before" and "after" results on their cholesterol levels (CL).

Person	"Before" CL	"After" CL
1	240	190
2	215	170
3	190	165
4	320	250

It certainly looks like the drug helps to lower the CL a bit. Let's assume that all of the above numbers are normal and that the four test subjects are independent of each other. Find a two-sided 95% confidence interval for the difference in the "before" and "after" means.

15. (§6.6) Suppose $X_1, X_2, \ldots, X_6$ are iid normal with unknown mean and unknown variance σ^2. Further suppose that $\bar{X} = 1000$ and $S^2 = 25$. Find a 90% two-sided confidence interval for σ^2.

16. (§6.6) Consider iid normal observations $X_1, X_2, \ldots, X_7$ with unknown mean μ and unknown variance σ^2. What is the *expected width* of the usual 90% two-sided confidence interval for σ^2? You can keep your answer in terms of σ^2.

17. (§6.7]) Suppose $X_1, X_2, \ldots, X_{20}$ are iid $\text{Nor}(\mu_x, \sigma_x^2)$, $Y_1, Y_2, \ldots, Y_{20}$ are iid $\text{Nor}(\mu_y, \sigma_y^2)$, and all of the X_i's are independent of all of the Y_i's. Further suppose that $\bar{X} = 23$, $S_x^2 = 5184$, $\bar{Y} = -11$, and $S_y^2 = 900$. Find a 95% two-sided confidence interval on the ratio of the population variances, σ_x^2/σ_y^2.

18. (§6.8) A pollster asked a sample of 2000 people whether or not they were in favor of a particular proposal. Exactly 1200 people answered "yes." Find a two-sided 95% confidence interval for the percentage of the population in favor of the proposal.

19. (§6.8) Suppose that $X_1, X_2, \ldots, X_n$ are iid Bernoulli with unknown mean p, and that we have carried out a preliminary pilot investigation suggesting that $p \doteq 0.7$. How big would n have to be in order for a two-sided 95% confidence interval to have a half-length of 0.1? (Give the smallest such number.)

20. (§6.8) The probabilities that Tech School of Technology and Universal College University students can successfully integrate $\int_1^\infty e^{-x}\,dx = 1/e$ are p_x and p_y, respectively. A random sample of 800 TST students yielded 760 students who got the right answer. But, sadly, a sample of 800 UCU students revealed only 100 who succeeded. Find a 95% CI for the difference in success probabilities.

21. (§6.9) Suppose that $X_1, X_2, \ldots, X_n$ are iid continuous random variables with pdf $f(x) = \theta x^{\theta-1}$, for $0 < x < 1$ and $\theta > 1$. (This is a special case of the beta distribution.)

    (a) Find the Fisher information for θ.

    (b) Find the MLE of θ.

    (c) What is the approximate distribution of the MLE as the sample size n becomes large?

    (d) Find an approximate confidence interval for θ.

    (e) Consider the following data set of $n = 50$ randomly generated samples from a distribution having pdf $f(x) = \theta x^{\theta-1}$.

0.733	0.890	0.711	0.683	0.951	0.964	0.743	0.808	0.528	0.583
0.643	0.557	0.550	0.731	0.952	0.825	0.215	0.371	0.565	0.966
0.995	0.453	0.932	0.881	0.763	0.787	0.924	0.893	0.685	0.945
0.854	0.995	0.608	0.504	0.523	0.664	0.671	0.804	0.710	0.838
0.996	0.640	0.786	0.724	0.994	0.467	0.530	0.631	0.759	0.993

    Calculate an approximate 95% CI for θ. To save a little time, it turns out that $\prod_{i=1}^{50} x_i = 3.86406 \times 10^{-8}$.

22. Suppose that $I_n(\lambda)$ is the usual Fisher information for a parameter λ. Let $\eta = h(\lambda)$, where $h(\cdot)$ is a nice, differentiable function, so that $\lambda = h^{-1}(\eta)$. Under the mysterious assumptions alluded to in §5.2.4, it can be shown (you do not have to prove this) that the Fisher information for η is

$$\tilde{I}_n(\eta) \;=\; I_n(h^{-1}(\eta)) \left(\frac{dh^{-1}(\eta)}{d\eta} \right)^2 .$$

    For the rest of the problem, suppose $X_1, X_2, \ldots, X_n$ are iid $\text{Exp}(\lambda)$.

(a) Recall from §5.2.4 that the Fisher information for λ is $I_n(\lambda) = n/\lambda^2$. Find the Fisher information for the probability of survival past time t. i.e.,

$$\mathrm{P}(X > t) \;=\; e^{-\lambda t} \;=\; h(\lambda) \;=\; \eta.$$

(b) Recall from §5.2.5.3 that if $\hat{\lambda} = 1/\bar{X}$ is the MLE of λ, then the Invariance Property implies that the MLE of the survival probability is

$$\hat{\eta}_n \;\equiv\; e^{-\hat{\lambda}_n t} \;=\; e^{-t/\bar{X}}.$$

Find the approximate distribution of $\hat{\eta}_n$ for large n.

(c) Find an approximate confidence interval for $\eta = e^{-\lambda t}$. You may get an answer having a slightly different functional form than that of the CI for $e^{-\lambda t}$ given by Equation (6.5) — don't worry, they're both approximations.

(d) Suppose we have a sample of size $n = 100$ from which we obtain a sample mean of $\bar{X} = 25$. Calculate the 95% CI for $\eta = e^{-50\lambda}$ using the method from part (c).

Chapter 7

Hypothesis Testing

Hypothesis testing is an important subject area both for students interested in doing research and for practitioners conducting real-world analyses. The idea is to pose hypotheses and test their validity in a statistically rigorous way. For instance, "Is the mean lifetime of a manufactured part at least two years?" or "Is a new drug more efficacious than the medication currently on the market?" or "Is the mean tensile strength of item A greater than that of item B?"

We introduce and motivate this terrific and useful topic in §7.1, including commentary on how to formulate hypotheses properly and how to take into account certain types of errors that can occur during the testing process (e.g., we want to avoid rejecting a hypothesis that turns out to be correct). §7.2 discusses hypothesis tests related to the means of normal distributions in the very special (and possibly unrealistic) case in which the variances of the normals are somehow *known*. §7.3 does the same, but now when the variances are *unknown* — this case has tremendous applicability in practice. §7.4 presents a number of tests for other parameters such as the variance of a normal distribution and the success probability parameter of a Bernoulli distribution. §7.5 is an extended discussion on goodness-of-fit tests, which are concerned with hypothesis tests for entire distributions, rather than just unknown parameters (e.g., is my data coming from an exponential distribution?).

7.1 Introduction to Hypothesis Testing

The goal of this chapter is to study the properties of a population by collecting data and then using that data to make sound, statistically valid conclusions about the

population. We begin with a high-level view of hypothesis testing. Details regarding the use of hypothesis testing on specific distribution / parameter scenarios will follow in subsequent sections.

7.1.1 Our General Approach

Our game plan will be as follows:

1. State a hypothesis about a population.
2. Select a test statistic to test whether or not the hypothesis is true.
3. Calculate the test statistic based on observations that we take.
4. Interpret the results — does the test statistic suggest that we *reject* or *fail to reject* the hypothesis?

§7.1.1.1 State a Hypothesis

Hypotheses are simply statements or claims about parameter values or entire distributions. In order to provide evidence for or against the truth of a claim, we perform a **hypothesis test**, comprising a **null hypothesis (H_0)** and an **alternative hypothesis (H_1)** to cover the entire parameter space. The null hypothesis sort of represents the "status quo." It's not necessarily true, but we will grudgingly stick with it until "proven" otherwise.

Example: We currently believe that the mean weight μ of a filled package of chicken is μ_0 ounces. (We specify μ_0.) But since we have our suspicions, a reasonable test is

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0.$$

This is a **two-sided test**. We will reject the belief of the null hypothesis H_0 if the sample mean $\bar{X}$ (an estimator of μ) is "too big" or "too small." □

Example: We hope that a new brand of tires will last for a mean μ of more than μ_0 miles. (Again, we specify μ_0.) But we really need evidence before we can state that claim with reasonable certainty. Otherwise, we will stick with the old brand.

$$H_0 : \mu \leq \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0.$$

This is a **one-sided test**. We'll reject H_0 only if $\bar{X}$ is "too large." □

Example: We test to see if mean emissions from a certain type of car are less than μ_0 ppm. But we need evidence.

$$H_0 : \mu \geq \mu_0 \quad \text{vs.} \quad H_1 : \mu < \mu_0.$$

This is also a **one-sided test**. We'll reject the null hypothesis if $\bar{X}$ is "too small," and only then make the claim that the emissions are low. □

Idea: H_0 is the old, conservative "status quo" (think of it as "innocent until proven guilty"), while H_1 is the new, radical hypothesis, just itching to see the light of day. Although we might not be *tooooo* sure about the truth of H_0, we won't reject it in favor of H_1 unless we see substantial evidence in support of H_1. If there is insufficient evidence supporting H_1, then we "fail to reject" H_0. This roughly means that we grudgingly accept H_0.

§7.1.1.2 Select a Test Statistic

A **test statistic** is a random variable that we use to test whether or not H_0 might be true.

For instance, if we conduct a hypothesis test involving the mean μ, we would naturally compare the sample mean $\bar{X}$ with the hypothesized mean μ_0. The comparison is accomplished using a test statistic having a known sampling distribution, in this case,

$$Z_0 \;=\; \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \;\sim\; \text{Nor}(0,1) \quad (\text{if } \sigma^2 \text{ is known and } H_0 \text{ is true})$$

or

$$T_0 \;=\; \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \;\sim\; t(n-1) \quad (\text{if } \sigma^2 \text{ is unknown and } H_0 \text{ is true}).$$

We'll discuss this in much greater detail as the chapter proceeds.

§7.1.1.3 Calculate the Test Statistic

Here we collect sample data and then calculate the test statistic based on that data.

Example: Consider the time to metabolize a new drug. On average, the current drug takes $\mu_0 = 15$ min. Is the new drug faster?

Claim: The expected time μ for new drug is < 15 min. So the obvious test is

$$H_0 : \mu \geq 15 \quad \text{vs.} \quad H_1 : \mu < 15.$$

Suppose we have the following data: $n = 20$, $\bar{X} = 14.88$, and $S = 0.333$. Then the test statistic is

$$T_0 \;=\; \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \;=\; -1.61.$$

What would cause us to reject H_0? If $\bar{X} \ll \mu_0 = 15$, this would indicate that H_0 is probably wrong. Equivalently, we'd reject H_0 if T_0 is "significantly" $\ll 0$. So the question of the hour: Is $T_0 = -1.61$ small enough to reject H_0? □

§7.1.1.4 Interpret the Test Statistic

If H_0 is actually the true state of things, then $\mu = \mu_0$, and from our discussion on normal mean confidence intervals in the unknown variance case (§6.4), we have

$$T_0 \;=\; \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \;\sim\; t(n-1).$$

With this distributional result in mind, we'll examine the plausibility of the observed value of the test statistic T_0.

- If the value is deemed implausible, then we'll reject H_0 and select H_1.
- Otherwise, if the plausibility is reasonable, then we'll fail to reject H_0.

Example (continuing from previous example): If H_0 is true, then $T_0 \sim t(n-1) \sim t(19)$. So is it reasonable (or, at least, not outrageous) to have gotten $T_0 = -1.61$ from a $t(19)$ random variable?

If yes, then we will *fail to reject* ("grudgingly accept") H_0. If no, then we'll *reject* H_0 in favor of H_1.

Let's see... From the Table B.2, we have

$$t_{0.95,19} = -t_{0.05,19} = -1.729 \quad \text{and} \quad t_{0.90,19} = -t_{0.10,19} = -1.328.$$

In other words,

$$\mathrm{P}\big(t(19) < -1.729\big) \;=\; 0.05 \quad \text{and} \quad \mathrm{P}\big(t(19) < -1.328\big) \;=\; 0.10.$$

This means that

$$0.05 \;< \mathrm{P}\big(t(19) < \underbrace{-1.61}_{T_0}\big) \;<\; 0.10.$$

That is, if H_0 is true, there's only between a 5% and 10% chance that we'd see a value of $T_0 \leq -1.61$. That's not a very high probability, but it's not *soooo* small. Our reject / fail-to-reject decision will come down to what we perceive as being an implausible value for T_0 that's simply too far out in the tail to reasonably be a $t(19)$ observation.

For instance, we'd **reject** H_0 at "level" 0.10, since

$$T_0 \;=\; -1.61 \;<\; -t_{0.10,19} \;=\; -1.328.$$

But we'd **fail to reject** H_0 at "level" 0.05, since

$$T_0 \;=\; -1.61 \;>\; -t_{0.05,19} \;=\; -1.729.$$

More on this coming up next! □

7.1.2 The Errors of Our Ways

Where Can It All Go Wrong? When we conduct a hypothesis test, we typically end up either rejecting or failing to reject the null hypothesis H_0. Sometimes we make the correct decision, sometimes not. In fact, four things can happen — two good, two bad.

- If H_0 is actually true and we conclude that it's true — good! ☺
- If H_0 is actually false and we conclude that it's false — good! ☺

- If H_0 is actually true but we conclude that it's false — bad! This is called a **Type I error**. ☹

  Example: We incorrectly conclude that a new, inferior drug is better than the drug currently on the market. (Oops!) □

- If H_0 is actually false but we conclude that it's true — bad! This is called a **Type II error**. ☹

  Example: We incorrectly conclude that a new, superior drug is worse than the drug currently on the market. (Oopsie!) □

We note that Type I errors are usually considered to be "worse" than Type II errors.

	Decision	
State of nature	Accept H_0	Reject H_0
H_0 is true	Correct! ☺	Type I error ☹
H_0 is false	Type II error ☹	Correct! ☺

It is our duty to set upper bounds α and β on the respective Type I and II error probabilities that we are willing to put up with (though we will henceforth treat these bounds as equalities), i.e.,

$$\text{P(Type I error)} \;=\; \text{P(Reject } H_0 \mid H_0 \text{ true)} \;=\; \alpha$$

and

$$\text{P(Type II error)} \;=\; \text{P(Fail to Reject } H_0 \mid H_0 \text{ false)} \;=\; \beta.$$

We usually choose the quantities α and β to be pretty small (for instance, 0.01 or 0.05), and certainly such that $\alpha + \beta < 1$. But note that very small values of α and β come at a cost, e.g., a very small α makes it harder to reject H_0 even if it happens to be false.

Definition: The probability of a Type I error, α, is called the **size** or **level of significance** of the test.

Definition: The **power** of a hypothesis test is

$$1 - \beta \;=\; \text{P(Reject } H_0 \mid H_0 \text{ false)}.$$

It's good to have high power, in which case we say that the test is "powerful."

7.2 Hypothesis Tests for Normal Means (Variance Known)

In this section, we'll discuss hypothesis tests involving normal distribution means under various scenarios, all of which involve *known variance(s)*. In §7.2.1, we cover tests for the mean of a single normal distribution. §7.2.2 concerns the design of such tests in order to satisfy certain constraints on Type I and Type II errors. §7.2.3 gives us tests for comparing the means of *two* competing normal distributions.

7.2.1 One-Sample Tests

Suppose that $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$, where σ^2 is somehow *known* (this is not very realistic, but humor us for a little while).

As discussed in §7.1, a **two-sided test** (also known as a **simple test**) checks whether or not the parameter in question *equals* a particular given value. For instance, here's a two-sided test for the mean:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0,$$

where we pre-specify a hypothesized value μ_0 for the mean.

Remark: Of course, this particular simple H_0 can never be *precisely* true to an arbitrary number of decimal places. But that's not our concern — we are primarily interested in seeing if we have enough statistical significant evidence to categorically reject it.

In any case, we'll use $\bar{X}$ to estimate μ. If $\bar{X}$ is "significantly different" than μ_0 (either crazy high or crazy low), then we'll reject H_0. But how much is "significantly different"? In order to determine what "significantly different" means, we first define the **test statistic**,

$$Z_0 \;\equiv\; \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

If H_0 is true, then $\text{E}[\bar{X}] = \mu_0$ and $\text{Var}(\bar{X}) = \sigma^2/n$, and so $Z_0 \sim \text{Nor}(0, 1)$. Then we have

$$\text{P}(-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2}) \;=\; 1 - \alpha.$$

Thus, if H_0 is true, a value of Z_0 outside of the interval $[-z_{\alpha/2}, z_{\alpha/2}]$ would be highly unlikely, and certainly suspicious. So our hypothesis test decision is as follows.

Two-Sided Test for μ:

$$\text{Reject } H_0 \quad \text{iff} \quad |Z_0| > z_{\alpha/2}.$$

This assures us that

$$\begin{aligned} \text{P}(\text{Type I error}) &= \text{P}(\text{Reject } H_0 \,|\, H_0 \text{ true}) \\ &= \text{P}\big(|Z_0| > z_{\alpha/2} \,|\, Z_0 \sim \text{Nor}(0,1)\big) \\ &= \alpha. \end{aligned}$$

If $|Z_0| > z_{\alpha/2}$, then we're in the test's **rejection region** (aka the **critical region**). If $|Z_0| \leq z_{\alpha/2}$, then we're in the **acceptance region**.

A **one-sided test** relates to hypotheses in which a parameter tends to fall in a certain direction. For instance, consider the null hypothesis that the mean is *at most* μ_0,

$$H_0 : \mu \leq \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0.$$

We actually attack the problem in a way that mimics the two-sided test, but we just make our final decision a little differently. In fact, we consider the *same* test statistic $Z_0 = \sqrt{n}(\bar{X} - \mu_0)/\sigma$ as before. But now, a very large value of Z_0 outside the *one-sided* interval $(-\infty, z_\alpha]$ is highly unlikely if H_0 is true. Thus, in this case, we

$$\text{Reject } H_0 \quad \text{iff} \quad Z_0 > z_\alpha.$$

In other words, if $Z_0 > z_\alpha$, this suggests that $\mu > \mu_0$.

Similarly, for the other one-sided test, we have

$$H_0 : \mu \geq \mu_0 \quad \text{vs.} \quad H_1 : \mu < \mu_0.$$

A value of Z_0 outside the interval $[-z_\alpha, \infty)$ is highly unlikely if H_0 is true. Therefore,

$$\text{Reject } H_0 \quad \text{iff} \quad Z_0 < -z_\alpha.$$

If $Z_0 < -z_\alpha$, this suggests $\mu < \mu_0$.

Example: We examine the weights of the 25 college football players from §6.2. Suppose we somehow know that the weights are normally distributed with $\sigma^2 = 324$. In that old example, we found that the sample mean $\bar{X} = 280.0$.

We'll test the two-sided (simple) hypothesis that the mean weight is $\mu_0 = 272$, and we'll keep the probability of Type I error $= 0.05$. Then we have

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0.$$

Here we have

$$Z_0 \;=\; \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \;=\; \frac{280 - 272}{\sqrt{324/25}} \;=\; 2.22.$$

Since $|Z_0| = 2.22 > z_{\alpha/2} = z_{0.025} = 1.96$, we *reject* H_0 at level $\alpha = 0.05$.

Notice that a lower α results in a higher $z_{\alpha/2}$, in which case it would be "harder" to reject H_0. For instance, if $\alpha = 0.01$, then $z_{0.005} = 2.58$, and we would *fail to reject* H_0 for the scenario given in this example. □

Definition: The ***p*-value** of a test is the smallest level of significance α that would lead to rejection of H_0.

Remark: Researchers often report the p-values of any tests that they conduct. Doing so helps prevent unscrupulous cads from picking and choosing certain α values after the fact in order to manipulate a desired reject / accept conclusion.

Anyhow, let's derive an expression for the p-value for the two-sided test

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0.$$

We reject H_0 iff

$$|Z_0| \;>\; z_{\alpha/2} \;=\; \Phi^{-1}(1 - \alpha/2),$$

where $\Phi(x)$ is our old buddy, the Nor$(0,1)$ cdf. So we reject H_0 iff

$$\Phi(|Z_0|) \ > \ 1-\alpha/2 \quad \text{iff} \quad \alpha \ > \ 2(1-\Phi(|Z_0|)).$$

Thus, for the two-sided normal mean test in the case of known variance, the p-value is $p = 2(1-\Phi(|Z_0|))$.

Similarly, for the one-sided test,

$$H_0 : \mu \leq \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0,$$

we have $p = 1-\Phi(Z_0)$.

And for the other one-sided test,

$$H_0 : \mu \geq \mu_0 \quad \text{vs.} \quad H_1 : \mu < \mu_0,$$

it can easily be shown that $p = \Phi(Z_0)$. □

Example: For the previous two-sided numerical example,

$$p \ = \ 2(1-\Phi(|Z_0|)) \ = \ 2(1-\Phi(2.22)) \ = \ 0.0263. \quad \square$$

7.2.2 Test Design

Goal: Design a two-sided test for the known-variance normal mean test in which we impose the following constraints on the Type I and II errors:

$$\text{P(Type I error)} \ \leq \ \alpha \quad \text{and} \quad \text{P}\big(\text{Type II error} \,|\, \mu = \mu_1 > \mu_0\big) \ \leq \ \beta.$$

By "design," we simply mean to determine the number of observations n that we need for the two-sided test to satisfy a Type I error bound of α and a Type II error bound of β.

Remark: The bound β is for the *special case* that the true mean μ happens to equal a *user-specified* value $\mu = \mu_1 > \mu_0$. In other words, we're trying to "protect" ourselves against the possibility that μ actually happens to equal μ_1.

If we change the "protected" μ_1, we'll need to change n. Generally speaking, the closer μ_1 is to μ_0, the more work we'll need to do (i.e., higher n) — because it's harder to distinguish between two close cases.

Theorem: Suppose the difference between the actual and hypothesized means is

$$\delta \ \equiv \ \mu - \mu_0 \ = \ \mu_1 - \mu_0.$$

(Without loss of generality, we can assume that $\mu_1 > \mu_0$.) Then the α and β design requirements can be achieved by taking a sample of size

$$n \ \doteq \ \sigma^2 (z_{\alpha/2} + z_\beta)^2 / \delta^2.$$

Remark: In the proof that follows, we will get an expression for β that involves the standard normal cdf evaluated at a mess that contains n. We will then do an inversion to obtain the desired approximation for n.

Proof: Let's first look at the β value.

$$\begin{aligned}
\beta &= \mathrm{P}(\text{Type II error} \,|\, \mu = \mu_1 > \mu_0) \\
&= \mathrm{P}\big(\text{Fail to Reject } H_0 \,|\, H_0 \text{ false } (\mu = \mu_1 > \mu_0)\big) \\
&= \mathrm{P}\big(|Z_0| \le z_{\alpha/2} \,|\, \mu = \mu_1\big) \\
&= \mathrm{P}\big(-z_{\alpha/2} \le Z_0 \le z_{\alpha/2} \,|\, \mu = \mu_1\big) \\
&= \mathrm{P}\Big(-z_{\alpha/2} \;\le\; \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \;\le\; z_{\alpha/2} \,\Big|\, \mu = \mu_1\Big) \\
&= \mathrm{P}\Big(-z_{\alpha/2} \;\le\; \frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} + \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \;\le\; z_{\alpha/2} \,\Big|\, \mu = \mu_1\Big).
\end{aligned}$$

Since $\mu = \mu_1$ in this expression, we'll define

$$Z \;\equiv\; \frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} \;\sim\; \mathrm{Nor}(0,1),$$

so that

$$\begin{aligned}
\beta &= \mathrm{P}\Big(-z_{\alpha/2} \;\le\; Z + \frac{\sqrt{n}\delta}{\sigma} \;\le\; z_{\alpha/2}\Big) \quad (\text{where } \delta = \mu_1 - \mu_0) \\
&= \mathrm{P}\Big(-z_{\alpha/2} - \frac{\sqrt{n}\delta}{\sigma} \;\le\; Z \;\le\; z_{\alpha/2} - \frac{\sqrt{n}\delta}{\sigma}\Big) \\
&= \Phi\Big(z_{\alpha/2} - \frac{\sqrt{n}\delta}{\sigma}\Big) - \Phi\Big(-z_{\alpha/2} - \frac{\sqrt{n}\delta}{\sigma}\Big).
\end{aligned}$$

Now, note that $-z_{\alpha/2} \ll 0$ and $-\sqrt{n}\delta/\sigma < 0$ (since $\delta > 0$). These two facts imply that the second $\Phi(\cdot)$ term in the expression for β above is $\doteq 0$, so we only need to use the first term in that expression. This results in

$$\beta \;\doteq\; \Phi\Big(z_{\alpha/2} - \frac{\sqrt{n}\delta}{\sigma}\Big)$$

iff

$$\Phi^{-1}(\beta) \;=\; -z_\beta \;\doteq\; z_{\alpha/2} - \frac{\sqrt{n}\delta}{\sigma}$$

iff

$$\frac{\sqrt{n}\delta}{\sigma} \;\doteq\; z_{\alpha/2} + z_\beta \quad \text{iff} \quad n \;\doteq\; \sigma^2(z_{\alpha/2} + z_\beta)^2/\delta^2. \quad \text{Done! Whew!} \quad \square$$

Recap: If you want to test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, and

- You know σ^2,
- You want P(Type I error) $= \alpha$, and
- You want P(Type II error) $= \beta$, when $\mu = \mu_1 (\neq \mu_0)$,

then you have to take $n \doteq \sigma^2(z_{\alpha/2} + z_\beta)^2/\delta^2$ observations.

Similarly, if you're doing a *one-sided* test, it turns out that you need to take $n \doteq \sigma^2(z_\alpha + z_\beta)^2/\delta^2$ observations.

Example: Weights of football players are normal with $\sigma^2 = 324$. We wish to test $H_0 : \mu = 272$ vs. $H_1 : \mu \neq 272$. Suppose we want to protect against the case that μ happens to actually equal $\mu_1 = 276$, and we want $\alpha = 0.05$ and $\beta = 0.05$. How many observations should we take?

$$n \doteq \frac{\sigma^2}{\delta^2}(z_{\alpha/2} + z_\beta)^2 = \frac{324}{16}(1.96 + 1.645)^2 = 263.2.$$

In other words, we need around 263 observations. □

7.2.3 Two-Sample Tests

With an eye towards comparing the means of two competing normal populations, suppose we have the following samples of X's (from population 1) and Y's (from population 2):

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma_x^2) \quad \text{and} \quad Y_1, Y_2, \ldots, Y_m \overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma_y^2),$$

where the samples are independent of each other, and σ_x^2 and σ_y^2 are somehow *known*.

A natural question to ask is whether or not the two populations have the same mean? Here's the two-sided test to see if the means are different:

$$H_0 : \mu_x = \mu_y \quad \text{vs.} \quad H_1 : \mu_x \neq \mu_y.$$

Define the quantity

$$Z_0 = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}}.$$

If H_0 is true (i.e., the means are equal), then everything is known or can be observed, so that Z_0 becomes a test statistic,

$$Z_0 = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim \text{Nor}(0, 1).$$

Thus, as before,

Two-Sided Test for $\mu_x - \mu_y$:

Reject H_0 iff $|Z_0| > z_{\alpha/2}$.

Using more of the same reasoning as before, we get the following one-sided tests:

$$H_0 : \mu_x \leq \mu_y \quad \text{vs.} \quad H_1 : \mu_x > \mu_y$$
$$\text{Reject } H_0 \quad \text{iff} \quad Z_0 > z_\alpha.$$

$$H_0 : \mu_x \geq \mu_y \quad \text{vs.} \quad H_1 : \mu_x < \mu_y$$
$$\text{Reject } H_0 \quad \text{iff} \quad Z_0 < -z_\alpha.$$

It's so easy!! ☺

Example: It's summertime, and the livin' is easy! Let's compare 22 supposedly iid normal temperature readings taken every four days in Atlanta and Barcelona.

First of all, here are the Atlanta readings.

83	89	82	78	80	94	83	92	76	89	83
90	80	91	88	88	88	79	87	83	86	84

And here are Barcelona's.

88	82	84	92	95	103	106	88	73	84	91
84	102	91	95	100	96	101	96	81	92	89

We can easily calculate the sample means, $\bar{A} = 85.14$ and $\bar{B} = 91.50$. In addition, historical evidence allows us to assume that Atlanta's known variance is $\sigma_a^2 = 36$, while Barcelona's is $\sigma_b^2 = 64$.

Suppose we want to test $H_0 : \mu_a = \mu_b$ vs. $H_1 : \mu_a \neq \mu_b$ at level $\alpha = 0.05$. Then the test statistic is

$$Z_0 = \frac{85.14 - 91.50}{\sqrt{\frac{36}{22} + \frac{64}{22}}} = -2.98$$

Then $|Z_0| = 2.98 > z_{\alpha/2} = 1.96$, and so we *reject* H_0. We conclude that the cities have different mean temperatures. □

7.3 Hypothesis Tests for Normal Means (Variance Unknown)

We repeat the work from the previous section, except now we deal with the *unknown* variance case. Specifically, in §7.3.1, we test for the mean of a single normal distribution. The goal in §7.3.2 is to compare the means of two normal distributions when both variances are unknown.

It's time for t again!

7.3.1 One-Sample Test

Suppose $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$, where σ^2 is *unknown* — which is a more-realistic scenario than that of §7.2.1. The two-sided (aka simple) mean test for one

normal population is

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0.$$

As usual, we'll use $\bar{X}$ to estimate μ. If $\bar{X}$ is "significantly different" from μ_0, then we'll reject H_0. To determine what "significantly different" means, we'll also need to estimate σ^2. To this end, define the test statistic

$$T_0 \equiv \frac{\bar{X} - \mu_0}{S/\sqrt{n}},$$

where S^2 is our beloved old friend, the sample variance,

$$S^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \frac{\sigma^2\chi^2(n-1)}{n-1}.$$

If H_0 is true, then

$$T_0 = \frac{\frac{\bar{X}-\mu_0}{\sqrt{\sigma^2/n}}}{\sqrt{S^2/\sigma^2}} \sim \frac{\text{Nor}(0,1)}{\sqrt{\frac{\chi^2(n-1)}{n-1}}} \sim t(n-1).$$

So we have

Two-Sided Test for μ:

$$\text{Reject } H_0 \quad \text{iff} \quad |T_0| > t_{\alpha/2,n-1},$$

where Table B.2 gives us the quantile.

Using the same reasoning as in the previous section, the **one-sided tests** are:

$$H_0 : \mu \leq \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0$$
$$\text{Reject } H_0 \quad \text{iff} \quad T_0 > t_{\alpha,n-1}.$$

$$H_0 : \mu \geq \mu_0 \quad \text{vs.} \quad H_1 : \mu < \mu_0$$
$$\text{Reject } H_0 \quad \text{iff} \quad T_0 < -t_{\alpha,n-1}.$$

Example: Consider the following data set consisting of the average daily waiting times of customers at a local bank over a 40-day period. For various reasons, we can regard these daily averages as iid normal data, albeit with unknown mean and variance.

6.0	4.9	5.8	4.7	4.9	6.5	4.5	6.6	4.8	6.7
5.1	5.0	5.5	4.6	5.8	4.9	5.7	6.3	4.6	2.7
4.9	4.6	5.7	5.8	4.1	5.0	6.7	5.3	5.4	4.1
4.0	5.7	7.6	4.7	4.1	5.8	5.0	3.7	5.8	3.1

An easy calculation reveals that the sample mean $\bar{X} = 5.168$, and the sample variance is $S^2 = 1.003$.

Suppose we want to test whether or not the mean of the waiting times is $\mu_0 = 5.5$. Then the test statistic is

$$T_0 \equiv \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} = \frac{5.168 - 5.5}{\sqrt{1.003/40}} = -2.097.$$

Let's do a two-sided test at level $\alpha = 0.10$. Since $|T_0| = 2.097 > t_{\alpha/2,n-1} = t_{0.05,39} = 1.685$, we *reject* H_0. Thus, the mean waiting time is probably not equal to 5.5. □

Recall: The ***p*-value** of a test is the smallest level of significance α that would lead to rejection of H_0. We can calculate the p-value for this two-sided normal mean test with unknown variance. To begin, we reject H_0 iff

$$|T_0| > t_{\alpha/2,n-1} = F_{n-1}^{-1}(1 - \alpha/2),$$

where $F_{n-1}(t)$ is the cdf of the $t(n-1)$ distribution (and $F_{n-1}^{-1}(\cdot)$ is the inverse). This relationship holds iff

$$F_{n-1}(|T_0|) > 1 - \alpha/2 \quad \text{iff} \quad \alpha > 2\big(1 - F_{n-1}(|T_0|)\big).$$

So for the two-sided test for the case of unknown variance, we have

$$p = 2\big(1 - F_{n-1}(|T_0|)\big).$$

Example: Continuing the bank waiting time example, recall that we had $n = 40$ data points that resulted in the test statistic $T_0 = -2.097$. Then the p-value is $2\big(1 - F_{39}(2.097)\big) = 0.0425$, where the Excel function `t.dist` can be used to evaluate the cdf $F_{39}(2.097)$. Since $p = 0.0425 < 0.05$, we rejected the null hypothesis at level $\alpha = 0.05$. □

7.3.2 Two-Sample Tests

Suppose we have the following two-sample setup,

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma_x^2) \quad \text{and} \quad Y_1, Y_2, \ldots, Y_m \overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma_y^2),$$

where the samples are independent of each other, and σ_x^2 and σ_y^2 are *unknown*.

It is natural to ask: which population has the larger mean? In parallel to the analogous confidence-interval discussion in §6.5, we'll look at three cases, which we'll attack in slightly different ways:

1. **Pooled-t test:** $\sigma_x^2 = \sigma_y^2 = \sigma^2$, say.
2. **Approximate (Welch) t test:** $\sigma_x^2 \neq \sigma_y^2$.
3. **Paired-t test:** (X_i, Y_i) observations paired.

§7.3.2.1 Pooled-*t* Test

Let's start out with the fortuitous case in which the two unknown variances happen to equal each other, i.e., suppose that $\sigma_x^2 = \sigma_y^2 = \sigma^2$, so that σ^2 is the common, unknown variance.

Consider the two-sided test to see if the means are different:

$$H_0 : \mu_x = \mu_y \quad \text{vs.} \quad H_1 : \mu_x \neq \mu_y.$$

In order to carry out the test, we first calculate the sample means and variances from the two populations,

$$\bar{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \bar{Y} \equiv \frac{1}{m}\sum_{i=1}^{m} Y_i$$

$$S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{and} \quad S_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2.$$

As in Chapter 6, we combine the two-sample variances to obtain the **pooled variance estimator**,

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2},$$

which has more degrees of freedom than either of the constituents alone. In fact, if H_0 is true, it can be shown that

$$S_p^2 \sim \frac{\sigma^2\chi^2(n+m-2)}{n+m-2},$$

and then the test statistic

$$T_0 \equiv \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n+m-2).$$

Thus, we have a collection of pooled tests.

Pooled Tests for $\mu_x - \mu_y$:

two-sided	$H_0 : \mu_x = \mu_y$ $H_1 : \mu_x \neq \mu_y$	Reject H_0 iff $\lvert T_0\rvert > t_{\alpha/2,n+m-2}$
one-sided	$H_0 : \mu_x \leq \mu_y$ $H_1 : \mu_x > \mu_y$	Reject H_0 iff $T_0 > t_{\alpha,n+m-2}$
one-sided	$H_0 : \mu_x \geq \mu_y$ $H_1 : \mu_x < \mu_y$	Reject H_0 iff $T_0 < -t_{\alpha,n+m-2}$

Example: The Farmer in the Dell claims that the average weight of red delicious apples is less than that of golden delicious apples. Therefore, he wants to test $H_0 : \mu_r \geq \mu_g$ vs. $H_1 : \mu_r < \mu_g$.

Suppose he has the following data (units are in grams):

$$\begin{aligned} n &= 25, \quad \bar{R} = 106.5, \quad S_{\mathrm{r}}^2 = 6.2 \\ m &= 25, \quad \bar{G} = 112.9, \quad S_{\mathrm{g}}^2 = 5.6. \end{aligned}$$

We see that S_{r}^2 is pretty close to S_{g}^2, so we'll assume $\sigma_{\mathrm{r}}^2 \doteq \sigma_{\mathrm{g}}^2$. This justifies the use of the pooled variance estimator,

$$S_p^2 = \frac{(n-1)S_{\mathrm{r}}^2 + (m-1)S_{\mathrm{g}}^2}{n+m-2} = \frac{24(6.2)+24(5.6)}{48} = 5.9,$$

so that

$$T_0 = \frac{\bar{R}-\bar{G}}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} = \frac{-6.4}{\sqrt{5.9}\sqrt{\frac{1}{25}+\frac{1}{25}}} = -9.32.$$

Let's test at level $\alpha = 0.05$. Then $t_{\alpha,n+m-2} = t_{0.05,48} = 1.677$, and since $T_0 < -t_{\alpha,n+m-2}$, we *reject* H_0 (we didn't even really need a table for this because T_0 was so negative). In other words, we conclude that goldens weigh more on average than reds, just like the farmer thought. □

§7.3.2.2 Approximate *t* Test

All variances are usually not created equal, so suppose that $\sigma_x^2 \neq \sigma_y^2$, where both are unknown. As with our work with Welch's hypothesis test from §6.5.2, define

$$T_0^\star \equiv \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}} \approx t(\nu) \quad (\text{if } H_0 \text{ true}),$$

where the approximate degrees of freedom is given by

$$\nu \equiv \frac{\left(\frac{S_x^2}{n}+\frac{S_y^2}{m}\right)^2}{\frac{(S_x^2/n)^2}{n-1}+\frac{(S_y^2/m)^2}{m-1}}.$$

The following table summarizes how to carry out the various two- and one-sided tests.

Approximate (Welch) Tests for $\mu_x - \mu_y$:

two-sided	$H_0 : \mu_x = \mu_y$ $H_1 : \mu_x \neq \mu_y$	Reject H_0 iff $\lvert T_0^\star \rvert > t_{\alpha/2,\nu}$
one-sided	$H_0 : \mu_x \leq \mu_y$ $H_1 : \mu_x > \mu_y$	Reject H_0 iff $T_0^\star > t_{\alpha,\nu}$
one-sided	$H_0 : \mu_x \geq \mu_y$ $H_1 : \mu_x < \mu_y$	Reject H_0 iff $T_0^\star < -t_{\alpha,\nu}$

Example: Suppose that we're simulating two different queueing policies. For example, think of one long line feeding into several parallel servers (like the post office) vs. several short lines feeding into specific servers (like the grocery store). We simulated $n = 25$ replications of the first policy and have obtained 25 approximately iid normal realizations of average waits. Similarly, we simulated $m = 35$ iid trials of the second policy. Here is the summary data on the waiting times (in seconds).

Policy 1	$n = 25$	$\bar{X}$	$=$	103.6	S_x^2	$=$ 28.62
Policy 2	$m = 35$	$\bar{Y}$	$=$	100.2	S_y^2	$=$ 13.29

Note that those sample variances indicate that the true variances are not equal. Let's test at level $\alpha = 0.05$ the hypothesis that the two queueing policies result in the same expected waiting times.

First, we'll need to calculate the approximate degrees of freedom,

$$\nu \;=\; \frac{\left(\frac{S_x^2}{n} + \frac{S_y^2}{m}\right)^2}{\frac{(S_x^2/n)^2}{n-1} + \frac{(S_y^2/m)^2}{m-1}} \;=\; \frac{\left(\frac{28.62}{25} + \frac{13.29}{35}\right)^2}{\frac{(28.62/25)^2}{24} + \frac{(13.29/35)^2}{34}} \;=\; 39.49 \;\to\; 39.$$

Then the test statistic is

$$T_0^\star \;=\; \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \;=\; \frac{3.4}{\sqrt{\frac{28.62}{25} + \frac{13.29}{35}}} \;=\; 2.754.$$

Since $|T_0^\star| \leq t_{\alpha/2,\nu} = t_{0.025,39} = 2.023$, we reject H_0. Thus, we conclude that the two policies don't have the same mean — Policy 2's is smaller. □

§7.3.2.3 Paired-*t* Test

Again consider two competing normal populations. ***Pair***alleling the confidence interval discussion in §6.5.3, suppose we collect observations from the two populations in *pairs*. In ***pair***ticular, the random variables between *different* pairs are *independent*. The two observations within the *same* pair may *not* be independent — in fact, it's often good for them to be positively correlated (as explained in §6.5.3)!

Example: Consider an investigation involving sets of twins; for each set, one twin takes a new drug, and the other takes a placebo.

$$\text{independent} \begin{cases} \text{Pair } 1: & (X_1, Y_1) \\ \text{Pair } 2: & (X_2, Y_2) \\ \quad\vdots & \quad\vdots \\ \text{Pair } n: & \underbrace{(X_n, Y_n)} \\ & \text{not independent} \end{cases}$$

Before any discussion regarding hypothesis testing, we'll review some notation. First of all, define the pair-wise differences,

$$D_i \;\equiv\; X_i - Y_i, \quad i = 1, 2, \ldots, n.$$

Note that $D_1, D_2, \ldots, D_n \overset{\text{iid}}{\sim} \text{Nor}(\mu_d, \sigma_d^2)$, where

$$\mu_d \equiv \mu_x - \mu_y \quad \text{and} \quad \sigma_d^2 \equiv \sigma_x^2 + \sigma_y^2 - 2\text{Cov}(X_i, Y_i).$$

In addition, define the sample mean and variance of the differences,

$$\bar{D} \equiv \sum_{i=1}^{n} D_i/n \sim \text{Nor}(\mu_d, \sigma_d^2/n),$$

$$S_d^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2 \sim \frac{\sigma_d^2 \chi^2(n-1)}{n-1}.$$

If we are testing the equality of the means, i.e., $H_0 : \mu_d = \mu_x - \mu_y = 0$, then the relevant test statistic is

$$T_0 \equiv \frac{\bar{D}}{\sqrt{S_d^2/n}} \sim t(n-1).$$

Using precisely the same manipulations as in the single-sample normal mean problem with unknown variance from §7.3.1, we get the following decision rules.

Paired Tests for $\mu_x - \mu_y$:

two-sided	$H_0 : \mu_d = 0$ $H_1 : \mu_d \neq 0$	Reject H_0 iff $\lvert T_0 \rvert > t_{\alpha/2,n-1}$
one-sided	$H_0 : \mu_d \leq 0$ $H_1 : \mu_d > 0$	Reject H_0 iff $T_0 > t_{\alpha,n-1}$
one-sided	$H_0 : \mu_d \geq 0$ $H_1 : \mu_d < 0$	Reject H_0 iff $T_0 < -t_{\alpha,n-1}$

Example: Consider the example that we last visited in §6.5.3 in which we are interested in problem-solving times before and after people receive training.

Person	Before Training X_i	After Training Y_i	Difference D_i
1	20	10	10
2	40	25	15
3	5	5	0
4	35	20	15
5	20	15	5

Though the five people are independent, the before and after times for the same individual to solve a problem are almost certainly positively correlated.

Let's assume that all times are normal. We'll test $H_0 : \mu_b = \mu_a$, at level $\alpha = 0.10$. We see that $n = 5$, $\bar{D} = 9$, and $S_d^2 = 42.5$. This gives $|T_0| = |\sqrt{n}\bar{D}/S_d| = 3.087$. Meanwhile, $t_{0.05,4} = 2.13$, so we *reject* H_0.

Thus, we conclude that $\mu_b \neq \mu_a$ (training probably helps). □

7.4 A Potpourri of Tests for Other Parameters

This section discusses a variety of tests for parameters other than the mean.

- The variance σ^2 of a normal distribution (§7.4.1).
- The ratio of variances σ_x^2/σ_y^2 from two normals (§7.4.2).
- The Bernoulli success parameter p (§7.4.3).
- The difference of success parameters, $p_x - p_y$, from two Bernoullis (§7.4.4).

7.4.1 Normal Variance Test

This subsection discusses hypothesis tests on the *variance* of observations from a normal distribution. To this end, suppose that $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$, where μ and σ^2 are *unknown*. We first consider the two-sided test (where you specify the hypothesized σ_0^2),

$$H_0 : \sigma^2 = \sigma_0^2 \quad \text{vs.} \quad H_1 : \sigma^2 \neq \sigma_0^2.$$

Recall (yet again) that the sample variance is

$$S^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \frac{\sigma^2\chi^2(n-1)}{n-1}.$$

We'll use the test statistic

$$\chi_0^2 \equiv \frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1) \quad (\text{if } H_0 \text{ is true}).$$

Thus, we immediately have the following set of tests.

Tests for σ^2:

two-sided	$H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 \neq \sigma_0^2$	Reject H_0 iff $\chi_0^2 < \chi_{1-\alpha/2,n-1}^2$ or $\chi_0^2 > \chi_{\alpha/2,n-1}^2$
one-sided	$H_0 : \sigma^2 \leq \sigma_0^2$ $H_1 : \sigma^2 > \sigma_0^2$	Reject H_0 iff $\chi_0^2 > \chi_{\alpha,n-1}^2$
one-sided	$H_0 : \sigma^2 \geq \sigma_0^2$ $H_1 : \sigma^2 < \sigma_0^2$	Reject H_0 iff $\chi_0^2 < \chi_{1-\alpha,n-1}^2$

Note that we can look up the quantiles in Table B.3.

Example: Suppose we are interested in reducing the variability of the products we make at our bakery — customers love consistency! For this reason, we'd really like to know whether or not the variance of the sizes of our loaves of bread is ≤ 4 cm^2.

Because we want to prove beyond a shadow of a doubt that variability is low, we'll test

$$H_0 : \sigma^2 \geq 4 \quad \text{vs.} \quad H_1 : \sigma^2 < 4.$$

Suppose we examine $n = 100$ loaves, and we find that $\bar{X} = 32$ cm and $S^2 = 3.2$ cm^2. Then the test statistic

$$\chi_0^2 \;=\; (n-1)S^2/\sigma_0^2 \;=\; (99)(3.2)/4 \;=\; 79.2$$

(and isn't explicitly dependent on $\bar{X}$). In addition, using the Excel routine `chisq.inv(0.1,99)`, we calculate $\chi^2_{\alpha,n-1} = \chi^2_{0.90,99} = 81.4$. Since $\chi_0^2 < \chi^2_{1-\alpha,n-1}$, we *reject* H_0. Looks like that bread has low variability — and it really smells mouth-wateringly great! □

7.4.2 Two-Sample Test for Equal Variances

Now we'll test whether or not two populations have the same variance. We begin by sampling

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma_x^2) \quad \text{and} \quad Y_1, Y_2, \ldots, Y_m \overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma_y^2),$$

where we assume that all of the X's and Y's are independent. Of course, we'll estimate the variances σ_x^2 and σ_y^2 by the sample variances S_x^2 and S_y^2.

In order to carry out the two-sided test

$$H_0 : \sigma_x^2 = \sigma_y^2 \quad \text{vs.} \quad H_1 : \sigma_x^2 \neq \sigma_y^2,$$

or equivalently,

$$H_0 : \frac{\sigma_x^2}{\sigma_y^2} \;=\; 1 \quad \text{vs.} \quad H_1 : \frac{\sigma_x^2}{\sigma_y^2} \;\neq\; 1,$$

we'll use the test statistic

$$F_0 \;\equiv\; \frac{S_x^2}{S_y^2} \;\sim\; F(n-1, m-1) \quad (\text{if } H_0 \text{ true}).$$

Thus,

Tests for σ_x^2/σ_y^2:

two-sided	$H_0 : \sigma_x^2 = \sigma_y^2$ $H_1 : \sigma_x^2 \neq \sigma_y^2$	Reject H_0 iff $F_0 < F_{1-\alpha/2,n-1,m-1}$ or $F_0 > F_{\alpha/2,n-1,m-1}$
one-sided	$H_0 : \sigma_x^2 \leq \sigma_y^2$ $H_1 : \sigma_x^2 > \sigma_y^2$	Reject H_0 iff $F_0 > F_{\alpha,n-1,m-1}$
one-sided	$H_0 : \sigma_x^2 \geq \sigma_y^2$ $H_1 : \sigma_x^2 < \sigma_y^2$	Reject H_0 iff $F_0 < F_{1-\alpha,n-1,m-1}$

Note that we can look up the quantiles via the Excel function `f.inv`. But if we are on a desert island, we can use Tables B.4–B.7, though we should recall (see §5.3.4) that $F_{1-\gamma,a,b} = 1/F_{\gamma,b,a}$ for any appropriate γ, a, b.

Example: Suppose we want to test at level $\alpha = 0.05$ whether or not two processes have the same variance, i.e., $H_0 : \sigma_x^2 = \sigma_y^2$ vs. $H_1 : \sigma_x^2 \neq \sigma_y^2$. If the ratio of the sample variances is "too high" or "too low," then we will reject H_0.

Suppose we have the following data (note that we don't explicitly need the sample means): $n = 7$ observations, with $S_x^2 = 17.78$; and $m = 16$, with $S_y^2 = 12.04$. Then $F_0 = S_x^2/S_y^2 = 1.477$. Meanwhile, we get the critical points,

$$F_{1-\alpha/2,n-1,m-1} \;=\; 1/F_{\alpha/2,m-1,n-1} \;=\; 1/F_{0.025,15,6} \;=\; 1/5.269 \;=\; 0.190,$$

and $F_{\alpha/2,n-1,m-1} = F_{0.025,6,15} = 3.415$.

Since F_0 falls between the two critical points, we *fail to reject* H_0. □

7.4.3 Bernoulli Proportion Test

Our interest in this subsection is in testing hypotheses about the Bernoulli success parameter p.

§7.4.3.1 Approximate Bernoulli Test

Suppose that $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$. The two-sided test is of the form

$$H_0 : p = p_0 \quad \text{vs.} \quad H_1 : p \neq p_0,$$

where we specify the hypothesized p_0.

To carry out the test, we let $Y = \sum_{i=1}^n X_i \sim \text{Bin}(n, p)$. We'll use the test statistic

$$Z_0 \;\equiv\; \frac{Y - np_0}{\sqrt{np_0(1-p_0)}} \;=\; \frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

If H_0 is true (i.e., $p = p_0$), the Central Limit Theorem (CLT) implies that

$$Z_0 \;\approx\; \text{Nor}(0, 1).$$

Using the tried-and-true reasoning that we have applied many times so far, we get the following class of tests for p.

Approximate Tests for p:

two-sided	$H_0 : p = p_0$ $H_1 : p \neq p_0$	Reject H_0 iff $\lvert Z_0 \rvert > z_{\alpha/2}$
one-sided	$H_0 : p \leq p_0$ $H_1 : p > p_0$	Reject H_0 iff $Z_0 > z_\alpha$
one-sided	$H_0 : p \geq p_0$ $H_1 : p < p_0$	Reject H_0 iff $Z_0 < -z_\alpha$

Remark: In order for the CLT to work, we need n large (say at least 30), and both $np \geq 5$ and $nq \geq 5$ (so that p isn't too close to 0 or 1). If n isn't very big, we may have to use binomial tables (instead of the normal approximation). This gets a little tedious, so we won't go into it here!

Example: The probability that a student correctly answers a certain test question is p. Suppose that a random sample of 1000 students taking a nationwide test yields 820 correct answers to the question. Let's test at level $\alpha = 0.01$ the hypothesis $H_0 : p \leq 0.8$ vs. $H_1 : p > 0.8$. The (approximate) test statistic is

$$Z_0 \equiv \frac{Y - np_0}{\sqrt{np_0(1-p_0)}} = \frac{820 - 800}{\sqrt{1000(0.8)(0.2)}} = 1.58.$$

Since $Z_0 = 1.58 \leq z_\alpha = z_{0.01} = 2.33$, we *fail to reject* H_0. □

§7.4.3.2 Sample-Size Selection for the Bernoulli Proportion Test

Can we design a two-sided test $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$, such that

$$\text{P(Type I error)} = \alpha \quad \text{and} \quad \text{P(Type II error} \,|\, p \neq p_0) = \beta?$$

That is, can we specify the sample size for a two-sided test that will work when we require a Type I error bound α, and a Type II probability bound β? Yes! We'll now show that the necessary sample size is

$$n \doteq \left[\frac{z_{\alpha/2}\sqrt{p_0 q_0} + z_\beta \sqrt{pq}}{p - p_0} \right]^2,$$

where, to save space, we let $q \equiv 1 - p$, and $q_0 \equiv 1 - p_0$.

Note that n is a function of the unknown p. In practice, we'll choose some $p = p_1$ and ask: How many observations should we take if p happens to equal p_1 instead of p_0? Thus, we guard against the scenario in which p actually equals p_1.

Proof (similar to the normal mean design proof from §7.2.2):

$$\begin{aligned}
\beta &= \mathrm{P}(\text{Type II error}) \\
&= \mathrm{P}(\text{Fail to Reject } H_0 \,|\, H_0 \text{ false}) \\
&\doteq \mathrm{P}\big(|Z_0| \le z_{\alpha/2} \,|\, p \neq p_0\big) \quad \text{(by the CLT)} \\
&= \mathrm{P}\big(-z_{\alpha/2} \le Z_0 \le z_{\alpha/2} \,|\, p \neq p_0\big) \\
&= \mathrm{P}\left(-z_{\alpha/2} \;\le\; \frac{Y - np_0}{\sqrt{np_0(1-p_0)}} \;\le\; z_{\alpha/2} \,\middle|\, p \neq p_0\right) \\
&= \mathrm{P}\left(-z_{\alpha/2}\sqrt{\frac{p_0q_0}{pq}} \;\le\; \frac{Y - np_0}{\sqrt{npq}} \;\le\; z_{\alpha/2}\sqrt{\frac{p_0q_0}{pq}} \,\middle|\, p \neq p_0\right) \\
&= \mathrm{P}\left(-c \;\le\; \frac{Y - np}{\sqrt{npq}} + \frac{n(p-p_0)}{\sqrt{npq}} \;\le\; c \,\middle|\, p \neq p_0\right),
\end{aligned}$$

where

$$c \;\equiv\; z_{\alpha/2}\sqrt{\frac{p_0q_0}{pq}}.$$

Now notice that

$$Z \;\equiv\; \frac{Y - np}{\sqrt{npq}} \;\approx\; \mathrm{Nor}(0,1)$$

(since p is the true success probability). This gives

$$\begin{aligned}
\beta &\doteq \mathrm{P}\left(-c \;\le\; Z + \frac{n(p-p_0)}{\sqrt{npq}} \;\le\; c\right) \\
&= \mathrm{P}\left(-c - \frac{\sqrt{n}(p-p_0)}{\sqrt{pq}} \;\le\; Z \;\le\; c - \frac{\sqrt{n}(p-p_0)}{\sqrt{pq}}\right) \\
&= \mathrm{P}(-c-d \;\le\; Z \;\le\; c-d) \\
&= \Phi(c-d) - \Phi(-c-d),
\end{aligned}$$

where

$$d \;\equiv\; \frac{\sqrt{n}(p-p_0)}{\sqrt{pq}}.$$

Also notice that

$$-c-d \;=\; -z_{\alpha/2}\sqrt{\frac{p_0q_0}{pq}} - \frac{\sqrt{n}(p-p_0)}{\sqrt{pq}} \;\ll\; 0.$$

This implies that $\Phi(-c-d) \doteq 0$, and so $\beta \doteq \Phi(c-d)$. Thus,

$$-z_\beta \;\equiv\; \Phi^{-1}(\beta) \;\doteq\; c-d \;=\; z_{\alpha/2}\sqrt{\frac{p_0q_0}{pq}} - \frac{\sqrt{n}(p-p_0)}{\sqrt{pq}}.$$

After a little algebra, we finally(!) get

$$n \;\doteq\; \left[\frac{z_{\alpha/2}\sqrt{p_0q_0} + z_\beta\sqrt{pq}}{p-p_0}\right]^2.$$

Similarly, the sample size for the corresponding one-sided test is

$$n \doteq \left[\frac{z_\alpha\sqrt{p_0q_0} + z_\beta\sqrt{pq}}{p - p_0}\right]^2. \quad \text{Whew!} \quad \square$$

Example: Suppose that we're conducting a research study to determine whether or not a particular allergy medication works effectively. We'll assume that the drug either clearly works or doesn't work for each independent subject, so that we'll have legitimate Bernoulli trials. Our hypothesis is $H_0 : p = p_0 = 0.8$ vs. $H_1 : p \neq 0.8$.

In order to design our test (i.e., determine its sample size), let's set our Type I error probability to $\alpha = 0.05$. We'd like to protect against goofing up on the poor performance side, so let's set our Type II to error to $\beta = 0.10$, in the special case that $p = p_1 = 0.7$. Then

$$\begin{aligned} n &\doteq \left[\frac{z_{\alpha/2}\sqrt{p_0q_0} + z_\beta\sqrt{p_1q_1}}{p_1 - p_0}\right]^2 \\ &= \left[\frac{1.96\sqrt{(0.8)(0.2)} + 1.28\sqrt{(0.7)(0.3)}}{0.7 - 0.8}\right]^2 \\ &= 187.8 \to 188. \quad \square \end{aligned}$$

7.4.4 Two-Sample Test for Equal Proportions

Suppose that $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p_x)$ and $Y_1, Y_2, \ldots, Y_m \overset{\text{iid}}{\sim} \text{Bern}(p_y)$ are independent samples from two competing Bernoulli populations. Now we're interested in testing hypotheses about the difference in the success parameters, i.e., $p_x - p_y$.

The two-sided test is of the form

$$H_0 : p_x = p_y \quad \text{vs.} \quad H_1 : p_x \neq p_y.$$

In order to carry out the test, denote the respective sample means by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim \frac{\text{Bin}(n, p_x)}{n} \quad \text{and} \quad \bar{Y} \sim \frac{\text{Bin}(m, p_y)}{m}.$$

By the Central Limit Theorem and our confidence interval work from Chapter 6, we know that for large n and m,

$$\bar{X} \approx \text{Nor}\left(p_x, \frac{p_x(1-p_x)}{n}\right) \quad \text{and} \quad \bar{Y} \approx \text{Nor}\left(p_y, \frac{p_y(1-p_y)}{m}\right).$$

Then

$$\frac{\bar{X} - \bar{Y} - (p_x - p_y)}{\sqrt{\frac{p_x(1-p_x)}{n} + \frac{p_y(1-p_y)}{m}}} \approx \text{Nor}(0, 1).$$

Moreover, under the null hypothesis, we have $p \equiv p_x = p_y$, in which case

$$\frac{\bar{X} - \bar{Y}}{\sqrt{p(1-p)\left[\frac{1}{n} + \frac{1}{m}\right]}} \approx \text{Nor}(0, 1).$$

Of course, p on the left-hand side of the above equation is unknown, but under H_0, we assume that $p = p_x = p_y$. So we'll estimate p by the **pooled estimator,**

$$\hat{p} \equiv \frac{\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j}{n+m}.$$

Now plug this into the previous equation to get one last approximation — the test statistic that we can finally work with,

$$Z_0 \equiv \frac{\bar{X} - \bar{Y}}{\sqrt{\hat{p}(1-\hat{p})\left[\frac{1}{n} + \frac{1}{m}\right]}} \approx \text{Nor}(0,1) \quad (\text{under } H_0).$$

Then we immediately have the following set of tests.

Approximate Tests for $p_x - p_y$:

two-sided	$H_0 : p_x = p_y$ $H_1 : p_x \neq p_y$	Reject H_0 iff $\lvert Z_0 \rvert > z_{\alpha/2}$
one-sided	$H_0 : p_x \leq p_y$ $H_1 : p_x > p_y$	Reject H_0 iff $Z_0 > z_\alpha$
one-sided	$H_0 : p_x \geq p_y$ $H_1 : p_x < p_y$	Reject H_0 iff $Z_0 < -z_\alpha$

Example: Let's compare two restaurants based on customer reviews (either "yummy" or "nasty"). Burger Fil-A received 550 yummy ratings out of $n = 950$ reviews (+ 400 nasty ratings), while McWendy's received 675 yummy ratings (+ 350 nasties) out of $m = 1025$ reviews. We'll test at level $\alpha = 0.05$ the hypothesis

$$H_0 : p_b = p_m \quad \text{vs.} \quad H_1 : p_b \neq p_m.$$

First of all, we calculate

$$\bar{B} = \frac{550}{950} = 0.5789, \quad \bar{M} = \frac{675}{1025} = 0.6585, \quad \text{and} \quad \hat{p} = \frac{550+675}{950+1025} = 0.6203.$$

This gives us

$$Z_0 = \frac{\bar{B} - \bar{M}}{\sqrt{\hat{p}(1-\hat{p})\left[\frac{1}{n} + \frac{1}{m}\right]}} = \frac{0.5789 - 0.6585}{\sqrt{0.6203(0.3797)\left[\frac{1}{950} + \frac{1}{1025}\right]}} = -3.642.$$

Since $\lvert Z_0 \rvert > z_{0.025} = 1.96$, we easily reject H_0, and informally declare McWendy's the winner. □

7.5 Goodness-of-Fit Tests

At this point, let's suppose that we've guessed a reasonable distribution and then estimated the relevant parameters. Now we will conduct a formal test to see just how successful our toils have been — in other words, are our hypothesized distribution + relevant parameters acceptable?

7.5.1 χ^2 Goodness-of-Fit Test

In particular, we'll test

$$H_0 : X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{ pmf / pdf } f(x) \quad \text{vs.} \quad H_1 : \text{not } H_0,$$

at level of significance

$$\alpha \equiv \text{P(Reject } H_0 \,|\, H_0 \text{ true)} = \text{P(Type I error)}.$$

This is known as a **goodness-of-fit test**. A goodness-of-fit test procedure has the following main steps:

1. Divide the domain of $f(x)$ into k sets, say, $A_1, A_2, \ldots, A_k$ (distinct points if X is discrete or intervals if X is continuous).
2. Tally the actual number of observations O_i that fall in set A_i, $i = 1, 2, \ldots, k$. Assuming H_0 is true (i.e., $f(x)$ is the actual pmf / pdf), define $p_i \equiv \text{P}(X \in A_i)$, in which case $O_i \sim \text{Bin}(n, p_i)$, $i = 1, 2, \ldots, k$.
3. Determine the expected number of observations that would fall into each set if H_0 were true, say, $E_i = \text{E}[O_i] = np_i$, $i = 1, 2, \ldots, k$.
4. Calculate a test statistic based on the differences between the E_i's and O_i's. The **chi-squared goodness-of-fit test statistic** is[1]
$$\chi_0^2 \equiv \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$
5. A large value of χ_0^2 indicates a bad fit (so just do a one-sided test). Thus, we *reject* H_0 if $\chi_0^2 > \chi^2_{\alpha,k-1-s}$, where:
   - s is the number of unknown parameters from $f(x)$ that have to be estimated. For example, if $X \sim \text{Nor}(\mu, \sigma^2)$, then $s = 2$.
   - $\chi^2_{\alpha,\nu}$ is the usual $(1-\alpha)$ quantile of the $\chi^2(\nu)$ distribution that we can find in Table B.3, i.e., $\text{P}\big(\chi^2(\nu) < \chi^2_{\alpha,\nu}\big) = 1 - \alpha$.

   If $\chi_0^2 \leq \chi^2_{\alpha,k-1-s}$, we *fail to reject* (and so we grudgingly accept) H_0.

Remarks:

- The usual recommendations: For the χ^2 goodness-of-fit test to work, pick $n \geq 30$, and k such that $E_i \geq 5$ for all i.
- If the degrees of freedom, $\nu = k - 1 - s$, happens to be very big, then it can be shown that
$$\chi^2_{\alpha,\nu} \doteq \nu \left[1 - \frac{2}{9\nu} + z_\alpha \sqrt{\frac{2}{9\nu}}\right]^3,$$
where z_α is the appropriate standard normal quantile.
- Other goodness-of-fit tests with funny names: Kolmogorov–Smirnov, Anderson–Darling, Shapiro–Wilk, etc.

[1] Why might this test statistic remind you of Old MacDonald's farm?

7.5.2 Beginner Examples

In this section, we'll work through a number of simple examples illustrating the use of the χ^2 goodness-of-fit test.

Baby Example: Test H_0: X_i's are Unif(0,1). Suppose we have $n = 10000$ observations (uniforms are cheap!), divided into $k = 5$ intervals of equal length (so these are easy-to-use "equal-probability" intervals).

Interval	[0,0.2]	(0.2,0.4]	(0.4,0.6]	(0.6,0.8]	(0.8,1.0]
p_i	0.2	0.2	0.2	0.2	0.2
$E_i = np_i$	2000	2000	2000	2000	2000
O_i	1962	2081	2022	1982	1953

We can easily calculate the goodness-of-fit statistic, $\chi_0^2 \equiv \sum_{i=1}^k (O_i - E_i)^2/E_i = 5.511$.

Now let's take $\alpha = 0.05$, and note that there are no unknown parameters, so $s = 0$. Then $\chi^2_{\alpha,k-1-s} = \chi^2_{0.05,4} = 9.49$.

Since $\chi_0^2 \leq \chi^2_{\alpha,k-1-s}$, we fail to reject H_0. Thus, we'll pretend that the numbers are indeed uniform. □

Discrete Example: 1000 kids at a basketball camp are asked to take four free throws, and the numbers of successful shots are recorded, where we let X_i denote the number of shots kid i makes.

# made	0	1	2	3	4
Frequency	32	131	329	366	142

We'll test H_0: The observations are Bin$(4, p)$. Let's start by finding p's maximum likelihood estimator. The likelihood function is

$$\begin{aligned} L(p) &= \prod_{i=1}^n f(x_i) \\ &= \prod_{i=1}^n \binom{4}{x_i} p^{x_i}(1-p)^{4-x_i} \\ &= C\, p^{\sum_{i=1}^n x_i}(1-p)^{4n-\sum_{i=1}^n x_i}, \end{aligned}$$

where $C = \prod_{i=1}^n \binom{4}{x_i}$ is a constant with respect to p. Thus,

$$\ell\mathrm{n}(L(p)) = \ell\mathrm{n}(C) + \Big(\sum_{i=1}^n x_i\Big)\,\ell\mathrm{n}(p) + \Big(4n - \sum_{i=1}^n x_i\Big)\,\ell\mathrm{n}(1-p),$$

so that

$$\frac{d\,\ell\mathrm{n}(L(p))}{dp} = \frac{\sum_{i=1}^n x_i}{p} - \frac{4n - \sum_{i=1}^n x_i}{1-p} = 0.$$

After a little algebra, we obtain the MLE,

$$\hat{p} \;=\; \frac{1}{4n}\sum_{i=1}^{n} X_i \;=\; \frac{1}{4000}\sum_{i=1}^{1000} X_i \;=\; \frac{0(32)+1(131)+\cdots+4(142)}{4000} \;=\; 0.61375,$$

which makes sense since it's the overall average of all of the shots taken by all of the $n = 1000$ kids.

Our aim is to calculate the goodness-of-fit test statistic χ_0^2 associated with the $\text{Bin}(4, \hat{p})$ distribution. To this end, we'll make a little table, assuming that $\hat{p} = 0.61375$ is correct. By the Invariance Property of MLE's (this is why we learned it!), the expected number of kids making x shots is $E_x = n\widehat{\text{P}}(X = x) = n\binom{4}{x}\hat{p}^x(1-\hat{p})^{4-x}$, $x = 0, 1, \ldots, 4$ (assuming $\hat{p}$ is actually p).

x	0	1	2	3	4	Total
$\widehat{\text{P}}(X = x)$	0.0223	0.1415	0.3372	0.3572	0.1419	1
E_x	22.26	141.47	337.19	357.19	141.89	1000
O_x	32	131	329	366	142	1000

Thus, the test statistic is

$$\chi_0^2 \;=\; \sum_{x=0}^{4} \frac{(O_x - E_x)^2}{E_x} \;=\; \frac{(32-22.26)^2}{22.26} + \frac{(131-141.47)^2}{141.47} + \cdots \;=\; 5.46.$$

Let $k = 5$ denote the number of cells, and let $s = 1$ denote the number of parameters we had to estimate. Suppose the level $\alpha = 0.05$. Then we compare $\chi_0^2 = 5.46$ against $\chi^2_{\alpha,k-1-s} = \chi^2_{0.05,3} = 7.81$.

Since $\chi_0^2 < \chi^2_{\alpha,k-1-s}$, we fail to reject H_0. This means that we can sort of regard the number of shots that the kids make as $\text{Bin}(4, \hat{p})$. $\square$

Continuous Distributions: For the continuous case, we'll denote the intervals $A_i \equiv (a_{i-1}, a_i]$, for $i = 1, 2, \ldots, k$. For convenience, we choose the a_i's to ensure that we have **equal-probability intervals**, i.e.,

$$p_i \;=\; \text{P}(X \in A_i) \;=\; \text{P}(a_{i-1} < X \le a_i) \;=\; 1/k, \quad \text{for all } i.$$

In this case, we immediately have $E_i = n/k$ for all i, and then

$$\chi_0^2 \;=\; \sum_{i=1}^{k} \frac{(O_i - (n/k))^2}{n/k}.$$

The issue is that the a_i's might depend on unknown parameters.

Example: Suppose that we're interested in fitting a distribution to a series of interarrival times. Could they be *exponential*?

$$H_0 : X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda).$$

Let's do a χ^2 goodness-of-fit test with equal-probability intervals. This amounts to choosing a_i's such that the cdf

$$F(a_i) \;=\; \mathrm{P}(X \le a_i) \;=\; 1 - e^{-\lambda a_i} \;=\; \frac{i}{k}, \quad i = 0, 1, 2, \ldots, k.$$

After a wee bit of algebra, we obtain

$$a_i \;=\; -\frac{1}{\lambda}\ell\mathrm{n}\Big(1 - \frac{i}{k}\Big), \quad i = 0, 1, 2, \ldots, k.$$

Great, but λ is unknown, so we'll need to estimate $s = 1$ parameter. The good news is that we know from §5.2.5 that the MLE is $\widehat{\lambda} = 1/\bar{X}$. Thus, by the Invariance Property, the MLE's of the a_i's are

$$\widehat{a}_i \;=\; -\frac{1}{\widehat{\lambda}}\,\ell\mathrm{n}\Big(1 - \frac{i}{k}\Big) \;=\; -\bar{X}\,\ell\mathrm{n}\Big(1 - \frac{i}{k}\Big), \quad i = 0, 1, 2, \ldots, k. \quad \square$$

Example (continued): Consider the following $n = 100$ observations.

0.9448	0.8332	0.6811	0.9574	0.8351	1.4689	1.1664	0.6908	1.0641	1.0850
0.4271	0.6450	0.5775	0.7418	0.7728	0.8622	0.9768	1.4564	0.7689	1.2720
1.4364	0.3484	0.9015	0.1020	0.7372	0.5493	1.0788	0.8315	0.7200	1.6785
0.4792	1.2647	1.6250	0.5860	0.5293	1.0847	0.6903	0.3253	1.1239	1.2604
0.5390	0.4309	0.7868	0.3558	0.8677	1.0312	1.4873	1.2868	1.1287	0.5592
0.6776	1.3921	0.9592	0.4818	0.5250	1.2473	1.0830	0.4711	0.5856	0.9798
1.3020	0.3465	1.3818	0.6614	0.7619	1.1047	0.4816	0.4628	1.2492	0.7993
1.1733	1.0558	0.6630	0.7472	0.6149	1.2403	1.1319	0.9306	0.3116	0.7711
1.2537	0.8989	1.1378	0.5341	1.1170	0.5455	1.0143	0.9954	0.8152	0.8712
0.7990	1.6576	0.5394	1.0199	0.6693	0.9880	0.9220	0.7691	1.0232	0.5635

We'll do a goodness-of-fit test for the exponential distribution using $k = 10$ equal-probability intervals, so that $E_i = n/k = 10$, for all i. It turns out that the sample mean based on the 100 observations is $\bar{X} = 0.8778$. Then

$$\widehat{a}_i \;=\; -0.8778\,\ell\mathrm{n}\big(1 - 0.1i\big), \quad i = 0, 1, 2, \ldots, 10.$$

Finally, trust me that we've determined which interval each of the 100 observations belongs to, and then we just tally them up to get the O_i's...

Interval $(\hat{a}_{i-1}, \hat{a}_i]$	O_i	$E_i = n/k$
[0, 0.092]	0	10
(0.092, 0.196]	1	10
(0.196, 0.313]	1	10
(0.313, 0.448]	6	10
(0.448, 0.608]	17	10
(0.608, 0.804]	21	10
(0.804, 1.057]	23	10
(1.057, 1.413]	24	10
(1.413, 2.021]	7	10
(2.021, ∞)	0	10
	100	100

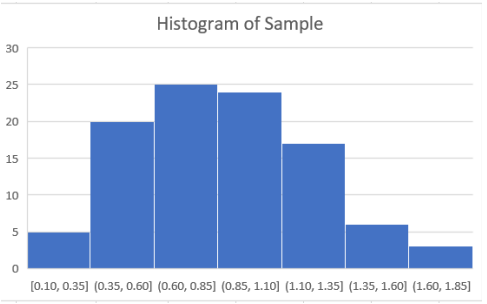
Then

$$\chi_0^2 = \sum_{i=1}^{k} (O_i - E_i)^2/E_i = 92.2 \quad \text{and} \quad \chi^2_{\alpha,k-1-s} = \chi^2_{0.05,8} = 15.51.$$

So we reject H_0. These observations ain't Expo. □

7.5.3 Mini-Project

Let's make things more interesting with an extended example / mini-project. We continue to consider the same sample of 100 iid observations from the previous example, but now with some more details. The sample mean for this dataset is (still) $\bar{X} = 0.8778$, and the sample standard deviation is $S = 0.3347$. It's often useful to graph a dataset, and here's what this little fella looks like...



For the remainder of this section, we'll do various goodness-of-fit tests to determine which distribution(s) the data could come from.

$$H_0 : X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} f(x),$$

where we'll consider the following possibilities:

- Exponential (rejected above).
- Gamma (which generalizes the exponential).
- Weibull (which generalizes the exponential in a different way).

In each case, we'll divide the data into $k = 10$ equal-probability intervals and perform a χ^2 goodness-of-fit test at level $\alpha = 0.05$. Along the way, we'll encounter a number of interesting issues that we'll need to deal with, but it'll all work out in the end. ☺

7.5.3.1 Exponential Fit

Bringing back bad memories, we tested in the previous example the null hypothesis that

$$H_0 : X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}(\lambda).$$

Recall that we failed miserably. But, in retrospect, this makes sense in light of the facts that:

- The graph doesn't look anywhere near exponential.
- The expected value and standard deviation of an Exp(λ) random variable are both $1/\lambda$, yet the sample mean $\bar{X} = 0.8778$ is $\gg$ the sample standard deviation $S = 0.3347$.

This motivates our need to look at other distributions in our quest to find a good data fit.

7.5.3.2 Gamma Fit

The gamma distribution with parameters r and λ has pdf

$$f(x) \;=\; \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, \quad x > 0.$$

Note that $r = 1$ yields the Exp(λ) as a special case, so the gamma has better potential to fit a dataset. We'll test

$$H_0 : X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Gam}(r, \lambda).$$

Way back in the good old days (§5.2.5), we found the MLE's for r and λ:

$$\hat{\lambda} \;=\; \hat{r}/\bar{X},$$

where $\hat{r}$ solves

$$g(r) \;\equiv\; n\,\ell\text{n}(r/\bar{x}) - n\psi(r) + \ell\text{n}\Big(\prod_{i=1}^{n} x_i\Big) \;=\; 0,$$

and where $\psi(r) \equiv \Gamma'(r)/\Gamma(r)$ is the *digamma function*.

Since the digamma is sometimes a little hard to find in the usual software packages, we'll incorporate the approximation

$$\Gamma'(r) \;\doteq\; \frac{\Gamma(r+h) - \Gamma(r)}{h} \quad \text{(for any small } h \text{ of your choosing)}.$$

So we need to find $\hat{r}$ that solves

$$g(r) \;\doteq\; n\,\ell\text{n}(r) - n\,\ell\text{n}(\bar{x}) - \frac{n}{h}\Big(\frac{\Gamma(r+h)}{\Gamma(r)} - 1\Big) + \ell\text{n}\Big(\prod_{i=1}^{n} x_i\Big) \;=\; 0. \tag{7.1}$$

Recall from §1.2.2 that we have a number of ways to attack this type of problem. In particular, the **bisection method** is an easy way to find a zero of any continuous function $g(r)$. It relies on the **Intermediate Value Theorem**, which states that if $g(\ell)g(u) < 0$, then there is a zero $r^\star \in [\ell, u]$. Using this fact, it's easy to hone in on a zero via sequential bisecting.

Let's try bisection out on our dataset of $n = 100$ observations, where we recall that the sample mean is $\bar{X} = 0.8778$. And trust me that $\ell\text{n}\big(\prod_{i=1}^{n} X_i\big) = -21.5623$. In

addition, we'll set the approximate differentiation term to $h = 0.01$. Then here's what Equation (7.1) simplifies to:

$$\begin{aligned} g(r) &\doteq 100\,\ell\mathrm{n}(r) - 100\,\ell\mathrm{n}(0.8778) - \frac{100}{0.01}\Big(\frac{\Gamma(r+0.01)}{\Gamma(r)} - 1\Big) - 21.5623 \\ &= 100\,\ell\mathrm{n}(r) - \frac{10000\,\Gamma(r+0.01)}{\Gamma(r)} + 9991.47 \;=\; 0. \end{aligned}$$

Following the bisection algorithm outlined in §1.2.2, we first need to initialize the search by noting that $g(5) = 0.5506$ and $g(7) = -3.0595$. So there's a zero in there somewhere just itching to be found! The algorithm is depicted in all of its glory in the table.

step	ℓ_j	$g(\ell_j)$	u_j	$g(u_j)$	r_j	$g(r_j)$
0	5.0000	0.5506	7.0000	−3.0595	6.0000	−1.5210
1	5.0000	0.5506	6.0000	−1.5210	5.5000	−0.5698
2	5.0000	0.5506	5.5000	−0.5698	5.2500	−0.0338
3	5.0000	0.5506	5.2500	−0.0338	5.1250	0.2519
4	5.1250	0.2519	5.2500	−0.0338	5.1875	0.1075
5	5.1875	0.1075	5.2500	−0.0338	5.2188	0.0365
6	5.2188	0.0365	5.2500	−0.0338	5.2344	0.0013
7	5.2344	0.0013	5.2500	−0.0338	5.2422	−0.0163
⋮						
14	5.2349	0.0000	5.2349	0.0000	$\boldsymbol{r^\star = 5.2349}$	0.0000

We see that the algorithm eventually gives $\hat{r} = r^\star = 5.2349$, and then $\hat{\lambda} = \hat{r}/\bar{X} = 5.9637$.

So now we can finally start our χ^2 goodness-of-fit toils, noting that we have $s = 2$ unknown parameters. We take the $n = 100$ observations and divide them into $k = 10$ equal-probability intervals, so that $E_i = n/k = 10$, for all i. The (approximate) endpoints of the intervals are implicitly given by $\widehat{F}(\hat{a}_i) = i/k$, $i = 0, 1, 2, \ldots, k$, where $\widehat{F}(x)$ is the cdf of the $\mathrm{Gam}(\hat{r}, \hat{\lambda})$ distribution.

Sadly, the gamma distribution's cdf doesn't have a closed form. But that's why we have Excel (or its friends) around, e.g.,

$$\hat{a}_i \;=\; \widehat{F}^{-1}(i/k) \;=\; \texttt{gammainv}(i/k, \hat{r}, \hat{\lambda}),$$

and this results in the following table.

interval $(\hat{a}_{i-1}, \hat{a}_i]$	O_i	$E_i = n/k$
[0.000, 0.436]	8	10
(0.436, 0.550]	12	10
(0.550, 0.644]	6	10
(0.644, 0.733]	9	10
(0.733, 0.823]	12	10
(0.823, 0.920]	8	10
(0.920, 1.032]	13	10
(1.032, 1.174]	14	10
(1.174, 1.391]	10	10
(1.391, ∞)	8	10
	100	100

We immediately find that

$$\chi_0^2 \;=\; \sum_{i=1}^{k} (O_i - E_i)^2/E_i \;=\; 6.2 \quad \text{and} \quad \chi^2_{\alpha,k-1-s} \;=\; \chi^2_{0.05,7} \;=\; 14.07.$$

So we fail to reject H_0. These may indeed be gamma! ☺

7.5.3.3 Weibull Fit

The Weibull distribution has cdf $F(x) = 1 - \exp\big[-(\lambda x)^r\big]$, for $x \geq 0$. Note that $r = 1$ yields the Exp(λ) as a special case. (Also note that the r and λ here aren't the same as for the gamma distribution discussed earlier.)

Let's start out by getting MLE's for the $s = 2$ unknown parameters (r and λ). After a little algebra (involving a couple of chain rules), the pdf is

$$f(x) \;=\; \lambda r (\lambda x)^{r-1} e^{-(\lambda x)^r}, \quad x \geq 0.$$

Thus, the likelihood function for an iid sample of size n is

$$L(r, \lambda) \;=\; \prod_{i=1}^{n} f(x_i) \;=\; \lambda^{nr} r^n \Big(\prod_{i=1}^{n} x_i \Big)^{r-1} \exp\Big[-\lambda^r \sum_{i=1}^{n} x_i^r \Big],$$

so that

$$\ell\mathrm{n}(L) \;=\; nr\,\ell\mathrm{n}(\lambda) + n\,\ell\mathrm{n}(r) + (r-1)\,\ell\mathrm{n}\Big(\prod_{i=1}^{n} x_i \Big) - \lambda^r \sum_{i=1}^{n} x_i^r.$$

At this point, we maximize with respect to r and λ by setting

$$\frac{\partial}{\partial r}\,\ell\mathrm{n}(L) \;=\; 0 \quad \text{and} \quad \frac{\partial}{\partial \lambda}\,\ell\mathrm{n}(L) \;=\; 0.$$

After more algebra — including the fact that $\frac{d}{dx}\, c^x = c^x \ell\mathrm{n}(c)$ — we get the simultaneous equations

$$\lambda \;=\; \Big(\frac{1}{n} \sum_{i=1}^{n} x_i^r \Big)^{-1/r} \quad \text{and}$$

$$g(r) \;=\; \frac{n}{r} + \ell\mathrm{n}\Big(\prod_{i=1}^{n} x_i\Big) - \frac{n\sum_i x_i^r\,\ell\mathrm{n}(x_i)}{\sum_i x_i^r} \;=\; 0.$$

The equation for λ looks easy enough, if only we could solve for r! ☹

But we can! Let's use **Newton's method** this time. It's usually a lot faster than bisection. Here's a reasonable implementation of Newton from §1.2.2.

1. Initialize $r_0 = \bar{X}/S$, where $\bar{X}$ is the sample mean and S^2 is the sample variance. Set $j \leftarrow 0$.
2. Update $r_{j+1} \leftarrow r_j - g(r_j)/g'(r_j)$.
3. If $|g(r_{j+1})|$ or $|r_{j+1} - r_j|$ or your budget is suitably small, then STOP and set the MLE $\hat{r} \leftarrow r_{j+1}$. Otherwise, let $j \leftarrow j+1$ and go to Step 2.

In order to use Newton, we need (after yet more algebra)

$$g'(r) \;=\; -\frac{n}{r^2} - \frac{n\sum_i x_i^r\,[\ell\mathrm{n}(x_i)]^2}{\sum_i x_i^r} + \frac{n\big[\sum_i x_i^r\,\ell\mathrm{n}(x_i)\big]^2}{\big[\sum_i x_i^r\big]^2}.$$

Let's try Newton on our dataset of $n = 100$ observations, where $r_0 = \bar{X}/S = 0.8778/0.3347 = 2.6227$. This results in the following table.

step	r_j	$g(r_j)$	$g'(r_j)$
0	2.6227	5.0896	−25.0848
1	2.8224	0.7748	−23.8654
2	2.8549	0.1170	−23.6493
3	2.8598	0.0178	−23.6174
4	2.8606	0.0027	−23.6126
5	2.8607		

Hence, $\hat{r} = r_5 = 2.8607$, and thus,

$$\widehat{\lambda} \;=\; \Big(\frac{1}{n}\sum_{i=1}^{n} X_i^{\hat{r}}\Big)^{-1/\hat{r}} \;=\; 1.0148.$$

We are now equipped to do a χ^2 goodness-of-fit test with equal-probability intervals. To get the endpoints, we note that $F(a_i) = i/k$, and then some algebra fun + the MLE Invariance Property yield

$$\hat{a}_i \;=\; \frac{1}{\widehat{\lambda}}\left[-\ell\mathrm{n}\Big(1 - \frac{i}{k}\Big)\right]^{1/\hat{r}} \;=\; 0.9854\,[-\ell\mathrm{n}(1 - 0.1\,i)]^{0.3496}, \quad i = 0, 1, 2, \ldots, 10.$$

Moreover, it turns out (see the next table) that

$$\chi_0^2 \;=\; \sum_{i=1}^{k} (O_i - E_i)^2/E_i \;=\; 5.0 \quad \text{and} \quad \chi^2_{\alpha,k-1-s} \;=\; \chi^2_{0.05,7} \;=\; 14.07.$$

interval $(\hat{a}_{i-1}, \hat{a}_i]$	O_i	$E_i = n/k$
[0, 0.4487]	8	10
(0.4487, 0.5833]	15	10
(0.5833, 0.6872]	9	10
(0.6872, 0.7792]	11	10
(0.7792, 0.8669]	8	10
(0.8669, 0.9557]	7	10
(0.9557, 1.0514]	10	10
(1.0514, 1.1637]	12	10
(1.1637, 1.3189]	11	10
(1.3189, ∞]	9	10
	100	100

So we fail to reject H_0, and we'll grudgingly allow that these observations are Weibull. □

The Big Reveal: I actually generated the observations from a Weibull distribution, with parameters $r = 3$ and $\lambda = 1$. So we did pretty well! ☺

7.6 Exercises

1. (§7.1) TRUE or FALSE? We reject the null hypothesis if we are given statistically significant evidence that it is false.
2. (§7.1) The quantity α is known as
   (a) P(Type I error)
   (b) P(Type II error)
   (c) level of significance
   (d) P(Reject H_0 | H_0 is true)
3. (§7.1) What does $1 - \beta$ represent in a typical hypothesis test?
   (a) power
   (b) P(Type II error)
   (c) P(Type I error)
   (d) 1 − confidence level
   (e) P(Fail to Reject H_1 | H_1 is true)
4. (§7.1) Suppose a drug manufacturer releases a new drug that is actually *less* effective than its current brand. What kind of error has the company just committed — Type I or II?
5. (§7.2.1) Suppose you conduct a hypothesis test and you end up with a p-value that is very close to zero. What would you do?

(a) Totally panic!
(b) Reject H_0.
(c) Fail to Reject H_0.
(d) Obtain more data.
(e) Increase/Reduce β appropriately.

6. (§7.2.1) TRUE or FALSE? The p-value of a test always greater than the Type I error, α.

7. (§7.2.1) The average Atlanta temperature in August is being studied. The past five years have resulted in the following average August temperatures: 81.6, 78.8, 80.8, 79.9, 81.3. Previous experience allows us to assume these averages are normal with variance $\sigma^2 = 5$ degrees^2. Is there reason to believe that the true mean monthly temperature is less than 80°? That is, test $H_0 : \mu \geq 80$ vs. $H_1 : \mu < 80$. Use $\alpha = 0.025$.

8. (§7.2.1) For certification purposes, the mean test score of a certain cohort of students must "provably" be greater than 60. To this end, four random students are tested and their sample average test score is 61. We'll assume that the scores are iid normal with a *known* standard deviation of 3.

   (a) Based on this sample, can we conclude that the mean score is > 60, with level of significance $\alpha = 0.05$?
   (b) What is the probability of rejecting $H_0 : \mu \leq 60$ if the true mean score is 65?

9. (§7.2.2) Consider a series of iid normal observations $X_1, X_2, \ldots, X_n$ with unknown mean μ and *known* variance $\sigma^2 = 100$. Suppose we are considering the one-sided test $H_0 : \mu \geq 90$ vs. $H_1 : \mu < 90$, at level $\alpha = 0.05$. How many observations should we take if we want the probability of a Type II error to be 0.10 when μ happens to equal 87?

10. (§7.2.3) Joe and Sam make pizzas! The thicknesses of their pizza crusts are assumed to be normally distributed, with (known) standard deviations of $\sigma_{\text{J}} = 0.10$ and $\sigma_{\text{S}} = 0.15$ inches, respectively. Here are random samples of the thicknesses of pizzas made by each chef (in inches).

Joe	1.33	1.14	1.25	1.25	1.22	1.21	1.16	1.18	1.42	0.99
Sam	0.92	1.37	0.96	1.31	1.19	1.23	1.24	1.22	1.21	1.10

    Do Joe and Sam produce pizzas with the same mean thickness? Use $\alpha = 0.05$.

11. (§7.3.1) Suppose that weekly profits at a certain manufacturing center are iid normal with unknown mean μ and unknown variance σ^2. After we take ten observations, we find that the sample mean $\bar{X} = 100$ and the sample standard deviation $S = 50$. Test to see if the mean is actually above 90, i.e., test $H_0 : \mu \leq 90$ vs. $H_1 : \mu > 90$. Use $\alpha = 0.025$.

12. (§7.3.2) TRUE or FALSE? When you compare two populations, you can use the pooled variance estimator if you believe that the variances of the two populations are approximately equal.

13. (§7.3.2) Suppose we want to compare the means of two normal populations, both of which have unknown variances. We take $n = 6$ observations from the first population and find that the sample mean and sample variance are $\bar{X} = 50$ and $S_x^2 = 120$. We take $m = 5$ observations from the second population and obtain a sample mean and sample variance of $\bar{Y} = 75$ and $S_y^2 = 100$. Because the sample variances are pretty close, go ahead and use the *pooled variance* estimator, S_p^2. Test the hypothesis that $\mu_x = \mu_y$ with $\alpha = 0.10$.

14. (§7.3.2) Suppose that $m = 10$ men and $w = 10$ women take a certain performance examination, with the resulting sample statistics $\bar{M} = 105$, $S_{\text{m}}^2 = 36$, $\bar{W} = 120$, and $S_{\text{w}}^2 = 100$. Looking at the sample variances, we'll have to assume that the variances of the two populations are unequal. Now test the null hypothesis that men and women perform about the same on the exam. Use $\alpha = 0.05$.

15. (§7.3.2) Suppose we conduct an experiment to test if people can throw farther right- or left-handed. We get 20 people to do the experiment. Each throws a ball right-handed once and a ball left-handed once (in random order), and we measure the distances. If we are interested in determining whether the expected length of a right-handed toss is more than that of a left-handed toss, which type of hypothesis test would we likely use?

    (a) z (normal) hypothesis test for differences

    (b) pooled-t hypothesis test for differences

    (c) paired-t hypothesis test for differences

    (d) χ^2 hypothesis test for differences

    (e) F hypothesis test for differences

16. (§7.3.2) TRUE or FALSE? A pooled-t test typically has fewer degrees of freedom than a paired-t test.

17. (§7.3.2) We are studying the times required to solve two elementary math problems. Suppose we ask four students to attempt both Problem A and Problem B (in some kind of randomized order). Assume all of the scores are normal and that the kids are independent. The results are presented below (in seconds).

Student	Problem A	Problem B
1	20	35
2	30	40
3	15	20
4	40	50

    Let's see if we can prove beyond a shadow of a doubt that B is harder (takes more time on average) than A. Specifically, test $H_0 : \mu_{\text{A}} \geq \mu_{\text{B}}$ vs. $H_1 : \mu_{\text{A}} < \mu_{\text{B}}$. Use $\alpha = 0.05$.

18. (§7.4.1) The heights of people in a certain population supposedly have a standard deviation of 2 inches. However, we have just taken a random sample of $n = 11$ people and obtained a sample standard deviation of 4.2 inches. What's going on? Use $\alpha = 0.01$ to test whether or not the claim of $H_0 : \sigma \leq 2$ is justified.

19. (§7.4.2) Suppose we want to test at level 0.05 whether or not two processes have the same variance. For the first system, we have $n = 12$ observations with $\bar{X} = -3.87$ and $S_x^2 = 67.78$; and for the second system, $m = 8$ with $\bar{Y} = 32.16$ and $S_y^2 = 12.04$. So do the systems have equal variance?

20. (§7.4.2) Consider the following random samples drawn from two normal populations, 20 X's and 10 Y's.

X's	0.81	0.69	2.71	4.12	1.88	−0.88	5.11	4.76	2.97	2.40
	3.57	−1.16	2.49	3.30	3.01	−3.10	−0.02	0.07	1.55	0.86
Y's	−2.90	−0.15	−0.20	0.63	−1.91	0.69	0.24	2.37	1.30	−0.01

Is there evidence to conclude that the variance of the X's is greater than that of the Y's? Use $\alpha = 0.05$.

21. (§7.4.3) Of $n = 1000$ randomly selected people in a certain population, $Y = 133$ were found to have COVID-19 antibodies. Test the null hypothesis that the actual rate of antibodies is at most 10%, i.e., $H_0 : p \leq 0.1$ vs. $H_1 : p > 0.1$. Use $\alpha = 0.01$.

22. (§7.4.4) Let's compare two colleges based on employer reviews of graduates (either winners or losers). Out of 500 Tech School of Technology students participating in the study, 475 received winner reviews, and 25 received loser reviews. On the other hand, out of 300 Justin Bieber Singing Academy students, only 150 received winner reviews while 150 received loser reviews. (Actually, some of the responding employers expressed regret that there was not a "total loser" category for some of the JBSA students.) Even though we all know how this is going to turn out, let's test at level $\alpha = 0.05$ the hypothesis

$$H_0 : p_{\text{TST}} = p_{\text{JB}} \quad \text{vs.} \quad H_1 : p_{\text{TST}} \neq p_{\text{JB}}.$$

First of all, we calculate the sample means and the pooled probability estimate,

$$\bar{T} = \frac{475}{500} = 0.95, \quad \bar{J} = \frac{150}{300} = 0.5, \quad \text{and} \quad \hat{p} = \frac{475 + 150}{500 + 300} = 0.7813.$$

This gives us

$$Z_0 \;=\; \frac{\bar{T} - \bar{J}}{\sqrt{\hat{p}(1-\hat{p})\left[\frac{1}{n} + \frac{1}{m}\right]}} \;=\; \frac{0.95 - 0.5}{\sqrt{0.7813(0.2187)\left[\frac{1}{500} + \frac{1}{300}\right]}} \;=\; 14.91.$$

Since $|Z_0| > z_{0.025} = 1.96$, we easily reject H_0, and declare Tech the winners and JBSA the losers. ☺

23. (§7.5.1) Suppose we observe 10000 numbers to obtain the following data.

Interval	$[0.00, 0.25)$	$[0.25, 0.50)$	$[0.50, 0.75)$	$[0.75, 1.0]$
Number observed	2440	2552	2428	2580

Conduct a χ^2 goodness-of-fit test to see if these numbers are approximately Unif(0,1). Use level of significance $\alpha = 0.05$.

24. (§7.5.2) Suppose we're conducting a χ^2 goodness-of-fit test to determine whether or not $n = 200$ iid observations are from a Pearson distribution — which is a cool distribution having $s = 3$ parameters that must be estimated. Let's suppose that I have done some of the work for you and have divided the observations into $k = 5$ *equal-probability intervals* and have already tallied the number of observations that have fallen in each interval.

i	O_i	E_i
1	46	
2	33	
3	47	
4	34	
5	40	

If the level $\alpha = 0.05$, do we accept or reject the Pearson fit?

25. (§7.5.2) How many free throws do I have to take before I finally make the basket? Here is data that I collected from 70 trials.

# of throws needed	Frequency
1	34
2	18
3	2
4	9
5	7
	70

In other words, I got it on the first try 34 times, it took me two tries 18 times, etc. Let's test at level $\alpha = 0.05$ the hypothesis H_0: The number of required throws is Geom(p).

26. (§7.5.3) Recall that $\psi(r) \equiv \Gamma'(r)/\Gamma(r)$ is the *digamma function*. By hook or by crook, solve $\psi(r) = 0$. Hint: There are actually an infinite number of such zeroes — the easiest one to find is in the interval $(-1, 0)$, so I suggest that you search there.

Appendix A

Tables of Probability Distributions

We provide a cavalcade of distributions and their main properties (expected value, variance, pmf/pdf, cdf, and mgf), at least when available in succinct, closed form. (We have tried to maintain notation that is more-or-less consistent with usage in the main text.)

- Table A.1: Discrete distributions
  - Bernoulli(p)
  - Binomial(n, p)
  - Hypergeometric(a, b, n)
  - Geometric(p)
  - Negative Binomial(r, p)
  - Poisson(λ)
- Table A.2: Continuous distributions
  - Uniform(a, b)
  - Exponential(λ)
  - Erlang$_k$(λ)
  - Gamma(r, λ)
  - Triangular(a, b, c)
  - Beta(a, b)
  - Weibull(a, b)
  - Cauchy
  - Normal(μ, σ^2) and Standard Normal(0,1)
  - Lognormal(ν, τ^2)

- Table A.3: Continuous sampling distributions
  - $\chi^2(k)$
  - Student $t(k)$
  - $F(n, m)$

Table A.1: Discrete Distributions

Distribution	E[X]	Var(X)	pmf $f(x)$	mgf $M_X(t)$
Bernoulli(p)	p	pq	$\begin{cases} 1 & \text{w.p. } p \text{ (“success”)} \\ 0 & \text{w.p. } q = 1-p \text{ (“failure”)} \end{cases}$	$pe^t + q$
Binomial(n, p)	np	npq	$\binom{n}{x} p^x q^{n-x}, \;\; x = 0, 1, \ldots, n$	$(pe^t + q)^n$
HyperGeom(a, b, n)	$\frac{na}{N}$, $N = a + b$	$n\left(\frac{a}{N}\right)\left(1 - \frac{a}{N}\right)\left(\frac{N-n}{N-1}\right)$, $N = a + b$	$\binom{a}{x}\binom{b}{n-x} / \binom{a+b}{n}$, $x = \max\{0, n-b\}, \ldots, \min\{a, n\}$	
Geometric(p)	$1/p$	q/p^2	$q^{x-1}p, \;\; x = 1, 2, \ldots$	$\frac{pe^t}{1-qe^t}, \;\; t < \ell\text{n}(1/q)$
NegBin(r, p)	r/p	rq/p^2	$\binom{x-1}{r-1} p^r q^{x-r}, \;\; x = r, r+1, \ldots$	$\left(\frac{pe^t}{1-qe^t}\right)^r, \;\; t < \ell\text{n}(1/q)$
Poisson(λ)	λ	λ	$e^{-\lambda}\lambda^x / x!, \;\; x = 0, 1, 2, \ldots$	$e^{\lambda(e^t - 1)}$

Table A.2: Continuous Distributions

Distribution	E[X]	Var(X)	pdf $f(x)$	cdf $F(x)$	mgf $M_X(t)$
Uniform(a,b)	$\frac{a+b}{2}$	$\frac{(a-b)^2}{12}$	$\frac{1}{b-a},\ \ a \le x \le b$	$\frac{x}{b-a}$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$
Exponential(λ)	$1/\lambda$	$1/\lambda^2$	$\lambda e^{-\lambda x},\ \ x > 0$	$1-e^{-\lambda x}$	$\frac{\lambda}{\lambda-t},\ \ t < \lambda$
Erlang$_k(\lambda)$	k/λ	k/λ^2	$\frac{\lambda^k e^{-\lambda x} x^{k-1}}{(k-1)!},\ \ x \ge 0$	$1-\sum_{i=0}^{k-1} \frac{e^{-\lambda x}(\lambda x)^i}{i!}$	$\left(\frac{\lambda}{\lambda-t}\right)^k,\ \ t < \lambda$
Gamma(r,λ)	r/λ	r/λ^2	$\frac{\lambda^r e^{-\lambda x} x^{r-1}}{\Gamma(r)},\ \ x \ge 0$		$\left(\frac{\lambda}{\lambda-t}\right)^r,\ \ t < \lambda$
Triangular(a,b,c)	$(a+b+c)/3$	mess	$\begin{cases} \frac{2(x-a)}{(b-a)(c-a)}, & a < x \le b \\ \frac{2(c-x)}{(c-b)(c-a)}, & b < x < c \end{cases}$		
Beta(a,b)	$a/(a+b)$	$\frac{ab}{(a+b)^2(a+b+1)}$	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1},\ 0 < x < 1$		
Weibull(a,b)	$\frac{1}{a}\Gamma(\frac{b+1}{b})$	$\frac{1}{a^2}[\Gamma(\frac{b+2}{b}) - \Gamma^2(\frac{b+1}{b})]$	$ab(ax)^{b-1}e^{-(ax)^b},\ \ x > 0$	$1-\exp[-(ax)^b]$	
Cauchy	undefined	infinite	$\frac{1}{\pi(1+x^2)},\ \ x \in \mathbb{R}$	$\frac{1}{2} + \frac{\arctan(x)}{\pi}$	doesn't exist
Normal(μ,σ^2)	μ	σ^2	$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right],\ \ x \in \mathbb{R}$	$\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(s-\mu)^2}{2\sigma^2}}\, ds$	$e^{(\mu t + \frac{1}{2}\sigma^2 t^2)}$
Standard Normal	0	1	$\phi(x) \equiv \frac{1}{\sqrt{2\pi}} e^{-x^2/2},\ \ x \in \mathbb{R}$	$\Phi(x) \equiv \int_{-\infty}^{x} \phi(s)\, ds$	$e^{t^2/2}$
Lognormal(ν,τ^2)	$e^{\nu+\frac{\tau^2}{2}}$	$e^{2\nu+\tau^2}(e^{\tau^2}-1)$	$\frac{1}{x\tau}\phi\left(\frac{\ell n(x)-\nu}{\tau}\right),\ \ x > 0$	$\Phi\left(\frac{\ell n(x)-\nu}{\tau}\right)$	doesn't exist

Table A.3: Sampling Distributions

Distribution	E[X]	Var(X)	pdf $f(x)$	cdf $F(x)$
$\chi^2(k)$	k	$2k$	$\frac{1}{2^{k/2}\Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-x/2}$, $x > 0$	Table B.3
Student $t(k)$	0, $k > 1$	$k/(k-2)$, $k > 2$	$\frac{\Gamma(\frac{k+1}{2})}{\sqrt{\pi k}\,\Gamma(\frac{k}{2})} \left(\frac{x^2}{k} + 1\right)^{-(k+1)/2}$, $x \in \mathbb{R}$	See Table B.2
$F(n, m)$	$m/(m-2)$, $m > 2$	$\frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}$, $m > 4$	$\frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \frac{(\frac{n}{m})^{\frac{n}{2}} x^{\frac{n}{2}-1}}{\left(\frac{nx}{m}+1\right)^{\frac{n+m}{2}}}$, $x > 0$	See Tables B.4–B.7

Appendix B

Quantile and cdf Tables

We provide various tables that might be of use in case you don't have a computer handy. Of course, these tables don't cover every possible probability or quantile, but they will give you sustenance if you are stranded on a desert island. If you happen to have your computer wherever you go, you can always use Excel, Minitab, R, or any other reasonable software package.

- Table B.1: $\mathrm{Nor}(0,1)$ cdf $\Phi(z) = \mathrm{P}\big(\mathrm{Nor}(0,1) \leq z\big)$.
- Table B.2: $t(k)$ distribution $(1-p)$ quantile $t_{p,k}$ such that $\mathrm{P}\big(t(k) > t_{p,k}\big) = p$.
- Table B.3: $\chi^2(k)$ distribution $(1-p)$ quantile $\chi^2_{p,k}$ such that $\mathrm{P}\big(\chi^2(k) > \chi^2_{p,k}\big) = p$.
- Table B.4: $F(n,m)$ distribution 99% quantile $F_{0.01,n,m}$ such that $\mathrm{P}\big(F(n,m) > F_{0.01,n,m}\big) = 0.01$.
- Table B.5: $F(n,m)$ distribution 97.5% quantile $F_{0.025,n,m}$ such that $\mathrm{P}\big(F(n,m) > F_{0.025,n,m}\big) = 0.025$.
- Table B.6: $F(n,m)$ distribution 95% quantile $F_{0.05,n,m}$ such that $\mathrm{P}\big(F(n,m) > F_{0.05,n,m}\big) = 0.05$.
- Table B.7: $F(n,m)$ distribution 90% quantile $F_{0.1,n,m}$ such that $\mathrm{P}\big(F(n,m) > F_{0.1,n,m}\big) = 0.1$.

Table B.1: Nor(0, 1) cdf $\Phi(z) = \mathrm{P}\big(\mathrm{Nor}(0,1) \le z\big)$, e.g., $\mathbf{\Phi(1.47) = 0.9292}$. Note that the $(1-p)$-quantile $z_p \equiv \Phi^{-1}(1-p)$.

	0.00	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	**0.9292**	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table B.2: t distribution $(1-p)$ quantile $t_{p,k}$ such that $\mathrm{P}\big(t(k) > t_{p,k}\big) = p$, e.g., $t_{0.005,17} = 2.898$.

$k \backslash p$	**0.40**	**0.30**	**0.20**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**	**0.0005**
1	0.325	0.727	1.376	3.078	6.314	12.71	31.821	63.66	318.3	636.6
2	0.289	0.617	1.061	1.886	2.920	4.303	6.965	9.925	22.33	31.60
3	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841	10.21	12.92
4	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.262	0.546	0.889	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.257	0.534	0.863	1.333	1.740	2.110	2.567	**2.898**	3.646	3.965
18	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750	3.385	3.646
31	0.256	0.530	0.853	1.309	1.696	2.040	2.453	2.744	3.375	3.633
32	0.255	0.530	0.853	1.309	1.694	2.037	2.449	2.738	3.365	3.622
33	0.255	0.530	0.853	1.308	1.692	2.035	2.445	2.733	3.356	3.611
34	0.255	0.529	0.852	1.307	1.691	2.032	2.441	2.728	3.348	3.601
35	0.255	0.529	0.852	1.306	1.690	2.030	2.438	2.724	3.340	3.591
40	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	0.255	0.528	0.849	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.254	0.526	0.846	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.254	0.526	0.845	1.290	1.660	1.984	2.364	2.626	3.174	3.390
200	0.254	0.525	0.843	1.286	1.653	1.972	2.345	2.601	3.131	3.340

Table B.3: χ^2 distribution $(1-p)$ quantile $\chi^2_{p,k}$ such that $\mathrm{P}\left(\chi^2(k) > \chi^2_{p,k}\right) = p$, e.g., $\chi^2_{0.025,12} = 23.34$.

$k\backslash p$	**0.999**	**0.995**	**0.99**	**0.975**	**0.95**	**0.9**	**0.1**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
1	0.000	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879	10.83
2	0.002	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.60	13.82
3	0.024	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.34	12.84	16.27
4	0.091	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.14	13.28	14.86	18.47
5	0.210	0.412	0.554	0.831	1.145	1.610	9.236	11.07	12.83	15.09	16.75	20.52
6	0.381	0.676	0.872	1.237	1.635	2.204	10.64	12.59	14.45	16.81	18.55	22.46
7	0.598	0.989	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48	20.28	24.32
8	0.857	1.344	1.646	2.180	2.733	3.490	13.36	15.51	17.53	20.09	21.95	26.12
9	1.152	1.735	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67	23.59	27.88
10	1.479	2.156	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21	25.19	29.59
11	1.834	2.603	3.053	3.816	4.575	5.578	17.28	19.68	21.92	24.72	26.76	31.26
12	2.214	3.074	3.571	4.404	5.226	6.304	18.55	21.03	**23.34**	26.22	28.30	32.91
13	2.617	3.565	4.107	5.009	5.892	7.042	19.81	22.36	24.74	27.69	29.82	34.53
14	3.041	4.075	4.660	5.629	6.571	7.790	21.06	23.68	26.12	29.14	31.32	36.12
15	3.483	4.601	5.229	6.262	7.261	8.547	22.31	25.00	27.49	30.58	32.80	37.70
16	3.942	5.142	5.812	6.908	7.962	9.312	23.54	26.30	28.85	32.00	34.27	39.25
17	4.416	5.697	6.408	7.564	8.672	10.09	24.77	27.59	30.19	33.41	35.72	40.79
18	4.905	6.265	7.015	8.231	9.390	10.86	25.99	28.87	31.53	34.81	37.16	42.31
19	5.407	6.844	7.633	8.907	10.12	11.65	27.20	30.14	32.85	36.19	38.58	43.82
20	5.921	7.434	8.260	9.591	10.85	12.44	28.41	31.41	34.17	37.57	40.00	45.31
21	6.447	8.034	8.897	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40	46.80
22	6.983	8.643	9.542	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80	48.27
23	7.529	9.260	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18	49.73
24	8.085	9.886	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56	51.18
25	8.649	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93	52.62
26	9.222	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29	54.05
27	9.803	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64	55.48
28	10.39	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99	56.89
29	10.99	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34	58.30
30	11.59	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67	59.70
31	12.20	14.46	15.66	17.54	19.28	21.43	41.42	44.99	48.23	52.19	55.00	61.10
32	12.81	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49	56.33	62.49
33	13.43	15.82	17.07	19.05	20.87	23.11	43.75	47.40	50.73	54.78	57.65	63.87
34	14.06	16.50	17.79	19.81	21.66	23.95	44.90	48.60	51.97	56.06	58.96	65.25
35	14.69	17.19	18.51	20.57	22.47	24.80	46.06	49.80	53.20	57.34	60.27	66.62
40	17.92	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77	73.40
50	24.67	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49	86.66
60	31.74	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95	99.61
80	46.52	51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.3	124.8
100	61.92	67.33	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	140.2	149.4
200	143.8	152.2	156.4	162.7	168.3	174.8	226.0	234.0	241.1	249.4	255.3	267.5

Table B.4: F distribution 99% quantile $F_{0.01,n,m}$ such that $\mathrm{P}(F(n,m) > F_{0.01,n,m}) = 0.01$, e.g., $F_{0.01,9,5} = 10.16$.

$m\backslash n$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**25**	**30**	**60**	**120**
1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6209	6240	6261	6313	6339
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.42	99.43	99.45	99.46	99.47	99.48	99.49
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	27.05	26.87	26.69	26.58	26.50	26.32	26.22
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.37	14.20	14.02	13.91	13.84	13.65	13.56
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	**10.16**	10.05	9.89	9.72	9.55	9.45	9.38	9.20	9.11
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.30	7.23	7.06	6.97
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.06	5.99	5.82	5.74
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.26	5.20	5.03	4.95
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.71	4.65	4.48	4.40
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.31	4.25	4.08	4.00
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.01	3.94	3.78	3.69
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.76	3.70	3.54	3.45
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.57	3.51	3.34	3.25
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.41	3.35	3.18	3.09
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.28	3.21	3.05	2.96
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.16	3.10	2.93	2.84
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.07	3.00	2.83	2.75
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	2.98	2.92	2.75	2.66
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.91	2.84	2.67	2.58
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.84	2.78	2.61	2.52
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.79	2.72	2.55	2.46
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.73	2.67	2.50	2.40
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.69	2.62	2.45	2.35
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.64	2.58	2.40	2.31
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.60	2.54	2.36	2.27
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.45	2.39	2.21	2.11
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.27	2.20	2.02	1.92
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.10	2.03	1.84	1.73
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.93	1.86	1.66	1.53

Table B.5: F distribution 97.5% quantile $F_{0.025,n,m}$ such that $\mathrm{P}(F(n,m) > F_{0.025,n,m}) = 0.025$, e.g., $F_{0.025,9,5} = 6.68$.

$m\backslash n$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**25**	**30**	**60**	**120**
1	647.8	799.5	864.2	899.6	921.8	937.1	948.2	956.7	963.3	968.6	976.7	984.9	993.1	998.1	1001	1010	1014
2	38.51	39.00	39.17	39.25	39.30	39.33	39.36	39.37	39.39	39.40	39.41	39.43	39.45	39.46	39.46	39.48	39.49
3	17.44	16.04	15.44	15.10	14.88	14.73	14.62	14.54	14.47	14.42	14.34	14.25	14.17	14.12	14.08	13.99	13.95
4	12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.75	8.66	8.56	8.50	8.46	8.36	8.31
5	10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	**6.68**	6.62	6.52	6.43	6.33	6.27	6.23	6.12	6.07
6	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.37	5.27	5.17	5.11	5.07	4.96	4.90
7	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.67	4.57	4.47	4.40	4.36	4.25	4.20
8	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.20	4.10	4.00	3.94	3.89	3.78	3.73
9	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.87	3.77	3.67	3.60	3.56	3.45	3.39
10	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.62	3.52	3.42	3.35	3.31	3.20	3.14
11	6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.43	3.33	3.23	3.16	3.12	3.00	2.94
12	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.28	3.18	3.07	3.01	2.96	2.85	2.79
13	6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.15	3.05	2.95	2.88	2.84	2.72	2.66
14	6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	3.05	2.95	2.84	2.78	2.73	2.61	2.55
15	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.96	2.86	2.76	2.69	2.64	2.52	2.46
16	6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.89	2.79	2.68	2.61	2.57	2.45	2.38
17	6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.82	2.72	2.62	2.55	2.50	2.38	2.32
18	5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.77	2.67	2.56	2.49	2.44	2.32	2.26
19	5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.72	2.62	2.51	2.44	2.39	2.27	2.20
20	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.68	2.57	2.46	2.40	2.35	2.22	2.16
21	5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.64	2.53	2.42	2.36	2.31	2.18	2.11
22	5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.60	2.50	2.39	2.32	2.27	2.14	2.08
23	5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.57	2.47	2.36	2.29	2.24	2.11	2.04
24	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.54	2.44	2.33	2.26	2.21	2.08	2.01
25	5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.51	2.41	2.30	2.23	2.18	2.05	1.98
30	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.41	2.31	2.20	2.12	2.07	1.94	1.87
40	5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.29	2.18	2.07	1.99	1.94	1.80	1.72
60	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.17	2.06	1.94	1.87	1.82	1.67	1.58
120	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	2.05	1.94	1.82	1.75	1.69	1.53	1.43

Table B.6: F distribution 95% quantile $F_{0.05,n,m}$ such that $\mathrm{P}(F(n,m) > F_{0.05,n,m}) = 0.05$, e.g., $F_{0.05,9,5} = 4.77$.

$m\backslash n$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**25**	**30**	**60**	**120**
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.3	250.1	252.2	253.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.46	19.46	19.48	19.49
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.63	8.62	8.57	8.55
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.69	5.66
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	**4.77**	4.74	4.68	4.62	4.56	4.52	4.50	4.43	4.40
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.83	3.81	3.74	3.70
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.40	3.38	3.30	3.27
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.11	3.08	3.01	2.97
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.89	2.86	2.79	2.75
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.73	2.70	2.62	2.58
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.60	2.57	2.49	2.45
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.50	2.47	2.38	2.34
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.41	2.38	2.30	2.25
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.34	2.31	2.22	2.18
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.28	2.25	2.16	2.11
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.23	2.19	2.11	2.06
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.18	2.15	2.06	2.01
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.14	2.11	2.02	1.97
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	1.98	1.93
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.07	2.04	1.95	1.90
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.92	1.87
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.02	1.98	1.89	1.84
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.00	1.96	1.86	1.81
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.97	1.94	1.84	1.79
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.82	1.77
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.88	1.84	1.74	1.68
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.78	1.74	1.64	1.58
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.69	1.65	1.53	1.47
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.60	1.55	1.43	1.35

Table B.7: F distribution 90% quantile $F_{0.1,n,m}$ such that $\mathrm{P}(F(n,m) > F_{0.1,n,m}) = 0.1$, e.g., $F_{0.1,9,5} = 3.32$.

$m\backslash n$	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**25**	**30**	**60**	**120**
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.05	62.26	62.79	63.06
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.48
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.17	5.17	5.15	5.14
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.79	3.78
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	**3.32**	3.30	3.27	3.24	3.21	3.19	3.17	3.14	3.12
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.81	2.80	2.76	2.74
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.57	2.56	2.51	2.49
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.50	2.46	2.42	2.40	2.38	2.34	2.32
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.27	2.25	2.21	2.18
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.17	2.16	2.11	2.08
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.03	2.00
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.03	2.01	1.96	1.93
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.90	1.88
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.93	1.91	1.86	1.83
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.89	1.87	1.82	1.79
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.86	1.84	1.78	1.75
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.83	1.81	1.75	1.72
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.80	1.78	1.72	1.69
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.78	1.76	1.70	1.67
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.76	1.74	1.68	1.64
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.74	1.72	1.66	1.62
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.64	1.60
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.71	1.69	1.62	1.59
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.61	1.57
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.82	1.77	1.72	1.68	1.66	1.59	1.56
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.77	1.72	1.67	1.63	1.61	1.54	1.50
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.71	1.66	1.61	1.57	1.54	1.47	1.42
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.66	1.60	1.54	1.50	1.48	1.40	1.35
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.60	1.55	1.48	1.44	1.41	1.32	1.26

Bibliography

[1] W. J. Cook (2012). *In Pursuit of the Traveling Salesman: Mathematics at the Limits of Computation*, Princeton University Press, Princeton. [This book explains one of the most-famous problems in combinatorial optimization at an intuitive level. Although there is no discussion on probability and statistics in this text, readers will be entertained by the history of the problem and will certainly get some practice on elementary proof techniques.]

[2] H. Enderton (1977). *Elements of Set Theory*, Academic Press, San Diego. [This text covers set theory at an undergraduate level, including the development of Cantor's findings on cardinality.]

[3] I. S. Gradshteyn and I. M. Ryzhik (2014). *Table of Integrals, Series, and Products* (ed. D. Zwillinger), 8th Edition, Elsevier Inc., Amsterdam. [This is the most-comprehensive resource on integrals and related functions; and it makes for great bedtime reading!]

[4] W. W. Hines, D. C. Montgomery, D. Goldsman, and C. M. Borror (2003). *Probability and Statistics in Engineering*, 4th Edition, John Wiley and Sons, Hoboken, NJ. [For readers who would like more details on certain aspects of applied statistics.]

[5] P. L. Meyer (1970). *Introductory Probability and Statistical Applications*, 2nd Edition, Addison-Wesley, Reading, MA. [For readers who would like more details on certain aspects of applied probability.]

[6] G. Polya (2004). *How To Solve It: A New Aspect of Mathematical Method* (with a new foreword by John H. Conway), Princeton University Press, Princeton. [This is the "must-read" text if you want to learn how to prove things. A classic!]

[7] S. M. Ross (2019). *Introduction to Probability Models*, 12th Edition, Elsevier Inc., Amsterdam. [The is the go to book for masters-level stochastic processes and applied probability, and is encyclopedic in scope, with hundreds of interesting examples and problems.]

Index