Statistics for Calculus Students

Bjørn Kjos-Hanssen
Professor
University of Hawai'i at Mānoa
bjoernkh@hawaii.edu

Samuel D. Birns
Graduate Student
University of Hawai'i at Mānoa
sbirns@hawaii.edu

Contents

Preface

This book is a supplement to OpenIntro Statistics, which may be downloaded as a free PDF at **openintro.org**.

By choosing the title *Statistics for Calculus Students* we intended to summarize the following prerequisite situation.

(1) Students should have studied some calculus, say Calculus 1 and 2.

(2) Students could still be studying Calculus — for instance, Calculus 3 (multivariable) is not assumed.

(3) Although it is essential for a full understanding of statistics, *linear algebra* is not required to read this book.

In particular, the book is designed to be appropriate for MATH 372 (Elementary Probability and Statistics) at University of Hawai'i at Mānoa.
This is not a textbook on probability. While we review some basic concepts, we assume that students have some knowledge of probability as can be gained through Chapters 1–9 of Grinstead and Snell's *Introduction to Probability*[1]. Thus, we assume the student has seen some single-variable calculus-based probability, and some algebra-based statistics; and we intend to bridge the gap to single-variable calculus-based statistics.

Textbook overview

The chapters of this book are as follows:

1. **Introduction to data.** Data structures, variables, summaries, graphics, and basic data collection techniques.
2. **Probability.** The basic principles of probability.
3. **Distributions of random variables.** Introduction to the normal model and other key distributions.
4. **Foundations for inference.** General ideas for statistical inference in the context of estimating the population mean.
5. **Inference for numerical data.** Inference for one or two sample means using the t-distribution.
6. **Inference for categorical data.** Inference for proportions using the normal and χ^2 distributions, as well as simulation and randomization techniques.

[1] Free download available at http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/amsbook.mac.pdf.

7. **Introduction to linear regression.** An introduction to regression with two variables.
8. **Special topics.** Introduces maximum likelihood models and their applications to finding a hidden Markov model. Algebraic statistics, entropy, likelihood, Markov chains, and information criteria.

Examples, exercises, and appendices

Examples and Guided Practice throughout the textbook may be identified by their distinctive bullets:

● **Example 0.1** Large filled bullets signal the start of an example.

Full solutions to examples are provided and may include an accompanying table or figure.

⊙ **Guided Practice 0.2** Large empty bullets signal to readers that an exercise has been inserted into the text for additional practice and guidance. Students may find it useful to fill in the bullet after understanding or successfully completing the exercise. Solutions are provided for all Guided Practice in footnotes.[2]

There are exercises at the end of each chapter for practice or homework assignments. Odd-numbered exercise solutions are in Appendix A.

Google Sheets

Occasional references to technology are made, in which cases our software of choice is `Google Sheets`. This choice was made on the basis of Google Sheets being freely available and only a few clicks away from University of Hawai‘i students' Gmail accounts.

Acknowledgements

This project would not be possible without the passion and dedication of all those involved. The authors would like to thank faculty members in the Department of Mathematics at University of Hawai‘i at Mānoa for their involvement and ongoing contributions. We are also very grateful to the students who took our classes and provided us with valuable feedback over the last several years.

We are thankful to Outreach College at University of Hawai‘i at Mānoa for support under their Open Educational Resources grant program.

[2] Full solutions are located down here in the footnote!

Chapter 1

Introduction to data

1.1 Continuous and discrete

The distinction between continuous and discrete data is common in mathematics. *Continuous* refers to real numbers drawn from the set of all such, $\mathbb{R}$. *Discrete* refers to a finite or countably infinite set such as the natural numbers $\mathbb{N}$.

If salaries are measured in integer multiples of \$100, they can be considered discrete; in fact, a salary is presumably always an integer multiple of 1 cent. However, it is somewhat artificial to consider a salary as discrete, as only convenience stops us from using salaries involving fractions of a cent.

Example 1.1 Impossible salaries

Estelle has \$100,000 to spend on salaries for three employees. Since the employees are indistinguishable, other things being equal, she decides to give them each a salary of

$$\$\frac{100,000}{3}.$$

Example 1.2 Discrete time?

Time is usually measured in real numbers, although it is not known for sure in theoretical physics whether time is actually discrete[1]. This would mean that there is a limit to how small time steps one can take.

1.2 Categorical variables

Within categorical variables, the ordinal ones are ordered. Note that one can imagine variables whose values are structured, without being numeric or ordinal: for instance, consider `direction`:

> North, Northwest, West, Southwest, South, Southeast, East, Northeast.

These are naturally ordered in a circle or loop.

[1] https://physics.stackexchange.com/questions/35674/is-time-continuous-or-discrete

⊙ **Guided Practice 1.3** What could be another example of a structured variable, which is nevertheless not numerical or ordinal? Maybe another variable that is naturally ordered in a "wheel"?[2]

⊙ **Guided Practice 1.4** What kind of mathematical structure can you find in the variable "state of residence" among the U.S. states?[3]

1.3 Examining numerical data

Mean

The sample mean of a numerical variable is computed as the sum of all of the observations divided by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} \tag{1.5}$$

where $x_1, x_2, \ldots, x_n$ represent the n observed values.

n
sample size

This arithmetic mean has a very important property: suppose you want to load goods onto a ship, airplane, or car. If you know the mean weight of the goods, and you know how many there are of them, then you know the total weight. Some other kinds of "mean" are developed in Exercise 1.5. The geometric mean studied there could be similarly applied: if you know the geometric mean of the factors by which a stock price changes, and you know the number of changes, then you know the overall change.

A **mode** is represented by a prominent peak in the distribution. To be mathematically precise, we could let the mode be the value with the most occurrences. But it is common to have *no* observations with the same value in a data set, which makes this other definition useless for many real data sets. See Chapter 2 for a definition of mode in that case.

s^2
sample variance

We define as follows the sample **variance**, denoted by s^2:

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

We divide by $n-1$, rather than dividing by n. To see why, consult Chapter 2. The **sample standard deviation** is then $s = \sqrt{s^2}$.

The reason for using squares rather than, say, fourth powers in the definition of variance is related to Pythagoras Theorem and ideas about orthogonality. It turns out that for random variables X and Y that are independent (which is analogous to orthogonal), we have $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$, analogously to how if a and b are side lengths of orthogonal sides in a triangle, then the length of the hypothenuse is $\sqrt{a^2+b^2}$.

The standard deviation has an advantage over the variance in that, when used in physical examples, it has the same physical units as the original data. This also means

[2] How about a kids' favorite: color. Colors are ordered cyclically as

Red, Orange, Yellow, Green, Blue, Purple.

[3] For instance, consider whether two states are neighbors; this will give a graph structure as studied in discrete mathematics.

that it should make sense to add a fixed number of standard deviations to the data points (where it would make little sense too a number of variances).

Would it make sense to consider, say, twice the standard deviation 2σ instead of the standard deviation σ? Yes, and in fact there is a vigorous debate in some circles about the merits of $pi = 3.14\ldots$ versus $\tau = 2\pi$. In the normal probability density function, there is a factor $\sqrt{2\pi}\sigma$, so that arguably it would look simpler if we used $\sqrt{2\pi}\sigma$ instead of σ. At some point, we just have to assert that what constant to put in front is a matter of convention.

1.3.1 About those pie charts

A **pie chart** is shown in Figure 1.1 alongside a bar plot. The much maligned pie charts are actually useful when you want to see at a glance if a proportion is close to 50% or 25%.

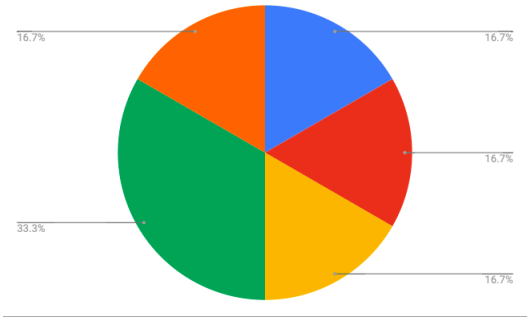


Figure 1.1: A pie chart for the data 1,1,1,2,1.

⊙ **Guided Practice 1.6** Could any other shapes than discs be as useful for "pie" charts?[4]

1.4 Sampling

This topic is treated extensively in *OpenIntro Statistics* and we confine ourselves to some remarks.

A *simple random sample* is one in which each set of n outcomes is equally likely to appear. For instance, if we sample $n = 2$ individuals from $\{1, 2, 3\}$, we would need to know that $\{1, 2\}$, $\{2, 3\}$, and $\{1, 3\}$ are equally likely. It is not enough that, say, 1, 2, and 3 each have probability $1/3$ of appearing in our set.

If two samples of individuals are drawn from the same distribution, but then we *treat* the individuals in two different ways, we are *causing* a change and can conclude that, most likely, any difference in the results for the two samples was *caused* by our treatment. For observational studies we must always worry that there may be other explanations. Suppose countries with a lot of pianos also have low rates of malnutrition. We cannot conclude that there is a causal relationship. In this example it seems silly to think there would be one.

[4] A hexagonal chart could be useful for identifying whether a proportion was greater than 1/6, for instance.

1.5 Exercises

1.1 Category of categories. Categorize the following statements as true or false and briefly explain why.

(a) A variable can be either categorical or numerical but not both.

(b) A variable can be either discrete or continuous but not both.

(c) There exists a categorical variable that can be expressed as an equivalent continuous variable.

1.2 Mean medians. Consider the following set of exam scores: 70, 55, 76, 64, 98, 71, 68, 70, 76, 59.

(a) Find the mean and median of the exam scores. Which more accurately describes the data set?

(b) Generalize this idea and discuss what conditions on a data set would result in a mean that does not fit the pattern of the given data.

(c) Similarly, in what data sets would the mean be higher than the median? Lower than the median?

1.3 Extra mean. Consider the exam scores from Problem 1.2.

(a) Assume the instructor added 10 points of extra credit to each student's score (and assume that exam scores can be greater than 100). What is the mean of the new scores?

(b) Assume that the instructor decided to make the exam out of 200 points instead of 100 and, consequently, doubled each student's raw score. What is the mean of the new scores? Which way makes the greatest change to the mean, the one in (a) or (b)?

(c) Generalize this pattern: if a constant c_1 is added to each of $x_1, \ldots, x_n$, what will the new mean be? If each of $x_1, \ldots, x_n$ is multiplied by a constant c_2, what will the new mean be?

1.4 Extra variance. Same as Problem 1.3, but with variance instead of mean.

1.5 Harmonic mean. There are several ways to define the "average" of a set of data. Given data $x_1, \ldots, x_n$, let $\bar{x}_A$ denote the arithmetic (or "usual") mean:

$$\bar{x}_A := \frac{x_1 + \cdots + x_n}{n}$$

Let $\bar{x}_G$ denote the geometric mean:

$$\bar{x}_G := \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_{n-1} \cdot x_n}$$

Let $\bar{x}_H$ denote the harmonic mean:

$$\bar{x}_H := \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

Show that $\bar{x}_A \geq \bar{x}_G \geq \bar{x}_H$ for any $x_1, \ldots, x_n$, with equality if and only if $x_i = x_j$ for all $i, j \leq n$.

1.6 Continuous averages. Let $\mathbb{R} = (-\infty, \infty)$ be the set of all real numbers. The average of a function $f : [a, b] \to \mathbb{R}$ can be defined via

$$\text{avg}_f := \frac{1}{b-a} \int_a^b f(x)dx$$

assuming this integral converges. Given data $x_1, \ldots, x_n$, find a function f such that $\text{avg}_f = \bar{x}_A$.

1.7 Invariant average. Let $f : [a, b] \to \mathbb{R}$ and assume $\int_a^b f(x)dx$ converges. Show that the properties of the mean from Problem 1.3(c) hold for avg_f from Problem 1.6.

1.8 A valuable theorem. Let $f : [a, b] \to \mathbb{R}$ be continuous and assume $\int_a^b f(x)dx$ converges.

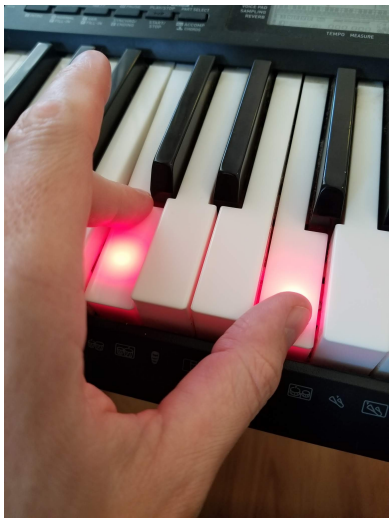
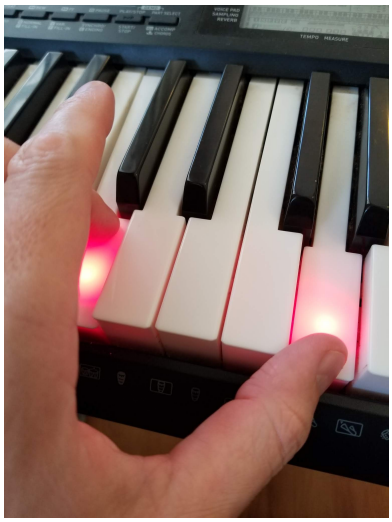
(a) Cite a well-known calculus theorem to conclude that there exists $c \in (a, b)$ such that $f(c) = \text{avg}_f$.

(b) How does the existence of c from part (a) agree with the intuition behind averages?

1.9 Musical means. Show that the geometric means (Exercise 1.5) of $\{\frac{4}{3}, \frac{3}{2}\}$ and $\{2^{5/12}, 2^{7/12}\}$ are the same, but that

$$\frac{4}{3} < 2^{5/12} < 2^{7/12} < \frac{3}{2}.$$

The relationship between these particular numbers is important in music theory[5]. For musical harmony, rational numbers like $\frac{4}{3}, \frac{3}{2}$ are desired, but for equally tempered scales, powers of $2^{1/12}$ are desired.

[5] See https://math.stackexchange.com/questions/11669/mathematical-difference-between-white-and-black-notes-in-a-11671.

Chapter 2

Probability

2.1 Law of large numbers

2.1.1 Probability and the real world

Philosophically, probability can be connected to the real world via a principle from 1843:

Cournot's principle

Very, very unlikely events simply do not happen.

Some philosophers and others may disagree with this principle; we shall not try to settle the debate. But note that if we just say

Very, very unlikely events will happen very, very rarely...

then it begs the question: what if they don't? What if 100 heads in a row occur when you toss a coin — and the same happens to your friend? And your friend's friend? This should happen rarely indeed, but at what point do you put your foot down and declare it just should not happen? If we do not adopt Cournot's principle, probability theory is in danger of losing touch with reality.

Probability

The **probability** of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.

Probability is defined as a proportion, and it always takes values between 0 and 1 (inclusively). It may also be displayed as a percentage between 0% and 100%.

Probability can be illustrated by rolling a die many times. Let $\hat{p}_n$ be the proportion of outcomes $X_1, \ldots, X_n$ that are `1` after the first n rolls:

$$\hat{p}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

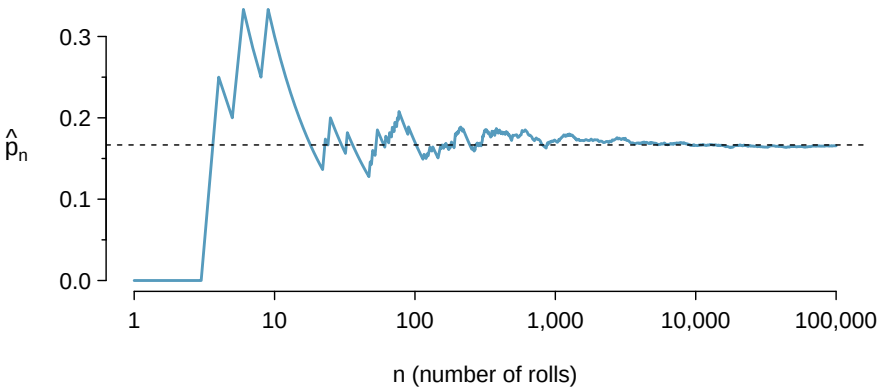


Figure 2.1: The fraction of die rolls that are 1 at each stage in a simulation. The proportion tends to get closer to the probability $1/6 \approx 0.167$ as the number of rolls increases.

As the number of rolls increases, $\hat{p}_n$ will converge to the probability of rolling a `1`, $p = 1/6$. Figure 2.1 shows this convergence for 100,000 die rolls. The tendency of $\hat{p}_n$ to stabilize around p is described by the **Law of Large Numbers**.

Law of Large Numbers

As more observations are collected, the proportion $\hat{p}_n$ of occurrences with a particular outcome converges to the probability p of that outcome.

Occasionally the proportion will veer off from the probability and appear to defy the Law of Large Numbers, as $\hat{p}_n$ does many times in Figure 2.1. However, these deviations become smaller as the number of rolls increases.

Above we write p as the probability of rolling a `1`. We can also write this probability as

$$\mathbb{P}(\text{rolling a } \texttt{1})$$

$\mathbb{P}(A)$
Probability of outcome A

As we become more comfortable with this notation, we will abbreviate it further. For instance, if it is clear that the process is "rolling a die", we could abbreviate $\mathbb{P}(\text{rolling a } \texttt{1})$ as $\mathbb{P}(\texttt{1})$.

The Law of Large Numbers (LLN) comes in two flavors. The **strong LLN** says that for an infinite sequence of observations, with probability 1 the sequence $\hat{p}_n$ converges to p:

$$\mathbb{P}(\lim_{n\to\infty} \hat{p}_n = p) = 1.$$

The **weak LLN** says that as $n \to \infty$, the probability that $|\hat{p}_n - p| \geq \epsilon$ goes to 0, no matter how small ϵ is:

$$\lim_{n\to\infty} \mathbb{P}(\hat{p}_n - p| < \epsilon) = 1.$$

Both are true. The strong LLN is stronger (it says more, and it implies the weak LLN) but the strong LLN requires consideration of infinite sequences of outcomes, hence in applications it is sometimes conceptual "overkill". For students familiar with ϵ-δ proofs, we offer:

Theorem 2.1. *The strong LLN implies the weak LLN.*

Proof. Let $\epsilon > 0$ be given. Since $\mathbb{P}(\lim_{n\to\infty} \hat{p}_n = p) = 1$, we have

$$\mathbb{P}(\text{there is an } m \text{ such that for all } n \geq m,\ |\hat{p}_n - p| < \epsilon) = 1.$$

Therefore for each $\delta > 0$ there is an m such that $\mathbb{P}(\text{for all } n \geq m,\ |\hat{p}_n - p| < \epsilon) \geq 1 - \delta$. In particular, for all $n \geq m$, $\mathbb{P}(|\hat{p}_n - p| < \epsilon) \geq 1 - \delta$. □

2.1.2 Disjoint or mutually exclusive outcomes

Events are **sets** of **outcomes**[1]. You have perhaps encountered basic set theory in earlier mathematics courses.

Two events A and B are called **disjoint** or **mutually exclusive** if they cannot both happen in the sense that they have no outcome in their intersection $A \cap B$ ("A and B"). In this case we have $A \cap B = \emptyset$, the empty set. Thus, with $\cup$ or union meaning "(inclusive) or", the **Addition Rule** says

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B),$$

and if $A \cap B = \emptyset$ then $\mathbb{P}(A \cap B) = \mathbb{P}(\emptyset) = 0$.

We often abbreviate set notation in probability statements. Instead of

$$\mathbb{P}(\{\omega : X(\omega) \leq 3\})$$

we may just write $\mathbb{P}(X \leq 3)$ or even $\mathbb{P}(1 \text{ or } 2 \text{ or } 3)$ if the only possible outcomes that are at most 3 are 1, 2, and 3. Here the outcomes ω are drawn from the **sample space** Ω which is a set consisting of all possible outcomes.

Addition Rule of disjoint outcomes

If A_1 and A_2 represent two disjoint outcomes, then the probability that one of them occurs is given by

$$\mathbb{P}(A_1 \text{ or } A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2)$$

If there are many disjoint outcomes $A_1, \ldots, A_k$, then the probability that one of these outcomes will occur is

$$\mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots + \mathbb{P}(A_k) \tag{2.1}$$

Statisticians rarely work with individual outcomes and instead consider *sets* or *collections* of outcomes. Let A represent the event where a die roll results in `1` or `2` and B represent the event that the result is a `4` or a `6`. We write A as the set of outcomes $\{\texttt{1}, \texttt{2}\}$ and $B = \{\texttt{4}, \texttt{6}\}$. These sets are commonly called **events**. Because A and B have no elements in common, they are disjoint events. A and B are represented in Figure 2.2.

[1] The *OpenIntro Statistics 3rd edition* text is not sufficiently clear on this important distinction, nor are many other statistics texts.

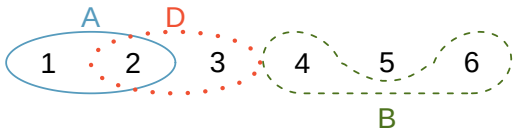


Figure 2.2: Three events, A, B, and D, consist of outcomes from rolling a die. A and B are disjoint since they do not have any outcomes in common. B and D are also disjoint.

Guided Practice 2.2 Can you modify D in Figure 2.2 to make A, B, D pairwise disjoint?[2]

General Addition Rule

If A and B are any two events, disjoint or not, then the probability that at least one of them will occur is

$$\mathbb{P}(A \text{ or } B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \text{ and } B) \tag{2.3}$$

where $\mathbb{P}(A \text{ and } B)$ is the probability that both events occur.

Example 2.4 S

uppose we toss two coins. Let A be the event that the 1st coin is heads, and B the event that the 2nd coin is heads. Then the probability that at least one coin is heads is $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) = \frac{1}{2} + \frac{1}{2} - \frac{1}{4} = \frac{3}{4}$.

2.1.3 Probability distributions

The **probability distribution** of a **random variable** $X : \Omega \to \mathbb{R}$ tells us what the probabilities of all events involving X are: $\mathbb{P}(X \leq 3)$, $\mathbb{P}(X = 5)$, etc.

For continuous random variables, the probability of $X \in \mathcal{A}$ is calculated as

$$\mathbb{P}(X \in \mathcal{A}) = \int_{x \in \mathcal{A}} f(x)\, dx$$

for a **probability density function** (pdf) f. In particular, the probability of any single particular outcome is zero. This may seem strange at first, but the idea can be seen as follows. If you measure somebody's height with infinite precision, the probability of anybody else having exactly the same height is zero.

The **cumulative distribution function** (cdf) F is defined by

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

[2] Yes, let $D = \{3, 5\}$, or more generally $D \subseteq \{3, 5\}$.

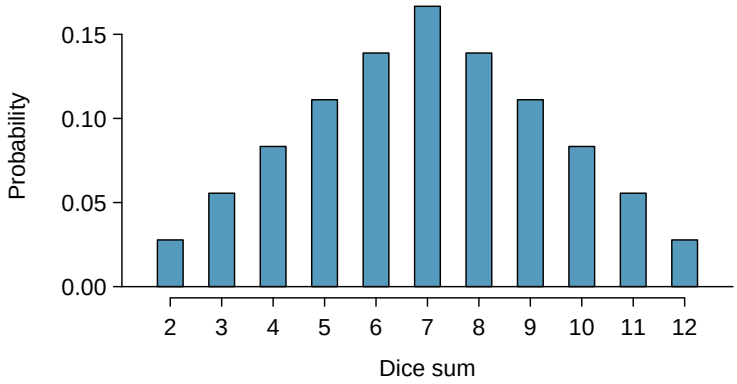


Figure 2.3: The probability distribution of the sum of two dice.

Rules for probability distributions

A probability distribution is a list of the possible outcomes with corresponding probabilities that satisfies three rules:

1. The outcomes listed must be disjoint.
2. Each probability must be between 0 and 1.
3. The probabilities must total 1.

In the bar plots Figure 2.3, the bar heights represent the probabilities of outcomes. If the outcomes are numerical and discrete, it is usually (visually) convenient to make a bar plot that resembles a histogram, as in the case of the sum of two dice.

2.1.4 Complement of an event

S
Sample space

Rolling a die produces a value in the set $\{1, 2, 3, 4, 5, 6\}$. This set of all possible outcomes is called the **sample space** (S) for rolling a die. We often use the sample space to examine the scenario where an event does not occur.

A^c
Complement of outcome A

Let $D = \{2, 3\}$ represent the event that the outcome of a die roll is 2 or 3. Then the **complement** of D represents all outcomes in our sample space that are not in D, which is denoted by $D^c = \{1, 4, 5, 6\}$. That is, D^c is the set of all possible outcomes not already included in D.

A complement of an event A is constructed to have two very important properties: (i) every possible outcome not in A is in A^c, and (ii) A and A^c are disjoint. Property (i) implies

$$\mathbb{P}(A \text{ or } A^c) = 1 \tag{2.5}$$

That is, if the outcome is not in A, it must be represented in A^c. We use the Addition Rule for disjoint events to apply Property (ii):

$$\mathbb{P}(A \text{ or } A^c) = \mathbb{P}(A) + \mathbb{P}(A^c) \tag{2.6}$$

Combining Equations (2.5) and (2.6) yields a very useful relationship between the probability of an event and its complement.

Complement

The complement of event A is denoted A^c, and A^c represents all outcomes not in A. A and A^c are mathematically related:

$$\mathbb{P}(A) + \mathbb{P}(A^c) = 1, \quad \text{i.e.} \quad \mathbb{P}(A) = 1 - \mathbb{P}(A^c) \tag{2.7}$$

In simple examples, computing A or A^c is feasible in a few steps. However, using the complement can save a lot of time as problems grow in complexity.

2.2 Independence and conditional probability

Conditional probability. Suppose we don't know the outcome of our probability experiment, we only know that the event $B \subseteq \Omega$ occurred. Then it makes sense to replace Ω by B, since the sample space should consist of all "possible" outcomes, and only the outcomes in B are still possible. The probabilities in this space should be given by

$$\mathbb{P}(A \text{ given } B) = \mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

since this represents the fraction of the "probability mass" in B that is in A. To put it another way, if we run the probability experiment associated with Ω many times, this is the long-term fraction of times $A \cap B$ occurs, divided by the long-term fraction of time B occurs. In other words, it is the long-term fraction of times that A occurs, *out of* the times that B occurs[3]

Once we have this, independence of A and B is given by

$$\mathbb{P}(A \mid B) = \mathbb{P}(A),$$

i.e., the knowledge that B occurs does not affect our probability that A occurs.

⊙ **Guided Practice 2.8** Under what conditions is this equivalent to

$$\mathbb{P}(B \mid A) = \mathbb{P}(B)?$$[4]

[3] Here "long-term" should be understood in the context of Cournot's Principle and the Law of Large Numbers.

[4] If $\mathbb{P}(B) = 0$ or $\mathbb{P}(A) = 0$ then they cannot be equivalent, as they will not both be defined. However, once both probabilities are nonzero, both equations are just expressing that $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.

		X		
		$X = 1$	$X = 0$	Total
Y	$Y = 1$	0.0382	0.8252	0.8634
	$Y = 0$	0.0010	0.1356	0.1366
	Total	0.0392	0.9608	1.0000

Table 2.4: Joint probability table. Compare to the `smallpox` table in the main text.

Multiplication Rule for independent processes

If A and B represent events from two different and independent processes, then the probability that both A and B occur can be calculated as the product of their separate probabilities:

$$\mathbb{P}(A \text{ and } B) = \mathbb{P}(A) \times \mathbb{P}(B) \tag{2.9}$$

Here and elsewhere in *OpenIntro Statistics* and in the present text, $\times$ is used to denote multiplication.
Similarly, if there are k events $A_1, \ldots, A_k$ from k independent processes, then the probability they all occur is

$$\mathbb{P}(A_1) \times \mathbb{P}(A_2) \times \cdots \times \mathbb{P}(A_k)$$

Sometimes we wonder if one outcome provides useful information about another outcome. The question we are asking is, are the occurrences of the two events independent? We say that two events A and B are independent if they satisfy Equation (2.9).

2.2.1 Marginal and joint probabilities

When random variables X and Y are **jointly distributed**, i.e., they are defined on the same sample space Ω with the same probabilities for events in Ω, we can speak of probabilities involving them both, such as

$$\mathbb{P}(X < Y)$$

This is an example of joint probability. For non-continuous (discrete) random variables, we can then discuss $M(x) = \mathbb{P}(X = x)$ for particular values of x. This is the **probability mass function** of X. When also discussing Y and possibly other random variables that are jointly distributed with X, it is also called the **marginal distribution** of X. The function $M(x, y) = \mathbb{P}(X = x \text{ and } Y = y)$ is then the **joint probability mass function** of X and Y. Since Y must take *some* value, we can compute

$$\mathbb{P}(X = x) = \sum_y \mathbb{P}(X = x \text{ and } Y = y).$$

If we make a table with x and y axes this hints at why the term "marginal" is used: each axis is a margin.

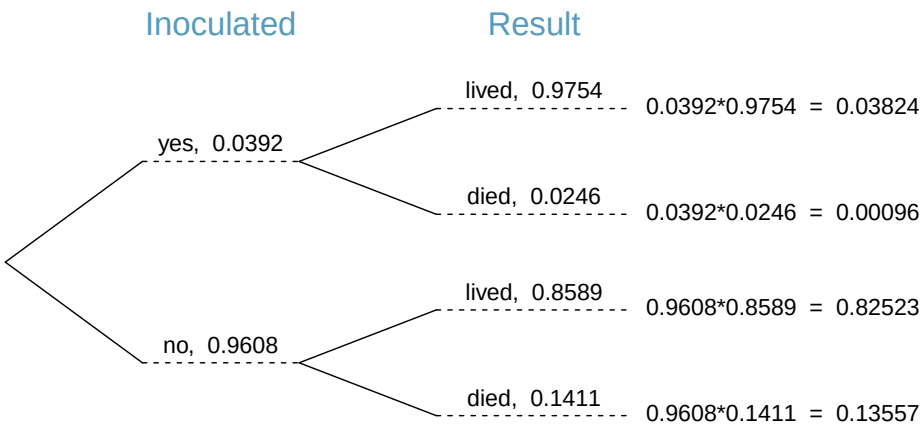


Figure 2.5: A tree diagram of the `smallpox` data set. To make this abstract, consider $\mathbb{P}(X = 1) = 0.0392$ to be "inoculated? yes, 0.0392", $X = 0$ to be "no, not inoculated", $\mathbb{P}(Y = 1 \mid X = 1) = 0.9754$ is then "lived, 0.9754" and so on.

Sum of conditional probabilities

Let $A_1, \ldots, A_k$ represent all the disjoint outcomes for a variable or process. Then if B is an event, possibly for another variable or process, we have:

$$\mathbb{P}(A_1|B) + \cdots + \mathbb{P}(A_k|B) = 1$$

The rule for complements also holds when an event and its complement are conditioned on the same information:

$$\mathbb{P}(A|B) = 1 - \mathbb{P}(A^c|B)$$

2.2.2 Independence considerations in conditional probability

If two events are independent, then knowing the outcome of one should provide no information about (*the probability of!*) the other. We can show this is mathematically true using conditional probabilities.

⊙ **Guided Practice 2.10** Can you think of a situation where we have events A, B, C such that A and B are independent, but nevertheless neither one is independent of C?[5]

Tree diagrams are a tool to organize outcomes and probabilities around the structure of the data. Figure 2.5 shows the general idea.

[5]Let A and B be the events that two spouses win on their lottery tickets (assume they are playing different lotteries, so that A and B can be independent). Then A and B both influence the event C that the family becomes richer than before.

2.2.3 Bayes' Theorem

It is a consequence of the definition of conditional probability and the fact that $\mathbb{P}(A) = \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B^c)$. And that more generally, if B_i are disjoint with $\bigcup_i B_i = \Omega$, then

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap B_i).$$

Even though that may sound easy enough, it takes some practice to successfully solve Bayes Theorem problems in a reasonable amount of time.

Bayes' Theorem: inverting probabilities

Consider the following conditional probability for variable 1 and variable 2:

$$\mathbb{P}(\text{outcome } A_1 \text{ of variable 1} \mid \text{outcome } B \text{ of variable 2})$$

Bayes' Theorem states that this conditional probability can be identified as the following fraction:

$$\frac{\mathbb{P}(B|A_1)\mathbb{P}(A_1)}{\mathbb{P}(B|A_1)\mathbb{P}(A_1) + \mathbb{P}(B|A_2)\mathbb{P}(A_2) + \cdots + \mathbb{P}(B|A_k)\mathbb{P}(A_k)} \tag{2.11}$$

where $A_2, A_3, \ldots, A_k$ represent all other possible outcomes of the first variable.

Bayes' Theorem is just a generalization of what we have done using tree diagrams. The numerator identifies the probability of getting both A_1 and B. The denominator is the marginal probability of getting B. This bottom component of the fraction appears long and complicated since we have to add up probabilities from all of the different ways to get B. We always completed this step when using tree diagrams. However, we usually did it in a separate step so it didn't seem as complex.

To apply Bayes' Theorem correctly, there are two preparatory steps:

(1) First identify the marginal probabilities of each possible outcome of the first variable: $\mathbb{P}(A_1), \mathbb{P}(A_2), \ldots \mathbb{P}(A_k)$.

(2) Then identify the probability of the outcome B, conditioned on each possible scenario for the first variable: $\mathbb{P}(B|A_1), \mathbb{P}(B|A_2), \ldots, \mathbb{P}(B|A_k)$.

Once each of these probabilities are identified, they can be applied directly within the formula.

2.2.4 Sampling from a small population

It is time to remind you of the notation

$$\bar{w} = www\ldots$$

used in repeated decimal expansions. In particular, you should find it useful to know the expansions in Table 2.6.

We now discuss **sampling without replacement** and **with replacement**.

Suppose we pick two marbles out of a jar of n marbles, without replacement. Suppose b many of the marbles in the jar are blue. The probability that the two marbles we picked

$1/1 = 1$	$1/2 = .5$	$1/3 = .\bar{3}$	$1/4 = .25$	$1/5 = .2$	$1/6 = .1\bar{6}$
$1/7 = .\overline{142857}$	$1/8 = .125$	$1/9 = .\bar{1}$	$1/10 = .1$	$1/11 = .\overline{09}$	$1/12 = .08\bar{3}$.

Table 2.6: Few people know by heart what $1/13$ is, but $1/n$ for smaller n are listed here.

are both blue is

$$\frac{b}{n} \cdot \frac{b-1}{n-1}. \tag{1}$$

If we replaced the first one before picking the second time,

$$\frac{b}{n} \cdot \frac{b}{n}. \tag{2}$$

⊙ **Guided Practice 2.12** True or false: If the population is large, so that n and b are both large, then the difference between (1) and (2) is negligible.[6]

2.3 Bayesian statistics

Suppose we have a Bernoulli distribution with parameter p, which is unknown. Perhaps we should take $p = 1/2$ as our starting assumption (this is consistent with the idea of maximum entropy from Section 8.2). Now suppose we observe $X = 1$ (heads). Then we can *update* our estimate of p as follows, in the spirit of Bayes' Theorem:

$$\mathbb{P}(p = p_0 \mid X = 1) = \frac{\mathbb{P}(X = 1 \mid p = p_0) \cdot \mathbb{P}(p = p_0)}{\mathbb{P}(X = 1)}$$

However, if we consider p to be a random variable, then any particular value will have probability 0 and we should be considering the **probability density function** (see Section 2.5) instead:

$$f_p(p_0 \mid X = 1) = \frac{\mathbb{P}(X = 1 \mid p = p_0) \cdot f_p(p_0)}{\mathbb{P}(X = 1)}. \tag{2.13}$$

So let us not assume $p = 1/2$ to start, but instead take $f_p(x) = 1$ for each $0 \le x \le 1$, which we write as

$$f_p(x) = 1_{[0,1]}(x).$$

This reflects us having no prior knowledge of what p might be. Now $\mathbb{P}(X = 1 \mid p = p_0)$ is supposed to be p_0 (even though $p = p_0$ is a probability-zero event, which leads to dividing by zero if we literally using it in conditional probability) because, if $\mathbb{P}(X = 1) = p = p_0$ if $p = p_0$. So we get

$$f_p(p_0 \mid X = 1) = p_0 \cdot 1_{[0,1]}(p_0)/(1/2) = 2p_0 \cdot 1_{[0,1]}(p_0).$$

If we now make another observation, we can repeat this process starting with our new pdf and get yet another one. Namely, let's say $f_0(x) = 1_{[0,1]}(x)$ and $f_1(x) = 2x \cdot 1_{[0,1]}(x)$. Then

[6] True. Since $b \le n$, we have

$$\left|\frac{b-1}{n-1} - \frac{b}{n}\right| = \frac{|n(b-1) - (n-1)b|}{n(n-1)} = \frac{n-b}{n(n-1)} \le \frac{1}{n-1}$$

which goes to zero as $n \to \infty$.

if our next observation is $X = 0$, we get

$$\begin{aligned} f_2(p_0) &= f_1(p_0 \mid X = 0) \\ &= \mathbb{P}(X = 0 \mid p = p_0) f_p(p_0) / \mathbb{P}(X = 0) = (1 - p_0) 2p_0 / (1/3) = 6p_0(1 - p_0). \end{aligned}$$

Here we have used

$$\begin{aligned} \mathbb{P}(X = 0) &= \int \mathbb{P}(X = 0 \mid p_0) f_1(p_0)\, dp \\ &= \int (1 - p_0) 2p_0\, dp = 2(1/2 - 1/3) = 1/3. \end{aligned}$$

We see that f_2 has its mode at $p_0 = 1/2$, which is reasonable since we have observed one heads and one tails. On the other hand, the variance of f_2 is smaller than that of f_0, which is also reasonable since we have more observations to go by:

$$\begin{aligned} \mathrm{Var}(p) = \mathbb{E}(p^2) - \mathbb{E}(p)^2 &= \int_0^1 p^2 (6p(1 - p))\, dp - (1/2)^2 \\ &= 6\left(\frac{1}{4} - \frac{1}{5}\right) - \frac{1}{4} = \frac{6}{20} - \frac{1}{4} = \frac{1}{20} < \frac{1}{3} - \frac{1}{4} = \frac{1}{12}. \end{aligned}$$

We will not go deeply into Bayesian statistics in this book, but suffice it to say that philosophically, there are at least two possible approaches to probability and statistics:

- **Frequentist statistics** holds that probabilities like p above are estimated from a sample $\hat{p} = \sum X_i / n$;
- **Bayesian statistics** holds that rather than an infinitely precise estimate value $\hat{p}$, we have a pdf representing our current beliefs about what p might be. This pdf becomes more sharply peaked as we gather more sample points.

While the frequentist stance avoids the tricky issue "what should our initial pdf, after seeing 0 values, be?", the Bayesian stance is well-suited to machine learning, whereby a machine can be continuously updating its beliefs about the world using the method of (2.13).

2.4 Random variables

2.4.1 Expectation

Just like probability is a long-term proportion, expectation is a long-term average.

We call a variable or process with a numerical outcome a **random variable**, and we usually represent this random variable with a capital letter such as X, Y, or Z. Actually a random variable could take values that are vectors, functions and several other things. The main requirement is that we have a function $X : \Omega \to \mathcal{A}$ where Ω is our sample space, and probability enters the picture as $\mathbb{P}(X \in A) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in \mathcal{A}\})$.

$\mathbb{E}(X)$
Expected value of X

We compute the average outcome of X and call this average the **expected value** of X, denoted by $\mathbb{E}(X)$. The expected value of a random variable is computed by adding each outcome weighted by its probability, for instance:

$$\begin{aligned} \mathbb{E}(X) &= 0 \times \mathbb{P}(X = 0) + 137 \times \mathbb{P}(X = 137) + 170 \times \mathbb{P}(X = 170) \\ &= 0 \times 0.20 + 137 \times 0.55 + 170 \times 0.25 = 117.85 \end{aligned}$$

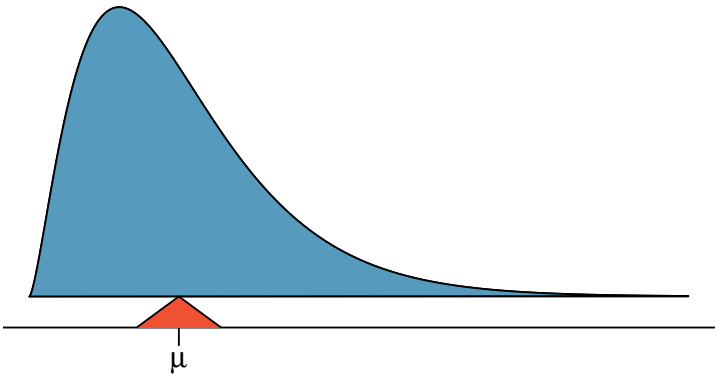


Figure 2.7: A continuous distribution balanced at its mean.

Expected value of a Discrete Random Variable

If X takes outcomes $x_1, \ldots, x_k$ with probabilities $\mathbb{P}(X = x_1), \ldots, \mathbb{P}(X = x_k)$, the expected value of X is the sum of each outcome multiplied by its corresponding probability:

$$\begin{aligned}\mathbb{E}(X) &= x_1 \times \mathbb{P}(X = x_1) + \cdots + x_k \times \mathbb{P}(X = x_k) \\ &= \sum_{i=1}^{k} x_i \mathbb{P}(X = x_i) \end{aligned} \tag{2.14}$$

The Greek letter μ may be used in place of the notation $\mathbb{E}(X)$.

It is also possible to compute the expected value of a continuous random variable (see Section 2.5):

$$\mu = \int x f(x) dx$$

where $f(x)$, the probability density function, represents a function for the density curve.

The **(sample) mean**, sometimes called the (sample) average of a data set $x_1, \ldots, x_n$ is distinct from the mean of a random variable.

However, we can "freeze" a data set and consider the uniform distribution on it.

Definition 2.2. *The **uniform distribution** on a finite set of n elements, $\{x_1, \ldots, x_n\}$, is given by $\mathbb{P}(X = x_i) = \frac{1}{n}$. If some of the x_i are identical (say k many are equal to a particular x_i) this makes that element have probability k/n.*

Example 2.15 The uniform distribution on $\{1.3, 2.0, -4\}$.

This distribution is such that if a variable X has this distribution then $\mathbb{P}(X = 1.3) = 1/3$, $\mathbb{P}(X = 2.0) = 1/3$, and $\mathbb{P}(X = -4) = 1/3$.

In this case the sample mean of the original data set becomes the mean of the new random variable. The sample standard deviation of the original data set does not quite become the standard deviation of the new random variable. Indeed, if we have just a single data point, the sample standard deviation is (rightly so) undefined, as it tells us nothing

about the underlying standard deviation; but the frozen distribution has standard deviation 0.

In the sample standard deviation from Chapter 1, we divide by $n-1$, rather than dividing by n. This way, the expectation of the sample variance turns out to be the original distribution's variance. Indeed, define the random variable S by

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Theorem 2.3. $\mathbb{E}(S^2) = \sigma^2$.

Proof. Let us make the simplifying assumption that $\mathbb{E}(X_i) = 0$. The value of

$$\mathbb{E}\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)$$

is obtained by

$$\begin{aligned}\mathbb{E}((X_i - \bar{X})^2) &= \mathbb{E}(X_i^2 - 2X_i\bar{X} + (\bar{X})^2) \\ &= \underbrace{\mathbb{E}(X_i^2)}_{\text{first term}} - 2\underbrace{\mathbb{E}(X_i\bar{X})}_{\text{second term}} + \underbrace{\mathbb{E}((\bar{X})^2)}_{\text{third term}}\end{aligned}$$

The first term is σ^2. The second term is

$$\mathbb{E}(X_i\bar{X}) = \frac{1}{n}\sum_j \mathbb{E}(X_iX_j) = \frac{1}{n}\,\mathbb{E}(X_iX_i) = \frac{1}{n}\sigma^2.$$

The third term is found using

$$\mathbb{E}(n^2(\bar{X})^2) = \mathbb{E}\left(\left(\sum_{i=1}^{n}X_i\right)^2\right) = \mathbb{E}\left(\sum_i X_i^2 + \sum_{i\neq j}X_iX_j\right)$$

$$= \mathbb{E}\left(\sum_i X_i^2 + 2\sum_{i<j}X_iX_j\right) = n(\sigma^2)$$

to be $\mathbb{E}((\bar{X})^2) = \sigma^2/n$. Thus

$$\mathbb{E}((X_i - \bar{X})^2) = \sigma^2 - 2\sigma^2/n + \sigma^2/n = \sigma^2\left(1 - \frac{1}{n}\right) = \sigma^2 \cdot \frac{n-1}{n}.$$

Consequently,

$$\mathbb{E}\left(\frac{1}{n-1}\sum_i(X_i - \bar{X})^2\right) = \sigma^2. \qquad \square$$

2.4.2 Variability in random variables

$\text{Var}(X)$
Variance of X

General variance formula

If X takes outcomes $x_1, \ldots x_k$ with probabilities $\mathbb{P}(X = x_1), \ldots, \mathbb{P}(X = x_k)$ and expected value $\mu = \mathbb{E}(X)$, then the variance of X, denoted by $\text{Var}(X)$ or the symbol σ^2, is

$$\sigma^2 = \sum_{j=1}^{k} (x_j - \mu)^2 \mathbb{P}(X = x_j)$$

The standard deviation of X, labeled σ or $\text{SD}(X)$, is the square root of the variance.

2.4.3 Linear combinations of random variables

Two important concepts concerning combinations of random variables have so far been introduced. First, a final value can sometimes be described as the sum of its parts in an equation. Second, intuition suggests that putting the individual average values into this equation gives the average value we would expect in total. This second point needs clarification – it is guaranteed to be true in what are called *linear combinations of random variables*.

A **linear combination** of two random variables X and Y is a fancy phrase to describe a combination

$$aX + bY$$

where a and b are some fixed and known numbers.

When considering the average of a linear combination of random variables, it is safe to plug in the mean of each random variable and then compute the final result. For a few examples of nonlinear combinations of random variables – cases where we cannot simply plug in the means – see the footnote.[7]

Linear combinations of random variables and the average result

If X and Y are random variables, then a linear combination of the random variables is given by

$$aX + bY \tag{2.16}$$

where a and b are some fixed numbers. To compute the average value of a linear combination of random variables, plug in the average of each individual random variable and compute the result:

$$a \times \mathbb{E}(X) + b \times \mathbb{E}(Y)$$

Recall that the expected value is the same as the mean, e.g. $\mathbb{E}(X) = \mu_X$.

[7] If X and Y are random variables, consider the following combinations: X^{1+Y}, $X \times Y$, X/Y. In such cases, plugging in the average value for each random variable and computing the result will not generally lead to an accurate average value for the end result.

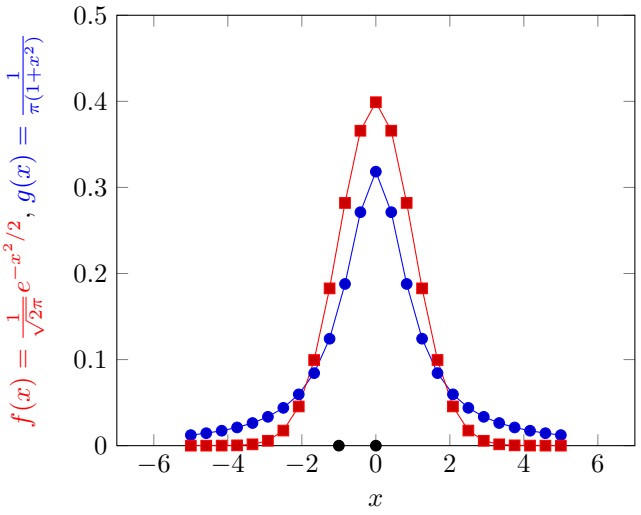


Figure 2.8: The Cauchy distribution has "heavier tails" than the normal distribution.

2.4.4 Variability in linear combinations of random variables

Variability of linear combinations of random variables

The variance of a linear combination of random variables may be computed by squaring the constants, substituting in the variances for the random variables, and computing the result:

$$\mathrm{Var}(aX + bY) = a^2 \times \mathrm{Var}(X) + b^2 \times \mathrm{Var}(Y)$$

This equation is valid as long as the random variables are independent of each other. The standard deviation of the linear combination may be found by taking the square root of the variance.

In the independent case, assuming $\mathbb{E}(X) = \mathbb{E}(Y) = 0$,

$$\mathrm{Var}(X \pm Y) = \mathbb{E}((X \pm Y)^2) = \mathbb{E}(X^2) \pm 2\,\mathbb{E}(XY) + \mathbb{E}(Y^2) = \mathbb{E}(X^2) + \mathbb{E}(Y^2)$$

is independent of the sign $\pm$. Here we are using the fact that the **covariance** $\mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y)$ is zero for independent random variables. This can be verified by integration $\iint dx\,dy$ which is covered in multivariable calculus. In the dependent case, thing are more complicated. For instance, $X - X$ and $X + X$ have quite different variances.

2.5 Continuous distributions and a review of calculus

The following subsections roughly correspond to chapters in a popular calculus text such as Stewart's *Calculus* or Hass, Weir and Thomas' *University Calculus*.

2.5.1 Limits

$\lim_{x\to a} f(x) = L$ means that for each $\epsilon > 0$, there is a $\delta > 0$ such that for all x, if $0 < |x-a| < \delta$ then $|f(x) - L| < \epsilon$. The concept is used to define derivatives and integrals, but is also used directly in statistics, in the Law of Large Numbers and the Central Limit Theorem.

2.5.2 Derivatives

The derivative $f'(x) = \frac{df}{dx}$ is defined by $\lim_{h\to 0} \frac{f(x+h)-f(x)}{h}$. The **Chain Rule** says that

$$\frac{d}{dx} f(g(x)) = f'(g(x)) \cdot g'(x).$$

The **Product Rule** says that

$$\frac{d}{dx} f(x)g(x) = f'(x)g(x) + f(x)g'(x).$$

Differentiation is a linear operation in the sense that

$$\frac{d}{dx} cf(x) = cf'(x), \quad \frac{d}{dx}(f(x) + g(x)) = f'(x) + g'(x),$$

where c is a constant.

2.5.3 Applications of derivatives

To find a maximum or minimum of $f(x)$, we solve the equation $f'(x) = 0$ for x. If $f''(x) \geq 0$ (as is the case with the function $f(x) = x^2$) then we will have a minimum, and if $f''(x) \leq 0$ (as with $f(x) = -x^2$) we will have a maximum.

For a continuous random variable with pdf f_X, we can naturally define a **mode** to be any point x with $f'_X(x) = 0$ and $f''_X(x) \leq 0$, i.e., a local maximum of f_X.

2.5.4 Integrals

Integrals can be understood via the Fundamental Theorem of Calculus: $f(b) - f(a) = \int_a^b f'(x)\,dx$ and $\frac{d}{dx}\int_a^x f(t)\,dt = f(x)$. In particular, the probability density function f and the cumulative distribution function F are related by $F' = f$.

For discrete random variables with probability mass function $m(x) = \mathbb{P}(X = x)$, we require

$$\sum_x m(x) = 1,$$

and analogously in the continuous case we need

$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1$$

when f_X is the pdf of X.

2.5.5 Applications of integrals

The mean of a random variable X with pdf f is $\int_{-\infty}^{\infty} x\, f(x)\, dx$.

The expectation of X^2 is $\int x^2 f(x)\, dx$, similarly. The general rule is that if f_X is the pdf of X, and g is any (deterministic, i.e., non-random) function, then

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx.$$

2.5.6 Inverse and trigonometric functions

The normal distribution (Exercise 2.1) and the exponential distribution use the exponential function e^x, which is related to the natural logarithm ln via

$$y = e^x \iff \ln y = x$$

where $y > 0$.

By the Chain Rule, if $y = f(f^{-1}(y))$ then

$$1 = \frac{dy}{dy} = f'(f^{-1}(y)) \cdot (f^{-1})'(y).$$

This implies that, since $\frac{d}{dx} e^x = e^x$, using $f(x) = e^x$,

$$\frac{d}{dy} \ln y = (f^{-1})'(y) = \frac{1}{f'(f^{-1}(y))} = \frac{1}{e^x} = \frac{1}{y}.$$

Properties of logarithms generally follow from analogous ones for exponentials, e.g., the general rule

$$\frac{\log_a x}{\log_a y} = \frac{\log_b x}{\log_b y}$$

follows from the special case $b = y$:

$$\frac{\log_a x}{\log_a y} = \frac{\log_y x}{\log_y y}$$

(since this shows that the left hand side does not depend on a) which follows, using $a^{\log_a z} = z$, from

$$\frac{\log_a x}{\log_a y} = \frac{\log_y x}{\log_y y} = \log_y x \iff \log_a y \log_y x = \log_a x$$
$$\iff a^{\log_a y \log_y x} = a^{\log_a x} \iff (a^{\log_a y})^{\log_y x} = x \iff y^{\log_y x} = x.$$

2.5.7 Techniques of integration

Integration by parts $\int u'(x)v(x)\, dx = [u(x)v(x)] - \int v'(x)u(x)\, dx$ is often useful. Improper integrals $\int_{-\infty}^{\infty} f(x)\, dx$ are used for pdfs f.

2.5.8 Applications of integration to probability

The probability that $a \le X \le b$ is

$$\int_a^b f_X(x)\, dx$$

where f_X is the pdf of X.

2.5.9 Infinite sequences and series

The geometric distribution is understandable after studying infinite series and in particular the geometric series

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}, \quad |x| < 1.$$

When we check the 68-95-99.7 rule in Section 3.1 we shall use Taylor series:

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} x^n.$$

2.6 Exercises

2.1 Normal distributions. The *normal distribution* is a continuous probability distribution defined by the probability density function $\varphi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for fixed constants μ and σ.

(a) Use that $\int_{-\infty}^{\infty} e^{-\frac{x^2}{2}}\,dx = \sqrt{2\pi}$ to show that φ is indeed a probability density function.

(b) Show that the mean, median, and mode of the normal distribution are all equal to μ.

(c) Show that the variance of the normal distribution is equal to σ^2 and therefore that the standard deviation of the normal distribution is equal to σ.

2.2 Skewed normals. Let φ be as in the previous problem with $\mu = 0$ and $\sigma = 1$ (the "standard normal" case). Define $\Phi(x) := \int_{-\infty}^{x} \varphi(t)dt$, the *cumulative distribution function of* φ. Fix $c \in \mathbb{R}$ and define the *skewed normal distribution with skewness parameter* c by

$$f_c(x) := \frac{\varphi(x)\Phi(cx)}{\Phi(0)}$$

Note that $f_c(x) = \varphi(x)$ when $c = 0$.

(a) Show that $\int_{-\infty}^{\infty} f_c(x) < \infty$.

(b) Show that f_c is indeed a probability density function. (Hint: This requires multivariable calculus. $\int_{-\infty}^{\infty} f_c(x)\,dx$ is the probability that for two independent standard normal random variables X and Y, we have $\mathbb{P}(Y \le cX)$:

$$\mathbb{P}(Y \le cX) = \int_{-\infty}^{\infty} \varphi(x) \int_{-\infty}^{cx} \varphi(y)\,dy\,dx.$$

This is $1/2 = \Phi(0)$ because $\{(x, y) : y \le cx\}$ is a half-plane and the joint pdf of X and Y is spherically symmetric.)

(c) Find the mean of $f_c(x)$. (Hint: the mean is

$$E(2X[Y \le cX]) = 2\frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{cx} xe^{-r^2/2}\,dy\,dx.$$

This is $\frac{1}{\pi} \int_0^{\infty} r^2 e^{-r^2/2}\,dr \int_{\tan^{-1}(c)-\pi}^{\tan^{-1}(c)} \cos\theta\,d\theta$ using polar coordinates in multiple integrals. We have $\int_{\tan^{-1}(c)-\pi}^{\tan^{-1}(c)} \cos\theta\,d\theta = [\sin\theta]_{\tan^{-1}(c)-\pi}^{\tan^{-1}(c)} = 2\frac{c}{\sqrt{1+c^2}}$ since if $\tan\theta = c$ then $\sin\theta = c/\sqrt{1+c^2}$ (draw a triangle). On the other hand $\int_0^{\infty} r^2 e^{-r^2/2}\,dr = \frac{1}{2}\int_{-\infty}^{\infty} x^2 e^{-x^2/2}\,dx = \frac{1}{2}\sqrt{2\pi}$ using the fact that $E(X^2) = 1$ for a standard normal X.)

(d) Show that the mode of $f_c(x)$ is unique (there is no simple analytic expression for it, however).

(e) Show that $g(x) = \frac{d}{dx}(\Phi(x))^2$ is also a probability density function. (Hint: $\int_{-\infty}^{\infty} g(x)\,dx = 1^2 - 0^2 = 1$.) How does it relate to f_c?

2.3 Cauchy distribution. The *Cauchy distribution* is a continuous probability distribution defined by the probability density function $f(x) = \frac{1}{\pi(1+x^2)}$

(a) Show that f is indeed a probability density function.

(b) Find the median and the mode of the Cauchy distribution.

(c) Show that the mean of the Cauchy distribution does not exist.

(d) Conclude that the weak (and therefore the strong) Law of Large Numbers fails for the Cauchy distribution.

2.4 Bayesian statistics. Suppose X is a Bernoulli random variable with unknown parameter p. Your initial assumption is that p is distributed uniformly on $[0, 1]$. Then you observe in sequence the values $X = 1, X = 0, X = 0$. What is your new p.d.f. for p?

Chapter 3

Distributions of random variables

3.1 Normal distribution

The probability density function of a standard normal random variable (one with mean $\mathbb{E}(X) = \mu = 0$ and standard deviation $\mathrm{SD}(X) = \sigma = 1$) is

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty.$$

In general, if X is normal with mean μ and variance σ^2 then

$$f_X(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-((x-\mu)/\sigma)^2/2}, \quad -\infty < z < \infty.$$

See Figure 3.1.

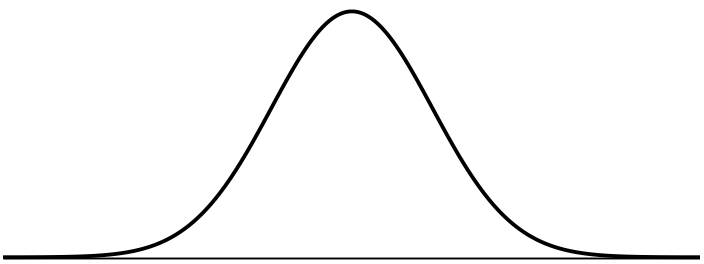


Figure 3.1: A normal curve.

Normal distribution facts

Variables appearing in nature are often nearly normal, but not exactly normal. This is due to the Central Limit Theorem stating that a sum of many (say n many) independent and identically distributed random variables is approximately normally distributed (and that as the number $n \to \infty$, this approximation becomes as good as you like).

The Z-score

The Z-score of an observation is the number of standard deviations it falls above or below the mean. We compute the Z-score for an observation x that follows a distribution with mean μ and standard deviation σ using

$$Z = \frac{x - \mu}{\sigma}$$

3.1.1 Normal probability table

Using Google Sheets'

`normdist`

and

`norminv`

is more practical than table look-up when a computer is handy.

Let us see how to reproduce the values in Table 3.2 using Google Sheets.

```
=normdist(0.01,0,1,TRUE)
```

gives the probability that $Z \leq 0.01$ for a standard normal Z (the TRUE referring to using the cumulative distribution function rather than the probability density function).

Here `NORMDIST(x, mean, standard_deviation, cumulative)` has

- x - The input to the normal function (whether probability density or cumulative distribution).
- mean - The mean of the normal distribution function.
- standard_deviation - The standard deviation (sigma) of the normal distribution.
- cumulative - Whether to use the cumulative distribution function rather than the probability density function.

We can also find the Z-score associated with a percentile. For example, to identify Z for the 80^{th} percentile, do `=norminv(0.8,0,1)` which yields 0.84.

	Second decimal place of Z									
Z	0.00	0.01	0.02	*0.03*	**0.04**	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	*0.6664*	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	**0.7995**	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 3.2: A section of the normal probability table. The percentile for a normal random variable with $Z = 0.43$ has been *highlighted*, and the percentile closest to 0.8000 has also been **highlighted**.

3.1.2 68-95-99.7 rule

This useful rule of thumb for the probability of falling within 1, 2, and 3 standard deviations of the mean in the normal distribution is telling us values of

$$\int_{-n}^{n} f_Z(z)\,dz$$

where $f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ is the standard normal probability density function and $n \in \{1, 2, 3\}$. That is,

$$\begin{aligned} \int_{-1}^{1} f_Z(z)\,dz &\approx 68\% \\ \int_{-2}^{2} f_Z(z)\,dz &\approx 95\% \\ \int_{-3}^{3} f_Z(z)\,dz &\approx 99.7\% \end{aligned}$$

As there is no elementary antiderivative for $e^{-z^2/2}$, these integrals are calculated by *numerical integration* that you may recall from a calculus course.

We may consider Taylor series as one way of approximating. Recalling that $e^x = \sum_{n=0}^{\infty} x^n/n!$, it follows that

$$e^{-x^2/2} = \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{2^n n!}.$$

Integrating this term by term, we have

$$\begin{aligned}\int e^{-x^2/2}\,dx &= \sum_{n=0}^{\infty}\frac{(-1)^n x^{2n+1}}{(2n+1)2^n n!} + \text{constant}\\ &\approx x - \frac{x^3}{6} + C.\end{aligned}$$

For the standard normal cdf Φ we want $\Phi(0) = 1/2$, which means

$$\Phi(x) \approx \frac{1}{\sqrt{2\pi}}\left(x - \frac{x^3}{6}\right) + \frac{1}{2}$$

for $x \approx 0$. To see how good this approximation is, let us compute

$$68\% \approx \Phi(1) - \Phi(-1) \approx \frac{1}{\sqrt{2\pi}}\left(2 - \frac{2}{6}\right) = \frac{1}{\sqrt{2\pi}}\frac{5}{3} = .6649$$

which is not bad.

Let us also try using Simpson' Rule: with $f = f_Z$, $\Delta x = 1$, and $(x_0, x_1, x_2) = (-1, 0, 1)$, we have

$$\begin{aligned}\int_{-1}^{1} f_Z(z)\,dz &\approx \frac{\Delta x}{3}\left(f(x_0) + 4f(x_1) + f(x_2)\right)\\ &= \frac{1}{3}\frac{1}{\sqrt{2\pi}}(e^{-(-1)^2/2} + 4e^{-0^2/2} + e^{-(1)^2/2})\\ &= \frac{1}{3\sqrt{2\pi}}(2e^{-1/2} + 4)\\ &= \frac{\sqrt{2}}{3\sqrt{\pi e}}(1 + 2\sqrt{e})\\ &\approx .6932\ldots\end{aligned}$$

and with $(x_0, x_1, x_2, x_3, x_4) = (-2, -1, 0, 1, 2)$, we have

$$\begin{aligned}\int_{-2}^{2} f_Z(z)\,dz &\approx \frac{\Delta x}{3}\left(f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + f(x_4)\right)\\ &= \frac{1}{3\sqrt{2\pi}}\left(e^{-4/2} + 4e^{-1/2} + 2e^{-0/2} + 4e^{-1/2} + e^{-4/2}\right)\\ &= \frac{1}{3\sqrt{2\pi}}\left(2e^{-2} + 8e^{-1/2} + 2\right)\\ &\approx .9472\ldots\end{aligned}$$

3.1.3 Normal probability plots

To make a normal probability plots, proceed as follows.

- Using the mean and (sample) standard deviation, standardize the data set.
- Construct the **empirical cdf** which is just the cdf of the uniform distribution on the data set we are working with.
- In Google Sheets, highlight two columns with the values of the empirical and normal cdfs (using `normdist` function with cdf set to 1 or True) at the values in our data set.
- Display a scatter plot of these two columns. That is our normal probability plot.

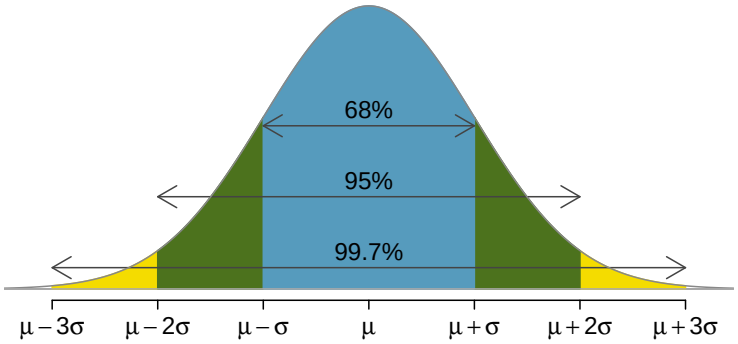


Figure 3.3: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

3.2 Bernoulli distribution

Bernoulli random variable, descriptive

A Bernoulli random variable has exactly two possible outcomes. We typically label one of these outcomes a "success" and the other outcome a "failure". We may also denote a success by `1` and a failure by `0`.

In a way, the Bernoulli distribution is too simple to be of much use by itself. On the other hand, sums of independent Bernoulli's give us the binomial distribution $\sum_i X_i$ which approximates the normal distribution and can be used for estimating proportions such as voter support and disease prevalence. The Bernoulli distribution is just complicated enough that its sample averages $\sum_{i=1}^{n} X_i/n$ are interesting.

Bernoulli random variables are often denoted as `1` for a success and `0` for a failure. In addition to being convenient in entering data, it is also mathematically handy. Suppose we observe ten trials:

```
0 1 1 1 1 0 1 1 0 0
```

Then the **sample proportion**, $\hat{p}$, is the sample mean of these observations:

$$\hat{p} = \frac{\text{\# of successes}}{\text{\# of trials}} = \frac{0+1+1+1+1+0+1+1+0+0}{10} = 0.6$$

This mathematical inquiry of Bernoulli random variables can be extended even further. Because `0` and `1` are numerical outcomes, we can define the mean and standard deviation of a Bernoulli random variable. If p is the true probability of a success, then the mean of a Bernoulli random variable X is given by

$$\begin{aligned}\mu = \mathbb{E}[X] &= \mathbb{P}(X=0)\times 0 + \mathbb{P}(X=1)\times 1 \\ &= (1-p)\times 0 + p\times 1 = 0 + p = p\end{aligned}$$

Similarly, the variance of X can be computed:

$$\begin{aligned}\mathrm{Var}(X) = \sigma^2 &= \mathbb{P}(X=0)(0-p)^2 + \mathbb{P}(X=1)(1-p)^2 \\ &= (1-p)p^2 + p(1-p)^2 = p(1-p)\end{aligned}$$

The standard deviation is $\mathrm{SD}(X) = \sigma = \sqrt{p(1-p)}$.

Bernoulli random variable, mathematical

If X is a random variable that takes value 1 with probability of success p and 0 with probability $1-p$, then X is a Bernoulli random variable with mean and standard deviation

$$\mu = p \qquad\qquad \sigma = \sqrt{p(1-p)}$$

In general, it is useful to think about a Bernoulli random variable as a random process with only two outcomes: a success or failure. Then we build our mathematical framework using the numerical labels `1` and `0` for successes and failures, respectively.

3.3 Geometric distribution

We now derive the formulas for the mean (expected) number of trials needed to find the first success and the standard deviation or variance of this distribution.

Geometric Distribution

If the probability of a success in one trial is p and the probability of a failure is $1-p$, then the probability of finding the first success in the n^{th} trial is given by

$$(1-p)^{n-1}p \tag{3.1}$$

The mean (i.e. expected value), variance, and standard deviation of this wait time are given by

$$\mu = \frac{1}{p} \qquad \sigma^2 = \frac{1-p}{p^2} \qquad \sigma = \sqrt{\frac{1-p}{p^2}} \tag{3.2}$$

Note that the variance σ^2 of a random variable X has the relatively simple formula $\sigma^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$, as we verify as follows:

$$\begin{aligned}\sigma^2 &= \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2 - 2X\,\mathbb{E}(X) + (\mathbb{E}(X))^2) \\ &= \mathbb{E}(X^2) - 2(\mathbb{E}(X))^2 + (\mathbb{E}(X))^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.\end{aligned}$$

We can verify the important property $\mathbb{P}(-\infty < X < \infty) = 1$ of probability distributions in the geometric case as follows. First recall the geometric series

$$\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}, \quad |x| < 1$$

Then

$$\begin{aligned}\mathbb{P}(1 \le X < \infty) = \sum_{k=1}^{\infty} \mathbb{P}(X = k) &= \sum_{k=1}^{\infty} (1-p)^{k-1} p \\ &= \sum_{t=0}^{\infty} (1-p)^t p = p \cdot \frac{1}{1-(1-p)} = 1.\end{aligned}$$

Theorem 3.1. *The mean of a geometric random variable X is $1/p$.*

Proof. First let us recall (or learn) another series from Calculus:

$$\alpha := \sum_{n=0}^{\infty} n x^n = \sum_{n=0}^{\infty} (n+1) x^{n+1} = \sum_{n=0}^{\infty} n x^{n+1} + \sum_{n=0}^{\infty} x^{n+1} = \alpha \cdot x + \left(\frac{1}{1-x} - 1 \right),$$

so we can solve for α and get

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

Therefore,

$$\begin{aligned}\mathbb{E}(X) = \sum_{k=1}^{\infty} k \mathbb{P}(X = k) &= \sum_{k=1}^{\infty} k(1-p)^{k-1} p \\ &= \frac{p}{1-p} \sum_{k=0}^{\infty} k(1-p)^k = \frac{p}{1-p} \cdot \frac{1-p}{p^2} = \frac{1}{p}. \quad \square\end{aligned}$$

Theorem 3.2. *The variance of a geometric random variable X is $\frac{1-p}{p^2}$.*

Proof. Let

$$\beta := \sum_{n=0}^{\infty} n^2 x^n = \sum_{n=0}^{\infty} (n+1)^2 x^{n+1} = \sum_{n=0}^{\infty} (n^2 + 2n + 1) x^{n+1} = x \left(\beta + 2\alpha + \frac{1}{1-x} \right)$$

Solving this for β and using $\sigma^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$ we get our formula for σ; see Exercise 3.5. $\square$

3.4 Binomial distribution

The **binomial distribution** describes the probability of having exactly k successes in n independent Bernoulli trials with probability of a success p.

The quantity

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

is read **n choose k.**[1] The exclamation point notation (e.g. $k!$) denotes a **factorial** expression (Table 3.4).

[1] Other notation for n choose k includes ${}_nC_k$, C_n^k, and $C(n,k)$.

n	$n!$
0	1
1	1
2	2
3	6
4	24
5	120
6	720
7	5040
8	40320

Table 3.4: How many of these do you remember? An urban legend says that undergraduates know that $3! = 6$, graduate students that $4! = 24$, postdocs that $5! = 120$, and professors that $6! = 720$.

Theorem 3.3. $\binom{n}{k}$ *equals the number of subsets* F *of an* n*-element set* $[n] := \{1, \ldots, n\}$ *such that* F *has* k *elements.*

Proof idea. The number of sequences $(a_1, \ldots, a_k)$ where all the a_i are distinct and coming from $[n]$, is $n!/(n-k)!$. Indeed, we have n choices for a_1, then $n-1$ choices for a_2 given a choice of a_1, and so on. A set F is represented by $k!$ such sequences. □

Example 3.3 Understanding binomial coefficients.

We have $\binom{4}{2} = 6$ because the 2-element subsets of $[4]$ are the following six:

$$\{1,2\}, \{1,3\}, \{1,4\}, \{2,3\}, \{2,4\}, \{3,4\}.$$

We have $\binom{4}{1} = 4$ because the 1-element subsets of $[4]$ are the following four:

$$\{1\}, \{2\}, \{3\}, \{4\}.$$

We have $\binom{4}{0} = 1$ because there is exactly one 0-element subset of $[4]$:

$$\emptyset.$$

We have $\binom{4}{3} = 4$ because there are exactly four 3-element subsets of $[4]$:

$$\{1,2,3\}, \{1,2,4\}, \{1,3,4\}, \{2,3,4\}.$$

Finally, we have $\binom{4}{4} = 1$ as there is only one 4-element subset of $[4]$:

$$\{1,2,3,4\}.$$

Finally, $\sum_{i=0}^{4} \binom{4}{i} = 16 = 2^4$ collects all these 16 subsets.

Binomial distribution

Suppose the probability of a single trial being a success is p. Then the probability of observing exactly k successes in n independent trials is given by

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{3.4}$$

Additionally, the mean, variance, and standard deviation of the number of observed successes are

$$\mu = np \qquad \sigma^2 = np(1-p) \qquad \sigma = \sqrt{np(1-p)} \tag{3.5}$$

The binomial formula is cumbersome when the sample size (n) is large, particularly when we consider a range of observations. In some cases we may use the normal distribution as an easier and faster way to estimate binomial probabilities.

Normal approximation of the binomial distribution

The binomial distribution with probability of success p is nearly normal when the sample size n is sufficiently large that np and $n(1-p)$ are both at least 10. The approximate normal distribution has parameters corresponding to the mean and standard deviation of the binomial distribution:

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

The 10% condition. When drawing a sample from a larger population, the **10% condition** helps our samples to be independent. If our sample has more than 10% of the individuals, then knowing that 5% has a certain property may mean that the other 5% cannot have this property. See also Section 6.1.1.

Caution: The normal approximation may fail on small intervals

The normal approximation to the binomial distribution tends to perform poorly when estimating the probability of a small range of counts, even when the conditions are met.

3.5 More discrete distributions

3.5.1 Negative binomial distribution

The geometric distribution describes the probability of observing the first success on the n^{th} trial. The **negative binomial distribution** is more general: it describes the probability of observing the k^{th} success on the n^{th} trial.

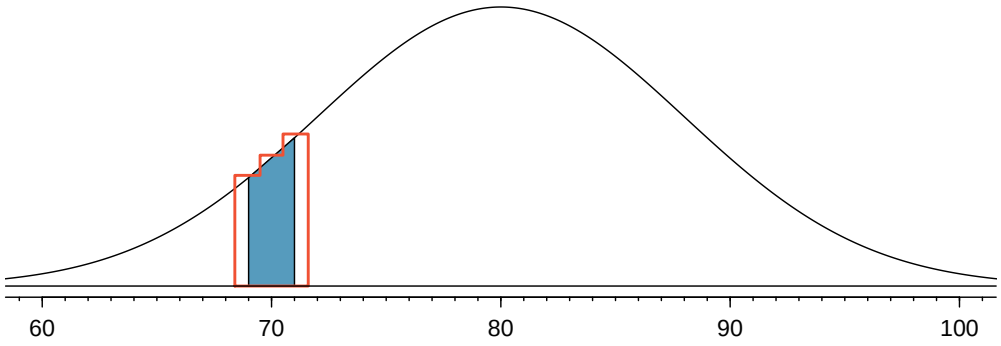


Figure 3.5: A normal curve with the area between 69 and 71 shaded. The outlined area represents the exact binomial probability.

Negative binomial distribution

The negative binomial distribution describes the probability of observing the k^{th} success on the n^{th} trial:

$$\mathbb{P}(\text{the } k^{th} \text{ success on the } n^{th} \text{ trial}) = \binom{n-1}{k-1} p^k (1-p)^{n-k} \tag{3.6}$$

where p is the probability an individual trial is a success. All trials are assumed to be independent.

To check that if X is negative binomial(k, p), then

$$1 = \mathbb{P}(k \leq X < \infty),$$

we have to get into *binomial series*, a topic often neglected in calculus courses, so we shall refrain.

The justification for the probability mass function of the negative binomial distribution is clear: if the kth success occurs on the nth trial then there is some choice of $k-1$ of the first $n-1$ trials to also be successes, and the remaining factors are as in the binomial distribution.

3.5.2 Poisson distribution

Poisson distribution

Suppose we are watching for events and the number of observed events follows a Poisson distribution with rate λ. Then

$$\mathbb{P}(\text{observe } k \text{ events}) = \frac{\lambda^k e^{-\lambda}}{k!}$$

where k may take a value 0, 1, 2, and so on, and $k!$ represents k-factorial, as described on page 37. The letter $e \approx 2.718$ is the base of the natural logarithm. The mean and standard deviation of this distribution are λ and $\sqrt{\lambda}$, respectively.

The rigorous explanation of the Poisson distribution is as follows. If $\lambda = np$ where $n \to \infty$ and p remains constant (so $p = \lambda/n \to 0$) then the binomial distribution with parameters n and p converges to Poisson in the following sense. For fixed k (so k is relatively small),

$$\binom{n}{k} p^k (1-p)^{n-k} \to e^{-\lambda} \frac{\lambda^k}{k!}$$

$$e^{-\lambda} = \lim \left(1 - \frac{\lambda}{n}\right)^n = (1-p)^n$$

$$(1-p)^{-k} \approx 1^{-k} = 1$$

So it remains to show

$$\frac{n!}{(n-k)!} p^k \sim \lambda^k = (np)^k$$

where $a \sim b$ means $\lim_{n\to\infty} \frac{a}{b} = 1$. This simplifies to

$$\frac{n!}{(n-k)!} = n(n-1)(n-2)\dots(n-(k+1)) \sim n^k$$

which is clearly true as $n \to \infty$ and k stays fixed.

The important fact that

$$1 = \mathbb{P}(0 \le X < \infty)$$

follows from the Taylor series

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

The expectation of X that is Poisson(λ) is λ, just like the binomial(n, p) has expectation np. We verify as follows, using $t = k - 1$:

$$\begin{aligned} \mathbb{E}(X) &= \sum_{k=0}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=1}^{\infty} k \frac{\lambda^k e^{-\lambda}}{k!} = \sum_{t=0}^{\infty} (t+1) \frac{\lambda^{t+1} e^{-\lambda}}{(t+1)!} \\ &= \sum_{t=0}^{\infty} \frac{\lambda^{t+1} e^{-\lambda}}{t!} = \lambda \sum_{t=0}^{\infty} \frac{\lambda^t e^{-\lambda}}{t!} = 1. \end{aligned}$$

To find the standard deviation we need

$$\begin{aligned}\mathbb{E}(X^2) &= \sum_{k=0}^{\infty} k^2 \frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=1}^{\infty} k^2 \frac{\lambda^k e^{-\lambda}}{k!} = \sum_{t=0}^{\infty} (t+1)^2 \frac{\lambda^{t+1} e^{-\lambda}}{(t+1)!} \\ &= \sum_{t=0}^{\infty} (t+1) \frac{\lambda^{t+1} e^{-\lambda}}{t!} = \lambda(\mathbb{E}(X)+1) = \lambda(\lambda+1)\end{aligned}$$

hence $\sigma^2 = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \lambda(\lambda+1) - \lambda^2 = \lambda$. Thus, Poisson has the unusual property that its variance equals its mean.

3.6 Applications

The Poisson distribution is a popular model for risky events studied by actuaries, in the case where we know how many events happen on average but we do not have values of n and p to set up a binomial or normal distribution. Click on the camera icon above for a video showcasing Guided Practice 3.7 and featuring an author of the present text.

Guided Practice 3.7 Car accidents. Suppose car accidents happen on average 1 every 10,000 miles. What is the probability that after driving 10,000 miles we have no car accidents?[2]

Example 3.8 Sunny Day cards.

The following example may come up in playing with Kindergarteners. Suppose there are five cards, two of which shows hair clips, and one each show scissors, hair brush, and comb. Suppose you pick two cards at random. What is the probability that exactly one of them shows hair clips? This is known as the hypergeometric distribution: the probability that $k = 1$, where $K = 2$, $N = 5$, and $n = 2$, is

$$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} = \frac{\binom{2}{1}\binom{5-2}{2-1}}{\binom{5}{2}} = 60\%.$$

The probability of two hair clips is 10% and thus the probability of no hair clips is 30%. In general the probability, when picking n items from N items without replacement, of picking k of the K many with a particular property Q, is

$$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

because there are $\binom{N}{n}$ total sets of n that you can choose, all equally likely; and those sets in which you choose k many with the property Q are determined by which ones you chose with property Q ($\binom{K}{k}$ many sets of k) and which ones you chose without property Q ($\binom{N-K}{n-k}$ many choices).

Example 3.9 Banking.

A bank portfolio contains a loan that has gone into default. Each time period a proportion p_i of the outstanding balances are paid. Interest is not added to the

[2] We have $\lambda = 1$ and $\mathbb{P}(X = 0) = e^{-\lambda}\lambda^0/0! = 1/e$.

outstanding balances (as that would be cruel, perhaps), but the bank loses money as the borrowers take a long time to repay. The proportion of funds recovered by the bank is, in terms of a discount rate $d = 1/(1+r)$ (where $r > 0$ is an interest rate),

$$s = p_0 + (1-p_0)p_1 d + (1-p_0)(1-p_1)p_2 d^2 + \cdots = \sum_{i=0}^{\infty} d^i p_i \prod_{j=0}^{i-1}(1-p_j)$$

Suppose p_i are independent random variables. Then

$$\begin{aligned}\mathbb{E}(s) &= \sum_{i=0}^{\infty} d^i \,\mathbb{E}\left(p_i \prod_{j=0}^{i-1}(1-p_j)\right) \\ &= \sum_{i=0}^{\infty} d^i \,\mathbb{E}(p_i) \prod_{j=0}^{i-1}(1-\mathbb{E}(p_j)).\end{aligned}$$

Calculate $\mathbb{E}(s)$ when each $\mathbb{E}(p_i) = 1/2$ and $r = 1/2$.[3]

Example 3.10 Earned premium.

An insurance company offers flood insurance at a cost (premium) of $c = \$100$. The insurance is good for one year. Each month there is a probability p of a flood leading to a $\$100,000$ payout. As the months go by, the insurance company considers that it is earning the premium in a linear fashion: after k months, $c\frac{k}{12}$ has been earned. Assume that no more than one payout can be made in a year. Find the probability distribution of the amounth earned after 6 months. For which value of p is the expected amount earned after 6 months equal to 0?[4]

[3] We have $\prod_{j=0}^{i-1}(1-\mathbb{E}(p_j)) = 2^{-i}$ and $d = 2/3$, so

$$\mathbb{E}(s) = \sum_{i=0}^{\infty}(2/3)^i(1/2)(1/2)^i = (1/2)\sum_{i=0}^{\infty}(1/3)^i = (1/2)\frac{1}{1-\frac{1}{3}} = \frac{3}{4}.$$

[4] \$50 will have been earned if there is no flood. If there has been a flood, then $\$100 - \$100,000$ will have been earned. Thus, if X is the amount earned then $P(X=k) = (1-p)^6 \cdot 1_{k=\$50} + (1-(1-p)^6)\cdot 1_{k=-\$99,900}$. The expectation is $0 = E(X) = \$50(1-p)^6 - \$99,900\cdot(1-(1-p)^6)$. This gives, for $\alpha = (1-p)^6$, that $0 = 50\alpha - 99900(1-\alpha)$ and hence $99900 = 99950\alpha$, $\alpha = \frac{99900}{99950} = \frac{9990}{9995} = \frac{1998}{1999}$. Hence $p = 1 - \left(\frac{1998}{1999}\right)^{1/6} \approx 0.000083$.

3.7 Exercises

3.1 Skewed but real.

(a) Give an example of real-life data that can be expressed as a normal distribution.

(b) Give an example of real-life data that can be expressed as a skewed normal distribution (Exercise 2.2).[5]

(c) Give an example of real-life data whose probability density function is constant-power function $f(x) = ax^{-k}$. Such a probability distribution is known as a *power law distribution*.

3.2 Three sigma. Suppose you are flipping a coin.

(a) What is the probability that you will see the first heads on the third flip? The tenth flip? The hundredth flip?

(b) What is the least n such that seeing the first heads after n flips results in a Z-score greater than 3? Use this answer to comment on the discussion on the first page of Chapter 2.

3.3 Negative binomial. Consider a Bernoulli random variable with probability of success p.

(a) Find the probability that $k-1$ successes will be observed in $n-1$ trials.

(b) Use part (a) to find the probability that $k-1$ successes will be observed in $n-1$ trials and that the nth trial will be a success.

(c) Relate your answer to part (b) to the formula for the negative binomial distribution. Why or why not might this be surprising?

3.4 Tens. Let X be normally distributed with mean μ and standard deviation σ.

(a) Assume X takes on only positive values. Write an inequality that expresses that n is at least Z standard deviations greater than μ, for some constants μ and Z.

(b) Rewrite the inequality in part a in terms of Z^2.

(c) Rewrite the inequality in part c assuming X is a binomial variable, so that $\mu = np$ and $\sigma^2 = np(1-p)$ for some p. Using part a, intuitively describe why this assumption is reasonable.

(d) Assume that $np, np(1-p) > 10$. Conclude that $Z^2 < 10$.

(e) Calculate $\Phi(\sqrt{10}) - \Phi(-\sqrt{10})$ (where $\Phi(x)$ is defined in Chapter 1, Exercise 10). State why $np, np(1-p)$ are both required to be greater than 10 in 3.4.2.

3.5 Verifying the variance. Complete the proof of Theorem 3.2 as follows.

(a) Solve for β to find $E(X^2)$.

(b) Use $\sigma^2 = E(X^2) - (E(X))^2$ to find σ^2.

[5] Andel, J., Netuka, I. and Zvara, K. (1984) On threshold autoregressive processes. Kybernetika, 20, 89-106

Chapter 4

Foundations for inference

4.1 Variability in estimates

The standard deviation of a sample mean is an important calculation. If the X_i are independent then we can use the fact that for independent X, Y, $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$. So if $\text{Var}(X_i) = \sigma^2$ for all i then

$$\text{Var}^2\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}^2(X_i) = \frac{1}{n^2}\cdot n\cdot\sigma^2 = \sigma^2/n.$$

Thus the standard deviation of the sample mean is $\sigma/\sqrt{n}$.

Theoretically, the **sampling distribution** is the probability distribution of the sample mean of a distribution. In practice, this means that if you observe many (say m many) sample means $\sum_{i=1}^{n} X_{i,j}/n$, $1 \le j \le m$, from a random variable X with a given probability distribution, the observed distribution of these means will be close to the theoretical sampling distribution. Note that these means all have the same value of n, and indeed the sampling distribution depends on n.

Sampling distribution

The sampling distribution represents the distribution of the point estimates based on samples of a fixed size from a certain population. It is useful to think of a particular point estimate as being drawn from such a distribution. Understanding the concept of a sampling distribution is central to understanding statistical inference.

The standard deviation of the sample mean is called the **standard error (SE)** of the estimate.

SE
standard error

Standard error of an estimate

The standard deviation associated with an estimate is called the *standard error*. It describes the typical error or uncertainty associated with the estimate.

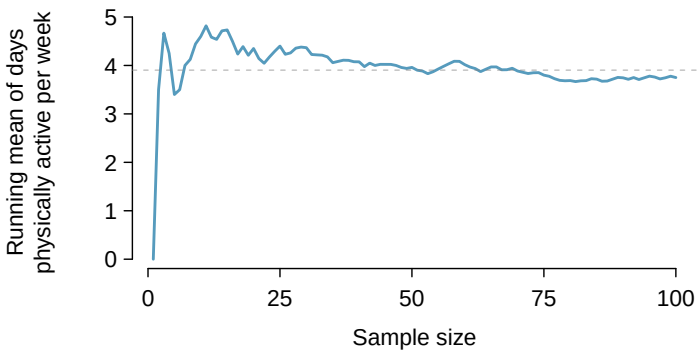


Figure 4.1: The mean computed after adding each individual to the sample. The mean tends to approach the true population average as more data become available.

Computing SE for the sample mean

Given n independent observations from a population with standard deviation σ, the standard error of the sample mean is equal to

$$\text{SE} = \frac{\sigma}{\sqrt{n}} \tag{4.1}$$

A reliable method to ensure sample observations are independent is to conduct a simple random sample consisting of less than 10% of the population.

Example 4.2 Standard example

Suppose you have two Bernoulli random variables X and Y that are independent and have the same parameter p. We can use $\frac{X+Y}{2}$ as our estimate of p. This will tend to be closer to p than if we only used X. For one thing, X can only take the values 0 and 1, whereas $\frac{X+Y}{2}$ can also take the value 1/2. In this case $\sigma = \sqrt{p(1-p)}$ and the standard error is $\sqrt{p(1-p)/2}$. Of course, if we don't know what p is then we don't know this standard error, either, so we use $\text{SE} = \sqrt{\hat{p}(1-\hat{p})/n}$ instead, where $\hat{p} = \frac{X+Y}{2}$. For small n like $n = 2$ here, this SE may be zero, so we should not divide by it. For large n and moderate $0 < p < 1$, however, this will not be a problem.

4.2 Confidence intervals

A point estimate provides a single plausible value for a parameter. However, a point estimate is rarely perfect; usually there is some error in the estimate. Instead of supplying just a point estimate of a parameter, a next logical step would be to provide a plausible *range of values* for the parameter.

A plausible range of values for the population parameter is called a **confidence interval**.

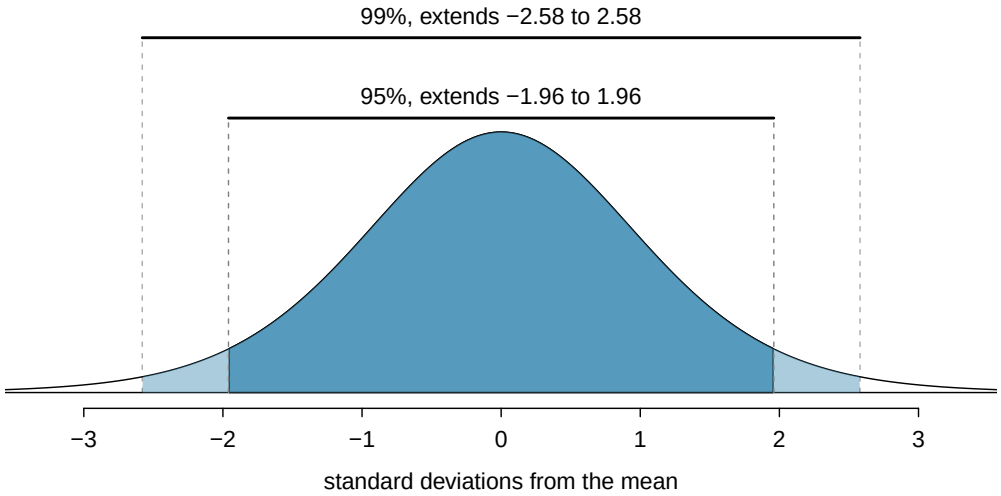


Figure 4.2: The area between -$z^\star$ and $z^\star$ increases as $|z^\star|$ becomes larger. If the confidence level is 99%, we choose $z^\star$ such that 99% of the normal curve is between -$z^\star$ and $z^\star$, which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: $z^\star = 2.58$.

Central Limit Theorem, informal description

If a sample consists of at least 30 independent observations and the data are not strongly skewed, then the distribution of the sample mean is well approximated by a normal model.

The normal approximation is crucial to the precision of these confidence intervals. Section 4.4 provides a more detailed discussion about when the normal model can safely be applied. When the normal model is not a good fit, we will use alternative distributions that better characterize the sampling distribution.

Conditions for $\bar{x}$ being nearly normal and SE being accurate

Important conditions to help ensure the sampling distribution of $\bar{x}$ is nearly normal and the estimate of SE sufficiently accurate:

- The sample observations are independent.
- The sample size is large: $n \geq 30$ is a common rule of thumb.
- The population distribution is itself not too different from the normal distribution in terms of skew.

Additionally, the larger the sample size, the more lenient we can be with the sample's skew.

How to verify sample observations are independent

If the observations are from a simple random sample and consist of fewer than 10% of the population, then they may be considered to be approximately independent.

Subjects in an experiment are considered independent if they undergo random assignment to the treatment groups.

If a sample is from a seemingly random process, e.g. the lifetimes of wrenches used in a particular manufacturing process, checking independence is more difficult. In this case, use your best judgement.

It is too much to ask but it may be worth remembering that 99% confidence corresponds to 2.58 standard deviations and 90% to 1.65.

Confidence interval for any confidence level

If the point estimate follows the normal model with standard error SE, then a confidence interval for the population parameter is

$$\text{point estimate} \pm z^\star \, \text{SE}$$

where $z^\star$ corresponds to the confidence level selected.

Figure 4.2 provides a picture of how to identify $z^\star$ based on a confidence level. We select $z^\star$ so that the area between -$z^\star$ and $z^\star$ in the normal model corresponds to the confidence level. That is, if the area is 95%, say, then we have 95% confidence.

Margin of error

In a confidence interval, $z^\star \times \text{SE}$ is called the **margin of error**.

4.3 Hypothesis testing

If the null hypothesis asserts that $p = 1/2$ then it is making a very bold assertion. Even if $p = 0.500000001$, the null hypothesis is strictly speaking false. This is why we do not "accept" null hypotheses; we cannot gather enough evidence to conclude an assertion with infinite precision. In this setting, a Type 1 error amounts to declaring that p is not $1/2$ when actually it is, which is offensive since p took the trouble to have infinitely many digits correct! A type 2 error is to believe that $p = 1/2$ when it's not, so maybe actually $p = 0.500000001$. This is potentially not so offensive.

Null and alternative hypotheses

The **null hypothesis** (H_0) often represents either a skeptical perspective or a claim to be tested. The **alternative hypothesis** (H_A) represents an alternative claim under consideration and is often represented by a range of possible parameter values.

		Test conclusion	
		do not reject H_0	reject H_0 in favor of H_A
Truth	H_0 true	No error	Type 1 Error
	H_A true	Type 2 Error	No error

Table 4.3: Four different scenarios for hypothesis tests.

The null hypothesis often represents a skeptical position or a perspective of no difference. The alternative hypothesis often represents a new perspective, such as the possibility that there has been a change.

Note that the main text uses *sample standard deviation* to refer to the standard deviation of the original distribution; so this is not the same as the standard deviation of the sample *mean*, which is also called the standard error.

4.3.1 Decision errors

A **Type 1 Error** is rejecting the null hypothesis when H_0 is actually true. A **Type 2 Error** is failing to reject the null hypothesis when the alternative is actually true.

If we reduce how often we make one type of error, we generally make more of the other type.

Hypothesis testing is built around rejecting or failing to reject the null hypothesis. That is, we do not reject H_0 unless we have strong evidence. But what precisely does *strong evidence* mean? As a general rule of thumb, for those cases where the null hypothesis is actually true, we do not want to incorrectly reject H_0 more than 5% of the time. This corresponds to a **significance level** of 0.05. We often write the significance level using α (the Greek letter *alpha*): $\alpha = 0.05$. We discuss the appropriateness of different significance levels in Section 4.3.4.

α
significance level of a hypothesis test

If we use a 95% confidence interval to evaluate a hypothesis test where the null hypothesis is true, we will make an error whenever the point estimate is at least 1.96 standard errors away from the population parameter. This happens about 5% of the time (2.5% in each tail). Similarly, using a 99% confidence interval to evaluate a hypothesis is equivalent to a significance level of $\alpha = 0.01$.

A confidence interval is, in one sense, simplistic in the world of hypothesis tests. Consider the following two scenarios:

- The null value (the parameter value under the null hypothesis) is in the 95% confidence interval but just barely, so we would not reject H_0. However, we might like to somehow say, quantitatively, that it was a close decision.
- The null value is very far outside of the interval, so we reject H_0. However, we want to communicate that, not only did we reject the null hypothesis, but it wasn't even close. Such a case is depicted in Figure 4.4.

In Section 4.3.2, we introduce a tool called the *p-value* that will be helpful in these cases. The p-value method also extends to hypothesis tests where confidence intervals cannot be easily constructed or applied.

4.3.2 Formal testing using p-values

The p-value is a way of quantifying the strength of the evidence against the null hypothesis and in favor of the alternative. Formally the *p-value* is a conditional probability.

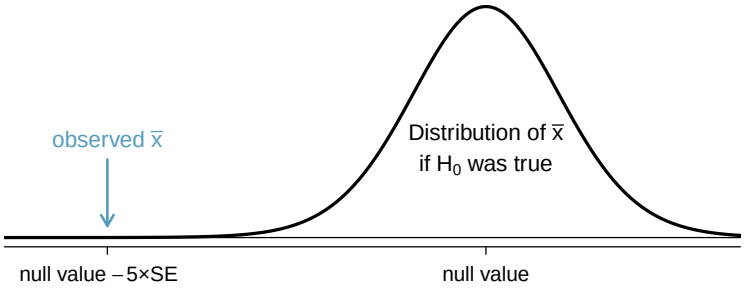


Figure 4.4: It would be helpful to quantify the strength of the evidence against the null hypothesis. In this case, the evidence is extremely strong.

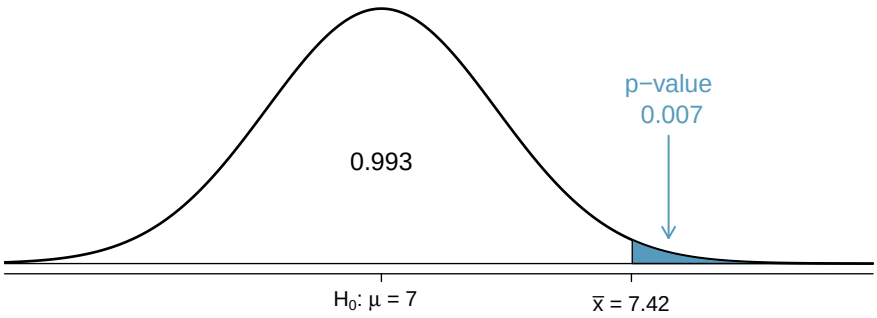


Figure 4.5: Example of p-value for a 1-sided hypothesis test.

p-value

The **p-value** is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis is true. We typically use a summary statistic of the data, in this chapter the sample mean, to help compute the p-value and evaluate the hypotheses.

TIP: Always write the null hypothesis as an equality

We will find it most useful if we always list the null hypothesis as an equality, like:

$$H_0: \quad \mu = 7$$

while the alternative always uses an inequality, like one of the following:

$$\begin{aligned} H_A: &\quad \mu \neq 7, \quad \text{2-sided,} \\ H_A: &\quad \mu > 7, \quad \text{1-sided,} \\ H_A: &\quad \mu < 7, \quad \text{1-sided.} \end{aligned}$$

The p-value is the probability, according to H_0, of observing something at least as

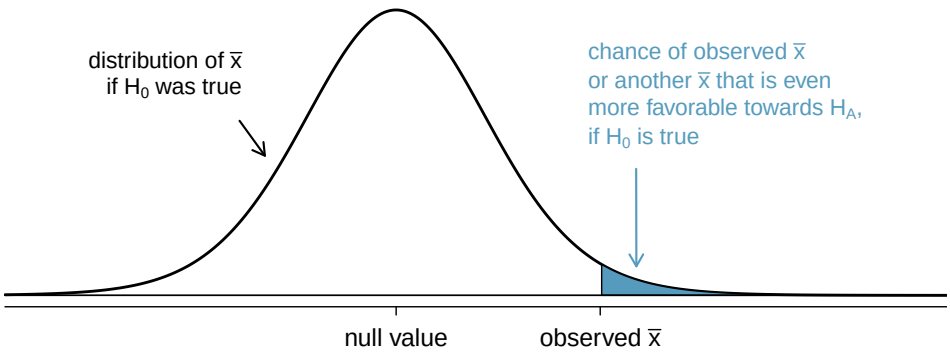


Figure 4.6: To identify the p-value, the distribution of the sample mean is considered as if the null hypothesis was true. Then the p-value is defined and computed as the probability of the observed $\bar{x}$ or an $\bar{x}$ even more favorable to H_A under this distribution.

favorable to H_A as what was actually observed.

If the p-value is less than the significance level (say p-value $= 0.007 < 0.05 = \alpha$), we reject the null hypothesis.

p-value as a tool in hypothesis testing

The smaller the p-value, the stronger the data favor H_A over H_0. A small p-value (usually < 0.05) corresponds to sufficient evidence to reject H_0 in favor of H_A.

What's so special about 0.05?

The main text gives a video that doesn't really explain it. In fact, we can think of two reasons for 0.05: it's approximately 2 standard deviations (1.96) and Fisher suggested using 0.05, about a hundred years ago.

4.3.3 Two-sided hypothesis testing with p-values

To compute a p-value for a two-sided test, consider Figure 4.7.

Caution: One-sided hypotheses are allowed only *before* seeing data

After observing data, it is tempting to turn a two-sided test into a one-sided test. Avoid this temptation. Hypotheses must be set up *before* observing the data. If they are not, the test should be two-sided.

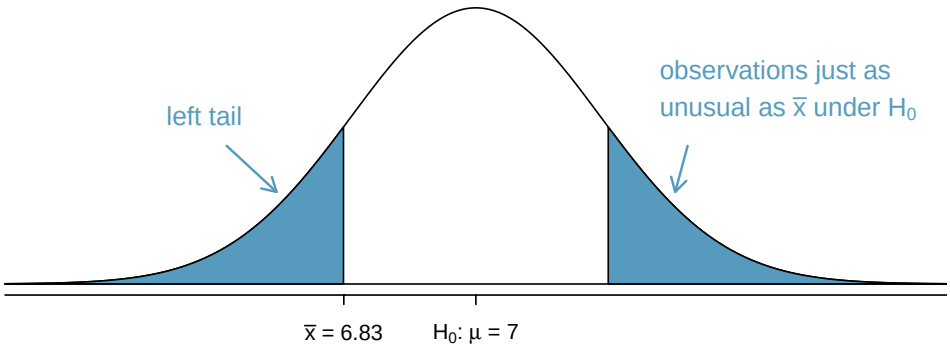


Figure 4.7: H_A is two-sided, so *both* tails must be counted for the p-value.

4.3.4 Choosing a significance level

Choosing a significance level for a test is important in many contexts, and the traditional level is 0.05. However, it is often helpful to adjust the significance level based on the application. We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.

If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring H_A before we would reject H_0.

If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false.

4.4 Examining the Central Limit Theorem

The *OpenIntro* text gives an informal statement of the CLT. Here we shall not be so advanced as to prove the CLT; but we shall give a precise statement of it.

The CDF of the standardized sample mean $(\overline{X} - \mu)/\sigma$ of an IID sequence X_i with $\mathbb{E}(X_i) < \infty$ and $\mathbb{E}(X^2) < \infty$ converges pointwise to the CDF of the standard normal distribution.

So

$$\lim_{n\to\infty} \mathbb{P}((\overline{X}_n - \mu)/\sigma \leq z) = \Phi(z)$$

where

$$\Phi(z) = \mathbb{P}(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-z^2/2}\, dz.$$

The normal model for the sample mean tends to be very good when the sample consists of at least 30 independent observations and the population data are not strongly skewed. The Central Limit Theorem provides the theory that allows us to make this assumption.

Central Limit Theorem, informal definition

The distribution of $\bar{x}$ is approximately normal. The approximation can be poor if the sample size is small, but it improves with larger sample sizes.

The Central Limit Theorem states that when the sample size is small, the normal approximation may not be very good. However, as the sample size becomes large, the normal approximation improves. We will investigate three cases to see roughly when the approximation is reasonable.

We consider three data sets: one from a *uniform* distribution, one from an *exponential* distribution, and the other from a *log-normal* distribution. These distributions are shown in the top panels of Figure 4.8. The uniform distribution is symmetric, the exponential distribution may be considered as having moderate skew since its right tail is relatively short (few outliers), and the log-normal distribution is strongly skewed and will tend to produce more apparent outliers.

The left panel in the $n = 2$ row represents the sampling distribution of $\bar{x}$ if it is the sample mean of two observations from the uniform distribution shown. The dashed line represents the closest approximation of the normal distribution. Similarly, the center and right panels of the $n = 2$ row represent the respective distributions of $\bar{x}$ for data from exponential and log-normal distributions.

Example 4.3 Would the normal approximation be good in all applications where the sample size is at least 30?

Not necessarily. For example, the normal approximation for the log-normal example is questionable for a sample size of 30. Generally, the more skewed a population distribution or the more common the frequency of outliers, the larger the sample required to guarantee the distribution of the sample mean is nearly normal.

Caution: Examine data structure when considering independence

Some data sets are collected in such a way that they have a natural underlying structure between observations, e.g. when observations occur consecutively. Be especially cautious about independence assumptions regarding such data sets.

Caution: Watch out for strong skew and outliers

Strong skew is often identified by the presence of clear outliers. If a data set has prominent outliers, or such observations are somewhat common for the type of data under study, then it is useful to collect a sample with many more than 30 observations if the normal model will be used for $\bar{x}$.

4.4.1 Exponential, log-normal, uniform

We explain the distributions in Figure 4.8.

The **exponential distribution** with parameter λ is an important distribution omitted in *OpenIntro*. It is the continuous-time analogue of the geometric distribution. It has pdf

$$f_X(x) = \lambda \cdot e^{-\lambda x}, \quad 0 \le x < \infty.$$

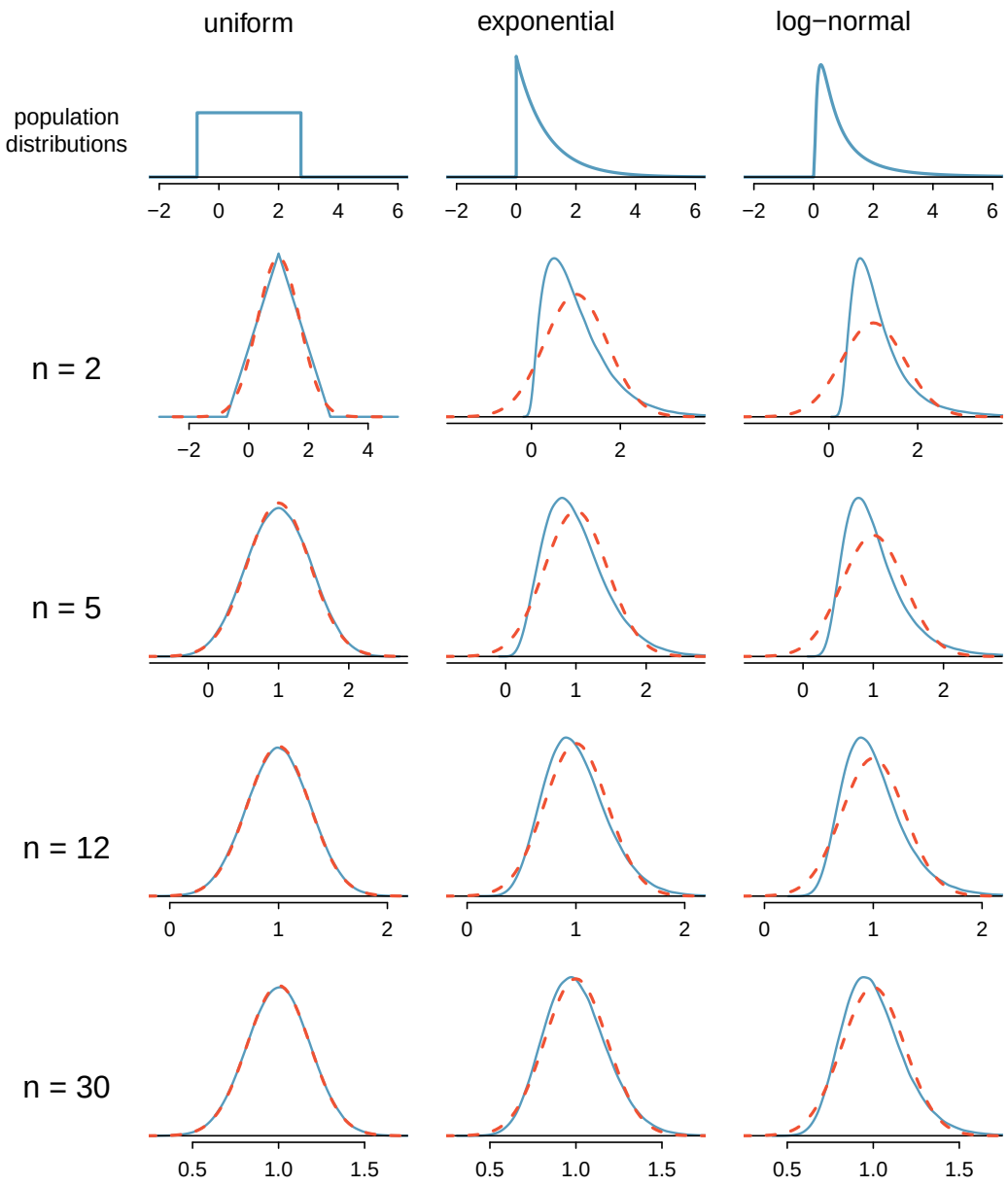


Figure 4.8: Sampling distributions for the mean at different sample sizes and for three different distributions. The dashed red lines show normal distributions.

You can verify that it has a memory-less property shared with the geometric distribution, namely

$$\mathbb{P}(X \geq x + y \mid X \geq x) = \mathbb{P}(X \geq y).$$

The **log-normal** distribution is important in finance. It can be used to model stock prices, which are always nonnegative.

Definition 4.1. *X is a log-normal random variable with parameters μ and σ if $X = e^W$ where the logarithm W is normal with parameters μ and σ.*

The **uniform** distribution on an interval $[a, b]$ has a constant pdf on that interval:

$$f_X(x) = \frac{1}{b-a}, \quad a \leq x \leq b.$$

4.5 Inference for other estimators

The sample mean is not the only point estimate for which the sampling distribution is nearly normal. For example, the sampling distribution of sample proportions closely resembles the normal distribution when the sample size is sufficiently large. In this section, we introduce a number of examples where the normal approximation is reasonable for the point estimate. Chapters 5 and 6 will revisit each of the point estimates you see in this section along with some other new statistics.

We make another important assumption about each point estimate encountered in this section: the estimate is unbiased. A point estimate is **unbiased** if the sampling distribution of the estimate is centered at the parameter it estimates. That is, an unbiased estimate does not naturally over or underestimate the parameter. Rather, it tends to provide a "good" estimate. The sample mean is an example of an unbiased point estimate, as are each of the examples we introduce in this section.

Finally, we will discuss the general case where a point estimate may follow some distribution other than the normal distribution. We also provide guidance about how to handle scenarios where the statistical techniques you are familiar with are insufficient for the problem at hand.

4.5.1 Confidence intervals for nearly normal point estimates

In Section 4.2, we used the point estimate $\bar{x}$ with a standard error $\text{SE}_{\bar{x}}$ to create a 95% confidence interval for the population mean:

$$\bar{x} \pm 1.96 \times \text{SE}_{\bar{x}} \tag{4.4}$$

We constructed this interval by noting that the sample mean is within 1.96 standard errors of the actual mean about 95% of the time. This same logic generalizes to any unbiased point estimate that is nearly normal. We may also generalize the confidence level by using a place-holder $z^\star$.

General confidence interval for the normal sampling distribution case

A confidence interval based on an unbiased and nearly normal point estimate is

$$\text{point estimate} \pm z^\star \, \text{SE} \tag{4.5}$$

where $z^\star$ is selected to correspond to the confidence level, and SE represents the standard error. The value $z^\star \, \text{SE}$ is called the *margin of error*.

Generally the standard error for a point estimate is estimated from the data and computed using a formula. For example, the standard error for the sample mean is

$$\text{SE}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

The variance of a sample mean $\sum_{i=1}^{n} X_i/n$, where the X_i are iid with variance σ^2, is σ^2/n, by the following calculation (which uses the fact, proved in advanced courses, that the variance of a sum of independent random variables is the sum of their variances):

$$\begin{aligned} \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) &= \frac{1}{n^2}\,\text{Var}\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i) \\ &= \frac{n}{n^2}\,\text{Var}(X_1) = \frac{\sigma^2}{n}. \end{aligned}$$

Consequently, the standard deviation of the sample mean is $\sigma/\sqrt{n}$. When σ is unknown we approximate it by s, and hence our estimated standard deviation of the sample mean (the standard error of the mean) is $s/\sqrt{n}$.

4.5.2 Hypothesis testing for nearly normal point estimates

Just as the confidence interval method works with many other point estimates, we can generalize our hypothesis testing methods to new point estimates. Here we only consider the p-value approach, introduced in Section 4.3.2, since it is the most commonly used technique and also extends to non-normal cases.

Hypothesis testing using the normal model

1. First write the hypotheses in plain language, then set them up in mathematical notation.
2. Identify an appropriate point estimate of the parameter of interest.
3. Verify conditions to ensure the standard error estimate is reasonable and the point estimate is nearly normal and unbiased.
4. Compute the standard error. Draw a picture depicting the distribution of the estimate under the idea that H_0 is true. Shade areas representing the p-value.
5. Using the picture and normal model, compute the *test statistic* (Z-score) and identify the p-value to evaluate the hypotheses. Write a conclusion in plain language.

A Z-score is an example of a **test statistic**. In most hypothesis tests, a test statistic is a particular data summary that is especially useful for computing the p-value and evaluating the hypothesis test. In the case of point estimates that are nearly normal, the test statistic is the Z-score.

Test statistic

A *test statistic* is a summary statistic that is particularly useful for evaluating a hypothesis test or identifying the p-value. When a point estimate is nearly normal, we use the Z-score of the point estimate as the test statistic. In later chapters we encounter situations where other test statistics are helpful.

4.5.3 Non-normal point estimates

We may apply the ideas of confidence intervals and hypothesis testing to cases where the point estimate or test statistic is not necessarily normal. There are many reasons why such a situation may arise:

- the sample size is too small for the normal approximation to be valid;
- the standard error estimate may be poor; or
- the point estimate tends towards some distribution that is not the normal distribution.

For each case where the normal approximation is not valid, our first task is always to understand and characterize the sampling distribution of the point estimate or test statistic. Next, we can apply the general frameworks for confidence intervals and hypothesis testing to these alternative distributions.

4.5.4 Statistical significance versus practical significance

When the sample size becomes larger, point estimates become more precise and any real differences in the mean and null value become easier to detect and recognize. Even a

very small difference would likely be detected if we took a large enough sample. Sometimes researchers will take such large samples that even the slightest difference is detected. While we still say that difference is **statistically significant**, it might not be **practically significant**.

For instance, if a coin somehow has an inherent probability $\frac{1}{2} + 10^{-10}$ of landing heads, with a large enough sample size we may be able to convince ourselves that $p \neq 1/2$. If this sample size is larger than the number of times we would ever toss the coin in practice, it is not clear that we have established a practical significance.

4.6 Exercises

4.1 Sidedness.

(a) Give an example where a one-sided test would be more appropriate.

(b) Give an example where a two-sided test would be more appropriate.

4.2 Infinite precision. Using that the null hypothesis is intuitively a position that cannot be proven, only disproven, explain why a null hypothesis is expressed as an equality.

4.3 Moment generating function. Let X be a random variable. Define the *moment generating function* $M_X(t) := E(e^{tX})$. Assume this function is defined for some $t \in \mathbb{R}$.

(a) If X and Y are independent variables and c is a constant, show that $M_{X+Y}(t) = M_X(t)M_Y(t)$ and $M_{cX}(t) = M_X(ct)$ for all t where these expressions are defined.

(b) Use the power series of e^t to show that $M_X(t) = \sum_0^\infty \frac{E(X^n)}{n!} t^n$.

(c) Use part b to show that $M_X^{(n)}(0) = E(X^n)$ for all n.

4.4 Standard error. Let $x_1, \ldots, x_n$ be independent observations of a population with mean μ and standard deviation σ.

1. Calculate the variance of the total $X = x_1 + \cdots + x_n$.
2. Use part a to calculate the variance and standard deviation of X/n.
3. Discuss how your answer to part b is related to the standard error.

4.5 Cumulant generating function. Let X be a random variable. Define the *cumulant generating function* $K_X(t) := \log M_X(t)$, where $M_X(t)$ is the moment generating function in the previous problem.

(a) If X and Y are independent variables and c is a constant, show that $K_{X+Y}(t) = K_X(t)+K_Y(t)$ and $K_{cX}(t) = K_X(ct)$ for all t. Conclude that if $X_1, \ldots, X_n$ are random variables,

$$K_{\frac{X_1+\cdots+X_n}{\sqrt{n}}}(t) = K_{X_1}\left(\frac{t}{\sqrt{n}}\right) + \cdots + K_{X_n}\left(\frac{t}{\sqrt{n}}\right).$$

(b) Use the power series of $\log(t+1)$ to rewrite $K_X(t)$ as a power series such that the coefficient of the first term is $E(X)$ and the coefficient of the second term is $\mathrm{Var}(X)$. *Hint: recall that* $\mathrm{Var}(X) = E(X^2) - E(X)^2$.

(c) Define $K_n(X) := K_X^{(n)}(0)$ and let X_i be random variables. Use parts (a) and (b) to show that

$$K_m\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) = \frac{K_m(X_1) + \cdots + K_m(X_n)}{n^{\frac{m}{2}}}.$$

(d) Use part (c) to show that if the cumulants of X_i are all bounded, then for all $m > 2$,

$$K_m\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right) \to 0 \text{ as } n \to 0.$$

(e) Conclude that the central limit theorem holds for random variables with bounded moment generating functions.

4.6 Geometric mean of lognormal. Calculate the geometric mean (Exercise 1.5) of a log-normal random variable (Definition 4.1) with parameters $\mu = 0$ and $\sigma = 1$.

Chapter 5

Inference for numerical data

5.1 One-sample means with the t-distribution

The main text's so-called "Central Limit Theorem for normal data" has the following more precise version: The sampling distribution of the mean is normal when the sample observations are independent and come from a normal distribution. This is true for any sample size.

5.1.1 Introducing the t-distribution

We define precisely what the t-distribution is here in the supplement. It is simply the distribution of $(\overline{X} - \mu)/(S/\sqrt{n})$ where S is the standard error. This depends on n and is said to have $n - 1$ degrees of freedom (df). (Note that if $n = 1$, we have $S = 0$ with probability 1, so there is no such thing as the t-distribution with 0 degrees of freedom.)

The pdfs get complicated at high df, but the t-distribution with 1 degree of freedom is also known as the Cauchy distribution and has pdf

$$f_T(t) = \frac{1}{\pi}\frac{1}{1+t^2}.$$

⊙ **Guided Practice 5.1** Check that $\int_{-\infty}^{\infty} f_T(t)\,dt = 1$.[1]

We can think about this as follows. We write out the t statistic in the $n = 2$, $n - 1$ df case as:

$$\begin{aligned} T = (\overline{X} - \mu)/(S/\sqrt{n}) &= \frac{\frac{1}{2}(X_1 + X_2) - \mu}{\frac{1}{\sqrt{n}}\sqrt{\frac{1}{2-1}\sum_{i=1}^{2}(X_i - \bar{X})^2}} \\ &= \frac{\frac{1}{2}(X_1 + X_2) - \mu}{\frac{1}{\sqrt{2}}\sqrt{\sum_{i=1}^{2}(X_i - \bar{X})^2}} \\ &= \frac{\frac{1}{2}(X_1 + X_2) - \mu}{\frac{1}{\sqrt{2}}\sqrt{2(\frac{X_1 - X_2}{2})^2}} \\ &= \frac{\frac{1}{2}(X_1 + X_2) - \mu}{|X_1 - X_2|} \end{aligned}$$

[1] Recall that the derivative of $\arctan(t)$ is $1/(1+t^2)$.

For simplicity, take $\mu = 0$. Then

$$T^2 = \frac{\frac{1}{4}(X_1 + X_2)^2}{(X_1 - X_2)^2}$$

and $T^2 \leq t^2$ iff

$$\frac{1}{4}(X_1 + X_2)^2 \leq t^2(X_1 - X_2)^2$$

We would now have to perform a double integral over a region in (X_1, X_2)-plane, but we shall leave that for a course that requires multivariable calculus.

When using a t-distribution, we use a T-score (same as Z-score)

To help us remember to use the t-distribution, we use a T to represent the test statistic, and we often call this a **T-score**. The Z-score and T-score represent how many standard errors the observed value is from the null value. In general we can say that a *score* is an expression of the form $\frac{X - \mathbb{E}(X)}{\mathrm{SD}(X)}$; it will be a Z-score for a normally distributed random variable X and a T-score for a random variable X having a t-distribution.

5.2 Paired data

Mathematically, observations are paired if they are of the form $(X(\omega), Y(\omega))$ for the same outcome ω. So for instance if ω is a textbook, $X(\omega)$ could be its price in the UCLA bookstore, and $Y(\omega)$ its price at Amazon:

Definition 5.1. *Two sets of observations $\mathcal{A}$ and $\mathcal{B}$ are paired if for each $a \in \mathcal{A}$ there is exactly one ω such that $a = X(\omega)$ and $Y(\omega) \in \mathcal{B}$, and vice versa.*

We can then consider the random variable $W(\omega) = X(\omega) - Y(\omega)$ in order to compare the means of the data in $\mathcal{A}$ and $\mathcal{B}$. For arbitrary sets $\mathcal{A}$ and $\mathcal{B}$ (not paired) this would not be possible. Note X and Y are not assumed independent, in fact as the UCLA/Amazon example indicates they typically will not be.

Thus, when drawing n samples $\omega_1, \ldots, \omega_n$ or in another expression $(X_1(\omega), \ldots, X_n(\omega))$, we assume the X_i are independent, as are the Y_i, but X_i is not independent of Y_i.

When the data is not paired, we again have

$$(X_1(\omega), \ldots, X_n(\omega)) \text{ and } (Y_1(\omega'), \ldots, Y_m(\omega'))$$

where $n = m$ is allowed, but not required, and there is no implied relation between $X_1(\omega)$ and $Y_1(\omega')$ or in general between any particular index of the X's and any particular index of the Y's. Here $X_i(\omega) = X(\omega_i)$ where ω is really a vector of outcomes. In the paired case, $\omega = \omega'$. In both cases, when outcomes are distinct we assume they are selected independently.

5.3 Difference of two means

The variance of a difference of (independent) means $\overline{X} - \overline{Y}$ is

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

So far this concerns the normal case. The t-distribution is used when the sample sizes are small enough that we cannot trust that the sample standard deviation is close enough to the actual standard deviation. Thus we use the standard error,

$$\text{SE} = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}.$$

5.3.1 Confidence interval for a difference of means

Using the t-distribution for a difference in means

The t-distribution can be used for inference when working with the standardized difference of two means if (1) each sample meets the conditions for using the t-distribution and (2) the samples are independent.

We can quantify the variability in the point estimate, $\bar{x}_1 - \bar{x}_2$, using the following formula for its standard error:

$$\text{SE}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

We usually estimate this standard error using standard deviation estimates based on the samples:

$$\begin{aligned} \text{SE}_{\bar{x}_1 - \bar{x}_2} &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ &\approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \end{aligned}$$

Because we will use the t-distribution, we also must identify the appropriate degrees of freedom. This can be done using computer software. An alternative technique is to use the smaller of $n_1 - 1$ and $n_2 - 1$, which is the method we will typically apply in the examples and guided practice. This technique for degrees of freedom is conservative with respect to a Type 1 Error; it is more difficult to reject the null hypothesis using this df method.

5.3.2 The Welch t-test

Welch discovered that in the case of unequal variances, it is approximately correct to use a t-distribution with a peculiar count for degrees of freedom.

The search for an exact test is ongoing and is called the *Behrens-Fisher problem*. The problem is that, whereas Gosset (Student) showed that $(\overline{X} - \mu)/S$ has a distribution that does not depend on the unknown σ, the analogous quantity in the 2-sample case $(\overline{X}_1 - \overline{X}_2 - 0)/\sqrt{S_1^2/n_1 + S_2^2/n_2}$ has not that property. So we would want some function $f(\overline{X}_1, \overline{X}_2, S_1, S_2)$ which used the information given in a powerful way, and had a known distribution under the null hypothesis.

Welch's number of degrees of freedom (which may not be an integer) is

$$\nu := \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2 \nu_1} + \frac{s_2^4}{n_2^2 \nu_2}}$$

where $\nu_i = n_i - 1$. The claim is now that the quantity $\mu := \min(\nu_1, \nu_2)$ is a conservative count compared to ν. In other words, we are more likely to reject our null hypothesis with ν. That is, for a given value t, the probability of something as extreme as t is lower (hence more likely to be below $\alpha = 5\%$ say) according to the $t(\nu)$ distribution than according to the $t(\mu)$ distribution. In other words, $t(\nu)$ is closer to the normal distribution, which has very low probability of given extreme values t (very thin tails). In other words yet, $\mu \leq \nu$. Let us check that. This means

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2\nu_1} + \frac{s_2^4}{n_2^2\nu_2}} \geq \min(\nu_1, \nu_2).$$

Let us assume $\nu_1 \leq \nu_2$. Then it suffices to show

$$\frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2\nu_1} + \frac{s_2^4}{n_2^2\nu_2}} \geq \nu_1$$

$$\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2 \geq \left(\frac{s_1^4}{n_1^2\nu_1} + \frac{s_2^4}{n_2^2\nu_2}\right)\nu_1.$$

And indeed,

$$\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2 \geq \frac{s_1^4}{n_1^2} + \frac{s_2^4}{n_2^2} \geq \frac{s_1^4}{n_1^2} + \frac{s_2^4\nu_1}{n_2^2\nu_2} = \left(\frac{s_1^4}{n_1^2\nu_1} + \frac{s_2^4}{n_2^2\nu_2}\right)\nu_1.$$

We see that when $n_1 = n_2$, the difference between the Welch df and the conservative df is at most 1, basically since $2xy \leq x^2 + y^2$.

Distribution of a difference of sample means

The sample difference of two means, $\bar{x}_1 - \bar{x}_2$, can be modeled using the t-distribution and the standard error

$$\mathrm{SE}_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{5.2}$$

when each sample mean can itself be modeled using a t-distribution and the samples are independent. To calculate the degrees of freedom, use statistical software or the smaller of $n_1 - 1$ and $n_2 - 1$.

5.3.3 Hypothesis tests based on a difference in means

We use $\min(n_1 - 1, n_2 - 1)$ as our df; this is conservative in the sense that it makes it harder to reject a null hypothesis. We can discuss the relationship with Welch's test.

We need the individual distributions of n_1 and n_2 points to be normal, in theory, in order to apply the t-test.

The critical value for the t-test is $t^\star_{\mathrm{df}}$.

5.3.4 Examining the standard error formula

The formula for the standard error of the difference in two means is similar to the formula for other standard errors. Recall that the standard error of a single mean, $\bar{x}_1$, can be approximated by

$$SE_{\bar{x}_1} = \frac{s_1}{\sqrt{n_1}}$$

where s_1 and n_1 represent the sample standard deviation and sample size.

The standard error of the difference of two sample means can be constructed from the standard errors of the separate sample means:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{5.3}$$

This special relationship follows from probability theory.

⊙ **Guided Practice 5.4** Prerequisite: Section 2.4. We can rewrite Equation (5.3) in a different way:

$$SE_{\bar{x}_1 - \bar{x}_2}^2 = SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2$$

Explain where this formula comes from using the ideas of probability theory.[2]

5.3.5 Pooled standard deviation estimate

Occasionally, two populations will have standard deviations that are so similar that they can be treated as identical. For example, historical data or a well-understood biological mechanism may justify this strong assumption. In such cases, we can make the t-distribution approach slightly more precise by using a pooled standard deviation.

The **pooled standard deviation** of two groups is a way to use data from both samples to better estimate the standard deviation and standard error. If s_1 and s_2 are the standard deviations of groups 1 and 2 and there are good reasons to believe that the population standard deviations are equal, then we can obtain an improved estimate of the group variances by pooling their data:

$$s_{pooled}^2 = \frac{s_1^2 \times (n_1 - 1) + s_2^2 \times (n_2 - 1)}{n_1 + n_2 - 2}$$

where n_1 and n_2 are the sample sizes, as before. To use this new statistic, we substitute s_{pooled}^2 in place of s_1^2 and s_2^2 in the standard error formula, and we use an updated formula for the degrees of freedom:

$$df = n_1 + n_2 - 2$$

The benefits of pooling the standard deviation are realized through obtaining a better estimate of the standard deviation for each group and using a larger df parameter for the t-distribution. Both of these changes may permit a more accurate model of the sampling distribution of $\bar{x}_1 - \bar{x}_2$, if the standard deviations of the two groups are equal.

[2] The standard error squared represents the variance of the estimate. If X and Y are two random variables with variances σ_x^2 and σ_y^2, then the variance of $X - Y$ is $\sigma_x^2 + \sigma_y^2$. Likewise, the variance corresponding to $\bar{x}_1 - \bar{x}_2$ is $\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2$. Because $\sigma_{\bar{x}_1}^2$ and $\sigma_{\bar{x}_2}^2$ are just another way of writing $SE_{\bar{x}_1}^2$ and $SE_{\bar{x}_2}^2$, the variance associated with $\bar{x}_1 - \bar{x}_2$ may be written as $SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2$.

Caution: Pool standard deviations only after careful consideration

A pooled standard deviation is only appropriate when background research indicates the population standard deviations are nearly equal. When the sample size is large and the condition may be adequately checked with data, the benefits of pooling the standard deviations greatly diminishes.

There are several flavors of t-test: one-sample, two-sample, paired, and *pooled*. The *pooled* flavor is used when, on the one hand, the sample sizes are small enough that we do not have a good idea what the standard deviations are; but on the other hand, we have reason, somehow, to believe that the standard deviations of the two groups should come out to be the same. Now where does the formula

$$s^2_{\text{pooled}} = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \qquad (*)$$

and degrees of freedom n_1+n_2-2, come from? One answer is that it is an **unbiased estimator** of σ^2, in other words, its expectation is σ^2.

$$\mathbb{E}(s^2_{\text{pooled}}) = \frac{(n_1-1)\,\mathbb{E}(s_1^2) + (n_2-1)\,\mathbb{E}(s_2^2)}{n_1+n_2-2} = \frac{(n_1-1)\sigma^2 + (n_2-1)\sigma^2}{n_1+n_2-2} = \sigma^2.$$

But, we could have just used s_1^2, or say $(s_1^2+s_2^2)/2$, if that were our only goal.[3] Notice that if n_2 is much greater than n_1 then $(*)$ is close to s_2^2. So among all statistics of the form $\alpha s_1^2 + \beta s_2^2$, where $\alpha+\beta=1$, perhaps s^2_{pooled} minimizes

$$\begin{aligned}
\mathbb{E}((\alpha s_1^2 + \beta s_2^2 - \sigma^2)^2) &= \mathbb{E}(\alpha^2 s_1^4 + \beta^2 s_2^4 + \sigma^4 + 2\alpha\beta s_1^2 s_2^2 - 2\alpha s_1^2\sigma^2 - 2\beta s_2^2\sigma^2) \\
&= \alpha^2 f(n_1) + \beta^2 f(n_2) + \sigma^4 + 2\alpha\beta\sigma^4 - 2\alpha\sigma^4 - 2\beta\sigma^4 \\
&= \alpha^2 f(n_1) + \beta^2 f(n_2) + (1 + 2\alpha\beta - 2\alpha - 2\beta)\sigma^4 \\
&= \alpha^2 f(n_1) + \beta^2 f(n_2) + (2\alpha\beta - 1)\sigma^4
\end{aligned}$$

where[4]

$$\begin{aligned}
f(n) = \mathbb{E}(s^4) &= \left[\frac{n}{n-1}\right]^2 \frac{n-1}{n^3}((n-1)\mu_4 + (n^2-n+3)\mu_2^2) \\
&= \frac{1}{n(n-1)}((n-1)\mu_4 + (n^2-n+3)\mu_2^2)
\end{aligned}$$

for sample size n where μ_n is the nth central moment. For normal distributions $\mu_4 = 3\sigma^4$ and $\mu_2 = \sigma^2$ so

$$f(n) = \frac{1}{n(n-1)}(3(n-1) + n^2 - n + 3)\sigma^4 = \frac{n+2}{n-1}\sigma^4$$

So we consider

$$g(\alpha) = \alpha^2\frac{n_1+2}{n_1-1} + (1-\alpha)^2\frac{n_2+2}{n_2-1} + 2\alpha(1-\alpha) - 1$$

[3] You may wonder whether it would make sense to use a weighted average of the standard deviations s_1, s_2 in place of the squared root of a weighted average of s_1^2 and s_2^2. However, the expectation of such a quantity will not be what we want, as $\mathbb{E}(s_1)$ is a more complicated thing that $\mathbb{E}(s_1^2)$.

[4] Source: Wolfram MathWorld http://mathworld.wolfram.com/SampleVarianceDistribution.html, line 23.

and seek to minimize it. Let use write $=^!$ for "we would like to prove that", or more modestly, "we would like to investigate whether". We have:

$$\begin{aligned}
0 &=^! g'(\alpha) = 2\alpha\frac{n_1+2}{n_1-1} - 2(1-\alpha)\frac{n_2+2}{n_2-1} + 2(1-2\alpha) \\
0 &=^! \alpha\frac{n_1+2}{n_1-1} - (1-\alpha)\frac{n_2+2}{n_2-1} + (1-2\alpha) \\
0 &=^! \alpha\frac{3}{n_1-1} - (1-\alpha)\frac{3}{n_2-1} \\
0 &=^! \alpha\frac{1}{n_1-1} - (1-\alpha)\frac{1}{n_2-1} \\
\alpha &=^! \frac{n_1-1}{n_1-1+n_2-1}
\end{aligned}$$

So it is indeed $\frac{n_1-1}{n_1+n_2-2}$.

Example 5.5 Learning assistants.

The University of Hawai'i at Mānoa Department of Mathematics used *learning assistants* in MATH 203 (Calculus for Business and Social Sciences) in Spring 2017 and Fall 2017. Prior to that, data is available for the Fall 2008 semester, the Spring 2009 semester, the Fall 2009 semester, and so on until and including Fall 2016. It was found that the mean grade point (where A is 4.0, A- is 3.7, B+ is 3.3, B is 3.0, and so on) for the period before learning assistants was 2.29 with a standard error (standard deviation for the means) of 0.098. The mean for Spring 2017 and Fall 2017 was 2.385. This is $\frac{2.385-2.29}{0.098}$ standard errors above the null value. However, since we are using two semesters for 2017 and not just one, we need to compare to $0.098/\sqrt{2}$ instead of 0.098. This still doesn't make the impact of learning assistants on mean grades statistically significant, but it makes it closer to significance.

5.4 Exercises

5.1 The price of pairs. As in Section 5.2, let ω be a textbook, let $X(\omega)$ be its price in the UCLA bookstore. Let $Y(\omega)$ be its price on Amazon. Give an example of sets of observations that are paired and an example of sets of observations that are not paired, where an observation is the price of a textbook.

5.2 Tee of infinity. Informally describe the quantity $\lim_{n\to\infty} \frac{S}{\sqrt{n}}$, where S is the standard error. Conclude that the t-distribution approximates the standard distribution as the number of degrees of freedom approaches infinity.

5.3 Abnormality. Briefly describe why each of the following distributions might fail to be normal.

1. 200 homes in a 1000-home district are surveyed about their voting habits.
2. 10 homes in a 1000-home district are surveyed about their preference for one of two candidates.

5.4 Standard error of a difference. If p_1 and p_2 are two population proportions with sample sizes n_1 and n_2 respectively such that $\hat{p_1}$ and $\hat{p_2}$ are normal, verify directly that $SE_{\hat{p_1}-\hat{p_2}} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$.

Chapter 6

Inference for categorical data

6.1 Inference for a single proportion

6.1.1 Identifying when the sample proportion is nearly normal

A sample proportion can be described as a sample mean. If we represent each "success" as a 1 and each "failure" as a 0, then the sample proportion is the mean of these numerical outcomes:

$$\hat{p} = \frac{0 + 1 + 1 + \cdots + 0}{1042} = 0.82$$

The distribution of $\hat{p}$ is nearly normal when the distribution of 0's and 1's is not too strongly skewed for the sample size. The most common guideline for sample size and skew when working with proportions is to ensure that we expect to observe a minimum number of successes (1's) and failures (0's), typically at least 10 of each. The labels **success** and **failure** need not mean something positive or negative. These terms are just convenient words that are frequently used when discussing proportions.

Conditions for the sampling distribution of $\hat{p}$ being nearly normal

The sampling distribution for $\hat{p}$, taken from a sample of size n from a population with a true proportion p, is nearly normal when

1. the sample observations are independent and
2. we expected to see at least 10 successes and 10 failures in our sample, i.e. $np \geq 10$ and $n(1-p) \geq 10$. This is called the **success-failure condition**.

If these conditions are met, then the sampling distribution of $\hat{p}$ is nearly normal with mean p and standard error

$$\text{SE}_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \tag{6.1}$$

$\hat{p}$
sample proportion

p
population proportion

Typically we don't know the true proportion, p, so we substitute some value to check conditions and to estimate the standard error. For confidence intervals, usually the sample proportion $\hat{p}$ is used to check the success-failure condition and compute the standard error. For hypothesis tests, typically the null value – that is, the proportion claimed in the null hypothesis – is used in place of p. Examples are presented for each of these cases in Sections 6.1.2 and 6.1.3.

6.1.2 Confidence intervals for a proportion

To verify that a sampling distribution of $\hat{p}$ is nearly normal we check:

Observations are independent. For instance, we might be using a simple random sample and consist of fewer than 10% of population, which verifies independence.

Success-failure condition. The sample size must also be sufficiently large, which is checked using the success-failure condition. There are certain numbers of "successes" and "failures" in the sample, both greater than 10.

With the conditions met, we are assured that the sampling distribution of $\hat{p}$ is nearly normal. Next, a standard error for $\hat{p}$ is needed, and then we can employ the usual method to construct a confidence interval.

Constructing a confidence interval for a proportion

- Verify the observations are independent and also verify the success-failure condition using $\hat{p}$ and n.
- If the conditions are met, the sampling distribution of $\hat{p}$ may be well-approximated by the normal model.
- Construct the standard error using $\hat{p}$ in place of p and apply the general confidence interval formula.

Example 6.2 Meta-confidence intervals?

A student once asked, is there such a thing as a confidence interval for confidence intervals? To answer this, let's step back and think about what is random and what is not. The confidence interval is a random interval; if you will, the left and right endpoints are two random variables L and R. In the long run, 95% (or whatever confidence level we're working with) of these intervals will contain the parameter we are seeking confidence about. This parameter is not itself random, just unknown. So instead of seeking meta-confidence, we may just study the distribution of L and R.

6.1.3 Hypothesis testing for a proportion

To apply the normal distribution framework in the context of a hypothesis test for a proportion, the independence and success-failure conditions must be satisfied. In a hypothesis test, the success-failure condition is checked using the null proportion: we verify np_0 and $n(1-p_0)$ are at least 10, where p_0 is the null value.

Hypothesis test for a proportion

Set up hypotheses and verify the conditions using the null value, p_0, to ensure $\hat{p}$ is nearly normal under H_0. If the conditions hold, construct the standard error, again using p_0, and show the p-value in a drawing. Lastly, compute the p-value and evaluate the hypotheses.

6.1.4 Thinking deeper about standard errors

When the null hypothesis says $p = 1/2$, our score can be

$$\frac{\hat{p} - 1/2}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

but actually we may substitute $\hat{p} = 1/2$ here in the denominator in order to be fair to the null hypothesis, as mentioned in *OpenIntro Statistics*.

When we do not assume $p_1 = p_2$, our best (in some sense) guess for the standard error is

$$\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

This may be best in a maximum-likelihood sense (Chapter 8) but we should not expect it to be unbiased. For instance:

Guided Practice 6.3

$$\mathbb{E}(\hat{p}(1-\hat{p})) = \left(1 - \frac{1}{n}\right)(p - p^2)$$

[1]

And indeed, it is clear that when $n = 1$, $p(1-\hat{p}) = 0$ for both outcomes, and $\mathbb{E}(0) = 0$.

On the other hand, when we do assume $p_1 = p_2$, the best estimate of p_1 and p_2 is

$$\frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$$

which is just the proportion obtained by pooling the samples together.

Note that if n is small, a proportion $\hat{p} = \frac{1}{n}\sum_{i=1} X_i$ with X_i Bernoulli(p) is in no way close to normal. Thus the t-distribution is never used in connection with proportions (it is only used when we have a small n and n is the number of independent normals available). When n is large, we approximate a proportion using the normal distribution.

6.1.5 Possible pitfall for pooled proportions

When we're doing the standard error of $\hat{p}_1 - \hat{p}_2$ under the assumption that $p_1 = p_2$, we want to use the pooled proportion $\hat{p}$ and then the standard error

$$\sqrt{\frac{\hat{p}(1-\hat{p}}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$$

[1] To get started, rewrite $\hat{p} = \overline{X}$ and show that $\mathbb{E}((\overline{X})^2) = \frac{1}{n}(p + (n-1)p^2)$.

not just

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n_1+n_2}}$$

as the latter is the standard error of the pooled proportion $\hat{p}$, but we're interested in the difference $\hat{p}_1 - \hat{p}_2$, not in what the (possibly same for both) proportion actually is!

In fact, a similar remark pertains to the case when we do not assume $p_1 = p_2$. Consider Exercise 6.4. Let p_y be the male proportion and p_x female. We need the standard error of $\hat{p}_y - \hat{p}_x$ which will be

$$\sqrt{\frac{\hat{p}_y(1-\hat{p}_y)}{1924} + \frac{\hat{p}_x(1-\hat{p}_x)}{3666}}.$$

From this being 0.01 (since the width of the 95%, hence $1.96 \approx 2$ standard deviations, confidence interval is 0.02) and also $\hat{p}_y - \hat{p}_x = 0.04$ we can deduce what they each are (Exercise 6.5). One student in Fall 2018 instead considered $\hat{p} = \hat{p}_y - \hat{p}_x$ and then

$$\sqrt{\hat{p}(1-\hat{p})/n}$$

but this is treating $\hat{p}$ as a proportion. Note that $\hat{p} < 0$ is possible, so $\hat{p}$ is not quite the same kind of beast as $\hat{p}_y$ and $\hat{p}_x$.

6.2 Difference of two proportions

We would like to make conclusions about the difference in two population proportions: $p_1 - p_2$.

In our investigations, we first identify a reasonable point estimate of $p_1 - p_2$ based on the sample. You may have already guessed its form: $\hat{p}_1 - \hat{p}_2$. Next, in each example we verify that the point estimate follows the normal model by checking certain conditions. Finally, we compute the estimate's standard error and apply our inferential framework.

6.2.1 Sample distribution of the difference of two proportions

We must check two conditions before applying the normal model to $\hat{p}_1 - \hat{p}_2$. First, the sampling distribution for each sample proportion must be nearly normal, and secondly, the samples must be independent. Under these two conditions, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ may be well approximated using the normal model.

Conditions for the sampling distribution of $\hat{p}_1 - \hat{p}_2$ to be normal

The difference $\hat{p}_1 - \hat{p}_2$ tends to follow a normal model when

- each proportion separately follows a normal model, and
- the two samples are independent of each other.

The standard error of the difference in sample proportions is

$$\text{SE}_{\hat{p}_1-\hat{p}_2} = \sqrt{\text{SE}_{\hat{p}_1}^2 + \text{SE}_{\hat{p}_2}^2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \tag{6.4}$$

where p_1 and p_2 represent the population proportions, and n_1 and n_2 represent the sample sizes.

For the difference in two means, the standard error formula took the following form:

$$SE_{\bar{x}_1 - \bar{x}_2} = \sqrt{SE_{\bar{x}_1}^2 + SE_{\bar{x}_2}^2}$$

The standard error for the difference in two proportions takes a similar form. The reasons behind this similarity are rooted in the probability theory of Section 2.4, which is described for this context in Guided Practice 5.4 on page 64.

6.2.2 Confidence intervals for $p_1 - p_2$

In the setting of confidence intervals for a difference of two proportions, the two sample proportions are used to verify the success-failure condition and also compute the standard error, just as was the case with a single proportion.

6.2.3 Hypothesis tests for $p_1 - p_2$

Use the pooled proportion estimate when H_0 is $p_1 - p_2 = 0$

When the null hypothesis is that the proportions are equal, use the pooled proportion ($\hat{p}$) to verify the success-failure condition and estimate the standard error:

$$\hat{p} = \frac{\text{number of "successes"}}{\text{number of cases}} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2}$$

Here $\hat{p}_1 n_1$ represents the number of successes in sample 1 since

$$\hat{p}_1 = \frac{\text{number of successes in sample 1}}{n_1}$$

Similarly, $\hat{p}_2 n_2$ represents the number of successes in sample 2.

6.2.4 More on 2-proportion hypothesis tests (special topic)

When we conduct a 2-proportion hypothesis test, usually H_0 is $p_1 - p_2 = 0$. However, there are rare situations where we want to check for some difference in p_1 and p_2 that is some value other than 0. For example, maybe we care about checking a null hypothesis where $p_1 - p_2 = 0.1$.[2] In contexts like these, we generally use $\hat{p}_1$ and $\hat{p}_2$ to check the success-failure condition and construct the standard error.

In a hypothesis test where the null hypothesis is that $p_1 - p_2 = 0.03$, say, the sample proportions $\hat{p}_1$ and $\hat{p}_2$ are used for the standard error calculation rather than a pooled proportion. Actually one could imagine using something like

$$\hat{p}(1 - \hat{p})/n_1 + (\hat{p} + 0.03)(1 - \hat{p} - 0.03)/n_2 \qquad (*)$$

[2] We can also encounter a similar situation with a difference of two means, though no such example was given in Chapter 5 since the methods remain exactly the same in the context of sample means. On the other hand, the success-failure condition and the calculation of the standard error vary slightly in different proportion contexts.

Is using (*) a good idea? In other words, are these unbiased, or maximum likelihood estimators, or minimum-variance?

Actually, (*) does not really make sense. If anything we should use $\hat{p}_1$ and $\hat{p}_1 + 0.03$, but then that ignores the information in the sample of size n_2. The pooled $\hat{p}$ will give weight to one or the other depending on the relative sizes of n_1 and n_2, and so it does not make sense that it approximates p_1 in particular, or p_2 in particular.

Example 6.5 Curving exams.

Suppose the pass rates for the course MATH 134 in 2008, 2009, 2010 are .61, .63, .58, respectively, and the grand passing rate overall is .6. Suppose in 2009, $n = 100$ students took the course. Is the passing rate of .63 suspiciously close to the overall passing rate of .6? (Suspicion might arise that professors have decided approximately what they want the mean to be before the class even starts, regardless of how well students do.) The standard error is

$$\frac{\sqrt{(0.6)(1-0.6)}}{\sqrt{100}} = \frac{\sqrt{.24}}{10} = .049 = 4.9\%.$$

Our Z-score is

$$\frac{.63 - .6}{.049} = \frac{.03}{.049} = \frac{30}{49} = 0.61,$$

and the probability of obtaining a value within ± 0.61 of 0 for a standard normal distribution is calculated as

```
=normdist(0.61,0,1,TRUE)-normdist(-0.61,0,1,TRUE)
```

in Excel or Google Sheets which yields `0.458`. Thus, we do not reject the null hypothesis that 2009 was a year that just followed a normal distribution with mean 0.6; we do not have evidence of excessive curving.

Let us construct an example where curving does seem to have happened. Suppose in 2008, $\hat{p}_1 = .601$, with $n_1 = 1000$. Then suppose in 2009, $\hat{p}_2 = .599$, with $n_2 = 1000$ as well. Now we can use $p = .6$ as our grand proportion (estimated). The Z-score for a particular year, say 2008, is

$$\frac{.601 - .6}{\sqrt{(.6)(1-.6)/1000}} = \frac{.001}{\sqrt{1/240}} = \sqrt{240}/1000 = \sqrt{2.4}/100 = 0.0155$$

and the probability that a standard normal random variable Z will have $|Z| \leq 0.0155$ is quite small. If each year has a small p-value like this, we can be quite sure that curving occurs. If none of them do, we can be quite sure it doesn't. If some do and some don't, we are getting into a more subtle area: the ones that did may have just done so by coincidence.

Now let us try to construct an example for means rather than proportions. Suppose in 2008 the mean grade point (the grade point for A is 4, for A- is 3.7, for B+ is 3.3, for B is 3, and so on) is 2.9 with 100 students and a sample standard deviation in the grade points of 1.0 (a reasonable number that is sometimes, in fact, used by instructors to "curve the standard deviation"). Suppose in 2009, the mean grade point was 3.0, with the same number of students and sample standard deviation. Is this suspicious? If the true mean, in some sense, is 2.95, what is the Z-score for an observation of 3.0? Since the standard deviation is 1, an individual observation of 3.0 (a single B grade) is 0.05 standard deviations from the mean: The code is

```
=normdist(0.05,0,1,TRUE)-normdist(-0.05,0,1,TRUE)
```

which gives 4.0%. This might be suspicious if the class had just one student.

If 3.0 is the mean for a whole class of 100 students, the Z-score is

$$\frac{3.0 - 2.95}{s/\sqrt{n}} = \frac{0.05}{1/10} = 0.15$$

As n gets large, we expect the means to have smaller standard deviations, which means that a mean close to the true mean μ becomes less and less suspicious. But suppose that while the standard deviation of the individual grades is 1, the means are 2.951 in 2008 and 2.949 in 2009. Then assuming the true mean is 2.95, the Z-score of 2.951 is

$$\frac{2.951 - 2.95}{1/10} = .01$$

which gives a p-value 0.007978712629 via

```
=normdist(0.01,0,1,TRUE)-normdist(-0.01,0,1,TRUE)
```

The problem now is that the only way to get the means that close together (without making them actually equal) may be to increase the sample size; so let us try to make a more specific example. Since grade points are at least 0.3 apart, the sum of grade points in a class must be an integer multiple of 0.1. Thus in a class of 10^k students, the mean is an integer multiple of $10^{-(k+1)}$. On the other hand $\sqrt{10^k} = 10^{k/2}$ so the Z-score (when the grand mean is an integer, say) must be an integer multiple of $10^{k/2-k-1}$. When $k = 2$ this is 10^{-2}. As $k \to \infty$ this does go to zero, meaning that we can make meaningful checks for curving.

6.3 Testing for goodness of fit using χ^2

Sometimes we don't want to investigate a single parameter, but a whole distribution. Say you have a random variable X which takes values in $\{0, \ldots, k\}$ with certain probabilities $p_0, \ldots, p_k$. Now you draw n sample values and observe O_i many times that $X = i$. The expected number of times to observe $X = i$ would be np_i. So if $|O_i - np_i|$ is large we should reject the null hypothesis that the data were drawn from the suspected distribution given by the p_i. The χ^2-statistic gives us a way to package the numbers $|O_i - np_i|$ for $0 \le i \le k$ into a single number for this hypothesis testing task.

We do not study confidence intervals in this setting, although it could be an intriguing idea (imagine a "tube" around the histogram for our probability distribution, showing how sure we are about the values p_i).

6.3.1 The χ^2 test statistic

Squaring here enables a connection with normal random variables. In fact a χ^2 random variable with n degrees of freedom is $\sum_{i=1}^{n} Z_i^2$ where Z_i are independent standard normal random variables.

The test statistic χ^2, which is the sum of the Z^2 values, is generally used for these reasons. We can also write an equation for χ^2 using the observed counts and null counts:

χ^2
chi-square
test statistic

$$\chi^2 = \sum_{i=1}^{\#\text{counts}} \frac{(\text{observed count}_i - \text{null count}_i)^2}{\text{null count}_i}$$

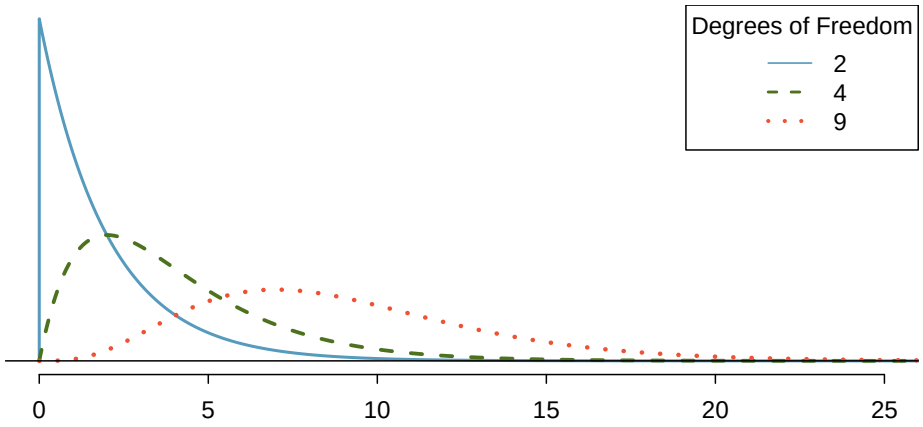


Figure 6.1: Three χ^2 distributions with varying degrees of freedom.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	2	2.41	**3.22**	**4.61**	5.99	7.82	9.21	10.60	13.82
	3	*3.66*	*4.64*	6.25	*7.81*	*9.84*	*11.34*	*12.84*	*16.27*
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32

Table 6.2: A section of a χ^2 table.

The final number χ^2 summarizes how strongly the observed counts tend to deviate from the null counts. In Section 6.3.3, we will see that if the null hypothesis is true, then χ^2 follows a new distribution called a χ^2 *distribution*. Using this distribution, we will be able to obtain a p-value to evaluate the hypotheses.

6.3.2 The χ^2 distribution and finding areas

The χ^2 **distribution** is sometimes used to characterize data sets and statistics that are always positive and typically right skewed. Recall the normal distribution had two parameters – mean and standard deviation – that could be used to describe its exact characteristics. The χ^2 distribution has just one parameter called **degrees of freedom (df)**, which influences the shape, center, and spread of the distribution.

Our principal interest in the χ^2 distribution is the calculation of p-values, which (as we have seen before) is related to finding the relevant area in the tail of a distribution. To do so, a new table is needed: the χ^2 **table**, partially shown in Table 6.2.

Note that we only consider 1-sided tests, where the area of the right tail is of interest. This is because there is only one way to get an extreme result: a large value for the χ^2 statistic. In the normal case we could get a positive or negative result, leading to subtleties related to 1-sidedness and 2-sidedness. The left "tail" would only be of interest if we were testing whether the fit is suspiciously good as in Example 6.5.

6.3.3 Finding a p-value for a χ^2 distribution

Chi-square test for one-way table

Suppose we are to evaluate whether there is convincing evidence that a set of observed counts $O_1, O_2, \ldots, O_k$ in k categories are unusually different from what might be expected under a null hypothesis. Call the *expected counts* that are based on the null hypothesis $E_1, E_2, \ldots, E_k$. If each expected count is at least 5 and the null hypothesis is true, then the test statistic below follows a χ^2 distribution with $k - 1$ degrees of freedom:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \cdots + \frac{(O_k - E_k)^2}{E_k}$$

The p-value for this test statistic is found by looking at the upper tail of this χ^2 distribution. We consider the upper tail because larger values of χ^2 would provide greater evidence against the null hypothesis.

TIP: Conditions for the χ^2 test

There are two conditions that must be checked before performing a χ^2 test:

Independence. Each case that contributes a count to the table must be independent of all the other cases in the table.

Sample size / distribution. Each particular scenario (i.e. cell count) must have at least 5 expected cases.

Failing to check conditions may affect the test's error rates.

When examining a table with just two bins, pick a single bin and use the one-proportion methods introduced in Section 6.1.

6.4 Testing for independence in two-way tables

Computing expected counts in a two-way table

To identify the expected count for the i^{th} row and j^{th} column, compute

$$\text{Expected Count}_{\text{row } i,\ \text{col } j} = \frac{(\text{row } i \text{ total}) \times (\text{column } j \text{ total})}{\text{table total}}$$

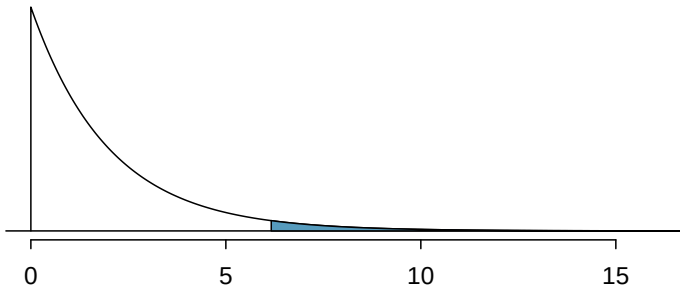


Figure 6.3: A p-value represented as an area under the curve of the pdf for a χ^2 with 2 df.

6.4.1 The χ^2 test for two-way tables

Computing degrees of freedom for a two-way table

When applying the χ^2 test to a two-way table, we use

$$df = (R-1) \times (C-1)$$

where R is the number of rows in the table and C is the number of columns.

TIP: Use two-proportion methods for 2-by-2 contingency tables

When analyzing 2-by-2 contingency tables, we end up with 1 degree of freedom, i.e., Z^2 where Z is standard normal. So then we use the two-proportion methods introduced in Section 6.2.

6.4.2 The χ^2 distribution with 1 degree of freedom

For $a \geq 0$,

$$\begin{aligned} F(a) = \mathbb{P}(Z^2 \leq a) &= \mathbb{P}(-\sqrt{a} \leq Z \leq \sqrt{a}) \\ &= \mathbb{P}(Z \leq \sqrt{a}) - \mathbb{P}(Z \leq -\sqrt{a}) \\ &= \Phi(\sqrt{a}) - (1 - \Phi(\sqrt{a})) \\ &= 2\Phi(\sqrt{a}) - 1; \\ f(a) &= 2f_Z(\sqrt{a}) \cdot \frac{1}{2\sqrt{a}} = \frac{1}{\sqrt{a}} \cdot \frac{1}{\sqrt{2\pi}} e^{-a/2}. \end{aligned}$$

Note that

$$\lim_{a \to 0^+} f(a) = \frac{1}{0 \cdot 1} = +\infty.$$

so the pdf has a vertical asymptote at 0.

For 2 degrees of freedom, we would have to consider

$$\mathbb{P}(Z_1^2 + Z_2^2 \leq a) = \iint \dots$$

which is done in advanced courses (that assume knowledge of multivariable calculus). Would we use $\chi^2(1)$ for flipping a coin? Let X be binomial with parameters n and p, then

$$\chi^2 = \frac{(X - np)^2}{np} + \frac{(n - X - n(1-p))^2}{n(1-p)} = \left(\frac{X - np}{\sqrt{np(1-p)}}\right)^2$$

so using χ^2 in this case would be no different from using the binomial approximation to the normal distribution.

⊙ **Guided Practice 6.6** Verify the computation just made.[3]

[3]The intermediate step is

$$\frac{(X - np)^2(1-p) + (np - X)^2 p}{np(1-p)}.$$

6.5 Exercises

6.1 One-half factorial. Define the gamma function $\Gamma(x) := \int_0^\infty t^{x-1}e^{-t}dt$, where x is a non-negative real number.

(a) Show that $\Gamma(x+1) = x\Gamma(x)$ for all x. Conclude that $\Gamma(n+1) = n!$ for all natural numbers n. *Hint: use integration by parts for the first part, and calculate* $\Gamma(1)$ *and apply induction for the second.*

(b) Calculate $\Gamma\left(\frac{1}{2}\right)$.

6.2 Moments with χ^2. Let a random variable X have the *chi-square density with r degrees of freedom*, or X *is* $\chi^2(r)$ for short, if its probability density function is of the form

$$f(x) = \frac{1}{\Gamma\left(\frac{r}{2}\right)2^{\frac{r}{2}}}x^{\frac{r}{2}-1}e^{-\frac{x}{2}} \tag{6.7}$$

(a) Using Exercise 6.1(b), show that for any r, f from (6.7) is indeed a probability density function.

(b) If $X = \chi^2(r)$, show that $M_X(t) = \frac{1}{(1-2t)^{\frac{r}{2}}}$, where $M_X(t)$ is the moment generating function of X.

(c) Use part (b) to show that if X is a normal random variable with expected value 0 and variance 1, then X^2 is $\chi^2(1)$.

(d) Use part (c) to show that if $X_1, \ldots, X_n$ are mutually independent normal random variables, each with expected value 0 and variance 1, then $\sum_{i=1}^n X_i^2$ is $\chi^2(n)$.

6.3 Difference of χ^2. Let X_1 and X_2 be independent, let X_1 be $\chi^2(r_1)$, and let $X_1 + X_2$ be $\chi^2(r)$ for some $r > r_1$. Show that X_2 is $\chi^2(r - r_1)$.

6.4 Gender and color preference. (This is Exercise 6.25 from *OpenIntro Statistics*, copied here because of Exercise 6.5 below.) A 2001 study asked 1,924 male and 3,666 female undergraduate college students their favorite color. A 95% confidence interval for the difference between the proportions of males and females whose favorite color is black ($p_{male} - p_{female}$) was calculated to be (0.02, 0.06). Based on this information, determine if the following statements are true or false, and explain your reasoning for each statement you identify as false.

(a) We are 95% confident that the true proportion of males whose favorite color is black is 2% lower to 6% higher than the true proportion of females whose favorite color is black.

(b) We are 95% confident that the true proportion of males whose favorite color is black is 2% to 6% higher than the true proportion of females whose favorite color is black.

(c) 95% of random samples will produce 95% confidence intervals that include the true difference between the population proportions of males and females whose favorite color is black.

(d) We can conclude that there is a significant difference between the proportions of males and females whose favorite color is black and that the difference between the two sample proportions is too large to plausibly be due to chance.

(e) The 95% confidence interval for ($p_{female} - p_{male}$) cannot be calculated with only the information given in this exercise.

6.5 Finding $\hat{p}_x$ and $\hat{p}_y$. In Exercise 6.4, let p_y be the male proportion and p_x female. We need the standard error of $\hat{p}_y - \hat{p}_x$ which will be

$$\sqrt{(\hat{p}_y(1-\hat{p}_y)/1924) + (\hat{p}_x(1-\hat{p}_x)/3666)}.$$

From this being 0.01 (since the width of the 95%, hence $1.96 \approx 2$ standard deviations, confidence interval is 0.02) and also $\hat{p}_y - \hat{p}_x = 0.04$, deduce what they each are.

6.6 American climate perspectives. A 2018 survey[4] found a value $\hat{p} = 79\%$ and reported a margin of error of 3.5%. The sample size was $n = 800$ and the confidence level was 95% as usual. Did they calculate their standard error using $\hat{p} = .79$ or using an assumed value $p = 1/2$? What would the margin of error be with the other choice?

[4] https://ecoamerica.org/wp-content/uploads/2018/10/november-2018-american-climate-perspectives-survey.pdf

Chapter 7

Introduction to linear regression

For data that seems to have a nonlinear trend, we can often transform the variables to obtain a linear trend. Linear relationships have the pleasing feature that the correlation coefficient $\rho = \rho_{X,Y}$ is ± 1 exactly when there is a perfect linear relationship between the variables X and Y[1].

7.1 Deriving formulas for linear regression

We seek that values b_0, b_1 that minimize the sum of squared errors $(y_i - (b_0 + b_1 x_i))^2$. So we take the derivatives with respect to b_0 and b_1 and set them equal to zero. For fixed b_1, let

$$f(b_0) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

and let us solve

$$0 = f'(b_0) = -2 \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))$$

Dividing by $-2n$,

$$0 = \overline{y} - b_0 - b_1 \overline{x}$$

which means $\overline{y} = b_0 + b_1 \overline{x}$. In other words, the point of means $(\overline{x}, \overline{y})$ lies on the regression line. Now for fixed b_0, we want to solve for b_1. It will be useful to use the notations

$$\begin{aligned} \overline{xy} &= \frac{1}{n} \sum_{i=1}^{n} x_i y_i, \quad \text{and} \\ \overline{x^2} &= \frac{1}{n} \sum_{i=1}^{n} x_i^2 \end{aligned}$$

Let

$$g(b_1) = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

[1] A similar result for, say, quadratic or exponential relationships is not known to us.

Setting the derivative equal to zero:

$$\begin{aligned} 0 &= g'(b_1) = -2\sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))(x_i) \\ 0 &= \overline{xy} - b_0\overline{x} - b_1\overline{x^2} \\ 0 &= \overline{xy} - (\overline{y} - b_1\overline{x})\overline{x} - b_1\overline{x^2} \\ 0 &= \overline{xy} - (\overline{y}\cdot\overline{x}) + b_1[(\overline{x})^2 - \overline{x^2}] \\ b_1 &= \frac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - (\overline{x})^2}. \end{aligned} \tag{7.1}$$

As taught in Calculus III, this is the beginning of the process of finding a maximum. Let us compare our value of b_1 to that stated in *OpenIntro Statistics*:

$$b_1 \stackrel{!}{=} \frac{1}{n-1}\sum(x_i - \overline{x})(y_i - \overline{y})/s_x^2 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{7.2}$$

and we see that our calculation is correct.

Let us furthermore compare this to the correlation coefficient r. We shall have[2]

$$b_1 = r s_y / s_x.$$

In other words,

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Example 7.3 Calculating a regression by hand and computer

For the data set $(-1,0),(0,0),(1,1)$, we have by forward reference to Example 8.6 on page 111 that

$$r = b_1 s_x / s_y = (1/2)\frac{\sqrt{(-1-0)^2 + (1-0)^2}}{\sqrt{2(0-1/3)^2 + (1-1/3)^2}} = \frac{1}{2}\frac{\sqrt{2}}{\sqrt{6/9}} = \frac{\sqrt{3}}{2}.$$

We can verify this in Google Sheets by `=correl(B1:B3,A1:A3)` where `A1:A3` contains the x-coordinates $-1, 0, 1$ and `B1:B3` contains the y-coordinates $0, 0, 1$.

Linear regression assumes that the relationship between two variables, x and y, can be modeled by a straight line:

β_0, β_1
Linear model parameters

$$y = \beta_0 + \beta_1 x \tag{7.4}$$

where β_0 and β_1 represent two model parameters (β is the Greek letter *beta*). These parameters are estimated using data, and we write their point estimates as b_0 and b_1. So far this leaves open whether b_0, b_1 are random variables or particular values of random variables.

Example 7.5 Influential point.

In the example $(-1,0),(0,0),(1,1)$, which point is most influential? We can investigate this by looking at how much b_0 and b_1 changes when the point is removed. We are not so interested in b_0. When none are removed the regression line is $y = \frac{x}{2} + \frac{1}{3}$.

In this case, at least, the least influential (in the sense of b_1) point $(0,0)$ also has the least influential x-value. See Table 7.1.

[2] In *OpenIntro Statistics*, the symbol R is used, but we shall use r. See https://stats.stackexchange.com/questions/134167/is-there-any-difference-between-r2-and-r2 for a discussion of this issue.

Point removed	New b_1	Change in b_1
(-1,0)	1	$\frac{1}{2}$
(0,0)	$\frac{1}{2}$	0
(1,1)	0	$-\frac{1}{2}$

Table 7.1: Influence on regression line for $(-1, 0), (0, 0), (1, 1)$ upon removing a point.

In general, we can investigate the difference between b_1 for three points $(x_1, y_1), (x_2, y_2), (x_3, y_3)$ and for just the two points $(x_1, y_1), (x_2, y_2)$.

7.2 Line fitting, residuals, and correlation

7.2.1 Residuals

Residuals are the leftover variation in the data after accounting for the model fit:

$$\text{Data} = \text{Fit} + \text{Residual}$$

Residual: difference between observed and expected

The residual of the i^{th} observation (x_i, y_i) is the difference of the observed response (y_i) and the response we would predict based on the model fit $(\hat{y}_i)$:[a]

$$e_i = y_i - \hat{y}_i$$

We typically identify $\hat{y}_i$ by plugging x_i into the model.

[a] A "hat" on y, as in $\hat{y}$, is used to signify that this is an estimate, i.e., it is a value predicted by our regression.

7.2.2 Describing linear relationships with correlation

Correlation: strength of a linear relationship

Correlation, which always takes values between -1 and 1, describes the strength of the linear relationship between two variables. We denote the correlation by r.

r
correlation

The (sample!) correlation for observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ is given by

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \frac{y_i - \bar{y}}{s_y}$$

where $\bar{x}$, $\bar{y}$, s_x, and s_y are the sample means and standard deviations for each variable.

7.3 Fitting a line by least squares regression

7.3.1 An objective measure for finding the best line

We choose the line that minimizes the sum of the squared residuals:

$$e_1^2 + e_2^2 + \cdots + e_n^2 \tag{7.6}$$

This corresponds to finding the closest vector in an n-dimensional space and so it fits well with other mathematical theory. For an analogy, the fact that the variance of a sum of independent random variables is equal to the sum of variances is directly due to the use of squares rather than some other mechanism. To demonstrate this with an example: imagine we just used the range, i.e., the difference between the maximum and minimum values of our random variables.

⊙ **Guided Practice 7.7** Is it true that

$$\max(X+Y) - \min(X+Y) = \max(X) - \min(X) + \max(Y) - \min(Y)$$

if X and Y are independent?[3]

⊙ **Guided Practice 7.8** Is it true that

$$E(|X+Y|) = E(|X|) + E(|Y|)$$

if X and Y are independent with $\mu_X = \mu_Y = 0$?[4]

7.3.2 Conditions for the least squares line

For the typical analysis of regression, the residuals are assumed to be normally distributed (with mean zero!), independent as we vary x and as we draw repeated observations for a single x, with the same variance σ^2 for all x. If we are wondering how stock prices depend on temperature for instance, and our data also have time associated with them (so for instance, on Monday the stocks rose, and the temperature remained constant, and so forth), we may question the independence assumption.

7.3.3 Finding the least squares line

We can also find the parameter estimates by applying two properties of the least squares line:

- The slope of the least squares line can be estimated by

$$b_1 = \frac{s_y}{s_x} r \tag{7.9}$$

  where r is the correlation between the two variables, and s_x and s_y are the sample standard deviations of the explanatory variable and response, respectively.

- If $\bar{x}$ is the mean of the horizontal variable (from the data) and $\bar{y}$ is the mean of the vertical variable, then the point $(\bar{x}, \bar{y})$ is on the least squares line.

b_0, b_1 Sample estimates of β_0, β_1

We use b_0 and b_1 to represent the point estimates of the parameters β_0 and β_1.

You might recall the **point-slope** form of a line from math class (another common form is *slope-intercept*). Given the slope of a line and a point on the line, (x_0, y_0), the equation for the line can be written as

$$y - y_0 = slope \times (x - x_0) \tag{7.10}$$

A common exercise to become more familiar with foundations of least squares regression is to use basic summary statistics and point-slope form to produce the least squares line.

TIP: Identifying the least squares line from summary statistics

To identify the least squares line from summary statistics:

- Estimate the slope parameter, b_1, using Equation (7.9).
- Noting that the point $(\bar{x}, \bar{y})$ is on the least squares line, use $x_0 = \bar{x}$ and $y_0 = \bar{y}$ along with the slope b_1 in the point-slope equation:

$$y - \bar{y} = b_1(x - \bar{x})$$

- Simplify the equation.

7.3.4 Using r^2 to describe the strength of a fit

A claim made in *OpenIntro Statistics* is that

$$\frac{s_y^2 - s_{RES}^2}{s_y^2} = r^2$$

where y is the response variable and s_{RES}^2 is the variance of the residuals. We shall prove it in Theorem 7.3.

For instance, this says that if there is a perfect linear relationship ($r^2 = 1$) then $s_{RES}^2 = 0$, which makes sense. And if $r^2 = 0$ then $s_{RES}^2 = s_y$, i.e., the residuals vary just as much as the response variable does, meaning the slope should be 0.

We say that r^2 is the proportion of variation that is explained, although this is perhaps more of a definition than a theorem.

Comparing to Theorem 7.2, we see again that total variation equals residual variation plus explained variation:

$$s_y^2 = s_{RES}^2 + r^2 s_y^2.$$

7.4 Outliers

Outliers can be interesting and require careful consideration. Do they truly belong to the data set or are they an artifact, a result of a mistake of some sort? For instance, if we have

[3] Yes: note that if $P(X = M_X) = \epsilon_1 > 0$ and $P(Y = M_Y) = \epsilon_2 > 0$ are the probabilities of achieving the maximum values then $P(X + Y = M_X + M_Y) = \epsilon_1\epsilon_2 > 0$.

[4] No. Imagine that X, Y can be either 0 or ± 1. Then $|X| + |Y| > |X + Y|$ will happen with positive probability, namely when $X = -Y$. Also, $|X + Y| \leq |X| + |Y|$ with probability 1.

a data set consisting of students' ages that looks like $\{21, 20, 201\}$, we can be sure that 201 is a mistake.

Leverage

Points that fall horizontally away from the center of the cloud tend to pull harder on the regression line, so we call them points with **high leverage**.

To explore the concept of leverage, let us consider the set of points $(0,0)$, $(1,0)$, $(n,1)$. Without the point $(n,1)$, we get the straight line $y=0$. With it, we have $\bar{x}=\frac{n+1}{3}$, $\bar{y}=1/3$,

$$\begin{aligned} s_x^2=\frac{1}{3-1}\sum\left(x_i-\bar{x}\right)^2 &= \frac{1}{2}\left(\left(\frac{n+1}{3}\right)^2+\left(\frac{n-2}{3}\right)^2+\left(\frac{2n-1}{3}\right)^2\right) \\ &= \frac{1}{18}\left((n+1)^2+(n-2)^2+(2n-1)^2\right) \\ &= \frac{1}{12}(2n-1)^2+\frac{1}{4}=\frac{1}{4}\left(\frac{(2n-1)^2}{3}+1\right) \end{aligned}$$

So we see that s_x is smallest at $n=1/2$ which is also when $(n,1)$ has the least leverage. We have

$$s_y^2=\frac{1}{2}\left(2\left(0-\frac{1}{3}\right)^2+\left(1-\frac{1}{3}\right)^2\right)=\frac{6}{18}=\frac{1}{3}.$$

We also calculate

$$\begin{aligned} \sum\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right) &= \left(0-\frac{n+1}{3}\right)\left(0-\frac{1}{3}\right)+\left(1-\frac{n+1}{3}\right)\left(0-\frac{1}{3}\right)+\left(n-\frac{n+1}{3}\right)\left(1-\frac{1}{3}\right) \\ &= \frac{n+1}{9}+\frac{n-2}{9}+2\frac{2n-1}{9} \\ &= \frac{1}{9}\left(n+1+n-2+2\left(2n-1\right)\right)=\frac{1}{9}\left(6n-3\right)=\frac{2n-1}{3}. \end{aligned}$$

Therefore

$$\begin{aligned} b_1=\frac{1}{n-1}\sum\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)/s_x^2 &= \frac{\frac{1}{2}\frac{2n-1}{3}}{\frac{1}{4}\left(\frac{(2n-1)^2}{3}+1\right)} \\ &= \left(2\left(-1+2n\right)\right)/\left(3\left(1+\left(-1+2n\right)^2/3\right)\right). \end{aligned}$$

⊙ **Guided Practice 7.11** Finish the calculation of the regression line in this case. Is the point $(n,1)$ influential?[5]

⊙ **Guided Practice 7.12** Examine the set of points $(-1,0),(0,0),(n,n)$. Is (n,n) influential?[6]

[5] As $n\to\infty$, the slope goes to 0, which is what it would be without the point $(n,1)$. So in that sense, $(n,1)$ is not very influential.

[6] You should find that (n,n) influences the slope of the regression line a great deal as $n\to\infty$.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-6.7142	5.4567	-1.23	0.2300
unemp	-1.0010	0.8717	-1.15	0.2617
				$df = 25$

Table 7.2: Output from statistical software for the regression line modeling the midterm election losses for the President's party as a response to unemployment.

7.5 Inference for linear regression

We can conduct hypothesis tests for the null hypothesis of the slope being zero, $\beta_1 = 0$, for instance.

7.5.1 Understanding regression output from software

Just like other point estimates we have seen before, we can compute a standard error and test statistic for b_1. We will generally label the test statistic using a T, since it follows the t-distribution.

We will discuss how the standard error is determined.

Example 7.13 What do the first and second columns of Table 7.2 represent?

The entries in the first column represent the least squares estimates, b_0 and b_1, and the values in the second column correspond to the standard errors of each estimate.

We previously used a t-test statistic for hypothesis testing in the context of numerical data. Regression is very similar. In the hypotheses we consider, the null value for the slope is 0, so we can compute the test statistic using the T (or Z) score formula:

$$T = \frac{\text{estimate} - \text{null value}}{\text{SE}} = \frac{-1.0010 - 0}{0.8717} = -1.15$$

We can look for the one-sided p-value using a probability table for the t-distribution.

Inference for regression

We usually rely on statistical software to identify point estimates and standard errors for parameters of a regression line. After verifying conditions hold for fitting a line, we can use the methods learned in Section 5.1 for the t-distribution to create confidence intervals for regression parameters or to evaluate hypothesis tests.

Caution: Don't carelessly use the p-value from regression output

The last column in regression output often lists p-values for one particular hypothesis: a two-sided test where the null value is zero. If your test is one-sided and the point estimate is in the direction of H_A, then you can halve the software's p-value to get the one-tail area. If neither of these scenarios match your hypothesis test, be cautious about using the software output to obtain the p-value.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	b_0	$\text{SE}(b_0)$	18.83	0.0000
family_income	b_1	$\text{SE}(b_1)$	t	0.0002
				$df = 48$

Table 7.3: A summary of least squares fit. The point is that "family_income" here refers to how y depends on family income. If there were more variables we would have rows for how y depends on those as well: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$.

TIP: Always check assumptions
If conditions for fitting the regression line do not hold, then the methods presented here should not be applied. The standard error or distribution assumption of the point estimate – assumed to be normal when applying the t-test statistic – may not be valid.

Now, what does the $t(n-2)$ distribution have to do with the slope estimator b_1?

Well, from Equation (7.1) on page 82, we essentially see that b_1 is a linear combination of normal random variables, with the unknown variance σ^2. (The x-values are not random and are thus treated as constants.) Since the variance is unknown, a t-distribution is used in lieu of a normal distribution.

The text *OpenIntro Statistics 3rd Edition* seems to neglect to clearly note that when b_1 is the estimated slope, under the null hypothesis that $\beta_1 = 0$, we have that

$$T = \frac{b_1 - \beta_1}{\text{SE}(b_1)} = \frac{b_1}{\text{SE}(b_1)}$$

is the variable that has a $t(n-2)$ distribution. Moreover, the regression output tables provide $b_0, b_1, \text{SE}(b_0), \text{SE}(b_1)$ as in Table 7.3.

7.5.2 Statistically significant slopes when correlation coefficient is bounded away from 1

An important statistical distinction is that between statistical significance and **effect size**. A result may be statistically significant without being "big". In this section we show that the p-value associated with a slope can go to zero while the correlation r^2 is bounded away from 1. But first, a different example.

Example 7.14

Imagine points $(2n, 2n+1)$ and $(2n+1, 2n)$. The slope b_1 is 1 in the limit. We have[7]

$$\text{SE}(b_1) = \frac{\sqrt{\sum (y_i - \hat{y}_i)^2/(n-2)}}{\sqrt{\sum (x_i - \overline{x})^2}}.$$

(Don't forget that denominator.) In our example, as the number of points increases, the numerator $\sqrt{\sum (y_i - \hat{y}_i)^2/(n-2)}$ may converge but the denominator

[7] See e.g. David E. Bock, Paul F. Velleman, and Richard D. De Veaux: *Stats: Data and Models*, 4th edition.

$\sqrt{\sum(x_i-\overline{x})^2}$ increases. Thus $b_1/\operatorname{SE}(b_1)$ goes to ∞. On the other hand, $r = b_1 s_x/s_y \approx s_x/s_y$. Now since $y \approx x$ this should all converge to 1.

Theorem 7.1. $\sum \hat{y} = \sum y$.

Proof. We have

$$\frac{1}{n}\sum y = \overline{y} = b_0 + b_1\overline{x} = \frac{1}{n}\left(nb_0 + b_1\sum x\right) = \frac{1}{n}\sum(b_0+b_1x) = \frac{1}{n}\sum \hat{y}. \qquad \square$$

The following Theorem 7.2 is often described as follows: the total variation is the sum of the explained variation and the residual variation. That is, define ESS (explained sum of squares), TSS (total sum of squares), RSS) (residual sum of squares) by

$$\begin{aligned} \text{ESS} &= \sum(\hat{y}-\overline{y})^2 \\ \text{TSS} &= \sum(y-\overline{y})^2 = (n-1)s_y^2 \\ \text{RSS} &= \sum(y-\hat{y})^2 \end{aligned}$$

The term "sum of squares" may seem unhelpful, until you learn that it is to be compared to **mean squares**, which is a sum of squares divided by its number of degrees of freedom.

Theorem 7.2 (Sum of squares decomposition)**.**

$$\text{TSS} = \sum(y-\overline{y})^2 = \sum(\hat{y}-\overline{y})^2 + \sum(y-\hat{y})^2 = \text{ESS} + \text{RSS}\,.$$

Proof. We need

$$\begin{aligned} \sum y^2 - 2y\overline{y} + \overline{y}^2 &= \sum \hat{y}^2 - 2\hat{y}\overline{y} + \overline{y}^2 + y^2 - 2y\hat{y} + \hat{y}^2 \\ \sum -2y\overline{y} &= \sum 2\hat{y}^2 - 2\hat{y}\overline{y} - 2y\hat{y} \\ -2n\overline{y}^2 &= \sum 2\hat{y}^2 - \sum 2\hat{y}\overline{y} - \sum 2y\hat{y} \\ -\sum y\overline{y} &= \sum \hat{y}^2 - \sum \hat{y}\overline{y} - \sum y\hat{y} \\ \sum \hat{y}\overline{y} + \sum y\hat{y} &= \sum y\overline{y} + \sum \hat{y}^2 \end{aligned}$$

Now by Theorem 7.1, $\sum \hat{y} = \sum y$, so we just need

$$\begin{aligned} \sum y\hat{y} &= \sum \hat{y}^2 \\ \sum y(b_0+b_1x) &= \sum(b_0+b_1x)^2 \\ b_0\overline{y} + b_1\overline{xy} &= b_0^2 + 2b_0b_1\overline{x} + b_1^2\overline{x^2} \\ (\overline{y}-b_1\overline{x})\overline{y} + b_1\overline{xy} &= (\overline{y}-b_1\overline{x})^2 + 2(\overline{y}-b_1\overline{x})b_1\overline{x} + b_1^2\overline{x^2} \\ b_1\overline{xy} &= (\overline{y}-b_1\overline{x})(-b_1\overline{x}) + 2(\overline{y}-b_1\overline{x})b_1\overline{x} + b_1^2\overline{x^2} \\ b_1\overline{xy} &= (\overline{y}-b_1\overline{x})b_1\overline{x} + b_1^2\overline{x^2} \\ b_1(\overline{xy} - \overline{x}\cdot\overline{y}) &= b_1^2\overline{x^2} - b_1^2\overline{x}^2 \end{aligned}$$

which upon dividing by b_1 is our formula for b_1 from Equation (7.1). $\square$

Theorem 7.3. $1 - r^2 = \frac{\sum(y-\hat{y})^2/(n-1)}{s_y^2}$.

Proof.

$$\begin{aligned}
1-r^2 &\stackrel{!}{=} \frac{\sum(y-\hat{y})^2/(n-1)}{s_y^2} \\
1-(b_1 s_x/s_y)^2 &\stackrel{!}{=} \frac{\sum(y-\hat{y})^2/(n-1)}{s_y^2} \\
s_y^2-(b_1 s_x)^2 &\stackrel{!}{=} \frac{\sum(y-\hat{y})^2}{n-1} \\
\sum(y-\overline{y})^2 - b_1^2\sum(x-\overline{x})^2 &\stackrel{!}{=} \sum(y-\hat{y})^2 \\
\sum(y-\overline{y})^2 - \sum(b_0+b_1x-(b_0+b_1\overline{x}))^2 &\stackrel{!}{=} \sum(y-\hat{y})^2 \\
\sum(y-\overline{y})^2 - \sum(\hat{y}-\overline{y})^2 &\stackrel{!}{=} \sum(y-\hat{y})^2 \\
\sum(y-\overline{y})^2 &\stackrel{!}{=} \sum(\hat{y}-\overline{y})^2 + \sum(y-\hat{y})^2
\end{aligned}$$

which is true by Theorem 7.2. □

Theorem 7.4.

$$r^2 = \frac{\text{ESS}}{\text{TSS}} = \frac{\sum(\hat{y}-\overline{y})^2}{\sum(y-\overline{y})^2}.$$

Proof. We have

$$1-r^2 = \frac{\sum(y-\hat{y})^2/(n-1)}{s_y^2} = \frac{[\sum(y-\overline{y})^2 - \sum(\hat{y}-\overline{y})^2]/(n-1)}{s_y^2} = 1-\frac{\sum(\hat{y}-\overline{y})^2/(n-1)}{s_y^2}.$$

□

Theorem 7.5. *The t statistic $t = b_1/\operatorname{SE}(b_1)$ is also given by*

$$t = r\sqrt{(n-2)/(1-r^2)}.$$

Note that $r/\sqrt{1-r^2}$ goes to ∞ only when $r \to 1$, but $\sqrt{n-2}$ goes to infinity anyway. So if r is a fixed value (there is a certain amount of spread around the regression line and it's not going away) then we should achieve statistical significance as $n \to \infty$. Namely, imagine that X and Y are correlated jointly normal random variables with a correlation coefficient ρ; for large n we will have $r \approx \rho$.

Proof of Theorem 7.5.

$$\begin{aligned}
\frac{b_1}{\mathrm{SE}(b_1)} &\stackrel{!}{=} \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \\
\frac{rs_y}{s_x\,\mathrm{SE}(b_1)} &\stackrel{!}{=} \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \\
\frac{s_y}{s_x\,\mathrm{SE}(b_1)} &\stackrel{!}{=} \frac{\sqrt{n-2}}{\sqrt{1-r^2}} \\
\frac{s_y^2}{s_x^2\,\mathrm{SE}(b_1)^2} &\stackrel{!}{=} \frac{n-2}{1-r^2} \\
\frac{s_y^2}{s_x^2 \frac{\sum(y-\hat{y})^2/(n-2)}{(n-1)s_x^2}} &\stackrel{!}{=} \frac{n-2}{1-r^2} \\
\frac{s_y^2}{\frac{\sum(y-\hat{y})^2/(n-2)}{(n-1)}} &\stackrel{!}{=} \frac{n-2}{1-r^2} \\
\frac{s_y^2}{\sum(y-\hat{y})^2/(n-1)} &\stackrel{!}{=} \frac{1}{1-r^2} \\
1-r^2 &\stackrel{!}{=} \frac{\sum(y-\hat{y})^2/(n-1)}{s_y^2}
\end{aligned}$$

which is true by Theorem 7.3. □

Consequently, we can also express r^2 in terms of t;

$$\begin{aligned}
t &= r\sqrt{(n-2)/(1-r^2)} && (7.15)\\
t^2 &= r^2(n-2)/(1-r^2) && (7.16)\\
(1-r^2)t^2 &= r^2(n-2) && (7.17)\\
t^2 &= r^2(n-2+t^2) && (7.18)\\
r^2 &= \frac{t^2}{t^2+n-2} && (7.19)
\end{aligned}$$

It is also instructive to express t^2 by

$$t^2 = \frac{r^2(n-2)}{1-r^2} = \frac{(n-2)\,\mathrm{ESS}\,/\,\mathrm{TSS}}{1-\mathrm{ESS}\,/\,\mathrm{TSS}} = \frac{(n-2)\,\mathrm{ESS}}{\mathrm{TSS}-\mathrm{ESS}} = \frac{\mathrm{ESS}\,/1}{\mathrm{RSS}\,/(n-2)} \tag{7.20}$$

Without explaining it in detail, we remark that in the equation

$$\underbrace{\mathrm{TSS}}_{n-1\text{ df}} = \underbrace{\mathrm{ESS}}_{1\text{ df}} + \underbrace{\mathrm{RSS}}_{n-2\text{ df}}$$

we have $n-1$ df for TSS, which breaks down into 1 df for ESS and $n-2$ df for RSS. The ratio in (7.20) has an F **distribution** $F(df_1, df_2)$ where $df_1 = 1$ and $df_2 = n-2$. Thus if we define a random variable $F = t^2$, we have for in the simple linear regression setting that

$$r^2 = \frac{F}{F + df_2/df_1}.$$

There is an underlying algebraic lemma here:

Lemma 7.6. *Let F, a, u_1, u_2, r be numbers. If $F = \frac{a/u_1}{b/u_2}$ and $r^2 = \frac{a}{a+b}$, then F and r^2 are related by*

$$r^2 = \frac{F}{F + u_2/u_1}.$$

Proof. We calculate:

$$\frac{F}{F + u_2/u_1} = \frac{\frac{a/u_1}{b/u_2}}{\frac{a/u_1}{b/u_2} + u_2/u_1} = \frac{a/u_1}{a/u_1 + b/u_1} = r^2. \qquad \square$$

We cannot recover a and b from F, r^2, u_1 and u_2, but we can recover their ratio a/b.

7.5.3 Multiple regression

In the multiple regression case $k > 1$ this generalizes to an $F(n-k-1, k)$ distribution[8] for

$$F = \text{RMS} \,/\, \text{EMS} = \frac{\text{RSS}\,/(n-k-1)}{\text{ESS}\,/(k)} = \frac{r^2(n-k-1)}{(1-r^2)k}$$

but RSS and ESS are calculated the same, as they are not really about the x's.

The standard error associated with a slope b_1 for a model $z = a + bx + cy$ is[9]

$$\text{SE}(b_1) = \frac{s_z}{s_x}\sqrt{\frac{1 - r^2_{z,(x,y)}}{(1 - r^2_{x,y})(n-k-1)}}$$

where $k = 2$ for the two explanatory variables x, y. Let us investigate how regression works for the model $z = a + bx + cy$ or equivalently $z = a + b_1 x + b_2 y$.

We seek a, b, c to minimize $f(a, b, c) = \sum(z_i - a - bx_i - cy_i)^2$. Setting the partial derivatives equal to zero we get

$$\begin{pmatrix} \overline{z} \\ \overline{xz} \\ \overline{yz} \end{pmatrix} = \begin{pmatrix} 1 & \overline{x} & \overline{y} \\ \overline{x} & \overline{x^2} & \overline{xy} \\ \overline{y} & \overline{xy} & \overline{y^2} \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

With some matrix algebra this gives, besides $a = \overline{z} - b\overline{x} - c\overline{y}$, and in terms of the shorthands $\text{cov}(x, y) = \overline{xy} - \overline{x}\,\overline{y}$ and $\text{var}(x) = \text{cov}(x, x)$, that

$$\text{slope}^{(2)}_{z,x} := b = \frac{\text{cov}(x,z)\,\text{var}(y) - \text{cov}(y,z)\,\text{cov}(x,y)}{\text{var}(x)\,\text{var}(y) - (\text{cov}(x,y))^2}$$

To understand this we introduce the notation $\text{slope}^{(1)}_{y,x} = \frac{\text{cov}(x,y)}{\text{var}(x)}$ which happens to be the slope for the simpler model $y = a + bx$. Then we can write

$$b = \frac{\text{slope}^{(1)}_{z,x} - \text{slope}^{(1)}_{z,y}\,\text{slope}^{(1)}_{y,x}}{1 - r^2_{x,y}}$$

since $r^2_{xy} = \frac{\text{cov}(x,y)^2}{\text{var}(x)\,\text{var}(y)}$. (We sometimes write a subscript x, y as simply xy.)

[8] See `https://www3.nd.edu/fwilliam/stats2/l02.pdf`

[9] See `nd.edu/fwilliam/stats1/x91.pdf`

Now in the formula for $\mathrm{SE}(b_1)$ above, the term $r_{z,(x,y)}$ represents the correlation between z and the best-fitting approximation to z by any linear combination of x and y. It is given[10] by

$$\begin{aligned} r^2_{z,(x,y)} &= \begin{pmatrix} r_{x,z} & r_{y,z} \end{pmatrix} \begin{pmatrix} r_{xx} & r_{xy} \\ r_{xy} & r_{yy} \end{pmatrix}^{-1} \begin{pmatrix} r_{xz} \\ r_{yz} \end{pmatrix} \\ &= \frac{r^2_{xz} + r^2_{yz} - 2r_{xy}r_{yz}r_{zx}}{1 - r^2_{xy}} \end{aligned}$$

where we have done some matrix algebra in the last step. Now let us consider

$$F_1 = t_1^2 = b_1^2 / \mathrm{SE}(b_1)^2$$

and similarly F_2.

Now[11] the appropriate F-statistic for testing whether $\beta_1 = \beta_2 = 0$ (where β is the "true" parameter being estimated by b) is

$$F = \frac{\mathrm{EMS}}{\mathrm{RMS}} = \frac{\mathrm{ESS}/d_1}{\mathrm{RSS}/d_2} = \frac{\sum(\hat{z} - \overline{z})^2/p}{\sum(z - \hat{z})^2/(n-p-1)}$$

where $p = 2$ is the number of explanatory variables. This has an $F(p, n-p-1)$ distribution. Doing the algebra we see that the numerator of F is

$$\mathrm{var}(x)(\mathrm{SE}(b_1))^2 \left[t_1^2 + t_2^2 + 2t_1 t_2 \frac{\mathrm{cov}(x,y)}{\mathrm{var}(y)} \right].$$

Thus the two F statistics F_1 and F_2 are combined nontrivially to form F, except if we ensure that $\mathrm{cov}(x,y) = 0$. We may note that $\mathrm{SE}(b_1) = \mathrm{SE}(b_2)s_y/s_x$ and $\mathrm{slope}^{(1)}_{xy}\,\mathrm{slope}^{(1)}_{yx} = r^2_{xy}$. Also F_1 gets the interesting form

$$(n-p-1)\frac{r^2_{zx} + r^2_{zy}r^2_{yx} - 2r_{xy}r_{yz}r_{zx}}{1 - r^2_{xy} - r^2_{yz} - r^2_{xz} + 2r_{xy}r_{yz}r_{xz}} = (n-p-1)\frac{(r_{zx} - r_{zy}r_{yx})^2}{1 - r^2_{xy} - r^2_{yz} - r^2_{xz} + 2r_{xy}r_{yz}r_{xz}}.$$

$$= (n-p-1)\frac{(r_{zx} - r_{zy}r_{yx})^2}{1 - (r_{xz} - r_{xy}r_{yz})^2 + r^2_{xy}r^2_{yz} - r^2_{xy} - r^2_{yz}}.$$

which mimics (7.16) if we take the y terms to be 0, $r_{xy} = r_{yz} = 0$.

We have

$$\frac{F}{F + (n-p-1)/p} = \frac{\mathrm{ESS}}{\mathrm{RSS} + \mathrm{ESS}}. \qquad (*)$$

Let us write ESS_x for the explained sum of squares in the case where we only use x (and not y) in our model. If we have independent variables x, y, then values of the effect size measure

$$\eta^2(x) := \frac{\mathrm{ESS}_x}{\mathrm{RSS} + \mathrm{ESS}}$$

can be added[12],

$$\eta^2(x) + \eta^2(y) = \frac{\mathrm{ESS}_x}{\mathrm{RSS} + \mathrm{ESS}} + \frac{\mathrm{ESS}_y}{\mathrm{RSS} + \mathrm{ESS}} \leq 1$$

[10] See `https://en.wikipedia.org/wiki/Multiple_correlation`.

[11] See `http://www.stat.yale.edu/Courses/1997-98/101/anovareg.htm`

[12] See `https://msu.edu/ levinet/eta%20squared%20hcr.pdf`.

as x and y cannot explain the same variation (see Subsection 7.5.4). But with partial η^2,

$$\eta_p^2(x) := \frac{\mathrm{ESS}_x}{\mathrm{RSS} + \mathrm{ESS}_x}$$

a sum like

$$\eta_p^2(x) + \eta_p^2(y) = \frac{\mathrm{ESS}_x}{\mathrm{RSS} + \mathrm{ESS}_x} + \frac{\mathrm{ESS}_y}{\mathrm{RSS} + \mathrm{ESS}_y}$$

gives no such control. It is the partial form that satisfies (*) given that $F = \mathrm{EMS}_x \,/\, \mathrm{RSS}$. If we want to completely throw out the independent variables other than x we could replace RSS by RSS_x.

7.5.4 Adding the variation explained by x and y

If x and y are independent in some sense then they explain different parts of the sum of squares.

Example 7.21 Perfect fit.

If actually $z = x + y$ is a perfect fit and $(0,0), (0,1), (1,0), (1,1)$ are the values for (x, y), then $\bar{z} = 1$ and $\mathrm{TSS} = (0-1)^2 + (2-1)^2 = 2$ but if we only consider x then we see 0 maps to $1/2$ and 1 maps to $3/2$ so maybe $z = x + 1/2$ is the best fit. The ESS for x is then $2(1/2 - 1)^2 + 2(3/2 - 1)^2 = 1$ and similarly for y.

This is reminiscent of the fact that if X and Y are independent random variables (or more generally just uncorrelated) then $\mathrm{Var}(aX + bY) = \mathrm{Var}(aX) + \mathrm{Var}(bY)$. In "found data" independence usually does not obtain, but in a controlled experiment it usually does.

The variation explained by x is

$$\mathrm{ESS}_x = \sum (\overline{z} - \mathrm{slope}_{z,x}^{(1)}\, x - \mathrm{intercept}_{z,x})^2$$

Here $\mathrm{intercept}_{z,x}$ is defined by

$$\overline{z} = \mathrm{slope}_{z,x}^{(1)}\, \overline{x} + \mathrm{intercept}_{z,x}$$

so

$$\mathrm{ESS}_x = \sum (\overline{z} - \mathrm{slope}_{z,x}^{(1)}\, x - [\overline{z} - \mathrm{slope}_{z,x}^{(1)}\, \overline{x}])^2$$

$$= \sum (-\mathrm{slope}_{z,x}^{(1)}\, x - [-\mathrm{slope}_{z,x}^{(1)}\, \overline{x}])^2 = (\mathrm{slope}_{z,x}^{(1)})^2 \,\mathrm{var}(x) \cdot n = \mathrm{cov}(z,x)^2 n / \,\mathrm{var}(x).$$

Similarly $\mathrm{ESS}_y = \mathrm{cov}(z,y)^2 n / \,\mathrm{var}(y)$. So

$$\mathrm{ESS}_x + \mathrm{ESS}_y = \frac{\mathrm{cov}(z,x)^2 n}{\mathrm{var}(x)} + \frac{\mathrm{cov}(z,y)^2 n}{\mathrm{var}(y)} = n \frac{(\mathrm{cov}(z,x)^2 \,\mathrm{var}(y) + \mathrm{cov}(z,y)^2 \,\mathrm{var}(x)}{\mathrm{var}(x)\,\mathrm{var}(y)}$$

On the other hand,

$$\mathrm{ESS} = \sum (\overline{z} - bx - cy - \mathrm{intercept}_{z,(x,y)})^2$$

where $\mathrm{intercept}_{z,(x,y)} = \overline{z} - b\overline{x} - c\overline{y}$, so

$$\mathrm{ESS} = \sum (\overline{z} - bx - cy - [\overline{z} - b\overline{x} - c\overline{y}])^2 = \sum (-bx - cy - [-b\overline{x} - c\overline{y}])^2$$

$$= \sum (b(x - \overline{x}) + c(y - \overline{y}))^2 = b^2 n \,\mathrm{var}(x) + c^2 n \,\mathrm{var}(y) + 2bcn \,\mathrm{cov}(x,y)$$

This does not immediately have the result

$$\text{cov}(x, y) = 0 \implies \text{ESS} = \text{ESS}_x + \text{ESS}_y \tag{7.22}$$

as there is a difference between $b = \text{slope}^{(2)}_{z,x}$ and $\text{slope}^{(1)}_{z,x}$:

$$b = \frac{\text{slope}^{(1)}_{z,x} - \text{slope}^{(1)}_{z,y}\,\text{slope}^{(1)}_{y,x}}{1 - r^2_{x,y}}$$

However, if $\text{cov}(x, y) = 0$ then $r_{x,y} = 0$ and also $\text{slope}^{(1)}_{y,x} = 0$. So we do get (7.22).

Example 7.23 Necessity of covariance zero.

It is possible to have $\text{cov}(x, y) < 0$ and $b, c > 0$, leading to $\text{ESS}_x + \text{ESS}_y > \text{ESS}$. Just consider the example

$$\{(1, 0), (1, 1), (0, 1)\}$$

with $z = x + y$ again.

7.6 ANOVA

We perform an Analysis of Variance (ANOVA) with three groups (i.e., $p = 3$). Group A consists of students who turned in Assessment Exam III on time, and reported that they had taken MATH 412 (Introduction to Abstract Algebra). Following the notation of[13], the scores on the Pilot Assessment Exam in Group A were

$$(Y_{1,1}, Y_{1,2}) = (5, 13)$$

with a group mean of

$$\overline{Y}_{1+} = 9.$$

Group B consists of students who turned in Assessment Exam III on time, and reported that they had *not* taken MATH 412. The scores on the Pilot Assessment Exam in Group B were

$$(Y_{2,1}, Y_{2,2}) = (6, 8)$$

with a group mean of

$$\overline{Y}_{2+} = 7.$$

Group C consists of students who did not turn in Assessment Exam III on time. The scores on the Pilot Assessment Exam in Group C were

$$(Y_{3,1}, \ldots, Y_{3,6}) = (2, 3, 6, 7, 8, 10)$$

with a group mean of

$$\overline{Y}_{3+} = 6.$$

Let $(n_1, n_2, n_3) = (2, 2, 6)$ and $n = \sum_{i=1}^{p} n_i = 10$. The grand mean is

$$\overline{Y}_{++} = \frac{2\overline{Y}_{1+} + 2\overline{Y}_{2+} + 6\overline{Y}_{3+}}{10} = \frac{18 + 14 + 36}{10} = 6.8.$$

[13] M.H. DeGroot and M.J. Schervish. *Probability and Statistics*. Pearson Education, 2011. ISBN: 9780321709707. URL: https://books.google.com/books?id=XHmitwAACAAJ, Section 11.6.

The residual sum of squares is

$$SS_{\text{within}} = \sum_{i=1}^{p}\sum_{j=1}^{n_i}(Y_{i,j} - Y_{i+})^2 = (5-9)^2 + (13-9)^2 \quad + \quad (6-7)^2 + (8-7)^2$$

$$+(2-6)^2 + (3-6)^2 + (6-6)^2 + (7-6)^2 + (8-6)^2 + (10-6)^2$$

$$= 16 + 16 + 1 + 1 + 16 + 9 + 0 + 1 + 4 + 16 = 64 + 9 + 4 + 3 + 0 = 80.$$

The "between-sample-means" or "factor" sum of squares is

$$SS_{\text{between}} = \sum_{i=1}^{p} n_i(\overline{Y}_{i+} - \overline{Y}_{++})^2$$

$$= 2(9-6.8)^2 + 2(7-6.8)^2 + 6(6-6.8)^2 = 2(2.2)^2 + 2(0.2)^2 + 6(0.8)^2 = 13.6.$$

The test statistic is

$$U^2 = \frac{SS_{\text{between}}/(p-1)}{SS_{\text{within}}/(n-p)} = \frac{13.6/(3-1)}{80/(10-3)} = \frac{6.8}{11.42857} = 0.595.$$

When compared to an $F(p-1, n-p) = F(2,7)$ distribution, we would need a value of 4.74 to reject the null hypothesis that the means are equal at the 95% level. While the data are consistent with the idea that punctuality, and taking MATH 412, both lead to better scores, we cannot reject the null hypothesis that these things have no effect.

7.6.1 2×2 ANOVA

Interaction between two variables that together produce an effect is studied in 2×2 ANOVA[14].

The idea of interaction between factors can be illustrated in several examples.

Example 7.24 Sandwich ingredients.

Suppose you make a sandwich using bread and some of the four ingredients peanut butter, herring, tomato, and avocado. Which combinations would yield a tasty sandwich? Perhaps each ingredient is tasty by itself, but some combinations are not.

Example 7.25 Addition mod 2.

Suppose you consider $X + Y + Z$ mod 2. No information about one or two of the variables gives you any information; you need all three, at which point you have complete information.

Example 7.26 Black cats crossing the street in front of you.

Suppose you are superstitious and worry if you see a black cat crossing the street in front of you. However, you need all the four aspects for the superstition to apply. Thus, a black cat crossing the street all alone is not a problem, neither is a black cat in front of you a problem if it's not crossing the street, and so on.

	1-term	2-term
D	Carter 5'10", JFK 6'0", Cleveland 5'11'	Obama 6'1", Clinton 6'2", LBJ 6'4"
R	GHW Bush 6'2", Ford 6'0", Hoover 6'0"	GW Bush 6'0", Reagan 6'2", Nixon 5'11"

Inches above 5'10"	1-term	2-term
Democrat	0,1,2	3,4,6
Republican	2,2,4	1,2,4

Means	1-term	2-term	
D	1.00	4.33	2.667
R	2.67	2.33	2.500
	1.833	3.333	2.583

Table 7.4: Height of U.S. Presidents by party and number of terms.

Since the GRE data has a missing group, let's apply the theory instead to presidents: Democrat or Republican, 2-term or 1-term, and look at their height. Let's take the last 3 of each.

Since 5'10" is the shortest, let's look at the number of inches above that (Table 7.4).

At first glance it seems that D vs. R makes very little difference, but 2-term presidents are taller than 1-term presidents. Really, though, most striking is that Democratic 2-term presidents are very tall, Democratic 1-term presidents very short, and Republicans somewhere in the middle regardless of number of terms.

With $n = 3$ and $p = q = 2$ being the number of levels of the factors, and with D vs. R being "factor A", we can write

$$X_{111}, X_{211}, X_{311} = 0, 1, 2, \quad X_{121}, X_{221}, X_{321} = 2, 2, 4$$

$$X_{112}, X_{212}, X_{312} = 3, 4, 6, \quad X_{122}, X_{222}, X_{322} = 1, 2, 4$$

(thus the convention for index order is backwards from the GRE example above)

$$\overline{X}_{11} = \frac{0+1+2}{3} = 1, \quad \overline{X}_{12} = \frac{2+2+4}{3} = 2.67, \quad \overline{X}_{21} = \frac{3+4+6}{3} = 4.33, \quad \overline{X}_{22} = 3.33$$

$$\overline{X}_{1\cdot} = 2.667, \quad \overline{X}_{2\cdot} = 2.500, \quad \overline{X}_{\cdot 1} = 1.833, \quad \overline{X}_{\cdot 2} = 3.333, \quad \overline{X}_{\cdot\cdot} = 2.583$$

Now we have five sums of squares, SS_T (total), SS_A (effect of factor A), SS_B, $SS_{A\times B}$ (interaction of A and B), SS_{within} (error, unexplained variance, within cells).

$$\begin{aligned}
SS_A &= nq\sum_{j=1}^{p}(\overline{X}_{j\cdot} - \overline{X}_{\cdot\cdot})^2 \\
SS_B &= np\sum_{k=1}^{q}(\overline{X}_{\cdot k} - \overline{X}_{\cdot\cdot})^2 \\
SS_{A\times B} &= n\sum_{j=1}^{p}\sum_{k=1}^{q}([\overline{X}_{jk} - \overline{X}_{\cdot\cdot}] - [\overline{X}_{j\cdot} - \overline{X}_{\cdot\cdot}] - [\overline{X}_{\cdot k} - \overline{X}_{\cdot\cdot}])^2 \\
SS_{\text{within}} &= \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{q}(X_{ijk} - \overline{X}_{jk})^2 \\
SS_T &= \sum_{i=1}^{n}\sum_{j=1}^{p}\sum_{k=1}^{q}(X_{ijk} - \overline{X}_{\cdot\cdot})^2
\end{aligned}$$

[14] Source: `https://www4.uwsp.edu/psych/stat/13/anova-2w.htm`

The degrees of freedom are as follows.

$$\begin{aligned} df_A &= p-1=1 \\ df_B &= q-1=1 \\ df_{A\times B} &= (p-1)(q-1)=pq-p-q+1=1 \\ df_{\text{within}} &= pq(n-1)=pqn-pq=N-pq \\ df_T &= npq-1=N-1 \end{aligned}$$

Note that

$$df_T = df_A + df_B + df_{A\times B} + df_{\text{within}}$$

It can also be shown that

$$SS_T = SS_A + SS_B + SS_{A\times B} + SS_{\text{within}}$$

For our F tests we take the relevant mean squares in the numerator and mean squares for error MS_{within} in the denominator as usual. As they nicely put it,

$$MS = SS/df \quad \text{and} \quad F = MS_{\text{source}}/MS_{\text{error}}.$$

Note the difference between saying there is an $A \times B$ effect (the interaction between A and B is significant, e.g., Democratic presidents are taller the more terms they serve) and saying that x and y are correlated in the regression section above (there, we could just set x and y to be independent by design, and we were only interested in predicting z from x and y, not in discovering that x and y were correlated or interacting).

In our example we get $SS_A = \frac{1}{12}$, $SS_B = \frac{27}{4}$, $SS_{\text{within}} = 14$, $SS_{A\times B} = 10 + \frac{1}{12}$. While computing $SS_{A\times B}$ by hand we notice that for a 2×2 anova like this the formula simplifies to

$$SS_{A\times B} = npq(\overline{X}_{jk} - \overline{X}_{..})^2$$

for any particular $1 \leq, j, k, \leq 2$, i.e., the quantity does not depend on j and k. This is an expression of the fact that $df_{A\times B} = 1$. To test for $A \times B$ interaction, we compute

$$F = \frac{SS_{A\times B}/df_{A\times B}}{SS_{\text{within}}/df_{\text{within}}} = \frac{(10+\frac{1}{12})/1}{14/8} = 5 + \frac{16}{21} = 5.7619.$$

and $F(1,8) = 5.7619$ gives a p-value of .04315. Thus, the interaction between political party and number of terms for a U.S. President is significant for height at the usual 5% level.

An example from marketing research. The data in Table 7.5 is taken from[15].

It seems clear, or plausible, that the marginals are not significant.

For the power × focus interaction they report $F(1,144) = 12.84, p < 0.01, \eta_p^2 = 0.08$. This fits with our relationship between F, df, and η_p^2 above.

- For the "experience" row they report $F(1,144) = 6.88, p = .01, \eta_p^2 = 0.07$. We find

$$F := t^2 = \left((2.17-1.16)/\sqrt{1.78^2/37 + 1.82^2/37}\right)^2 = 2.4^2 = 5.82.$$

[15] Derek D. Rucker et al. "The Experience versus the Expectations of Power: A Recipe for Altering the Effects of Power on Behavior". In: *Journal of Consumer Research* 41.2 (2014), pp. 381–396. DOI: 10.1086/676598. eprint: /oup/backfile/content_public/journal/jcr/41/2/10.1086_676598/2/41-2-381.pdf. URL: http://dx.doi.org/10.1086/676598, pages 386–387.

focus\ power	low	high	marginal of focus
experience	2.17	1.16	=1.665
expectations	1.39	2.31	=1.850
marginal of power	1.780	1.735	= 1.7575

focus\ power	low	high
experience	1.78	1.82
expectations	1.62	1.56

Table 7.5: Means and standard deviations of willingness to buy, on a Likert scale.

The degrees of freedom should be $2n - 2 = 2(n - 1) = 72$ which fits their reported η_p^2: if $F(1, 72) = 5.82$ then $\eta_p^2 = F/(F + 72) = 5.82/(5.82 + 72) = 0.07$. They state that they are using a between-subject design (p. 385) which suggests there may be $148/4 = 36$ individuals per group.

- For the "expectations" row they report $F(1, 144) = 5.03, p = .03, \eta_p^2 = 0.08$. We find

$$F := t^2 = \left((1.39 - 2.31)/\sqrt{1.62^2/37 + 1.56^2/37}\right)^2 = 6.20.$$

The degrees of freedom should be $2n - 2 = 2(n - 1) = 72$ which fits their reported η_p^2: if $F(1, 72) = 6.2$ then $\eta_p^2 = F/(F + 72) = 6.2/(6.2 + 72) = 0.08$.

We have

$$SS_{A\times B}/n = 4[(2.17 - 1.7575) - (1.665 - 1.7575) - (1.780 - 1.7575)]^2 = 0.931225$$

and hence $SS_{A\times B} = 34.455325$. In this case $df_{\text{within}} = 144 = N - pq$ where $N = 148$, $p = 2$, $q = 2$, and presumably $n = N/pq = 148/4 = 37$.

SS_{within} can be figured out from the standard deviations given in Table 7.5. So

$$SS_{\text{within}} = (37 - 1)(1.78^2 + 1.82^2 + 1.62^2 + 1.56^2) = 415.3968$$

should be the sum of all deviations between individual observations and their respective group means. The F statistic should then be

$$F = \frac{MS_{A\times B}}{MS_{\text{within}}} = \frac{SS_{A\times B}/df_{A\times B}}{SS_{\text{within}}/df_{\text{within}}} = \frac{34.45/1}{415.3968/144} = 11.94.$$

which is fairly close to the stated 12.84.

7.6.2 Permutation tests

Permutation tests were introduced in Section 6.2.2 of *OpenIntro Statistics* in the context of a difference of means.

Consider our example of the heights of U.S. Presidents. Is the mean height for Democrats large? Taking all the heights and considering all the permutations of them, there may seem to be 12! choices to consider. Doing so, we find that

$$179365520/(179365520 + 299634480) = 37\%$$

of the permutations give a larger mean. Thus we have a p-value of .37 for our alternative hypothesis that Democrats are taller. Not statistically significant.

Did we really have to consider all 12! permutations? Clearly not, since in this case all that matters is which of the $\binom{12}{6}$ sets of 6 Presidents was chosen. Searching through

all such sets instead, we find that 578 have a smaller mean height, and 346 a larger mean height, for a total of $924 = \binom{12}{6}$. The p-value is $346/924 = 37\%$, i.e., unchanged, but our computation much faster.

We can try to permutation tests also for more complicated test statistics such as $SS_{A\times B}/SS_{\text{between}}$ above. We can ignore the degrees of freedom since they will be the same for all the permutations, and we are not claiming to be able to compare with any limiting distribution anymore. Then what we do is look at all $12! = 479,001,600$ permutations of the 12 heights of U.S. presidents, and calculate the statistics for them (or, to save time, take a random sample). The percentile that our value finds itself (in this case a computer calculation showed it to be 4.1%) then becomes the p-value.

Permutation tests are in a way more satisfying than traditional t- and z-tests. They were introduced by Fisher about 100 years ago, but were impractical before the advent of computers. They remove the need to argue for approximate normality.

On the other hand, we still have to argue that our test statistic like $SS_{A\times B}/SS_{\text{within}}$ measures what we want it to measure. If we get a low p-value, we can argue that the labels "Democrat", "1-term" and so on were not assigned to the heights uniformly at random.

The fact that this is evident in the $SS_{A\times B}/SS_{\text{within}}$ statistic suggests that we can also rule out that they were assigned randomly and independently with given marginal distributions. In other words, there is an interaction of being Democrat and being 1-term on the variable *height*. To define interaction we first need conditional expectation. The conditional expectation of a random variable X given an event A is just

$$\sum_x x\mathbb{P}(X = x \mid A).$$

Now, the definition of P and Q having no **interaction** on X is $E(X \mid P, Q) + E(X) = E(X \mid P) + E(X \mid Q)$. In other words, the effect of P on X is independent of Q:

$$E(X \mid P) - E(X) = E(X \mid P, Q) - E(X \mid Q).$$

7.7 Exercises

7.1 Extreme correlations. Give specific numerical examples of observations with

(a) $r = 1$

(b) $r = 0$

(c) $r = -1$

[16]

7.2 Cauchy-Schwarz. Let $u_1, u_2, \ldots, u_n$ and $v_1, v_2, \ldots, v_n$ be real numbers.

(a) Rearrange the terms in $(u_1x + v_1)^2 + (u_2x + v_2)^2 + \cdots + (u_nx + v_n)^2$ to obtain a quadratic polynomial $ax^2 + bx + c$ in x.

(b) Note that the polynomial in part a has at most one real root and therefore that its discriminant $b^2 - 4ac$ is at most zero. Calculate this discriminant to conclude that $\left(\sum_{i=1}^n u_i v_i\right)^2 \leq \left(\sum_{i=1}^n u_i^2\right)\left(\sum_{i=1}^n v_i^2\right)$. This inequality (and its various generalizations) is known as the *Cauchy-Schwarz inequality.*

7.3 Theory of extreme correlations. Let $(x_1, y_1), (x_2, y_2), \ldots (x_n, y_n)$ be observations.

(a) Show that

$$\frac{1}{n-1}\sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x}\frac{y_i - \bar{y}}{s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Conclude that the term on the right is an equivalent formula for r.

(b) Apply the Cauchy-Schwarz inequality to $\left|\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})\right|$ to conclude that $|r| \leq 1$ and therefore that $-1 \leq r \leq 1$.

[16] By "specific numerical" is meant that, for instance, $\{(1, 2), (0, -3), (0, 0)\}$ could be a possible answer — although it will be wrong as r will not be exactly 0, 1 or -1 for this particular choice!

Chapter 8

Special topics

In this chapter we delve into some fun, modern, and more advanced topics in statistics: algebraic statistics, entropy, likelihood, Markov chains, and information criteria.

8.1 Algebraic statistics

Consider two Bernoulli random variables X and Y, both with parameter p. We can define four numbers $p_{00}, p_{01}, p_{10}, p_{11}$ by

$$p_{ij} = \mathbb{P}(X = i \text{ and } Y = j).$$

Let us write $q = p_{1,1}$. It might be that $\mathbb{P}(X = Y) = 1$, in which case $q = p$. It might also be that $\mathbb{P}(X = 1 - Y) = 1$, in which case $q = 0$. So it seems that q can vary, with $0 \leq q \leq p$. On the other hand, q determines the other parameters p_{00}, p_{01}, p_{10}, as can be seen from the following table:

	$X = 0$	$X = 1$	$X \in \{0, 1\}$
$Y = 0$	p_{00}	p_{10}	$p_{+0} = 1 - p$
$Y = 1$	p_{01}	p_{11}	$p_{+1} = p$
$Y \in \{0, 1\}$	$p_{0+} = 1 - p$	$p_{1+} = p$	$p_{++} = 1$

Indeed, the only solution is

	$X = 0$	$X = 1$	$X \in \{0, 1\}$
$Y = 0$	$p_{00} = 1 - p - (p - q)$	$p_{10} = p - q$	$p_{+0} = 1 - p$
$Y = 1$	$p_{01} = p - q$	$p_{11} = q$	$p_{+1} = p$
$Y \in \{0, 1\}$	$p_{0+} = 1 - p$	$p_{1+} = p$	$p_{++} = 1$

The value for p_{00} implies another constraint on q:

$$2p - 1 \leq q.$$

Indeed, these numbers then must satisfy some equations: they must be nonnegative, and:

$$p_{00} + p_{01} = 1 - p \quad \text{since } \mathbb{P}(X = 0) = 1 - p$$

$$p_{10} + p_{11} = p \quad \text{since } \mathbb{P}(X = 1) = p$$

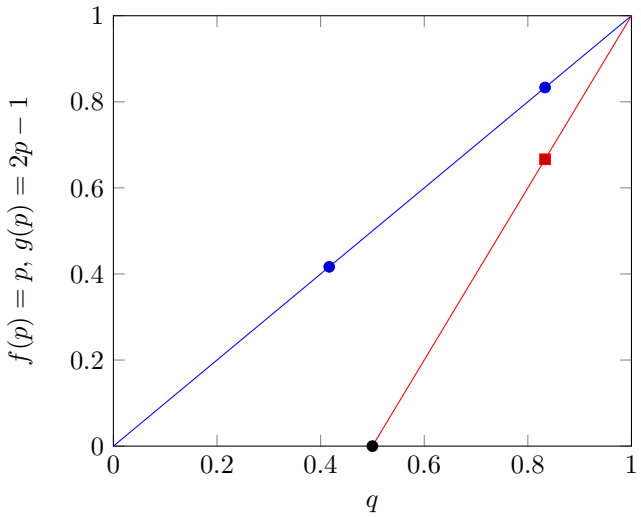
$$p_{00} + p_{10} = 1 - p \quad \text{since } \mathbb{P}(Y = 0) = 1 - p$$

$$p_{01} + p_{11} = p \quad \text{since } \mathbb{P}(Y = 1) = p$$

The fact that all the p_{ij} add up to 1 follows. How much variability remains? What are all the joint distributions X and Y can have? Well, let

$$q = \mathbb{P}(X = 1 \text{ and } Y = 1).$$

Then $\mathbb{P}(X = 1 \text{ and } Y = 0) = p - q = \mathbb{P}(X = 0 \text{ and } Y = 1)$, and $\mathbb{P}(X = 0 \text{ and } Y = 0) = 1 - p - (p - q)$. So we have a 1-dimensional family of joint distributions. When $q = 0$, we have $\mathbb{P}(X = 1 - Y) = 1$. When $q = p$, we have $\mathbb{P}(X = Y) = 1$. Moreover, the constraint on q is that $0 \le q \le p$ and $p - q \le 1 - p$, i.e., $q \ge 2p - 1$. In a sense, the joint distributions for fixed p form a parametrized curve in $\mathbb{R}^4$.[1]



8.2 Maximum entropy

If you receive a bit string of length n, where each possible string is equally likely (so in fact the probability is 2^{-n}), let us agree to say that you have received n "bits of information". This is consistent with defining the amount of information received when an event of probability $p = 2^{-n}$ happens is $-\log_2 p$.

If now the possible outcomes say $1, \dots, k$ of an experiment have probabilities $p_1, \dots, p_k$ with $\sum p_i = 1$, then the expected amount of information received is

$$H := \mathbb{E}(-\log_2 p_X) = \sum (-\log_2 p_i) p_i$$

This is called the entropy of the probability distribution or probability experiment under consideration. It can be shown that it is the greatest when $p_1 = p_2 = \cdots = p_k$.

For instance, if $k = 2$, this says that the two outcomes 0, 1 should have probabilities $1/2$ each. Thus, the maximum-entropy coin is a fair coin.

[1] For much more on algebraic statistics, see Seth Sullivant, Algebraic Statistics, *Graduate Studies in Mathematics* **194**, 2018.

Returning to Section 8.1, which value of $q = \mathbb{P}(X = 1, Y = 1)$ maximizes the entropy? Here,

$$H = -q\log_2 q - 2(p-q)\log_2(p-q) - (1-2p+q)\log_2(1-2p+q) = h(q) + 2h(p-q) + h(1-2p+q)$$

where $h(x) = -x\log_2 x$. Since $\log_2 x = \log(x)/\log(2)$ (where log could have any fixed base) it is safe to replace $\log_2$ by $\log_e = \ln$ when looking for maximum entropy. Then, taking dH/dq and setting it equal to 0, we obtain

$$0 = 1 + \ln q + 2(-1)(1 + \ln(p-q)) + 1 + \ln(1-2p+q)$$

since $(d/dq)(-q\ln(q)) = -(1 + \ln q)$. So

$$0 = 2 + \ln q - 2(1 + \ln(p-q)) + \ln(1-2p+q)$$

If we are lazy, Wolfram Alpha gives the solution $q = p^2$. In other words, no matter what p is, having X and Y be independent gives maximum entropy.

8.3 Maximum likelihood

A nice way to estimate parameters is to ask: which value of the parameter p maximizes the value of the probability density function (or probability mass function in the discrete case) at the value x that was observed?

Consider for instance a binomial random variable with parameters n (known) and p (unknown) where we have just observed a value $X = k$. The *likelihood function* is then, writing $\mathbb{P}_p$ for "probability using parameter value p",

$$L(p) = \mathbb{P}_p(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

To maximize the likelihood we set the derivative equal to zero. In fact, we might as well set the derivative of $\log L(p)$ (where $\log = \log_e = \ln$) equal to zero, since log is an increasing function.

$$\log L(p) = \log\binom{n}{k} + k\log p + (n-k)\log(1-p)$$

$$0 = \frac{d}{dp}\log L(p) = \frac{k}{p} - \frac{n-k}{1-p}$$

The solution is

$$k(1-p) - (n-k)p = 0, \quad k - np = 0, \quad p = \frac{k}{n},$$

which is what we would have guessed all along. However, maximum likelihood does not always give your first guess as result. For instance, recall from Chapter 2 that the sample variance is an unbiased estimator of the variance:

$$S^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_i - \bar{X})^2, \quad \mathbb{E}(S^2) = \sigma^2$$

It turns out that the maximum likelihood estimator of the variance is instead

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{k=1}^{n}(X_i - \bar{X})^2$$

Let us check this claim in the case $n = 1$. Of course, in this case S^2 is undefined and we naturally have no idea what the variance should be after observing just one data point. However, if we turn around and ask, which value of the variance would make what we just observed ($X_1 = x_1$) most likely, it is clear that the answer is: $\sigma^2 = 0$.

8.4 Hidden Markov chains

A hidden Markov chain has a finite set of **states** with probabilities of transitions between them. In each state there are probabilities of emitting (or writing) certain symbols. If we observe the sequence of emitted symbols over a period of time, can we determine the states the chain must have been in at various intervals of time?

Suppose a student can be in two states, "studious" (S) and "relaxing" (R).

- When studious, there is a 70% probability of emitting a homework set h and a 30% of emitting nothing, n.
- When relaxing, there is a 100% chance of emitting nothing.

Suppose also that the probability of moving

- from "studious" to "relaxing" is 90%, and
- from "relaxing" to "studious" is 50%.

Now we observe

$$nnnhh$$

Which sequence of states maximizes the probability of this "word" being emitted? Perhaps RRRSS? Let us for simplicity say that 1 means a homework set and 0 means nothing. Also, the states are s_0 (relaxing) and s_1 (studious). See Figure 8.1, where,

- p is the probability of emitting 1 in state s_0,
- q is the probability of emitting 1 in state s_1,
- γ is the probability of transitioning from s_0 to s_1, and
- δ is the probability of transitioning from s_1 to s_0.

The probability of RRRSS is then the probability that (starting in R),

- $1 - p$ of emitting n, then
- $1 - \gamma$ of staying in R, then
- $1 - p$ of emitting n, then
- $1 - \gamma$ of staying in R, then
- $1 - p$ of emitting n, then
- γ of moving to S, then
- q of emitting h, then
- $1 - \delta$ of staying in S, then

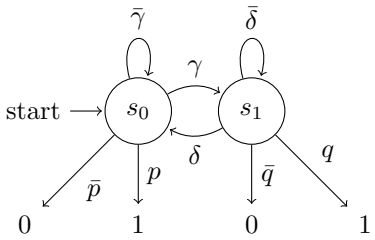


Figure 8.1: HMM with two states.

- q of emitting h.

We multiply these to find the overall probability:

$$(1-p)^3(1-\gamma)^2\gamma q^2(1-\delta). \tag{8.1}$$

This can be compared with the other 16-1 possible state sequences, R???? with each $? \in \{R, S\}$. On the other hand, given the outputted word and (possibly) the state sequence, we can find the maximum likelihood values of p, q, γ, δ.

We are interested here in the *maximum likelihood* associated with (8.1): $p = 0$, $\delta = 0$, $q = 1$, and to find γ, we solve

$$0 =^! \frac{d}{d\gamma}(1-\gamma)^2\gamma = \frac{d}{d\gamma}\gamma - 2\gamma^2 + \gamma^3 = 1 - 4\gamma + 3\gamma^2$$

which gives $x = 1/3$ and $x = 1$, of which $x = 1/3$ gives the maximum. So the maximum likelihood is

$$(1-p)^3(1-\gamma)^2\gamma q^2(1-\delta) = (2/3)^2(1/3) = \frac{4}{27}.$$

8.5 Overfitting and the complexity of a word

A word is a binary sequence like 011101. Some words are intuitively more complicated than others. For instance, 010101 may seem simpler than 011101. This can be made precise using theoretical computer science: Turing machines, finite automata and the like. Here we follow a suggestion of Prof. Alexander Shen from 2015 (which was developed in a paper[2]) to investigate such complexity of words in terms of Hidden Markov Models. One could ask, for which pairs (q, m) is there a HMM with q states which outputs x with probability at least 2^{-m}.

The function sending m to the least such q we call the **structure function** for x.

If we measure the probability only over a single path, we can solve the required optimization problem exactly using the following exercise.

⊙ **Guided Practice 8.2** A polynomial of the form

$$f(p) = p^a(1-p)^b$$

with $0 \le p \le 1$ is maximized when $p = a/(a+b)$.[3]

[2] Kjos-Hanssen. "Few paths, fewer words". In: *Experimental Mathematics* (2019).

[3] The derivative is

$$f'(p) = ap^{a-1}(1-p)^b - p^a b(1-p)^{b-1} = (a(1-p) - bp)p^{a-1}(1-p)^{b-1},$$

which equals zero, for $0 < p < 1$, when $a(1-p) = bp$, i.e., $p = a/(a+b)$.

⊙ **Guided Practice 8.3** A polynomial of the form

$$f(p_1, \ldots, p_k) = \prod_{i=1}^{k} p_i^{a_i}(1-p_i)^{b_i}$$

with $0 \leq p_i \leq 1$ is maximized when each $p_i = a_i/(a_i + b_i)$.[4]

We now exhibit a case where paths and words differ, in the sense that consider a single path and a single word give different results. For definiteness, we assume that a HMM has a designated start state s_0 and immediately emits a symbol before considering whether to then move to a different state.

Theorem 8.1. *There is a word x and a number of states q such that the maximum probability of emitting x by a HMM with q states is strictly larger than the maximum probability of emitting x by a HMM with q states along any particular single path.*

Proof. Let $x = 001$ and $q = 2$. Consider a general HMM with two states over the alphabet $\{0, 1\}$ as in Figure 8.1, where $\bar{\alpha} = 1 - \alpha$. Let $S(t)$ be the state after transitioning t times, a random variable. The probability of emitting the string 001 when starting in state s_0 is then

$$\begin{aligned} f(p,q,\gamma,\delta) &= \mathbb{P}(\text{emit } 001; S(1) = s_0 = S(2)) \\ &+ \mathbb{P}(\text{emit } 001; S(1) = s_0, S(2) = s_1) \\ &+ \mathbb{P}(\text{emit } 001; S(1) = s_1, S(2) = s_0) \\ &+ \mathbb{P}(\text{emit } 001; S(1) = s_1 = S(2)) \\ &= \bar{p}^2 p \bar{\gamma}^2 + \bar{p}^2 q \bar{\gamma}\gamma + \bar{p}\bar{q}p\gamma\delta + \bar{p}\bar{q}q\gamma\bar{\delta}. \end{aligned}$$

Here

$$\mathbb{P}(\text{emit } 001; S(1) = s_0 = S(2)) = \mathbb{P}(emit0 \mid s_0)\mathbb{P}(s_0 \mid s_0)\mathbb{P}(emit0 \mid s_0)\mathbb{P}(s_0 \mid s_0)\mathbb{P}(emit1 \mid s_0)$$

If we use the notation that the probability of observing a sequence

$$Y = y(0), y(1), \ldots, y(L-1)$$

of length L is given by

$$P(Y) = \sum_X P(Y \mid X)P(X),$$

where the sum runs over all possible hidden-node sequences

$$X = x(0), x(1), \ldots, x(L-1)$$

then this becomes

$$\begin{aligned} \mathbb{P}(\text{emit } 001; S(1) = s_0 = S(2)) &= \mathbb{P}(Y = (0,0,1) \text{ and } X = (s_0, s_0, s_0)) \\ &= \mathbb{P}(X = (s_0, s_0, s_0))\mathbb{P}(Y = (0,0,1) \mid X = (s_0, s_0, s_0)) \\ = \mathbb{P}(y(0) = 0 \mid x(0) = s_0) &\cdot \mathbb{P}(x(1) = s_0 \mid x(0) = s_0) \\ \cdot\mathbb{P}(y(1) = 0 \mid x(1) = s_0) &\cdot \mathbb{P}(x(2) = s_0 \mid x(1) = s_0) \\ &\cdot \mathbb{P}(y(2) = 1 \mid x(2) = s_0) = \bar{p}^2 p \bar{\gamma}^2. \end{aligned}$$

[4]Strictly speaking this exercise requires multivariable calculus. Consider all but one of the p_i to be constants and take the derivative with respect to the remaining one. Set all these derivatives equal to 0 simultaneously to find all plausible candidates for a maximum of f.

With $q = 2/3$, $\gamma = 2/3$, $\delta = 0$, and $p = 0$ the sum is $f = 8/27$.

On the other hand, each of the four terms of f, of the form given in Lemma 8.2, is bounded by $1/4$, simply because $x(1-x) \le 1/4$ for all $0 \le x \le 1$. □

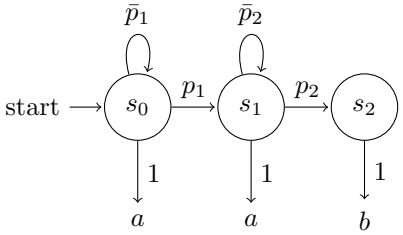
⊙ **Guided Practice 8.4** If $f(x,y) = f(y,x)$ for all x, y, then the maximum of f, if there is only one, must occur at a point where $x = y$.[5]

Let e be the base of the natural logarithm. When using probability, it is better to consider the alphabet $\{a, b\}$ than $\{0, 1\}$. Extending Theorem 8.1, we have:

Theorem 8.2. *Fix two distinct symbols a and b. For $x = a^{n-1}b$ and HMMs with $q = n-1$ states,*

$$\limsup_{n\to\infty} \max_{\vec{p}} \mathbb{P}(\textit{emit } x) \ge 1/e.$$

Proof. Let the $q-1$ many states be ordered left-to-right, and let p_i be the probability of moving to the right. Assume only the rightmost state allows a b to be output. For $n = 4$ the automaton looks like this:



Then

$$\mathbb{P}(\text{emit } a^{n-1}b) = \sum_{i=1}^{n-2} p_1 \dots p_{n-2}(1-p_i) = p_1p_2(1-p_1) + p_1p_2(1-p_2).$$

This function is symmetric in the variables p_i, $1 \le i \le n-2$, so by Guided Practice 8.4 it must be maximized at a point where $p_1 = \dots = p_{n-2}$. Moreover, that value of p_1 must be $1 - \frac{1}{n-1}$. So

$$\mathbb{P}(\text{emit } a^{n-1}b) = \sum_{i=1}^{n-2} \left(1 - \frac{1}{n-1}\right)^{n-2} \frac{1}{n-1} = (n-2) \cdot \frac{1}{n-1} \cdot \left(1 - \frac{1}{n-1}\right)^{n-2} \to \frac{1}{e}.$$

Indeed, considering the case $n = 4$,

$$\begin{aligned} \frac{\partial}{\partial p_1} \sum_{i=1}^{n-2} p_1 \dots p_{n-2}(1-p_i) &= \frac{\partial}{\partial p_1} \left(p_1p_2(1-p_1) + p_1p_2(1-p_2)\right) \\ &= (1-2p_1)p_2 + p_2(1-p_2) \\ &= (2-2p_1-p_2)p_2, \end{aligned}$$

which equals 0 if either $p_2 = 0$ (a solution we can disregard) or $1 = p_1 - \frac{1}{2}p_2$. Using that also $p_1 = p_2$, we get $p_1 = \frac{2}{3}$. □

[5] If the maximum occurs at (a, b) then it also occurs at (b, a). So if there is only one such point then $(a, b) = (b, a)$, i.e., $a = b$.

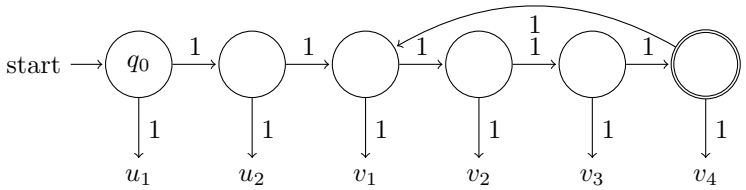


Figure 8.2: Figure for Theorem 8.3. The out-degree of a vertex is the number of edges whose tail is at that vertex.

On the other hand, the probability of x being omitted along any particular path stays bounded below $1/4$, because we must include both a probability p and its negation $1-p$, and $p(1-p) \leq 1/4$.

Words of the form $0^{n-1}1$ are in a sense of maximal complexity: any binary string other than $0^{n-1}1$ and $1^{n-1}0$ can have probability 1 along a single path through an HMM with $n-1$ states, by the following results.

Theorem 8.3. *The minimum number of states of an HMM such that x occurs with probability 1 along a single path is* $\min\{|u| + |v| : x = u\,v^p, \exists u, v, p\}$.

Here $p \in \mathbb{Q}$; for instance, $(abba)^{3/2} = abbaab$. The idea of Theorem 8.3 is shown in Figure 8.2: to always go to a fresh state except that we loop back when the second occurrence of v starts. The probability 1 condition expresses that the out-degree is always 1.

Theorem 8.4. *Any binary word of length n starting with 0, except $0^{n-1}1$, can be written as $u\,v^p$ with $|v| > 0$ and $p > 1$, $p \in \mathbb{Q}$.*

Proof. If the last bit of the word has occurred before, say $x = x_1 \ldots x_n$ where $x_n = x_k$, $k < n$, then we may write $x = x_1 \ldots x_{k-1}(x_k \ldots x_{n-1})^{1+\epsilon}$ where $\epsilon = 1/(n-k)$. □

Using $n-1$ states to "explain" a word of length n is a form of *overfitting*. That is, the word becomes a perfectly predicted by the model, but any word could be perfectly predicted by a model with that many states. The model does not really reveal much.

8.6 Akaike information criterion

Ideally, we want a model with few states (in general, few parameters), that still does fairly well at predicting the output. One way to make this precise is: find a model with a number of states and corresponding probability so that relatively few words can be predicted *that well* with *that few states*. For instance, a word like 000011110000 may be modeled well as "some zeros, then some ones, then some zeros" (three states), or "alternating some sequences zeros and some sequences of ones, with somewhat low probability of switching" (two states). The *Akaike information criterion*[6] says that we should try to minimize

$$\mathrm{AIC} = k - \log L$$

where k is the number of parameters, L is the likelihood, and $\log = \ln$.

[6]Hirotugu Akaike. "A new look at the statistical model identification". In: *IEEE Trans. Automatic Control* AC-19 (1974). System identification and time-series analysis, pp. 716–723. ISSN: 0018-9286. DOI: 10.1109/tac.1974.1100705. URL: https://doi.org/10.1109/tac.1974.1100705.

Example 8.5 Using AIC for $0^4 1^5 0^6$.

Suppose we observe the string $0^4 1^5 0^6$ in some experiment. We could have two states and two parameters p, q giving the transition probabilities. Or we could have two states and a single parameter p giving the transition probability in either direction. Then the question becomes whether having that extra parameter allows us to double the likelihood (or actually, multiply it by e). The AIC is only valid in the limit of many parameters so we should not take this too seriously. If there is only one parameter p then the probability of $0^4 1^5 0^6$ is

$$(1-p)^3 p (1-p)^4 p (1-p)^5 = (1-p)^{12} p^2 = ((1-p)^6 p)^2$$

which is maximized at $1-p = 6/7$, i.e., $p = 1/7$, giving probability at most $(6/7)^{12}(1/7)^2 = .0032$.

If there are two parameters, we get

$$(1-p)^3 p (1-q)^4 q (1-p)^5 = (1-p)^8 p (1-q)^4 q$$

which is maximized with $p = 1/9$ and $q = 1/5$, giving a maximal probability of

$$(8/9)^8 (1/9)(4/5)^4 (1/5) = .0035.$$

Thus this model performs a bit better, but far from e times better, and we should stick with just the single parameter p.

An Akaike coincidence, maybe. In the $0^{n-1}1$ example above, the probability was

- $1/e$ for $n-1$ states (so AIC $= k - \ln L = n - 1 - \ln(1/e) = n - 1 - (-1) = n$), and
- 1 for n states (so AIC $= k - \ln L = n - \ln 1 = n$),

meaning that according to Akaike, these explanations of $0^{n-1}1$ can be considered equally good in the limit! Is this a coincidence? With varying number of states we achieve varying probabilities, for varying reasons:

$n+1$	1
n	$\lim_n (1-\frac{1}{n})^{n-1}\binom{n-1}{1}(\frac{1}{n})^1 = 1/e$
$n-1$	$\lim_n (1-\frac{2}{n})^{n-2}\binom{n-1}{2}(\frac{2}{n})^2 = e^{-2}\frac{2}{2}$
$n+1-c$	$\lim_n (1-\frac{c}{n})^{n-c}\binom{n-1}{c}(\frac{c}{n})^c$

In fact the probability of generating $0^{n-1}1$ is exactly $1-\frac{c}{n}$ (taking the final edge labeled 1) times the probability that a binomial$(n-1, c/n)$ random variable U takes the value c:

$$\mathbb{P}(U=c) = \binom{n-1}{c}\left(\frac{c}{n}\right)^c \left(1-\frac{c}{n}\right)^{n-1-c}.$$

So we are in the limit looking at a Poisson random variable with $\lambda = \lim_n (n-1)\frac{c}{n} = c$. By Stirling's Formula,

$$\mathbb{P}(X=\lambda) = e^{-\lambda}\frac{\lambda^\lambda}{\lambda!} \approx \frac{1}{\sqrt{2\pi c}}.$$

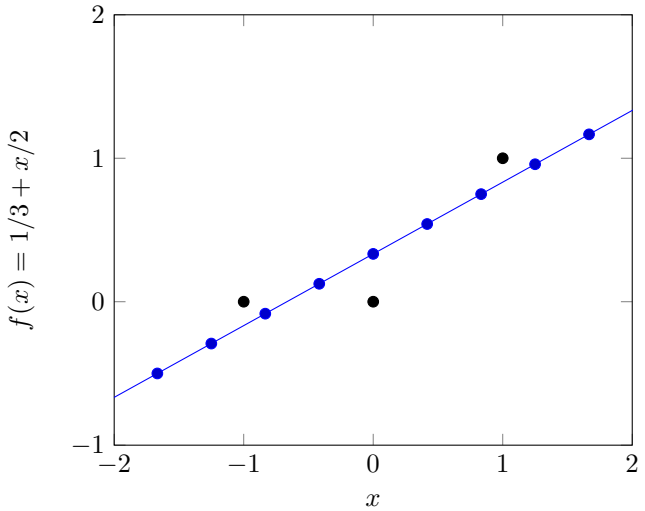
Note that $\sqrt{2\pi} = 2.5066 \approx e = 2.71828\ldots$ For $c = 1, 2, 3, \ldots$ this is

$$\frac{1}{e}, \frac{2}{e^2}, \frac{9}{2e^3}, \ldots$$

Thus $n-c$ states give "better" models the higher c is, as $n \to \infty$.

Example 8.6 The simplest nontrivial regression.

Let us consider a regression for the points $(-1,0),(0,0),(1,1)$.



The model has parameters β_0 and β_1 for the intercept and slope, and variance σ^2, and we get a likelihood calculation[7]. In principle we can now compare this model to one with only a slope, or only an intercept, according to AIC. This does go into multivariable calculus via the independence criterion $f(x,y) = f_X(x)f_Y(y)$ for pdfs, however. We have $\bar{x} = 0$, $\bar{y} = 1/3$, $\sum(x_i - \bar{x})(y_i - \bar{y}) = (-1)(0-1/3) + (0)(0-1/3) + (1)(1-1/3) = 1$ and $\sum(x_i - \bar{x})^2 = 2$, giving $\hat{\beta}_1 = 1/2$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 1/3$, and

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{3}\left(\left(0 - \left(\frac{1}{3} + \frac{1}{2}(-1)\right)\right)^2 + \left(0 - \left(\frac{1}{3} + \frac{1}{2}(0)\right)\right)^2 + \left(1 - \left(\frac{1}{3} + \frac{1}{2}\cdot 1\right)\right)^2\right) \\
&= \frac{1}{3}(1/36 + 1/9 + 1/36) = \frac{1}{3}\cdot\frac{1}{6} = \frac{1}{18}, \\
\hat{\sigma} &= \frac{1}{2\sqrt{3}}.
\end{aligned}$$

Then the log-likelihood is

$$-\frac{3}{2}\log 2\pi - 3\log\hat{\sigma} - \frac{3}{2} = -\frac{3}{2}(1 + \log(2\pi)) + 3(\log 2 + \frac{1}{2}\log 3)$$

The number of parameters of the model is 3; σ should be included since it is an estimated parameter (a quantity that helps index the family of possible probability distributions). So the AIC score is

$$\text{AIC} = 2k - \ln\hat{L} = 6 - (-\frac{3}{2}(1 + \log(2\pi)) + 3(\log 2 + \frac{1}{2}\log 3)) = 6.529$$

When dealing with pdfs the interpretation is less clear (gaining a state correspond to multiplying the pdf by a certain amount) than in the discrete case.

[7] https://www.stat.cmu.edu/~cshalizi/mreg/15/lectures/06/lecture-06.pdf

Bayesian information criterion This[8] is

$$\text{BIC} = \ln(n)k - 2\ln\hat{L}$$

For model selection with a fixed sample size n, we can divide this by $\ln(n)$ and get

$$k - 2\frac{\ln\hat{L}}{\ln n} = k - 2\log_n\hat{L} = k - \log_{\sqrt{n}}\hat{L}.$$

In other words, using the BIC rather than the AIC, one more parameter should multiply the likelihood by $\sqrt{n}$ rather than e. Thus, comparing the model for $0^{n-1}1$ with n states to the one with $n-1$ states, they are equally good for AIC, whereas the $n-1$ state model is better according to BIC when $e < \sqrt{n}$, which to nearest integer means $n \geq 8$.

The difference between AIC and BIC is sometimes described as follows. BIC assumes that the true model is among the ones being considered, whereas AIC assumes that the true model is unknowable, or cannot be known exactly.

⊙ **Guided Practice 8.7** Show that for each x and n, the equation $2\log_n x = \log_m x$ has the solution $m = \sqrt{n}$.[9]

8.7 Support vector machines

Suppose we consider clusters of points (x_1, x_2) and know that the points $(-1, 0)$ and $(0, 0)$ belong to one cluster (indicated by $y = -1$) and $(1, 1)$ to another (indicated by $y = 1$). We seek a straight line $\vec{w}\cdot\vec{x} - b = 0$ to separate the two clusters. To achieve proper separation we require $\vec{w}\cdot\vec{x} - b \geq 1$ whenever $y = 1$ and $\vec{w}\cdot\vec{x} - b \leq -1$ whenever $y = -1$. These two rules can be combined to: $y(\vec{w}\cdot\vec{x} - b) \geq 1$. For our three points this amounts to:

$$w_1 + b \geq 1, \quad b \geq 1, \quad w_1 + w_2 - b \geq 1.$$

We seek to minimize $\sqrt{w_1^2 + w_2^2}$ (since this maximizes the distance between the lines $\vec{w}\cdot\vec{x} - b = 1$ and $\vec{w}\cdot\vec{x} - b = -1$) subject to these constraints. This leads to $w_1 = w_2 = b = 1$ and the separating line is $x + y - 1 = 0$ which makes sense geometrically. This is a simple form of machine learning: the machine can now classify new points depending on which side of the separating line they lie on.

How did we get $w_1 = w_2 = b = 1$? Let $x = b - 1$, $y = w_1 + b - 1$, $z = w_1 + w_2 - b - 1$, then the constraints simplify to $x \geq 0, y \geq 0, z \geq 0$, at the cost of the objective function to be minimized becoming more complicated:

$$m(x, y, z) = (y - x)^2 + (z + 2x + 2 - y)^2.$$

The way this is minimized in multivariable calculus is that we take derivatives with respect to one variable at a time. It turns out that the (x, y, z) giving the global minimum value for m does not satisfy the constraints, and even going to a boundary case like $x = 0$, $y = 0$, $z = 0$, the new global minimum as a function of two variables does not satisfy the constraints. Combining boundaries such as setting $x = y = 0$ or $y = z = 0$ does not help, but setting $x = z = 0$ gives $y = 1$ which is our optimal solution. We say that

$$(x, y, z) = \underset{x\geq 0, y\geq 0, z\geq 0}{\arg\min} \; m(x, y, z).$$

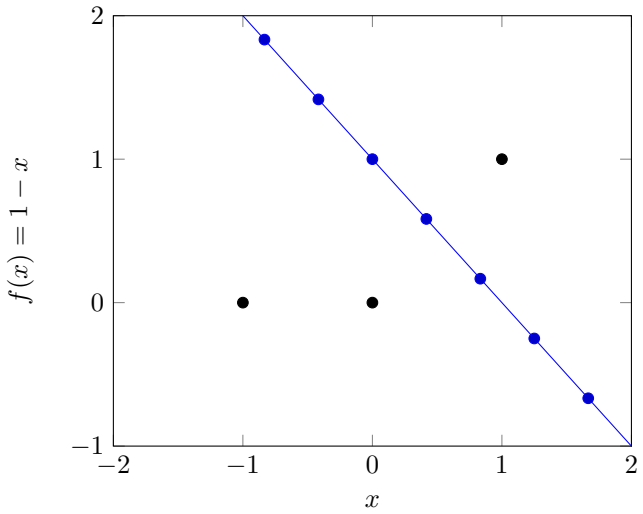
[8] Gideon Schwarz. "Estimating the dimension of a model". In: *Ann. Statist.* 6.2 (1978), pp. 461–464. ISSN: 0090-5364. URL: http://links.jstor.org/sici?sici=0090-5364(197803)6:2<461:ETDOAM>2.0.CO;2-5&origin=MSN.

[9] Using a property of logarithms, $2\log x/\log n = \log x/\log m$, so $2\log m = \log n$, and so $m = \sqrt{n}$.

Without discussing the precise assumption required, we have the following rule of thumb:

Theorem 8.5. *The minimum of a function $f(\vec{x})$ subject to constraints $\vec{g}(\vec{x}) \geq \vec{c}$ on $\vec{x}$ will be attained at a place where all the partial derivatives of f are zero; or else, where applying as few constraints $g_i(\vec{x}) = c_i$ as possible, all the remaining partial derivatives are zero.*

This is typically made precise, and proved, in multivariable calculus courses.



8.8 Shattering

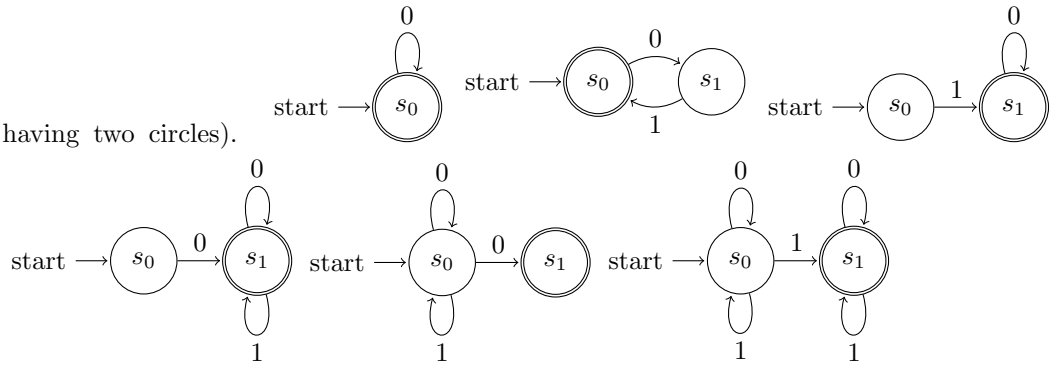
8.8.1 Support vector machines

Consider the set of points $F = \{(-1, 0), (0, 0), (1, 1)\}$. You may wonder why we separated (1,1) from (0,0) and (-1,0) and not some other combination, like separating (0,0) from (1,1) and (-1,0). In fact, such a separation can always be done. However, with four points it is not always possible. Consider the four points $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$. There is no (straight) line that separates $\{(1, 0), (0, 1)\}$ from $\{(0, 0), (1, 1)\}$. There is no way to get (1,0) and (0,1) to lie on one side of the separating line, and (0,0) and (1,1) on the other. This can be summarized by saying that our statistical model, involving finding parameters (w_1, w_2, b), has *Vapnik–Chervonenkis dimension* that is at least 3, but not at least 4. In other words, the VC dimension is exactly 3. A set of points of size 3 can be "shattered" but a set of points of size 4 sometimes cannot.

8.8.2 Discrete shattering

Shattering also occurs in discrete settings. Find a set of three binary strings $\{x_1, x_2, x_3\}$ of length 4 such that the following nondeterministic automata shatter the set. This means that for each subset F of $\{x_1, x_2, x_3\}$, there is an automaton M in the list below such that M accepts the strings in F, but not the strings in $\{x_1, x_2, x_3\} \setminus F$. Accepting means that there is a path through the automaton, starting at the state labeled "start", which consists of edges labeled according to the word x_i, and ends in the accepting state (the state

having two circles).



⊙ **Guided Practice 8.8** Show that $\{x_1, x_2, x_3\} = \{0000, 0101, 1000\}$ is shattered by the set of automata above.[10]

[10] For instance, 0000 and 0101 have in common that they start with 0, whereas 1000 does not.

8.9 Exercises

8.1 Statistics algebra. Let X and Y be Bernoulli random variables, both with parameter p, and let p_{ij} with $i, j \in \{0, 1\}$ be defined as in 8.1

(a) Explicitly calculate each value of p_{ij} to directly verify the equalities listed in 8.1.

(b) Use part a to show that $p_{00}p_{01} - p_{01}^2 = 0$. Thus, a probability distribution of two Bernoulli variables can be expressed as a point in $\mathbb{R}^4$ that lies on the curve $xy - zw = 0, z = w$.

(c) Show that the converse of part b is true in some sense: that is, show that solutions of the curve $xy - zw = 0, z = w$ can be expressed as probability distributions of two Bernoulli variables by showing that they satisfy the equalities in part a.

8.2 Change of parameter for likelihood. Let X be a binomial random variable with mean μ.

(a) Derive a formula $L(\mu)$ (so μ is acting as a parameter, instead of a probability p as in 8.3).

(b) Find the maximum likelihood of the formula in part a and explain why it might or might not be surprising.

8.3 Another variant on the likelihood parameter. Same as the previous problem, but replace mean μ with variance σ^2.

8.4 Akaike. Define the *Akaike information criterion with correction* by

$$AIC_C := AIC + \frac{2k^2 + 2k}{n - k - 1}$$

where n is the number of samples and k is the number of parameters.

(a) Show that as the number of samples increases, the AIC_C approaches the AIC.

(b) Derive the formula for the AIC_C of a linear regression model with n samples.

(c) For which sample sizes is the AIC_C a more accurate criterion than the BIC?

8.5 Equivalence of HMMS: alphabet. Show that any hidden Markov model is equivalent to one in which every state has exactly one transition for every symbol in the underlying language. (This is equivalent to the statement that any nondeterministic finite automaton is equivalent to a deterministic finite automaton.) Here, two hidden Markov models are equivalent if they accept the same symbols and output the same words with equal probabilities. Additionally, give an informal algorithm for finding such a hidden Markov model and give an upper bound on the number of states in the new Markov model.

8.6 Functional margin. Consider a cluster of points $\vec{x}_i = (x_{1i}, x_{2i})$ separated by a line $\vec{w} \cdot \vec{x} - b = 0$ for some b and $\vec{w} = (w_1, w_2)$. Let $y_i = 1$ if $\vec{x}_i$ is above this line and let $y_i = -1$ if $\vec{x}_i$ is below this line. Define the *functional margin of* (w, b) *with respect to* i by

$$\gamma_i := y_i(w_1x_1 + w_2x_2 + b)$$

for $\vec{x} = (x_1, x_2)$. Fix i.

(a) Show that γ_i is orthogonal (i.e. perpendicular) to the line $\vec{w} \cdot \vec{x} - b = 0$.

(b) Show that $\gamma_i(\vec{x}_i) > 0$.

8.7 Geometric margin. Again consider $\vec{x}_i$, $\vec{w}$, b, and y_i from the previous problem. Define the *geometric margin of* (w, b) *with respect to* i by

$$\hat{\gamma}_i := y_i \left(\frac{w_1x_1 + w_2x_2 + b}{\sqrt{w_1^2 + w_2^2}} \right)$$

for $\vec{x} = (x_1, x_2)$. Again as in the previous problem, fix i.

(a) Show that the same properties of γ_i in 6 also hold for $\hat{\gamma}_i$.

(b) Show that if w_1, w_2, and b are replaced with kw_1, kw_2 and kb for some constant k in the formula for $\hat{\gamma}_i$, the resulting expression will still be equal to $\hat{\gamma}_i$. In other words, show that $\hat{\gamma}_i$ is *invariant under rescaling of parameters*.

Appendix A

End of chapter exercise solutions

1 Introduction to data

1.1 (a) Possible answers: [1]Basically true. Categorical variables have no numerical meaning. Examples: Hair color, gender, field of study. [2]As a kind of counterexample, consider something like *What is your favorite number among* $\sqrt{2}$, $\sqrt{-1}$, *and* π*?* Here it doesn't make much sense to take the average of the responses, so it is not your typical numerical variable. But it does involve numbers.
(b) Possible answers: [1]Basically true. A discrete variable is a variable whose value is obtained by counting; a continuous variable is a variable whose value is obtained by measuring. [2]As a kind of counterexample, consider a *hybrid* variable X which has a 50% chance of $X = 3$, and a 50% of being a number $1 \le X \le 2$, uniformly distributed (so that, given that $1 \le X \le 2$, the probability that $X \le 5/4$ is 25%, for instance). (c) Possible answers: [1]Basically false. [2]As a kind of counterexample, consider *What is your favorite real number?* (where all real numbers are considered legitimate answers).

1.3 (a) New mean = 10+old mean = 70.7 + 10 = 80.7. (b) New mean = 2·old mean = 70.7·2 = 141.4. In formulae, $E(2X) = 2E(X) = 2(70.7) = 141.4$ where $E(X)$ is the mean of X.

(c) The means are

$$\begin{aligned} &= \frac{x_1 + c_1 + x_2 + c_1 + \cdots + x_n + c_1}{n} \\ &= \frac{nc_1 + (x_1 + \cdots + x_n)}{n} = c_1 + \frac{x_1 + \cdots + x_n}{n}. \end{aligned}$$

and $\frac{x_1 c_2 + \cdots + x_n c_2}{n} = c_2 \frac{x_1 + \cdots + x_n}{n}$, respectively. In formulae, $E(X + c_1) = E(X) + c_1$ and $E(c_2 X) = c_2 E(X)$. [1]

1.5 This uses Jensen's Inequality, which says that if $f''(x) \ge 0$ for all x (so f is *concave up* or *convex*) then a chord drawn between two points on the graph of f will lie above the graph of f. In symbols, $f(tx_1 + (1-t)x_2) \le tf(x_1) + (1-t)f(x_2)$ for al $0 \le t \le 1$. By iterating, we get for $a + b + c = 1$,

$$\begin{aligned} & f(ax + by + cz) \\ = \; & f\left(ax + (1-a)\left(\frac{b}{1-a}y + \frac{c}{1-a}z\right)\right) \\ \le \; & af(x) + (1-a)f\left(\frac{b}{1-a}y + \frac{c}{1-a}z\right) \\ \le \; & af(x) + (1-a)\left(\frac{b}{1-a}f(y) + \frac{c}{1-a}f(z)\right) \\ = \; & af(x) + bf(y) + cf(z) \end{aligned}$$

and similarly for $a_1 + \cdots + a_n = 1$, $f(\sum a_i x_i) \le \sum a_i f(x_i)$. In particular, since the function $f(x) = -\ln x$ satisfies $f''(x) = 1/x^2 \ge 0$ in its

[1]By the way, a solution to 1.2 is mean = (55 + 59 + 64 + 68 + 70 + 70 + 71 + 76 + 76 + 98)/10=70.7, median = (70+70)/2=70, mode = 70,76.

domain $x > 0$, and taking each $a_i = 1/n$, we have

$$-\ln(\overline{x}) \le \overline{-\ln x} := \frac{1}{n}\sum -\ln x_i,$$

which gives the arithmetic-geometric inequality by raising e to the power of both sides. For the harmonic inequality, consider instead

$$n/\sum(1/x_i) = \left(\sum(1/x_i)/n\right)^{-1}$$

which reduces the problem to the arithmetic-geometric inequality with x_i replaced by $1/x_i$.

1.7 (a) Let $g(x) = f(x) + c$. Then

$$\begin{aligned}\text{avg}_g &= \frac{1}{b-a}\int_a^b g(x)\,dx \\ &= \frac{1}{b-a}\int_a^b f(x) + c\,dx = \text{avg}_f + c.\end{aligned}$$

(b) Let $h(x) = c \cdot f(x)$. Then

$$\begin{aligned}\text{avg}_h &= \frac{1}{b-a}\int_a^b h(x)\,dx \\ &= \frac{1}{b-a}\int_a^b c\cdot f(x)\,dx = c\cdot\text{avg}_g.\end{aligned}$$

[2]

1.9 We have $(\frac{4}{3})(\frac{3}{2}) = 2 = 2^{12/12} = 2^{5/12}2^{7/12}$. The inequalities both follow from $2^{19} = 524,288 < 531,441 = 3^{12}$.

2 Probability

2.1 (a) Use the substitution $u = (x-\mu)/\sigma$. (b) What we need to verify: for the mean, $\int xf(x)\,dx = \mu$ (use the substitution $u = x - \mu$ in the integral). For the median, $P(X \le \mu = 1/2$. For the mode, $\varphi'(\mu) = 0$. (c) The variance is $E(X^2) - E(X)^2$ so we need to evaluate $E(X^2) = \int x^2\varphi(x)\,dx$.

2.3 (a) We have $\int f(x)\,dx = 1$ using the derivative of arctan.
(b) The median and mode are both 0, by symmetry $f(x) = f(-x)$ and since $f'(x) \le 0$ if and only if $x \ge 0$.
(c) Hint: the mean would be $\int xf(x)\,dx$ but this is similar to the harmonic series $\sum_n 1/n$. You can use the substitution $u = 1 + x^2$.
(d) Since the mean does not exist, we cannot claim that sample averages $\overline{x}_n$ from this distribution will converge to the mean. (If the mean were $+\infty$ we could perhaps argue that sample averages would diverge to $+\infty$, but since $-\infty$ is equally plausible as a mean here, all we can expect is "chaos" in the behavior of $\overline{x}_n$.

[2] By the way, a solution to 1.8 is: Since f is continuous, then there exists some number c such that $a < c < b$, and

$$f'(c) = \frac{f(b)-f(a)}{b-a} = \frac{\int_a^b f'(x)\,dx}{b-a}.$$

Let $g(x) = f'(x)$,then

$$g(c) = \frac{1}{b-a}\int_a^b g(x)\,dx = \text{avg}_g.$$

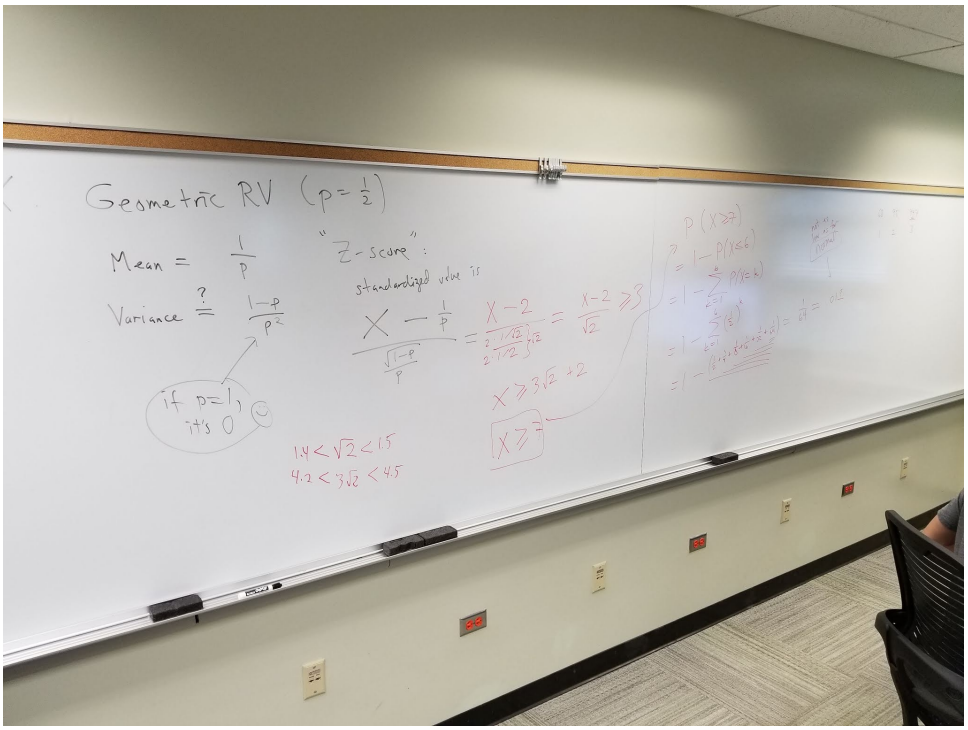


Figure A.1: A calculation related to the Likelihood Principle, from Fall 2018 in Webster Hall 101, UH Mānoa campus.

3 Distributions of random variables

3.1 (a) In general, normality is expected when dealing with variables for which the Central Limit Theorem holds approximately: sums of IID (independent, identically distributed) variables. A safe example is proportions of heads and tails when flipping a coin. Height of individuals may be an example if a person's height is approximately the result of a sum of independent events such as healthy meals, nights of sufficient sleep (for the individual or for their parents), etc.
(b) This is a tricky one; a library deep-dive into the given article is probably your best bet.
(c) We assume the student can look up things like *power-law distribution* online. One example is then *the frequencies of words in most languages.* [3]

3.3 (a) This is just $P(X = k-1)$ where X is binomial with parameters p and $n-1$. To wit:

$$\begin{aligned} & \binom{n-1}{k-1} p^{k-1}(1-p)^{(n-1)-(k-1)} \\ = \quad & \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k}. \end{aligned}$$

(b) Since the nth trial is independent of the pre-

[3] And here is a solution to 3.2 Three Sigma: (a) $P(X = 3) = 0.5(1-0.5)3-1 = 0.125 = 12.5\%$; $P(X = 10) = 0.5(1-0.5)10-1 = 0.00097 = 0.097\%$; $P(X = 100) = 0.5(1-0.5)100-1 = 0\%$. (b) $P(Z > 3) = 1-P(Z < 3) = 1-0.99865 = 0.00135$, $\implies$ 0.5(1-0.5)n-1=0.00135, $\implies$ $n \geq 9.5$. However, interestingly, if we consider the geometric distribution instead we get a different answer. This relates to a famous discussion about the Likelihood Principle https://en.wikipedia.org/wiki/Likelihood_principle (see Figure A.1).

vious trials, this is

$$\begin{aligned} & p \cdot \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k} \\ = & \binom{n-1}{k-1} p^k (1-p)^{n-k}. \end{aligned}$$

(c) The answer to (b) is exactly the probability that a negative binomial variable Y with parameters n and p will take value k. This is not surprising since the following two conditions are equivalent:

- the first success happens in trial n;
- the first $n-1$ trials are failures and the nth trial is a success.

3.5 We solve for β:

$$\beta = \frac{x}{1-x}\left(2\alpha + \frac{1}{1-x}\right)$$

Then

$$\begin{aligned} E(X^2) &= \sum_{k=1}^{\infty} k^2 P(X=k) \\ &= \sum_{k=1}^{\infty} k^2 (1-p)^{k-1} p = \frac{p}{1-p}\sum_{k=1}^{\infty} k^2 (1-p)^k \\ &= \frac{p}{1-p}\beta \\ &= \frac{p}{1-p}\frac{1-p}{p}\left(2\frac{1-p}{p^2} + \frac{1}{p}\right) \\ &= \frac{2-2p+p}{p^2} = \frac{2-p}{p^2} \end{aligned}$$

giving $\sigma^2 = \frac{2-p}{p^2} - (1/p)^2 = \frac{1-p}{p^2}$, as desired.

4 Foundations for inference

4.1 (a) If you are wondering whether a certain drug helps to cure cancer, you are not interested in proving that the drug just changes the outcomes for cancer. You really want it to help and not hurt. So a 1-sided test is appropriate.
(b) If you are hoping that a certain coin is fair, you don't necessarily care in which direction it may be biased. Too many heads and too many tails are not different outcomes for you, in any interesting sense. So you want a 2-sided test.

4.3 (a) If X and Y are independent then so are any deterministic (not random) functions of them $f(X)$, $g(Y)$. Moreover, whenever U and V are independent then $E(UV) = E(U)E(V)$. These facts are proved in more advanced texts. In particular, for any constant t, e^{tX} and e^{tY} are independent, and

$$\begin{aligned} M_{X+Y}(t) &= E(e^{t(X+Y)}) = E(e^{tX}e^{tY}) \\ &= E(e^{tX})E(e^{tY}) = M_X(t)M_Y(t). \end{aligned}$$

The other part is easier:

$$M_{cX}(t) = E(e^{t(cX)}) = E(e^{(ct)X}) = M_X(ct).$$

(b) The power series of e^x is $\sum x^n/n!$. Now assuming that $E(\sum) = \sum(E)$ (proved in advanced real analysis texts),

$$\begin{aligned} M_X(t) = E(e^{tX}) &= E\left(\sum (tX)^n/n!\right) \\ &= \sum E(X^n)t^n/n! \end{aligned}$$

(c) We have

$$M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)\Big|_{t=0}.$$

We then use $d/dt \sum^{\infty} = \sum^{\infty} d/dt$ (proved in advanced real analysis texts), and

$$\frac{d^n}{dt^n} t^n \Big|_{t=0} = n!$$

to finish.

4.5 (d) Suppose they are all bounded by b.
(e) We use the fact that the normal distribution with mean μ and variance σ^2 is the only distribution whose cumulants are $\mu, \sigma^2, 0, 0, 0, \ldots$.

5 Inference for numerical data

5.1 The price of *Statistics for Calculus Students* printed and bound, in the UCLA bookstore and on Amazon, is an example of paired data. The price of *Stats: Data and Models* by de Veaux, at UCLA, and *Probability: theory and examples* by Durrett, on Amazon, is an example of unpaired data.

5.3 (a) We are selecting a substantial proportion of the homes. Let's say we look at the proportion who plan to vote for the Democratic candidate (or, alternatively, the Republican candidate). If the sample was small compared to the population, but still large in absolute terms, we could argue that we have a sampling distribution to which the Central Limit Theorem applies. But 20% is too large a sample, creating dependencies. For instance, if all of the first 10% vote Republican it becomes unlikely that many of the other 10% vote Democrat.
(b) Now the proportion of homes selected is small enough for most purposes. However, the sample size of 10 is a bit small for the normal approximation to be reliable.

6 Inference for categorical data

6.1 (a) A hint is already given.
(b) We have

$$\Gamma(1/2) = \int_0^\infty t^{1/2-1} e^{-t}\, dt = \int_0^\infty \frac{dt}{e^t \sqrt{t}}.$$

Now try the substitution $t = u^2/2$.

6.3 Note that there are some "easy" solutions that don't quite work. But moment generating functions can be used:

$$M_{X_1+X_2}(t) = M_{X_1}(t) M_{X_2}(t)$$

Solve for $M_{X_2}(t)$.

6.5 Squaring both sides,

$$\begin{aligned} \frac{\hat{p}_y(1-\hat{p}_y)}{1924} + \frac{\hat{p}_x(1-\hat{p}_x)}{3666} &= 10^{-4} \\ \hat{p}_y - \hat{p}_x &= \frac{1}{25} \end{aligned}$$

$$\frac{(\hat{p}_x + \frac{1}{25})(1-\hat{p}_x-\frac{1}{25})}{1924} + \frac{\hat{p}_x(1-\hat{p}_x)}{3666} = 10^{-4}$$

$$\frac{(u + \frac{1}{25})(1-u-\frac{1}{25})}{1924} + \frac{u(1-u)}{3666} = 10^{-4}$$

This we can solve by hand; Wolfram Alpha gives[4] $u = 5093/10750 - \sqrt{7133687/2}/5375$, i.e., $u \approx 0.12240$ or $u \approx 0.82514$. Looking at the paper [5], we see that actually $p_y = 233/1924 = .1211$ and $p_x = 306/3666 = .0835$ which gives the standard error $0.0087 \approx 0.01$.

[4] https://www.wolframalpha.com/input/?i=%5Cfrac%7B(u%2B%5Cfrac1%7B25%7D)(1-u-%5Cfrac1%7B25%7D)%7D%7B1924%7D%2B%5Cfrac%7Bu(1-u)%7D%7B3666%7D+%3D+10%5E%7B-4%7D

[5] http://thegeeksverse.com/wp-content/uploads/2014/09/SexOrColorPref.pdf

7 Introduction to linear regression

7.1 (a) The points $\{(0,0),(1,1),(2,2)\}$ form one example. We need at least 3 points for r to be defined because of the $n-2$ in a denominator.
(b) Using

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

When $n=0$ or $n=1$, we have all $x_i=\bar{x}$ and $y_i=\bar{y}$, hence r is of the form "0/0" and undefined. When $n=2$, $x_1-\bar{x} = x_1-\frac{x_1+x_2}{2} = \frac{x_1-x_2}{2}$ and

$$\begin{aligned} r &= \frac{\frac{x_1-x_2}{2}\frac{y_1-y_2}{2}+\frac{x_2-x_1}{2}\frac{y_2-y_1}{2}}{\sqrt{2((x_1-x_2)/2)^2 2((y_1-y_2)/2)^2}} \\ &= \frac{\frac{x_1-x_2}{2}\frac{y_1-y_2}{2}+\frac{x_2-x_1}{2}\frac{y_2-y_1}{2}}{2\sqrt{((x_1-x_2)/2)^2((y_1-y_2)/2)^2}} \\ &= 2\frac{\frac{x_1-x_2}{2}\frac{y_1-y_2}{2}+\frac{x_2-x_1}{2}\frac{y_2-y_1}{2}}{\sqrt{((x_1-x_2))^2((y_1-y_2))^2}} \\ &= \frac{(x_1-x_2)(y_1-y_2)}{|(x_1-x_2)(y_1-y_2)|} \\ &= \text{sign}((x_1-x_2)(y_1-y_2)), \end{aligned}$$

where

$$\text{sign}(x) = \begin{cases} 1 & \text{if } x>0, \\ -1 & \text{if } x<0, \\ \text{undefined} & \text{if } x=0. \end{cases}$$

In any case, this is a long-winded way of saying that there can be no example with $n=2$, either. How about $n=3$? We get the algebraic equation

$$\begin{aligned} & (2x_1-x_2-x_3)(2y_1-y_2-y_3) \\ + \; & (2x_2-x_3-x_1)(2y_2-y_3-y_1) \\ + \; & (2x_3-x_1-x_2)(2y_3-y_1-y_2)=0. \end{aligned}$$

with the constraint that not all $x_i=\bar{x}$, and not all $y_i=\bar{y}$.[6] We might as well assume $(x_3,y_3)=(0,0)$, which gives

$$\begin{aligned} & (2x_1-x_2)(2y_1-y_2) \\ + \; & (2x_2-x_1)(2y_2-y_1) \\ + \; & (-x_1-x_2)(-y_1-y_2)=0. \end{aligned}$$

If we also assume $(x_2,y_2)=(1,1)$, this becomes

$$\begin{aligned} & (2x_1-1)(2y_1-1) \\ + \; & (2-x_1)(2-y_1) \\ + \; & (x_1+1)(y_1+1)=0. \end{aligned}$$

Let us rewrite it in variables without subscripts:

$$(2x-1)(2y-1)+(2-x)(2-y)+(x+1)(y+1)=0.$$

This simplifies (Wolfram Alpha) to $(2x-1)y = x-2$. So we can take $x=-1$ and $y=1$.
(c) This is similar to (a). Take $\{(0,0),(1,-1),(2,-2)\}$ for instance. [7]

7.3 (a) Square both sides, multiply through to clear denominators, and use your algebra.

[6] It may be tempting to propose $\{(0,0),(1,0),(2,0)\}$ as an example of $r=0$, based on the idea that when the slope is 0, there is no positive or negative correlation, so the correlation is 0. But technically, that is incorrect, as r will be undefined when there is no variation in the y-values. What is correct is that $b_1=0$ for this example.

[7] Here is a solution to 7.2 (Cauchy-Schwarz): (a) Expanding the terms

$$(u_1x+v_1)^2+\cdots+(u_nx+v_n)^2$$

Yields

$$u_1^2x^2+2u_1v_1x+v_1^2+\cdots+u_n^2x^2+2u_nv_nx+v_n^2$$

Factoring common powers of x gives

$$(u_1^2+\cdots+u_n^2)x^2+(2u_1v_1+\cdots+2u_nv_n)x+(v_1^2+\cdots+v_n^2)$$

(b) The discriminant D tells us

- there are two real roots if $D>0$;
- there is a real root if $D\geq 0$;
- there are no real roots if $D<0$.

The polynomial has at most most one root, therefore the discriminant is less than or equal to 0. So since $D=b^2-4ac$,

$$\begin{aligned} (2u_1v_1+...+2u_nv_n)^2-4(u_1^2+...+u_n^2)(v_1^2+...+v_n^2)\leq 0, \\ 4(u_1v_1+...+u_nv_n)^2\leq 4(u_1^2+...+u_n^2)(v_1^2+...+v_n^2), \\ (u_1v_1+...+u_nv_n)^2\leq (u_1^2+...+u_n^2)(v_1^2+...+v_n^2), \end{aligned}$$

as desired.

8 Hidden Markov models

8.1 (a) The probability that $X = 1$ and $Y = 0$, for instance, is the probability that $X = 1$, minus the probability that $X = 1$ and $Y = 1$.

8.3 (a) Here we have to consider the equation $p(1-p)n = \sigma^2$ and replace occurrences of p by expressions involving σ.

8.5 Here is a solution to 8.2(a). $\mu = np$ and $p = \mu/n$. So $P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$ and $L(\mu) = \binom{n}{i} (\mu/n)^i (1-\mu/n)^{n-i}$. (b) Let $x_1, \ldots, x_m$ be a random sample. Then starting with $L(\vec{x}, \mu)$ and taking logs, and differentiating with respect to μ, and setting it equal to 0, we get $\mu = \sum_{i=1}^{m} x_i/m$, which is not surprising as that is our sample mean.

8.7 Given

$$\gamma_i = \frac{(\bar{w^T}\bar{x} + b)}{\|(wTx + b)\|}$$

$$\bar{w^T}\bar{x} - b = 0 \implies \bar{w^T}\bar{x} = b \implies (\bar{w^T}\bar{x})^2 - b^2 = 0$$

$$\gamma_i \cdot (\text{Line}) = \gamma_i \cdot (w^T x - b) = 0.$$

Therefore the two are perpendicular because the dot product is zero.

Index