Stat 111: Introduction to Statistical Inference[1]

Joseph K. Blitzstein and Neil Shephard

February 19, 2024

[1]©2023 by Joseph K. Blitzstein and Neil Shephard

Contents

Chapter 1

Introduction

1.1 Overview

Welcome to Stat 111! The field of **statistics** centers around three main goals:

1. *exploring and describing* data and a phenomenon of interest,
2. *predicting* one variable using other observed variables, and
3. drawing *causal* conclusions about the effect of changing one variable on another.

This course will delve into the principles and methods for using **statistical inference** to achieve these goals. Statistical inference is a framework for:

1. *statistical model building*,
2. *learning* from data, and
3. making principled *decisions* under uncertainty.

Combining the goals of statistics with the principles of statistical inference gives the subject called **statistical science**. We take a **three-pronged approach** to studying statistical science in Stat 111, comprising:

1. developing *statistical theory*,
2. using *simulation* methods, and
3. analyzing real *data*.

A hallmark of statistics is that we aim not only to make statements based on data, but also to assess how *confident* we should be about those statements. It is human nature to see patterns in data that

are actually just artifacts of random noise, to confuse correlation for causality, or to be overconfident in one's predictions through underestimating variability or biases in data. Careful statistical inference helps us avoid fooling ourselves (and others), by insisting that we quantify the uncertainty in our statements.

We will use core ideas from probability throughout the course, since probability provides both a language and a toolbox for quantifying uncertainty. In particular, we assume a good understanding of the first 11 chapters of the book *Introduction to Probability* by Blitzstein and Hwang, which we refer to as the "Stat 110 book".

Three motivating questions to keep in mind throughout the course are:

1. *What is truth?* The real world is messy and complicated. Statisticians therefore use *models* to shed light on some aspect of reality. We will define "statistical model" mathematically later, but for now think of it as a simplified, approximate probability distribution for the data.

   Statisticians often oscillate between drawing conclusions from a model under the assumption the model is true, and criticizing and evaluating the model to see whether it needs to be replaced by a better model. Thus we recognize that we will never find the true model, but hope to build a better model if we find that our original model is inadequate for its intended purpose.

2. *What is good?* It is easy to come up with a statistical model or method, but doing so is not worth much if it is not accompanied by an evaluation of how *good* the model or method is. But before we can assess how good a model or method is, we need to define "goodness". There is no single standard definition, so we will study several criteria and measures by which models and methods can be evaluated.

3. *What is beautiful?* Statistics is often portrayed as an ugly litany of cryptic formulas and mechanical procedures. In this course we will try to show that statistics is replete with beautiful, surprising, and powerful ideas, from which one can derive methods that have countless applications. For example, a remarkable fact about statistics is that the *same data* that provides an estimator can also be used to assess the accuracy of the estimator.

   To quote a Statistics concentrator alumna, from her evaluation of Stat 110:

   > *I learned that Stats is my one true love. ... It's the bomb. Plus it makes you think about the world differently... it's like, so many things are unknown, but we can always know the extent to which we do not know them. That is crazy to think about.*

1.2 Big picture

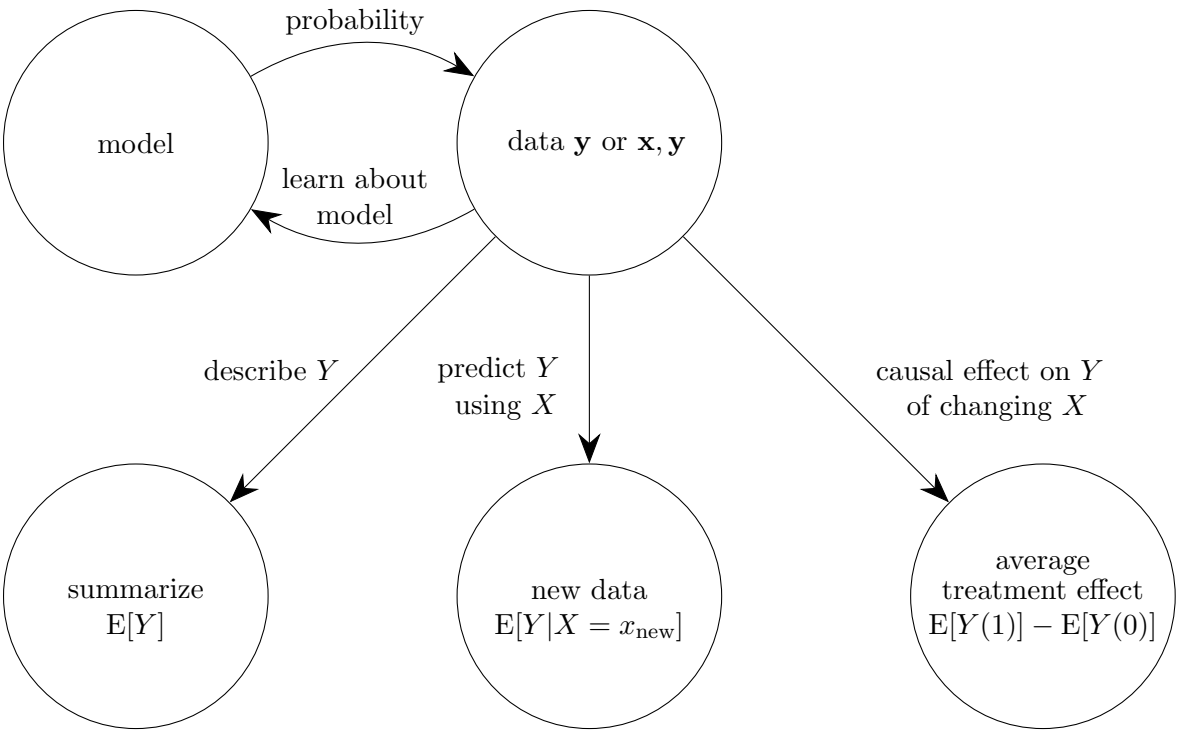


Figure 1.1: Roadmap of the relationships between some of the most fundamental concepts in statistics.

Figure 1.1 illustrates schematically the main tasks of statistics. In *probability*, we often specify a joint distribution for a collection of random variables, such as letting $Y_1, \ldots, Y_n$ be i.i.d. Then we can address questions like "What is the probability of $Y_1 \geq 1$? What are the mean and standard deviation of Y_1? What are the mean and standard deviation of $\max(Y_1, \ldots, Y_n)$?".

In *statistics*, we often then let $\mathbf{Y} = (Y_1, \ldots, Y_n)$ "crystallize" to observed data $\mathbf{y} = (y_1, \ldots, y_n)$, and we explicitly recognize that the joint distribution is unknown. In some problems we may be content to summarize the data $y_1, \ldots, y_n$. It may be much easier to understand and interpret a few key summaries than the entire dataset, especially if n is large. But often we are interested more broadly in the underlying *phenomenon* that gave rise to the data. So we can develop methods to *describe* the phenomenon, by learning about the joint distribution (or important aspects of the joint distribution, e.g. $\mathrm{E}[Y_1]$) from the data.

Another common goal is *prediction*: we may want to use the observed data $\mathbf{x} = (x_1, \ldots, x_n)$ to predict some other data $\mathbf{y}$.

Often we want to infer *causation* between $\mathbf{x}$ and $\mathbf{y}$ rather than mere correlation. We can then consider what will happen to an individual's Y value if we intervene by changing their X value.

Figure 1.2 displays the main goals of statistics and the core ideas in statistical inference. Together, they form the core of statistical science. References are given to which chapter focuses on each goal and idea.

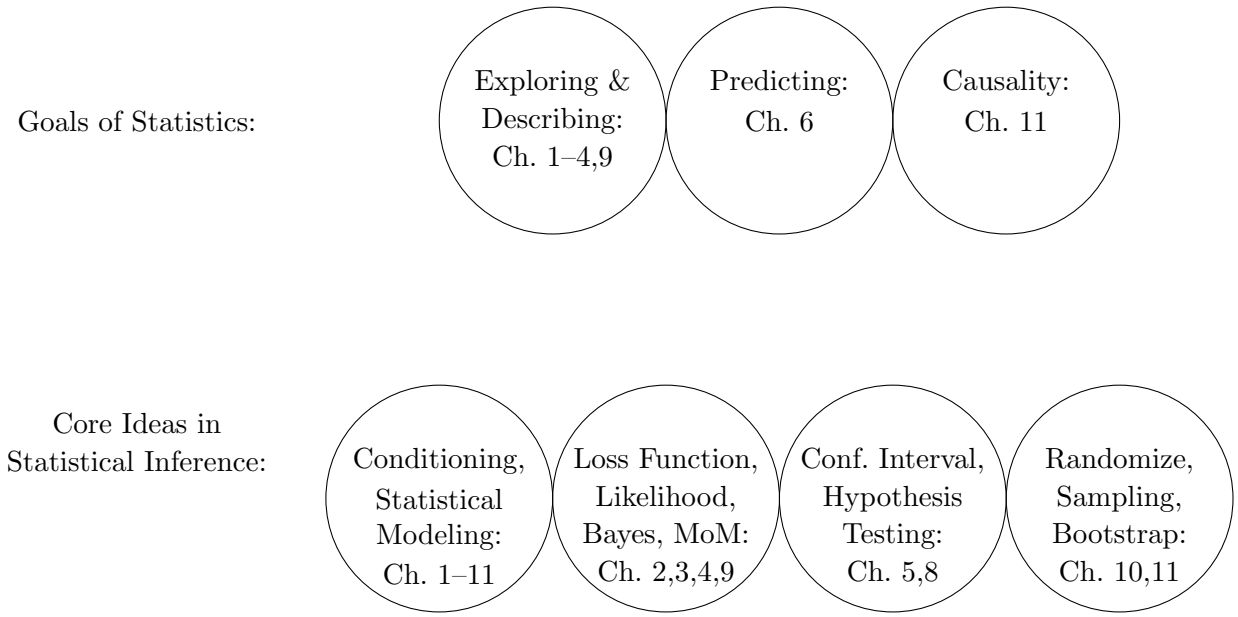


Figure 1.2: The core goals of statistics together with the main tools developed in statistical inference supporting them.

To run through Figure 1.1 in a concrete example, suppose that traffic engineers are studying the rate at which car accidents occur per month at a particular busy intersection. The engineers collect data over a year and record the results as $n = 12$ counts $y_1, \ldots, y_n$, where y_j is the number of accidents in the jth month, for $j = 1, \ldots, n$.

Notation 1.2.1. Typically, in statistics we use capital letters for random variables and the corresponding lowercase letters for the corresponding observed values of the random variables. That is, think of $Y_1, \ldots, Y_n$ as random variables that will "crystallize" into observed data $y_1, \ldots, y_n$. We often need to go back and forth between a pre-data point of view (before the data have been observed, so $Y_1, \ldots, Y_n$ are still random) and a post-data point of view (after we have observed numerical values $y_1, \ldots, y_n$).

When we have only one set of random variables, it is typical in statistics to use the letter Y. An old, not especially funny joke goes as follows:

How can you tell the difference between a statistician and a probabilist?
A probabilist uses the letter X, while a statistician uses the letter Y.

The reason for this difference is that statisticians often want to talk about using X to predict Y, or about estimating the causal impact of X on Y.

- Let $Y_1, \ldots, Y_n$ be the random variables that will "crystallize" into the observed values $y_1, \ldots, y_n$. A *statistical model* is a collection of possible joint distributions for $Y_1, \ldots, Y_n$.

  The model could be very general, e.g., we could assume that the Y_j are i.i.d. but not assume a specific named distribution. Or the model could be much more specific, such as assuming that the Y_j are i.i.d. Poisson random variables. There tend to be tradeoffs involved in how weak or strong to make our modeling assumptions.

- *A parametric statistical model* is a collection of possible joint distributions for $Y_1, \ldots, Y_n$, indexed by some parameter θ. For example, suppose that the times of the accidents occur according to a Poisson process of rate θ accidents per month. Then the model is

$$P(Y_1 = y_1, \ldots, Y_n = y_n; \theta) = \prod_{j=1}^{n} \frac{e^{-\theta}\theta^{y_j}}{y_j!} = \frac{e^{-n\theta}\theta^{y_1+\cdots+y_n}}{y_1! \ldots y_n!},$$

  indexed by θ in the *parameter space* $(0, \infty)$.

- *An estimand* is a particular quantity that we wish to learn. Here a natural estimand is the rate of accidents per month, which is denoted by θ.

- *Probability* focuses on problems such as "Find $\mathrm{E}[Y_1], \mathrm{Var}(Y_1)$, and $P(Y_1 = 0)$ (in terms of θ)." Based on the model, probability lets us determine how likely various events are and what the typical values of our random variables are.

- *Exploring* provides interpretable summaries of the data, either visually or numerically. For example, we can plot the y_j's as a function of time j to get a sense of whether it looks like there is a trend in traffic accidents over time (it will, however, be challenging to know if a pattern is meaningful or just random fluctuations. There could also be seasonal effects that would not be apparent from looking at only one year of data). As simple numerical summaries, we might want to compute the sample mean, the sample variance, and the sample probability of a zero:

$$\bar{y} = \frac{1}{12}\sum_{j=1}^{12} y_j,$$

$$s^2 = \frac{1}{11}\sum_{j=1}^{12} (y_j - \bar{y})^2,$$

$$\hat{P}(Y_1 = 0) = \frac{1}{12}\sum_{j=1}^{12} I(y_j = 0),$$

where $I(y_j = 0)$ is the indicator of $y_j = 0$ (defined to be 1 if $y_j = 0$, and 0 otherwise). These summaries are analogs of $\mathrm{E}[Y_1], \mathrm{Var}(Y_1)$, and $P(Y_1 = 0)$, respectively. They are actual numbers that can be computed from the data, rather than being theoretical quantities that can be calculated using probability in terms of θ.

- *Describing* goes in the reverse direction to probability, addressing the fact that θ is typically unknown and needs to be estimated from the data. So we can consider questions such as "Given the data $y_1, \ldots, y_n$, how should we estimate the rate θ? How should we estimate $P(Y_1 = 0)$? How confident should we be about our estimates?" With more fine-grained data about when accidents occur, we can also assess how good the Poisson process assumption is, e.g., what if accidents are more likely to occur during rush hour traffic than at other times?

  Various strategies for estimating an estimand are possible, and it may be far from obvious which is better (or how even to define "better"). For example, it would be natural to estimate $P(Y_1 = 0)$ with the estimate $\hat{P}(Y_1 = 0)$ defined above, since $\hat{P}(Y_1 = 0)$ is the proportion of months in which there was no accident. But if we assume $Y_1 \sim \mathrm{Pois}(\theta)$ then an alternative strategy would be create some estimate $\hat{\theta}$ for θ and then plug $\hat{\theta}$ in for θ in $P(Y_1 = 0) = e^{-\theta}$, which would give $e^{-\hat{\theta}}$. Later chapters will discuss extensively the principles for constructing and evaluating procedures for estimating unknown quantities.

- *Prediction* considers the question of how to use observed data to predict not yet observed data (such as data in the future). Before deciding how much time, attention, and money to devote to making the intersection safer, we may want to predict how many accidents will occur at the intersection in future months. Given that we have observed $y_1, y_2, \ldots, y_{12}$, what is our best guess for y_{13}, and how confident should we be about this guess?

- *Causality* asks what the effect on a variable will be if we intervene to *change* another variable. Instead of passively watching car crashes, can we intervene to make the intersection safer? Suppose that we observe an alarmingly high value of y_1 and then, before observing any more data, install a stop sign at the intersection. Imagine that we then observe that y_2 is much smaller than y_1.

  That is happy news, but we don't know that the stop sign *caused* the accident rate to decrease. Maybe $y_2 < y_1$ just due to chance, to changes in weather conditions, or people becoming more

cautious there after hearing how dangerous the intersection was. In deciding what policy interventions to make, we want to assess the *causal* impact of those interventions, but it is far from obvious what kinds of data will allow us to validly make causal claims.

☣ **1.2.2.** The above tasks are related but conceptually distinct. When working on a statistics problem, it is crucial to be clear on which of these tasks we are trying to do at any particular time.

In the remaining sections of this chapter, we give a very brief overview of two major approaches to statistical inference (*frequentist inference* and *Bayesian inference*) and the statistical goals (exploration and description, prediction, and causality) introduced above. In the remaining chapters of this book we discuss each of these topics in much greater depth.

1.3 Learning and deciding: frequentist and Bayesian inference

There are two major kinds of tasks for inference of unknown quantities, both of which we will delve into in Stat 111: *learning* and *deciding*. Let θ be an estimand, that is a unknown quantity of interest we wish to learn, e.g. the mean of a random variable or a parameter of a distribution.

1. *Learning about θ.*

   (a) *Point estimation.* We can come up with an *estimator* $\hat{\theta}$, as a function of the $Y_1, ..., Y_n$. Ideally, we will be able to prove that $\hat{\theta}$ will be close to θ with high probability. As discussed under the question "What is good?", we may need to think hard about the criteria for judging how good $\hat{\theta}$ is. Once we have chosen the criteria, we can try to use mathematical or computational tools to evaluate how good the estimator is.

   (b) *Interval estimation.* Rather than just giving a guess like "I think θ is probably close to 3", it may be much more informative to provide an *interval* that contains θ with high probability. This yields an *interval estimate* such as $(2.7, 3.3)$ rather than the *point estimate* $\hat{\theta} = 3$. If θ is multi-dimensional, we can try to construct a *region* that contains θ with high probability, e.g., in two dimensions we can provide a rectangle rather than an interval.

2. *Deciding about θ.* In some applications, the goal is to make a decision. A common example of this is to test a hypothesis, e.g., to assess whether it is plausible that $\theta = 0$ or the data lead us to resoundingly reject this hypothesis.

   Or we may wish to decide between two rival hypotheses, e.g., to choose between $\theta = \theta_0$ and $\theta = \theta_1$, where θ_0 and θ_1 are pre-determined values. In making such a decision, it may be logical

not only to consider whether $\theta = \theta_0$ or $\theta = \theta_1$ seems to be more consistent with the data, but also to consider the *costs* of wrong decisions. For example, the cost incurred from deciding in favor of θ_1 when θ_0 is true may be much higher than the cost from deciding in favor of θ_0 when θ_1 is true.

Learning and deciding can be approached from either *frequentist* or *Bayesian* perspectives. We will explore both of these schools of thought in detail in this book.

The frequentist approach focuses on coming up with procedures that work well in the long run; this requires considering drawing new datasets over and over again. That is, we look at the performance of the procedure in (hypothetical) repeated sampling.

In contrast, the Bayesian approach focuses on the data at hand. We model θ as a random variable, as a language for expressing our uncertainty about θ. After specifying a *prior distribution* for θ, we use Bayes' rule to update our probabilities for θ based on the data, thus obtaining the *posterior distribution* of θ given the data. Once we have the posterior distribution we can, for example, estimate θ using the posterior mean (the mean of the posterior distribution) or posterior median (the median of the posterior distribution), or create an interval that has, say, 95% chance of containing θ, given the data.

1.4 Exploring and describing Y

In *exploratory and descriptive statistics*, we attempt to give useful numerical or visual summaries of datasets, explore relationships between variables, or learn about the parameter of a model. Well-chosen summaries can help us make sense of the data, find errors or anomalies in the data, and see relationships or trends that warrant further investigation. Here is a (far from comprehensive) list of some useful summaries of a dataset $y_1, y_2, \ldots, y_n$.

1. There are many possible *visualizations* of the data, such as histograms and scatter plots. A bad visualization can be extremely misleading; a good visualization can help immensely in understanding trends and relationships in data. A good rule of thumb when starting to analyze a dataset is to *plot the data*.

2. Widely used summaries for the average and the spread of the dataset $y_1, \ldots, y_n$ are the *sample mean* and *sample standard deviation*.

Definition 1.4.1 (Sample mean, sample standard deviation)**.** The *sample mean* of $y_1, \ldots, y_n$ is

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j.$$

The *sample standard deviation* s is the square root of the *sample variance*

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2.$$

Later in the book we will discuss why it is conventional (though not necessarily optimal) to divide by $n-1$ rather than n in the definition of s^2.

3. Suppose that for each individual we observe both an x variable and a y variable, so our data are pairs $(x_1, y_1), \ldots, (x_n, y_n)$. The *sample covariance, sample correlation, and linear regression* are widely used summaries of the relationship between the x variable and the y variable.

   Definition 1.4.2 (Sample covariance, sample correlation, linear regression)**.** The *sample covariance* is

   $$s_{x,y} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})$$

   and the *sample correlation* is

   $$r_{x,y} = \frac{s_{x,y}}{s_x s_y}.$$

   The *linear regression* of y on x is

   $$b_{y\sim x} = \frac{s_{x,y}}{s_x^2},$$

   where s_x and s_y are the sample standard deviations of $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, respectively.

   Linear regression is discussed in detail in Chapter 6.

4. In many problems it is insightful to look at the data sorted into increasing order rather than just keeping the data in the original order in which they were collected.

   Definition 1.4.3 (order statistics)**.** The *order statistics* of $y_1, y_2, \ldots, y_n$ are the same data points, sorted in increasing order:

   $$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}.$$

   Some examples of quantities based on order statistics are the *sample minimum* $y_{(1)}$, the *sample maximum* $y_{(n)}$, the *range* $y_{(n)} - y_{(1)}$, and the *sample median* $y_{((n+1)/2)}$ (if n is odd; there are different conventions about what to do if n is even).

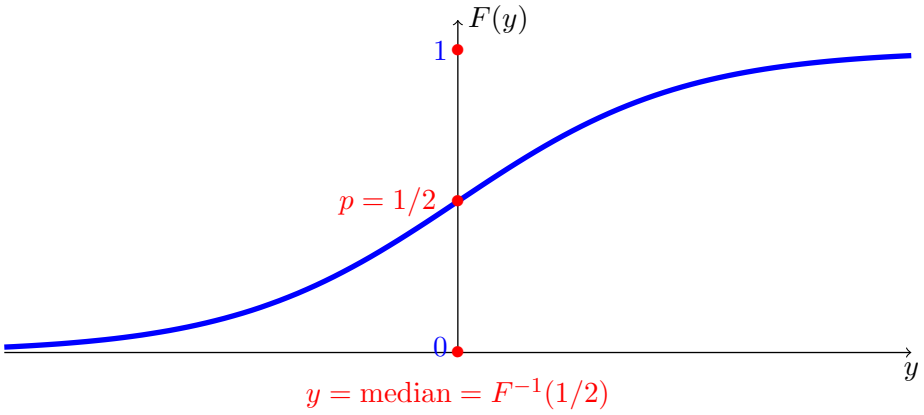


Figure 1.3: The cumulative distribution function $F(y)$ drawn against y. At the median, the function F hits the value 1/2.

For a continuous random variable Y with a strictly increasing cumulative distribution function (CDF) F, the *median* is $F^{-1}(1/2)$, i.e., there is a 50% chance that Y is below the median and a 50% chance that it is above the median. This is illustrated in Figure 1.3.

Similarly, if n is odd and the y_j do not have any repeated value, then 50% of the data points other than the sample median are below the sample median, and 50% are above the sample median. So the sample median is the value in the middle, after sorting the data. Note how the *sample median* of a dataset is analogous to the *theoretical median* of a distribution. Furthermore, if we assume that the $y_1, \ldots, y_n$ are generated as i.i.d. draws from F, then the sample median of $y_1, \ldots, y_n$ is a natural way to estimate the theoretical median $F^{-1}(1/2)$ (we will discuss the properties of this estimation procedure later).

More generally, let F be any CDF. If F^{-1} exists (in the ordinary sense from algebra of the inverse of a function), then $F^{-1}(p)$ may be a useful summary of the distribution for any fixed $p \in (0, 1)$. The *quantile function* generalizes this notion to work for any CDF. The quantile function plays an important role in modern statistics.

Definition 1.4.4 (Quantile function)**.** Let F be a CDF. The *quantile function* of F is defined by

$$F^{-1}(p) = Q(p) = \min\{y : F(y) \geq p\}.$$

The value $Q(p)$ is called the *p-quantile* of the distribution. For a random variable Y, we will use Q_Y to denote the quantile function of Y.

In English, this finds the smallest y such that the CDF at y attains at least a value of p. An example is illustrated in Figure 1.4, with a Bernoulli random variable Y with parameter 0.7.

This CDF only takes on the values $0, 0.3$, and 1. But, for example, $Q(0.6)$ is still well-defined: it is the smallest y value for which $F(y)$ is at least 0.6, which is $y = 1$.

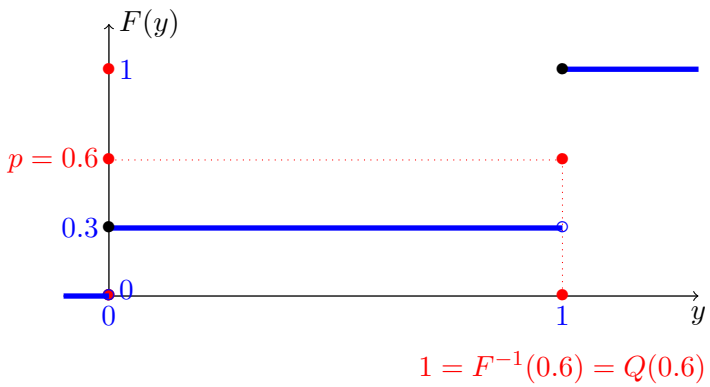


Figure 1.4: Distribution function of a binary random variable with $P(Y = 0) = 0.3$ and $P(Y = 1) = 0.7$. Here we compute the 0.6-quantile, $F^{-1}(0.6) = Q(0.6)$, which is 1.

An illustration in the continuous case is given in Figure 1.5. For this CDF, Q is just the inverse of F in the usual sense. For any fixed $p \in (0, 1)$, the value $Q(p)$ is the value on the horizontal axis such that the corresponding value of the function is p. Algebraically, this just says that

$$F(F^{-1}(p)) = p.$$

Note though that for a general CDF F (that may not be invertible), by definition of the quantile function we have

$$F(F^{-1}(p)) \geq p,$$

but this inequality may be strict. For example, for the distribution from Figure 1.4,

$$F(F^{-1}(0.6)) = F(1) = 1 > 0.6.$$

The corresponding sample quantity is the *p-sample quantile*, defined to be a value such that approximately proportion p of the sample is less than or equal to that value. There are various different conventions for the precise definition of the sample quantile of $y_1, \ldots, y_n$; typing `help(quantile)` in `R` lists 9 different conventions! If n is large then it is unlikely to matter which of these conventions is used. We will use a simple convention.

Definition 1.4.5 (Sample quantile)**.** The *p-sample quantile* of the dataset $y_1, \ldots, y_n$ is the order statistic $y_{(\lceil np \rceil)}$, where $\lceil \cdot \rceil$ is the ceiling function (which says to round up to the next integer, e.g., $\lceil 1.2 \rceil = 2$). We will denote the p-sample quantile by $\hat{Q}(p)$:

$$\hat{Q}(p) = y_{(\lceil np \rceil)}.$$

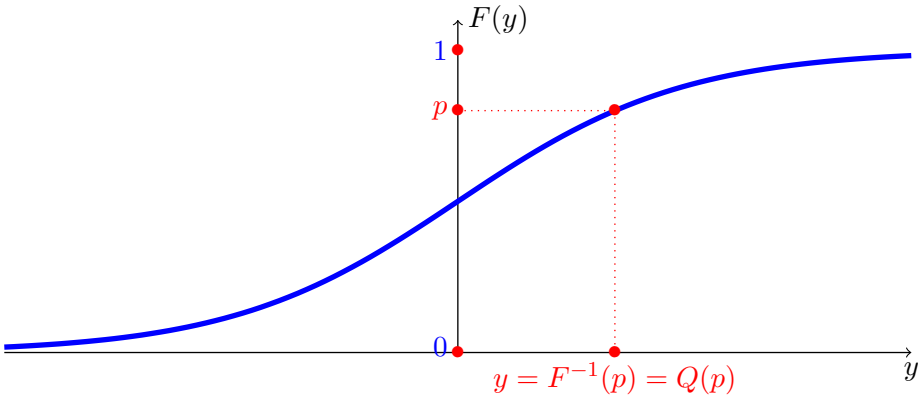


Figure 1.5: The cumulative distribution function $F(y)$ drawn against y, showing the p-quantile $Q(p)$.

The *sample first quartile*, *sample median*, and *sample third quartile* are $\hat{Q}(0.25), \hat{Q}(0.5)$, and $\hat{Q}(0.75)$, respectively.

For example, for $n = 100$ the 25% sample quantile is $y_{(25)}$, which makes sense since if there are no ties then 25% of the data points are less than or equal to $y_{(25)}$, while 75% of the data points are greater than $y_{(25)}$.

5. From probability, we know that the CDF is a useful way to specify a distribution. The analogous notion for a dataset is the *empirical CDF*.

Definition 1.4.6 (Empirical CDF)**.** The *empirical CDF* (ECDF) of the data set is the CDF of a random variable (r.v.) obtained by choosing one of the n data points uniformly at random. If $y_1, \ldots, y_n$ are distinct, this is the CDF of the discrete r.v. Y with probability mass function (PMF)

$$P(Y = y_j) = 1/n$$

for all j. We can write the ECDF mathematically as

$$\hat{F}(y) = \frac{1}{n} \sum_{j=1}^{n} I(y_j \leq y),$$

where $I(y_j \leq y)$ is the indicator of $y_j \leq y$.

An ECDF is illustrated in Figure 1.6. Note that the ECDF is always a step function, jumping every time it reaches one of the data points. Currently we are using the ECDF as a summary of the data, rather than looking at a statistical model. But if we do have a model, under which the

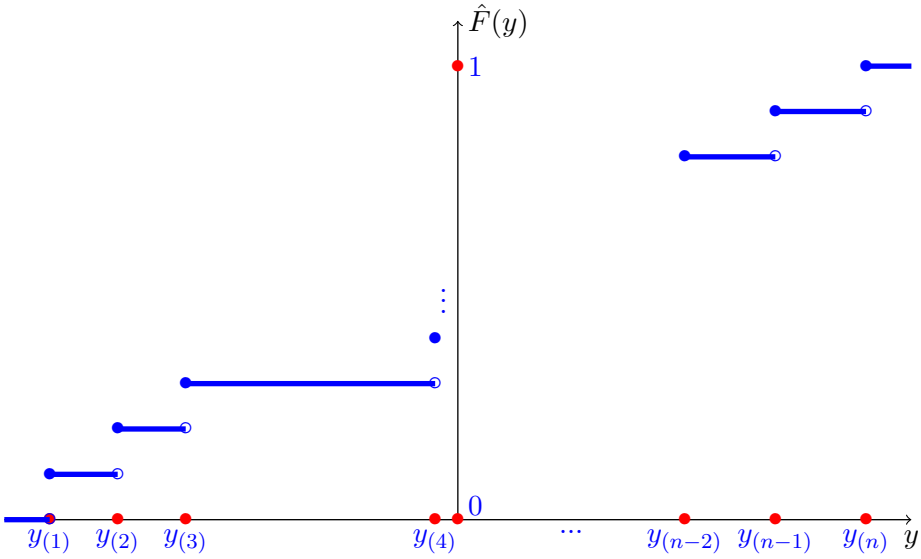


Figure 1.6: Empirical distribution function. On the x-axis we display the ordered data $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$.

data are realizations of i.i.d. r.v.s $Y_1, Y_2, \ldots$ with CDF F, then the strong law of large numbers implies that for each $y \in \mathbb{R}$, we have

$$\lim_{n\to\infty} \hat{F}(y) = E[I(Y_1 \leq y)] = P(Y_1 \leq y) = F(y) \text{ with probability 1.}$$

So the ECDF converges to the true CDF, as the sample size grows.

Furthermore, ECDFs give us a new way to think about sample quantile: the p-sample quantile can also be written as the quantile function of the ECDF, evaluated at p:

$$\hat{Q}(p) = \min\{y : \hat{F}(y) \geq p\}.$$

That is, $Q(p)$ is the theoretical quantile of the *distribution*, while $\hat{Q}(p)$ is the empirically computed quantile of the *sample*.

Example 1.4.7 (Births)**.** Ethel Burns, a researcher at Oxford who studies childbirth, collected data on the duration of births for babies born without a C-section at John Radcliffe Hospital in Oxford, U.K. over a seven day period. The data measures the time spent in the delivery room. This dataset was published and analyzed in Davison (2003).

Write the data, measured in hours, as

$$\mathbf{y} = (y_1, \ldots, y_n),$$

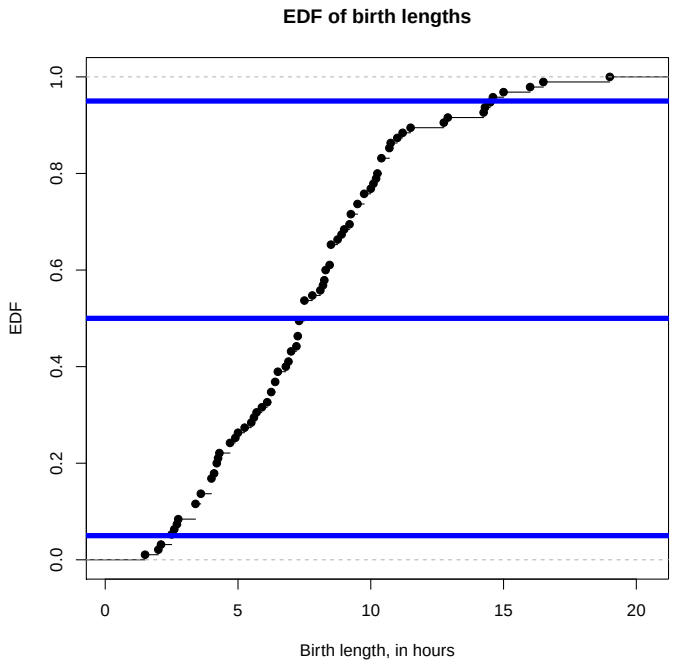


Figure 1.7: Empirical CDF for durations of births. Blue lines pick out the 0.05, 0.50, and 0.95 sample quantiles.

where y_j denotes the duration of birth for the jth individual and $n = 95$ is the *sample size*. Two simple but useful *descriptive* summaries are the sample mean and standard deviation:

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j \approx 7.72, \quad s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(y_j - \overline{y})^2} \approx 3.57.$$

The empirical CDF of the dataset is plotted in Figure 1.7 with three horizontal blue lines at levels 0.05, 0.5, and 0.95. These lines cut the ECDF at the 0.05, 0.5, and 0.95 sample quantiles, which are

$$\hat{Q}(0.05) \approx 2.57, \quad \hat{Q}(0.5) \approx 7.50, \quad \hat{Q}(0.95) \approx 14.53.$$

All the results for this example are produced by the code given in Section 1.8.5. Here a reasonable description might be: based on a sample of births at an Oxford hospital, roughly 90% of non-C-section labors last between 2.5 and 14.5 hours, with a typical duration being 7.5 hours.

However, it is unclear how confident we should be about these numbers, how accurately the durations were measured, and how *generalizable* these results should be to future births in the same hospital or births at a different hospital. What results would we get if we used a different sample of births from the hospital? Our dataset is pretty small, and based on only one hospital in one narrow time frame.

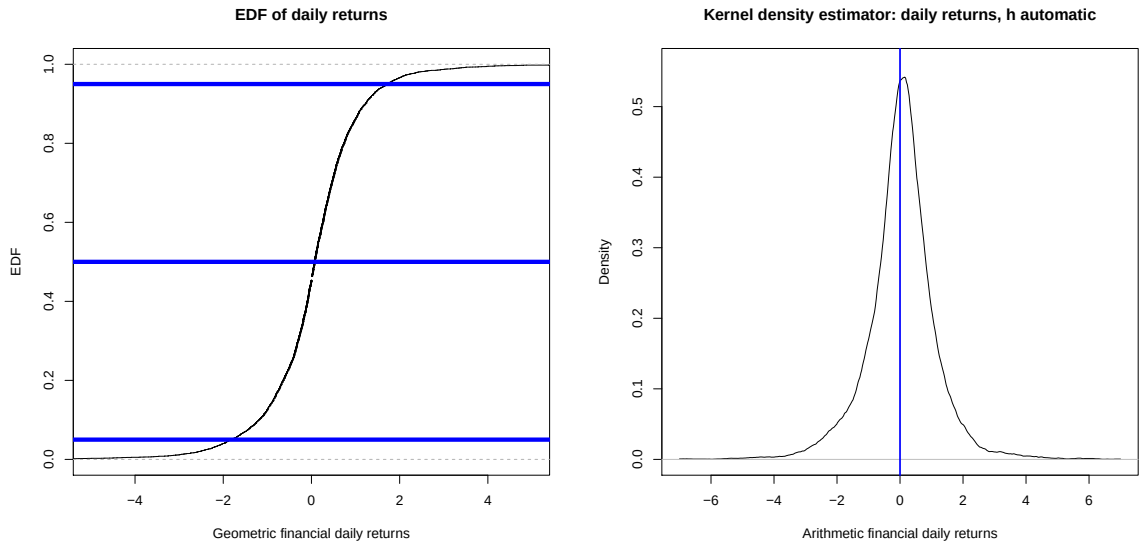


Figure 1.8: LHS: daily arithmetic returns for SPDR from 1 Feb 1993 to 31 Oct 2017. Plotted over the range of returns from -5 to 5; some data points are outside that interval. Blue lines pick out the 0.05, 0.5, and 0.95 quantiles from the ECDF. RHS: kernel density estimate for SPDR daily arithmetic returns, plotted over the interval $[-7, 7]$ with bandwidth $h = 0.21$.

Example 1.4.8 (S&P 500)**.** The Standard & Poor's depository receipt S&P 500 exchange traded fund (denoted SPDR) is an investment fund which approximates the value of a portfolio made from the 500 largest US listed (i.e., traded rather than privately held) companies. The index is weighted by market value, so companies with a high market value, such as Apple, have large weights. Most US personal pensions and university endowments have significant investments linked to the S&P 500. What numbers do we tend to see? Here we give a broad-brush *descriptive* answer.

Let P_j denote the price of the SPDR, on business day j, for $j = 1, 2, \ldots, n+1$. Our data start on the 1st of February 1993 and run until the 31st of October 2017. In total, $n + 1 = 6236$.

From this index we define daily scaled *arithmetic returns* as

$$y_j = 100 \left(\frac{P_{j+1} - P_j}{P_j} \right), \quad j = 1, 2, \ldots, n,$$

which are the percentage daily changes. We will now describe the behavior of the data $\mathbf{y} = (y_1, \ldots, y_n)$. For simplicity, we will ignore the serial dependence in the arithmetic returns.

The ECDF of the returns is given in the left-hand side of Figure 1.8. Again the blue lines cut off the 0.05, 0.5, and 0.95 quantiles. The sample size here is much larger than in the birth duration example, so the ECDF looks less jagged. Some descriptive statistics for these daily returns are given below (as before, $\overline{y}$ is the sample mean and s is the sample standard deviation):

$$\overline{y} \approx 0.042, \quad s \approx 1.16, \quad \hat{Q}(0.05) \approx -1.81, \quad \hat{Q}(0.5) \approx 0.067, \quad \hat{Q}(0.95) \approx 1.72.$$

Here the sample mean $\overline{y}$ is lower than the sample median $\hat{Q}(0.5)$, while the ranges around the median are $\hat{Q}(0.95) - \hat{Q}(0.5) \approx 1.66$ and $\hat{Q}(0.5) - \hat{Q}(0.05) \approx 1.87$. This hints that returns are somewhat skewed, with a longer left-hand tail.

A simple descriptive summary of the data is as follows: Around 90% of daily scaled arithmetic (percentage) returns are between -1.81 and 1.72. The fund rises on most days, with a typical daily return being 0.07. There are some really aggressive daily falls, but also some very large rises.

For this dataset the ECDF is not very informative graphically, so in the right-hand side of Figure 1.8 we plot a variant of the histogram, known as a *kernel density estimate (KDE)* and given by

$$\tilde{f}(y) = \frac{1}{nh} \sum_{j=1}^{n} I\left(y_j \in \left(y - \frac{h}{2}, y + \frac{h}{2}\right]\right).$$

This counts the number of data points in the interval $(y - h/2, y + h/2]$, which is centered at y, and divides the answer by the interval's length $h > 0$ times the sample size n. In statistics, objects like h are often called *bandwidths*. The choice of h needs to be determined by the data.

For the SPDR arithmetic return data, $\tilde{f}(y)$ is drawn in the right-hand side of Figure 1.8 using $h \approx 0.21$ (this choice of h is called the *Silverman rule of thumb*). The blue vertical line is placed at a zero return. It suggests the mode in the kernel density estimate is slightly above 0.

1.5 Predicting Y from X

An old adage, sometimes attributed to the physicist Niels Bohr, is that

Prediction is hard, especially about the future.

☣ **1.5.1.** Statisticians and computer scientists use the word "predicting" more broadly than this colloquial phrasing indicates (indeed statisticians are even out of keeping with the meaning of the Medieval Latin "predictionem"). Think of it is as: having seen some data X, what are the likely values of Y? For example, let X be the indicator of Greg having watched the movie *Barbie* (i.e., $X = 1$ if Greg has watched this movie, and $X = 0$ otherwise), and Y be the indicator of Greg having watched the movie *Oppenheimer*. Suppose that we know that Greg has seen *Barbie*. Then a prediction is the probability that he has seen *Oppenheimer*, given the information that he has seen *Barbie*.

It doesn't matter which movie came out first or, if Greg saw both movies, the order in which he saw them; we're just using known information to predict something we do not know. An important special case of prediction is where X is an aspect of the past and Y is an aspect of the future. Statisticians call this case *forecasting*. For example, given the values of a stock price for the past n days, we could try to forecast what the stock price will be tomorrow.

Statistics and machine learning provide many useful techniques for prediction. One notable class of methods is known as *regression models*. In a regression model, we have *predictor* variables $X_1, \ldots, X_k$ and an *outcome* variable Y, and we try to use the predictor variables to predict the outcome variable. Think of $X_1, \ldots, X_k, Y$ being the variables for one individual, where we have observed that individual's values of $X_1, \ldots, X_k$ and want to predict that individual's value of Y.

Regression is closely related to the notion of *conditional expectation* (see Chapter 9 of the Stat 110 book): we can think of

$$\mathrm{E}[Y|X_1, \ldots, X_k]$$

as the best prediction of Y as a function of $X_1, \ldots, X_k$. Of course, to make this statement precise we first need to define "best", which again returns to the question of what is good.

Example 1.5.2. Let X be how much money is spent on advertising for a movie and Y be how much money the movie brings in at the box office. Consider the following predictive model:

$$\mathrm{E}[Y|X; \beta_0, \beta_{Y\sim X}] = \beta_0 + \beta_{Y\sim X} X. \tag{1.1}$$

The coefficients β_0 and $\beta_{Y\sim X}$ are the parameters of the model. The subscript $Y \sim X$ means that X is being used to predict Y (so the meaning of $\sim$ in this notation does *not* mean "is distributed as"). This model is known as a *linear regression model*. Here "linear" refers to the fact that the model is linear in the parameters.

In the calculations below we will suppress $\beta_0, \beta_{Y\sim X}$ from the notation to remove clutter. So the model is

$$\mathrm{E}[Y|X] = \beta_0 + \beta_{Y\sim X} X.$$

By Adam's Law (also known as the Law of Total Expectation or iterated expectations),

$$\mathrm{E}[Y] = \mathrm{E}[\mathrm{E}(Y|X)] = \beta_0 + \beta_{Y\sim X}\mathrm{E}[X].$$

We will explore linear regression models much more in Chapter 6, but we can already note a few things.

- We can estimate $\theta = (\beta_0, \beta_{Y\sim X})$ from the data if we observe pairs $(x_1, y_1), \ldots, (x_n, y_n)$; we will study methods for doing so later. Once we have estimates $\hat{\beta}_0$ and $\hat{\beta}_{Y\sim X}$ for β_0 and $\beta_{Y\sim X}$, we have a very natural way to predict the Y value y_{new} for a new movie that spent x_{new} dollars on advertising:

$$\hat{y}_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_{Y\sim X}\, x_{\text{new}}.$$

We would prefer to use the theoretical prediction

$$\mathrm{E}[Y|X = x_{\text{new}}] = \beta_0 + \beta_{Y\sim X} x_{\text{new}},$$

but since β_0 and $\beta_{Y\sim X}$ are unknown, we replace them by their estimated values.

- The linear predictive regression model (1.1) implies, again by Adam's Law, that

$$\begin{aligned}\mathrm{Cov}(Y, X) &= \mathrm{E}[\mathrm{E}(XY|X)] - \mathrm{E}[X]\mathrm{E}[Y] = \mathrm{E}[X\mathrm{E}(Y|X)] - \mathrm{E}[X]\{\beta_0 + \beta_{Y\sim X}\mathrm{E}[X]\} \\ &= \beta_0\mathrm{E}[X] + \beta_{Y\sim X}\mathrm{E}[X^2] - \beta_0\mathrm{E}[X] - \beta_{Y\sim X}\mathrm{E}[X]^2 \\ &= \beta_{Y\sim X}\mathrm{Var}(X),\end{aligned}$$

so

$$\beta_{Y\sim X} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}.$$

Later in the course, if we assume our data comes from a statistical model with i.i.d. pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$, then we will see why the most widely used estimator for $\beta_{Y\sim X}$ is

$$\hat{\beta}_{Y\sim X} = \frac{\sum_{j=1}^n (X_j - \bar{X})(Y_j - \bar{Y})}{\sum_{j=1}^n (X_j - \bar{X})^2} = \frac{S_{X,Y}}{S_X^2}.$$

This formula may look complicated at first, but it is just the *sample* version of the above result for $\beta_{Y\sim X}$: note that $\beta_{Y\sim X}$ is the theoretical covariance of X and Y over the theoretical variance of X, while $\hat{\beta}_{Y\sim X}$ is the sample covariance of X and Y over the sample variance of X (since the factors of $n - 1$ cancel).

- Even if we get a large, positive estimate for $\beta_{Y\sim X}$, this model would *not* allow us to conclude that increasing our advertising expenditure would *cause* the movie to make more money at the box office.

1.6 Causal impact on Y of manipulating X

Prediction is a somewhat passive activity: Jose observed X; now what Y is Jose likely to observe? Statistical ideas of causality are more active. Jane *intervenes*, moving X from x to $\tilde{x}$; now what is the likely move in Y in response? It is crucial to distinguish between prediction and causation!

A famous saying is that *correlation does not imply causation*. See Figure 1.9 for a webcomic from Randall Munroe about this saying.

Example 1.6.1. Tech companies that rely on advertising, such as Google, would like to know if updating a website from design A to design B, will get a higher click-through rate for advertisers. This is a causal question.

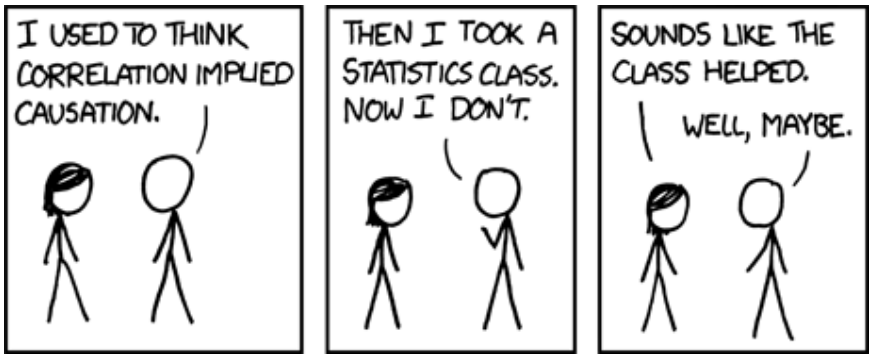


Figure 1.9: This comic by Randall Munroe, entitled "Correlation", is from `xkcd.com`, produced here under "Creative Commons Attribution - NonCommercial 2.5 License", detailed at `https://xkcd.com/license.html`.

Achieving valid *causal* inferences is important, challenging, and sometimes controversial. Causal inference plays a large role in modern social and medical science as well as in technology and marketing. Technology companies often investigate causal inference using *randomized experiments*, which we will discuss in this course. The most popular kind of experiment is called *A/B testing* in data science, though such experiments have been considered in statistics for over a century as an example of randomized experiments.

To get a better understanding about why prediction is fundamentally different from causality, let us establish some notation for thinking about causal effects. We will do so in the context of a new medical procedure.

For simplicity, suppose $X \in \{0, 1\}$ takes the value 1 if the patient has undergone the new medical procedure and 0 if they have not. In causal studies, X is called the *assignment*. The event $X = 1$ is called receiving the *treatment* and the event $X = 0$ is called receiving the *control*. We will write the outcome Y as 1 if the person's care is a success and 0 if it is not. (For simplicity, we assume that the outcome is binary, but the ideas extend naturally beyond this case.)

Example 1.6.2. In pharmaceutical drug trials, the treatment is usually a new drug regime and the control is the current best standard of care. The outcome is usually a particular measurement of the patient's health, e.g., based on HIV RNA viral load, blood pressure, or blood sugar level.

To connect the assignment and outcome to causal effects, write $Y(1)$ as the medical outcome which would happen if the person was treated and $Y(0)$ as the outcome if the *same* person had been in control. This notation will be more extensively explained in Chapter 11.

Definition 1.6.3 (Potential outcomes, treatment effect)**.** The pair

$$\{Y(0), Y(1)\}$$

are called the *potential outcomes*. The word "potential" is used because prior to running the study these outcomes are only potential. The random variable

$$\tau = Y(1) - Y(0)$$

is the *treatment effect* (or *causal effect*) on the outcome of moving a person from control to treatment. Across a population, moving everyone from control to treatment, the

$$\mathrm{E}[\tau] = \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)].$$

is the population's *average treatment effect*.

A major challenge is that even after the study is over we never see both $Y(0)$ and $Y(1)$: we only see one of them as the individual is either under treatment or control, not both. So we cannot compute τ directly, which makes it tricky to estimate $\mathrm{E}[\tau]$.

What data do we have to infer $\mathrm{E}[\tau]$? We have the assignment X and an outcome, which is the potential outcome Y corresponding to whether the person was treated or not treated. Mathematically, the outcome is written as

$$\begin{aligned} Y &= \begin{cases} Y(1), & \text{if } X = 1 \\ Y(0), & \text{if } X = 0 \end{cases} \\ &= Y(X) \\ &= XY(1) + (1 - X)Y(0). \end{aligned}$$

The potential outcome we do not see, which is $Y(1 - X)$, is called a *counterfactual*. It is impossible to observe a counterfactual (unless we could somehow go into a parallel world where the individual received no treatment if they did receive treatment in our world, and vice versa).

☣ **1.6.4.** The notation $Y(X)$ is shorthand for the r.v. that is $Y(1)$ if $X = 1$ and $Y(0)$ if $X = 0$; it does not denote Y times X. Similarly, $Y(1 - X)$ is $Y(0)$ if $X = 1$ and $Y(1)$ if $X = 0$. So $Y(X)$ is the observed outcome and $Y(1 - X)$ is the counterfactual outcome.

☣ **1.6.5.** In trying to learn $\mathrm{E}[\tau] = \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)]$, it may be tempting to use the power of prediction, using

$$\mathrm{E}[Y|X = 1] - \mathrm{E}[Y|X = 0],$$

since this is the difference between the average outcome if we see $X = 1$ and the average outcome if we see $X = 0$. However, this is *not* valid without strong further assumptions.

Intuitively, $\mathrm{E}[Y|X=1]-\mathrm{E}[Y|X=0]$ is the difference in means between the treated group and the control group, whereas $\mathrm{E}[Y(1)]-\mathrm{E}[Y(0)]$ is the difference in means between a hypothetical scenario where *everyone* gets the treatment and one where *everyone* gets the control. These differences are very different conceptually. To see in more detail what is going on mathematically, suppose again that X is a binary random variable. Then

$$\mathrm{E}[YX] = \mathrm{E}[YX|X=1]P(X=1) + \mathrm{E}[YX|X=0]P(X=0) = \mathrm{E}[Y|X=1]P(X=1),$$

so

$$\mathrm{E}[Y|X=1] = \frac{\mathrm{E}[YX]}{P(X=1)} = \frac{\mathrm{E}[YX]}{\mathrm{E}[X]} = \frac{\mathrm{E}[Y(1)X]}{\mathrm{E}[X]},$$

since X being Bernoulli implies that

$$P(X=1) = \mathrm{E}[X] \text{ and } YX = Y(1)X.$$

Hence,

$$\mathrm{E}[Y|X=1] = \frac{\mathrm{E}[Y(1)]\mathrm{E}[X]}{\mathrm{E}[X]} + \frac{\mathrm{Cov}(Y(1),X)}{\mathrm{E}[X]} = \mathrm{E}[Y(1)] + \frac{\mathrm{Cov}(Y(1),X)}{\mathrm{E}[X]}.$$

Similarly,

$$\mathrm{E}[Y|X=0] = \mathrm{E}[Y(0)] - \frac{\mathrm{Cov}(Y(0),X)}{1-\mathrm{E}[X]}.$$

Thus, the prediction quantity $\mathrm{E}[Y|X=1]-\mathrm{E}[Y|X=0]$ is not the same as the causal quantity $\mathrm{E}[\tau]$ in general, unless we have a strong additional assumption such as

$$\mathrm{Cov}(Y(1),X) = 0,\ \mathrm{Cov}(Y(0),X) = 0. \tag{1.2}$$

How can the covariances in (1.2) be forced to equal 0? They will be made to be 0 if the assignment is independent of the potential outcomes. This can be ensured by drawing the assignment using *random* numbers on the computer, which makes it independent of everything else, including the potential outcomes. Chapter 11 will go into further depth about randomized experiments and causal inference.

1.7 Recap

Statistics focuses on exploration and description, prediction, and causality. More abstractly, exploring and describing are about informally investigating a phenomenon or variable Y. Prediction is about pinning down Y given knowledge of X. Causality is about pinning down how Y will change if we change X.

Formula or idea	Description or name
description, prediction, causality	**Goals of statistics**
conditioning, modeling loss function, likelihood, Bayes, MoM confidence interval, testing sampling, randomization, resampling	**Statistical inference core ideas**
statistical theory, simulation, real data	**Three-pronged approach**
	Description:
$\bar{y}$, s	sample mean and sample standard deviation
$s_{x,y}$, $r_{x,y}$, $b_{y\sim x}$	sample covariance, correlation, and linear regression
$y_{(\lceil np \rceil)}$, $\hat{F}(y)$	sample p-quantile, empirical CDF
$\mu(x) = \mathrm{E}[Y\|X = x]$	**Prediction**:
$\mu(x) = \theta_0 + \theta_1 x$	linear regression model
	Causality:
$Y(0), Y(1)$	potential outcomes
$\tau = Y(1) - Y(0)$, $\mathrm{E}[\tau]$	treatment effect and average treatment effect
X	assignment
$Y = Y(X)$	outcome
$Y(1 - X)$	counterfactual

Table 1.1: Main ideas and notation in Chapter 1.

This book focuses on statistical inference. The main topics in statistical inference are statistical model building, statistical learning, conditioning, randomizing, and deciding. The main concepts covered in this chapter are listed in Table 1.1.

This book will approach statistical science via mathematical theory, simulation, and real data. The theory will often be based on Stat 110 techniques and results, with the help of some calculus (especially Taylor approximations) and, in the starred (optional) sections, linear algebra.

Simulation from models is a central tool in modern statistics, which we will explore time after time. But simulation plays many other roles in statistics, as we will see when we study bootstrapping, Bayesian inference, and randomized experiments.

Finally, the analysis of data brings statistics alive, demonstrating its importance in the world we

live in. Data is complicated and its context is crucial. Building great models for particular problems is not easy. It is not pure math, nor is it poetry. But improving at it (it is never possible to be perfect in statistics) is possible through practice and is eternally interesting. We conclude with a quotation from the statistician John W. Tukey:

The best thing about being a statistician is that you get to play in everyone's backyard.

1.8 A short introduction to R

This section provides a short introduction to R, introducing some of the core features which are used in statistical inference. R is an unusual language, created by statisticians for doing statistics and data science. A convenient interface for using R, RStudio Desktop, provides a code editor, a data viewer, and various useful features. R and RStudio can be downloaded for free via

`https://rstudio.com/products/rstudio/download/`

Good practice is to write your code in a "source file", and run the file through R to get output, rather than typing your commands directly into the console (which makes it easy to forget steps you have taken, and hard to reproduce your work and reuse your code). In RStudio there are several useful formats for your source file: an R Script, an R Markdown document, or a Quarto document. The R Script option is simple: just a handy place to store R code. The R Markdown and Quarto options are fancier, allowing you to weave together text, images, math, and code.

In this book, we teach the use of R by including a section towards the end of each chapter about some aspect of R, where we introduce useful functions, show how to compute various familiar quantities, and show code for various examples. There are also sections about R at the end of each chapter of the Stat 110 book, and there are many free books and tutorials on R available online. As mentioned in this chapter, we emphasize *simulation* and *real data* in addition to theory, so we also discuss in the R sections how to run simulations and analyze data to gain insights into the statistical ideas in this book.

We will jump right in with an example: let's simulate and plot some data.

Example 1.8.1 (Predictive regression)**.** Suppose that we want to sample n i.i.d. random variables $X_1, \ldots, X_n$ from t_4 (the Student t-distribution with 4 degrees of freedom; see Chapter 10 of the Stat 110 book). We call the X_j's *predictors*. Let $n = 400$. Place the results all together as a vector $\mathbf{x} = (x_1, \ldots, x_n)$. Then we will sample *outcomes*

$$Y_j | X_j = x_j \overset{\text{indep.}}{\sim} \mathcal{N}(0.2 + 0.9x_j, 0.3^2), \quad j = 1, \ldots, n.$$

Put the outcomes together as the vector $\mathbf{y} = (y_1, \ldots, y_n)$. The predictors and outcomes together make up a *predictive regression*, which will be a main focus of Chapter 6.

The code for this is

Code for Example 1.8.1

```
set.seed(111)  # sets the seed of the random number generator
n = 400 # sample size
x = rt(n,4) # draw n values from t with 4 degrees of freedom
y = rnorm(n,0.2+0.9*x,0.3) # get draws from N(0.2+0.9x_j,0.3^2)
```

We will now explain the above code.

- `R` ignores anything on a line after a #, allowing us to add comments to the code.
- `set.seed(111)` sets the *seed* (or initialization) of the random number generator to the number 111. This means `R` generates the same random numbers each time the code is run. This is terrible if we want random results each time, but fantastic if we want to make it easy to replicate our results. Replication is important in statistics!
- `n = 400` sets the sample size n equal to 400. It would have been equivalent to write `n <- 400`. The arrow `<-` is suggestive of the fact that we are giving the value 400 to n.
- In `R`, for many named distributions (e.g., Normal, Exponential, Beta, Poisson, Geometric) there are 4 key functions for working with the distribution, with prefixes d, p, q, r in front of an abbreviated name for the distribution. For example, the commands `dnorm`, `pnorm`, and `qnorm` provide the Normal density, CDF, and quantile function, respectively, and `rnorm` provides random generation from a Normal (which is extremely useful for simulation purposes). Then `dnorm(1,0.2,0.3)` is the density function at 1 for a Normal with mean 0.2 and standard deviation 0.3, while `qnorm(0.975, 0.2, 0.3)` is the 0.975-quantile of the same Normal r.v.

  You can type `help(distributions)` for a list of named distributions that are built in to `R`, and `help(dnorm)` for some documentation and examples of `dnorm` (and, it turns out, `pnorm`, `qnorm`, and `rnorm`).

  In `rnorm(n,0.2+0.9*x,0.3)`, the number of random numbers to generate, n, is the first argument. The means are the second argument. Interestingly, `rnorm`) allows the mean of the Normal to vary over the n draws: `0.9*x` has individual means $0.9x_j$. We could also allow the variances to vary, but by putting 0.3 as the third argument, we are setting all of the standard deviations equal to 0.3.

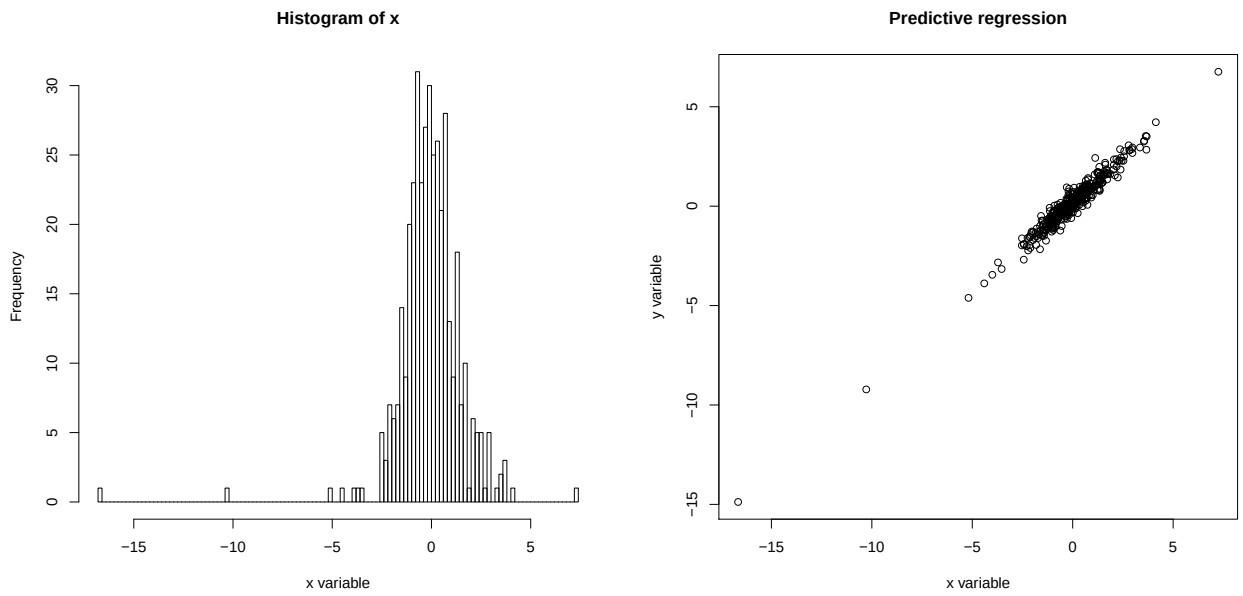


Figure 1.10: Left hand side: histogram of `x`. Right hand side: cross plot of `y` against `x` from a predictive regression.

☣ **1.8.2.** Note that R parameterizes the Normal using mean and SD rather than mean and variance. A very common mistake is to mix up SD and variance, either in a mathematical calculation or in R code!

- `rt(n,4)` generates n random draws from the t distribution with 4 degrees of freedom.

The cross plot of the elements of `x` and `y` and a histogram (which counts the number of data points in bins — more on this later) are produced by

Cross-plot and histogram

```
plot(x,y,xlab="x variable",ylab="y variable",
     main="Predictive regression") # plot Y against X
hist(x,breaks=120,xlab="x variable") # histogram of x
```

The cross plot of `y` drawn against `x` is shown on the right hand side of Figure 1.10 and was set up using generic axis labeling. The left-hand side of Figure 1.10 shows a histogram of `x` (with 120 bins, plotting the number of occurances in the data in each bin). You can save your graphs manually in RStudio, or in the code by surrounding your plot function by instructions about how to save the plot:

Saving plot to PDF file

```
pdf("yourfile.pdf") # opens a file with that name
  plot(x,y,xlab="x var",ylab="y var",main="Main")  # do the plotting
dev.off() # close the file
```

The plots can be tailored to your preferences in many ways, with many choices to be made over colors, plotting symbols, adding text, and additional lines. Whole courses can be taken on `R` and graphics! A neat alternative package for creating graphics is `ggplot2`, which is part of the `tidyverse` constellation of packages. We will briefly introduce `ggplot2` in a later chapter, but for our purposes in this book the built-in plotting features in `R` suffice.

It is important to know where files are being saved on your system. `R` has a simple way of identifying the relevant directory and a more complicated version with finer control. First the simple one:

Find current working directory

```
getwd() # find working directory: outputs get saved to this directory
```

This command tells you path to the current working directory. To change your working directory to a different path, you can use a command like

Set working directory

```
setwd("C:/Users/Neil/Rfiles") # sets working directory to given path
```

in Windows or like

Set working directory

```
setwd("~/Rfiles") # sets working directory to given path
# the ~ denotes your home folder
```

on a Mac.

Subsection 1.8.1 uses some simulated data to illustrate how descriptive statistics are computed. Sometimes it is helpful to bind data points or results together, storing the results in vectors or matrices. This is outlined in Section 1.8.2.

`R` has its own quirks, which can lead to common mistakes. Some of these are described in Subsection 1.8.3. The final subsection collects the code which produced the output used in various examples in this chapter.

1.8.1 Descriptive statistics

Returning to Example 1.8.1, let's compute some summary statistics. Summary statistics of the `y` data can be produced using:

Core descriptive statistics

```
c1 = sum(y) # sum up elements of y
c2 = mean(y) # sample average of elements of y
c3 = cov(x,y)/var(y) # sample cov divided by sample var of y.
c4 = sd(y) # sample standard deviation of elements of y
```

```
c5 = median(y) # sample median of elements of y
c6 = quantile(y,probs=c(0.1,0.5,0.9)) # 0.1, 0.5, 0.9 quantiles of y
```

The details of each of these summaries was given early in the chapter. Most of the R commands for these summaries are named intuitively, e.g., `sum` computes the sum of a vector, `mean` computes the sample mean of a dataset, and `sd` computes the standard deviation of a dataset.

A less intuitive but vitally important command is `c`, the *concatenate* function. This function puts a list of values together into a vector. For example, consider the code

Combining results

```
> c(0.1,0.5,0.9)
[1] 0.1 0.5 0.9
```

Here we used `c(0.1,0.5,0.9)` to create the vector $(0.1, 0.5, 0.9)$, which we needed for specifying which quantiles we wanted `R` to compute. If we want to create a vector of consecutive integers, such as $(3, 4, 5, 7)$, then it is simpler to use `3:7` rather than `c(3,4,5,6,7)`.

1.8.2 Vectors and matrices

We have written the code slightly differently now. In interactive mode in R, the symbol `>` shows R is waiting for new input. We have then typed `c(0.1,0.5,0.9)` and the output is `[1] 0.1 0.5 0.9`. The output is to the screen as we have not directed `c(0.1,0.5,0.9)` to anywhere. Sometimes, when we want you to read the output as we saw it in `R`, we will use the symbol `>`, other times it will be suppressed to make the code easier to read and easier to copy/paste into `R`.

Thinking in terms of *vectors* is crucial for using `R` effectively. The dimension of a vector `y` is found by the command `length(y)`. Operations on vectors tend to be performed *componentwise*. For example, in math if $\mathbf{y}$ is a vector then it's not entirely clear what the notation $\mathbf{y}^3$ means, but in `R` if `y` is a vector then `y^3` cubes each component of `y` separately. If we want to cube the entries of `y` then it is much better simply to use `y^3` than to write a for loop that iterates over the entries of `y`.

As another example, if for some reason we wanted to calculate the sum $\sum_{j=1}^{n}(1/(y_j^2+3))$ for a vector $\mathbf{y} = (y_1, \ldots, y_n)$, we could enter the vector as `y` in R and then use the very compact code `sum(1/(y^2+3))`. In math, adding a vector (of length greater than 1) to a scalar (a number) is undefined. In `R`, the `y^2+3` is an easy way to add 3 to each entry of the vector `y^2`.

Once we have a vector `y`, we can access the jth component using `y[j]`. Furthermore, if we want, say, the 7th through 12th components of `y`, we can use `y[7:12]`. If we want, say, entries 4, 6, 8 of `y`, we can use `y[c(4,6,8)]`. Also, in math if $\mathbf{y}$ is a vector then there isn't a standard, convenient notation for the vector obtained from $\mathbf{y}$ by removing the ith entry; we would have to write something

like $(y_1, y_2, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)$. In R though we can use `y[-i]` as a compact way to get the vector with the ith entry dropped. R "knows" that the - sign here is saying to exclude `i`, rather than referring to a negative number.

Multiple vectors can be combined into a matrix. We can build a matrix by combining column vectors through `cbind()` or by combining row vectors via `rbind()`.

Building vectors and matrices

```
> d1 = c(0.1,0.5,0.9)
> d2 = c(-2,-3,4)
> cbind(d1,d2)
     d1  d2
[1,] 0.1 -2
[2,] 0.5 -3
[3,] 0.9  4
> rbind(d1,d2)
   [,1] [,2] [,3]
d1  0.1  0.5  0.9
d2 -2.0 -3.0  4.0
```

Notice that the matrix in `cbind(d1,d2)` has automatically picked up column names from the names of the vectors `d1` and `d2`. Likewise, with `rbind(d1,d2)` the row names are automatically created. The dimensions of a matrix `A` in R is found using the command `dim(A)`. We access the individual entries of a matrix by indicating the row number and column number, such as using `A[3,1]` to get the entry in row 3, column 1 of matrix `A`.

Accessing features of vectors and matrices

```
> d1 = c(0.1,0.5,0.9); d2 = c(-2,-3,4)
> X = cbind(d1,d2)
> dim(X) # number of rows and columns in X
[1] 3 2
> X[1,2]  # 1,2-th element
d2
-2
> d2[2] # 2nd element of d2
[1] -3
```

In the above code, the semicolon in the first line lets us put two commands together in one line. Some programming languages require a semicolon at the *end* of each line; in R that is optional.

A beautiful aspect of R is that it is so easy to take *slices* of a matrix. For example, `X[,2]` produces the 2nd column of `X`, while `X[1:2,]` extracts the first two rows of `X`.

Accessing groups of elements of vectors and matrices

```
> X[,2]
```

```
[1] -2 -3  4
> X[1:2,]
       d1  d2
[1,] 0.1 -2
[2,] 0.5 -3
```

1.8.3 Looping

R has a simple looping setup. Here we estimate the standard deviation of the sample average of n data points which are t_4 simulations. The simulation is replicated, that is repeated, $B = 2500$ times.

When loops are used in statistics an output is usually stored at each step in the loop. Storage is usually a vector or a matrix. A vector of size B is set up as `rep(0,B)`, this is a vector of length B filled with 0s. If you want to declare a $B \times 3$ matrix you can use the command `matrix(0,B,3)`. The dimensions of a matrix `Q` are found by the command `dim(Q)`. As mentioned earlier, the dimension of a vector `b` is accessed through `length(b)`.

Looping over a calculation, storing results

```
B=2500 # preparing for the loop
xRes = rep(0,B) # B-dim vector, filled with 0
for (b in (1:B)){ # loop indexed by b, loop B times
   xRes[b] = mean(rt(n,4)) # store mean of n, t variables
}

print(sd(xRes)) # print results
```

A cumbersome aspect of R is printing results. `print(sd(xRes))` is simple enough. But if we wanted both the mean and the standard deviation, the command would be

`print(c(mean(xRes),sd(xRes)))`, i.e., the coder has to combine results before printing.

An interesting function used above is `rep(0,B)`, which generates a B dimensional vector of 0s. `rep` is a cousin of the function `replicate`. This replicates a quantity, such as a vector or a matrix. So

Repeating the same calculation many times

```
>replicate(5,rep(2,4))
     [,1] [,2] [,3] [,4] [,5]
[1,]    2    2    2    2    2
[2,]    2    2    2    2    2
[3,]    2    2    2    2    2
[4,]    2    2    2    2    2
```

5 replications of the `rep(2,4)`, where `rep(2,4)` is a 4 dimensional vector of 2s. Replication is a common action in statistics, e.g.,

More repeating the same calculation

```
>X = replicate(5,rnorm(6,2.3,1.2));
>X;
          [,1]      [,2]       [,3]      [,4]      [,5]
[1,] 2.441597 4.059272 2.05213814 2.229331 3.322457
[2,] 2.023846 1.557728 2.79631899 1.953087 3.377449
[3,] 3.283864 3.207129 0.90955891 2.528354 2.046194
[4,] 1.877304 3.146190 2.91562934 2.645579 1.307190
[5,] 2.402246 3.212862 0.09907067 2.864678 1.971153
[6,] 2.066790 2.733752 2.48633237 1.818538 3.312970
```

produces 5 samples each of size 6 from a normal distribution with mean 2.3 and standard deviation of 1.2.

Common early mistakes using R

In R you should try to think in terms of vectors and matrices whenever possible, rather than running a lot of loops. In many examples, R code will run much faster *and* be more compact and readable by operating directly on vectors and matrices rather than by using loops.

To illustrate, suppose y is a vector and we want to calculate the sum. Then compare `sum(y)`, with the looped version

```
n = length(y);
d1 = 0;
for (j in (1:n)){
  d1 = d1+ y[j];
}
```

Writing all this code instead of `sum(y)` makes the code harder to read and slower to run!

Another easy to make mistake is to write `print(a,b)`; hoping to get `a,b` printed out. As stated above you need to connect them: `print(c(a,b))`; to get the expected result.

If A and B are matrices, then R views the code `A*B` as an element-by-element multiplication, not a matrix multiplication. This element-by-element multiplication is convenient for some purposes, but it is easy to forget that this will be the result. You can produce matrix multiplication using the somewhat ugly notation `A%*%B` (if the product is defined).

With distributions, be sure to check carefully how it is parameterized in R. For the most part, the conventions in R are consistent with how the distributions are parameterized in the Stat 110 book. For example, there are two widely-used conventions on how to define the Geometric distribution, one where the support starts at 0 and one where the support starts at 1. In both the Stat 110 book and in R, the support of the geometric starts at 0 (so the Geometric counts the number of failures before the

first success). But note that, as mentioned earlier, R parameterizes the Normal in terms of mean and standard deviation, whereas it is traditional in statistics to use mean and variance as the parameters of the Normal.

1.8.4 R Markdown

Markdown is a formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com>.

The use of Markdown scripts has become popular in research for it allows researchers to fully document their work, integrating text, math, R code, and output. It is comparable to a full lab notebook for a laboratory scientist. This makes the research more *reproducible* — that is, another researcher should be able to perfectly reproduce, that is check, all of your work using your Markdown file plus any data you used.

A listing of a file called intro.rmd of R Markdown is given below (it is a very slightly edited version of the default file generated by RStudio), which, for simplicity, uses the cars and pressure datasets which are default datasets in R (i.e., you do not need to load them into R; they are already there).

Example of R Markdown document: intro.rmd

````
---
title: "Something Written in R Markdown"
author: "Joe Blitzstein and Neil Shephard"
date: "2023-01-23"
output: pdf_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax
for authoring HTML, PDF, and Word documents. For more details on using
R Markdown  see <http://rmarkdown.rstudio.com>.
R Markdown is a useful tool for writing up statistical work.

When you click the **Knit** button a document will be generated that
includes both  content as well as the output of any embedded R code
chunks within the document.  You can embed an R code chunk like this:

```{r cars}
summary(cars)
```
````

```
You can include math easily if you know some basic LaTeX syntax.
For example, one of my favorite results in probability, Adam's law,
states that
$$
{\mathrm E}[Y] = {\mathrm E}[{\mathrm E}[Y|X]]
$$
for any random variables $X$ and $Y$.

## Including Plots

You can also embed plots (here I made the plot small by setting
${\tt out.width}$  at 50 percent). For example:
```{r pressure, echo=FALSE,out.width="50%"}
plot(pressure,type="l")
```

The `echo = FALSE` parameter was added to the code chunk to prevent
printing of the R code that generated the plot. If we want the code
to be printed too, we can change that to
`echo = TRUE`.
```

Here we asked for a PDF output, so it produces `intro.pdf`. This PDF was directly pasted into our text below.

Something Written in R Markdown

Joe Blitzstein and Neil Shephard

2023-01-23

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com. R Markdown is a useful tool for writing up statistical work.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

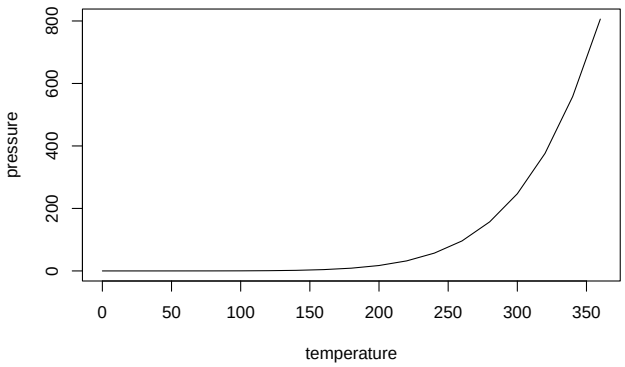
You can include math easily if you know some basic LaTeX syntax. For example, one of my favorite results in probability, Adam's law, states that

$$\mathrm{E}[Y] = \mathrm{E}[\mathrm{E}[Y|X]]$$

for any random variables X and Y.

Including Plots

You can also embed plots (here I made the plot small by setting `out.width` at 50 percent). For example:



The `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot. If we want the code to be printed too, we can change that to `echo = TRUE`.

1.8.5 Code for various examples in this chapter

Here we list the code which produces all the numerical results in this chapter.

Example 1.4.7 focused on the duration of births data.

Code for births Example 1.4.7

```
load("births.rda") # read data
write.csv(births, file = "births.csv") # save as csv file

mean(births$time); sd(births$time)  # sample average and sample sd
quantile(births$time,probs=c(0.05,0.5,0.95))  # sample quantiles
Fn <- ecdf(births$time) # compute ECDF
mean(births$time <= 5.0)  # ECDF at y=5
mean(births$time <= 10.0) # ECDF at y=10
mean(births$time <= 15.0) # ECDF at y=15

## plots comment out saving to PDF files, so appear on screen

#pdf("birthsHist.pdf")  # not used, histogram
  hist(births$time,xlim=c(0,20),breaks=40,xlab="Hours",freq=FALSE,
       main="Histogram of duration of births in hours")
#dev.off()


#pdf("births.pdf") # plot ECDF
  plot(Fn,ylab="EDF",xlab="Birth length, in hours",
    main="EDF of birth lengths")
  abline(h=c(0.05,0.5,0.95),col="blue",lwd=5)
#dev.off()
```

Example 1.4.8 focused on the S&P 500 daily returns. The data used in this example were originally downloaded using the `Quantmod R` package from `Yahoo Finance`.

Code for S&P 500 Example 1.4.8

```
load("sp500.rda")
write.csv(xRet, file = "sp500.csv")

mean(xRet); sd(xRet) # sample average and sample sd
quantile(xRet,probs=c(0.05,0.5,0.95)) # sample quantiles
Fn <-ecdf(xRet); # compute ECDF

## plots comment out saving to pdf files, so appear on screen

#pdf("xRetHist.pdf") # plot histogram
   hist(xRet,xlim=c(-6,6),breaks=100,xlab="Arithmetic returns",
   freq=FALSE, main="Histogram of arithmetic returns")
#dev.off()
```

```
#pdf("xRet.pdf") # plot ECDF
   plot(Fn,main="EDF of daily returns",xlim=c(-5,5),
    xlab="Geometric financial daily returns",
        ylab="EDF")
   abline(h=c(0.05,0.5,0.95),col="blue",lwd=5)
#dev.off()

h = 1.06*(length(xRet)^-0.2)*sd(xRet); # Silverman's bandwidth rule
#pdf("xRetDenSilverman.pdf") # kernel density estimator
   plot(density(xRet,bw=h,kernel="rectangular",from=-7,to=7),
    ylim=c(0,0.56),
         main="Kernel density estimator: daily returns, h automatic",
         xlab="Arithmetic financial daily returns")
   abline(v=c(0.0),col="blue",lwd=2)
#dev.off()
```

Chapter 2

Models, Likelihood, Estimation, and Method of Moments

2.1 Statistical models

In statistics, we are often interested in a model for observed data $\mathbf{y} = (y_1, \ldots, y_n)$, where y_j is the data point for the jth individual. In general the y_j may be vectors, e.g., for each individual we may observe a bunch of variables, but for simplicity we will assume for now that y_j is a number.

Definition 2.1.1 (Statistical model)**.** A *statistical model* views $\mathbf{y}$ as a realization of the random variables $\mathbf{Y} = (Y_1, \ldots, Y_n)$ from their joint cumulative distribution function (CDF) $F_{\mathbf{Y}}$. The model specifies a collection of possibilities for $F_{\mathbf{Y}}$. Before making the observation, we have a random vector $\mathbf{Y}$. After making the observation, $\mathbf{Y}$ crystallizes into the data $\mathbf{y}$. We say that the model *generated* the data, and often we want to use the data to learn about the model.

2.1.1 What is an estimand?

Definition 2.1.2 (Estimand)**.** An *estimand* is an aspect of $F_{\mathbf{Y}}$ that we wish to learn about from the data that we will observe.

Example 2.1.3. Some common examples of estimands include:

- $\mathrm{E}[Y_1]$ and $\mathrm{Var}(Y_1)$;
- the p-quantile $F_{Y_1}^{-1}(p)$;
- the cumulative distribution function evaluated at y, which we write as $F_{Y_1}(y)$.

In statistical science, estimands are often denoted by Greek letters, such as θ, ψ, μ (especially for a mean), σ (especially for a standard deviation), λ (especially for a rate), ρ (especially for a

correlation), β (especially for a coefficient in a regression model), τ (especially for a treatment effect), and π (especially for a probability).

☣ **2.1.4.** A review by Kahan, Morris, White, Carpenter, and Cro (2021) of the results of 50 different randomized trials published in leading medical journals found that none of the articles were explicit about the estimand in their trial and only in 74% of cases was it possible to infer the estimand from the text. In your statistical work, always make it clear what your estimand is!

2.1.2 Parametric statistical models and parameters

Reality is complicated, so we often use *parametric models* as approximations that hopefully provide useful insights into all of $F_{\mathbf{Y}}$ or parts of it. Parametric models can be used to encapsulate scientific theory, knowledge, or assumptions about how the data were generated. With the help of a good model, we can use data to address scientific questions, explore and describe, as well as predict or draw causal conclusions.

Definition 2.1.5 (Parametric model)**.** A *parametric statistical model* is a family of probability distributions for $\mathbf{Y}$, indexed by a finite-dimensional parameter θ. The distributions in a model are usually specified by their joint CDFs or by their joint densities (joint PMFs in the discrete case, joint PDFs in the continuous case). The *parameter space*, denoted by Θ, is the set of all allowable values of θ. Thus each $\theta \in \Theta$ picks out a single probability distribution for $\mathbf{Y}$.

If in the above definition we instead allow θ to be infinite-dimensional, then we have a *nonparametric model*.

☣ **2.1.6.** The term "nonparametric" sometimes causes confusion since it sounds like it's saying there are no parameters, when in fact it means the parameter is infinite-dimensional! Furthermore, even though in theory there is an infinite gulf between a finite value and an infinite value, in practice the borderline between parametric and nonparametric models can be blurry. For example, OpenAI's `GPT-3` is the autoregressive language model behind `ChatGPT`. It is a model with 175 billion parameters. Since 175 billion is finite, this is a parametric model. But 175 billion is so large that the model may be thought of as approximately nonparametric.

Example 2.1.7. Here are several examples of models for one data point. Having specified a model for one data point, we also automatically get a model for i.i.d. observations, as discussed in Section 2.1.3.

- We often refer to "the Poisson distribution", but strictly speaking there are *infinitely* many Poisson distributions: for each $\theta > 0$ there is a Poisson distribution with mean θ: the Pois(θ)

distribution. The collection $\{\text{Pois}(\theta) : \theta > 0\}$ is a 1-dimensional parametric model, with parameter space $\Theta = (0, \infty)$.

- The $\text{Bin}(n, p)$ model, if both parameters are unknown, is a 2-dimensional model. However, typically with a Binomial model the sample size n is treated as known (either by design or by conditioning on it), and then we regard $\text{Bin}(n, p)$ as a 1-dimensional model, with parameter p and parameter space $(0, 1)$.

- To specify a Normal distribution, we can specify the mean μ and variance σ^2. So the family of all Normal distributions, $\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R},\, \sigma > 0\}$, is a 2-dimensional parametric model. Here $\theta = (\mu, \sigma^2)$ and $\Theta = \mathbb{R} \times (0, \infty)$, where $\times$ denotes the Cartesian product of two sets, i.e.,

$$A \times B = \{(a, b) : a \in A,\, b \in B\}.$$

  Alternatively, we could take our parameters to be (μ, σ) rather than (μ, σ^2).

- Suppose that our population consists of cats and dogs, and we are studying the weight Y of a random animal from the population. Let p be the proportion of cats in the population. Suppose that the weight of a random cat is $\mathcal{N}(\mu_1, \sigma_1^2)$ and the weight of a random dog is $\mathcal{N}(\mu_2, \sigma_2^2)$. Then Y is a *mixture* of Normals, not a Normal. Let Φ be the $\mathcal{N}(0, 1)$ CDF. Then by the law of total probability (LOTP), the CDF of Y is

$$\begin{aligned} F_Y(y) = P(Y \le y) &= P(Y \le y|\text{cat})P(\text{cat}) + P(Y \le y|\text{dog})P(\text{dog}) \\ &= p\,\Phi\left(\frac{y - \mu_1}{\sigma_1}\right) + (1 - p)\,\Phi\left(\frac{y - \mu_2}{\sigma_2}\right). \end{aligned}$$

  The parameter is $\theta = (p, \mu_1, \sigma_1, \mu_2, \sigma_2)$ and the parameter space is $\Theta = (0, 1) \times (0, \infty)^4$ since an animal's weight must be positive.

- Consider the family of *all* CDFs, i.e., all functions $F_{\mathbf{Y}}$ satisfying the properties of a valid CDF. This is a nonparametric model since no finite-dimensional θ that would suffice to specify $F_{\mathbf{Y}}$.

- Consider the family of *all* CDFs with median θ. Such a CDF can be written as

$$F_{Y;\theta}(y) = F^*(y - \theta)$$

  where F^* is a CDF with median 0. Think of F^* as nonparametric. Models such as $F_{Y;\theta}$ with a combination of a finite-dimensional θ and a nonparametric component, are called *semiparametric*. All semiparametric models are nonparametric since overall the parameter space is

infinite-dimensional, but in some problems the goal is to estimate the parametric part of the model (θ in this example).

Notation 2.1.8. To make explicit the fact that the CDF in a parametric model depends on θ, we sometimes write the CDF as $F_{\mathbf{Y};\theta}$. Then the CDF evaluated at $\mathbf{y}$ is denoted by $F_{\mathbf{Y};\theta}(\mathbf{y})$ or $F_{\mathbf{Y}}(\mathbf{y};\theta)$. If θ is being modeled as a random variable, as in the Bayesian approach, then ; is replaced by the conditioning bar | in the notation, yielding $F_{\mathbf{Y}|\theta}$ for the CDF (which technically is, from a Bayesian point of view, the conditional CDF of $\mathbf{Y}$ given θ). Then the CDF evaluated at $\mathbf{y}$ is written as $F_{\mathbf{Y}|\theta}(\mathbf{y})$ or $F_{\mathbf{Y}}(\mathbf{y}|\theta)$.

Overall, the choice between a frequentist approach to inference and a Bayesian approach can make a big difference in one's results, but the choice of punctuation between a semicolon and a bar should not matter much. In both $F_{\mathbf{Y}}(\mathbf{y};\theta)$ and $F_{\mathbf{Y}}(\mathbf{y}|\theta)$, we are looking at the distribution of the data, treated the parameter θ as fixed.

We will most often focus on parametric models in this book, but we will also introduce quite a few methods for working with nonparametric models. At first glance, it might seem extremely difficult to do inference for a nonparametric model: how can we estimate an infinite-dimensional parameter using only a finite amount of data? But in fact we have *already* encountered a useful nonparametric method: the empirical CDF $\hat{F}$ is often a good estimator for the true CDF F_Y.

If nonparametric methods work well, why not *always* use them rather than making strong parametric assumptions? There is no such thing as a free lunch; tradeoffs are everywhere in statistics. Other things being equal, it is better to have weaker assumptions than stronger assumptions, but sometimes methods based on weaker assumptions require much more data to achieve the same level of accuracy, if the stronger assumptions are approximately true.

Semiparametric methods are also quite common in statistics. For example, suppose that the estimand is the mean, the standard deviation, or the p-quantile of $F_{\mathbf{Y}}$, but that a parametric model for $F_{\mathbf{Y}}$ is not assumed. Then the estimand is a 1-dimensional but the overall parameter space is infinite-dimensional.

2.1.3 When the $Y_1, \ldots, Y_n$ are i.i.d.

Often it is scientifically plausible to assume that $Y_1, \ldots, Y_n$ are independent and identically distributed. Under the i.i.d. assumption,

$$Y_j \overset{\text{i.i.d.}}{\sim} F_{Y_1;\theta}, \quad j = 1, ..., n,$$

where $F_{Y_1;\theta}$ is the CDF of each Y_j. By independence, the joint CDF is the product of the marginal CDFs:

$$F_{\mathbf{Y}}(\mathbf{y};\theta) = \prod_{j=1}^{n} F_{Y_1}(y_j;\theta).$$

Example 2.1.9. The heights $Y_1, \ldots, Y_n$ of n individuals are measured. Assume that they are i.i.d. (Of course, we can make up situations where i.i.d. would not be a reasonable assumption. For example, if we know in advance that persons 1 and 2 are identical twins, it would not make sense to treat their heights as independent. If we know in advance that persons 1 through 5 are professional basketball players, while persons 6 through n are children, then it would not make sense to treat their heights as identically distributed.) Let's also assume a Normal model with parameter $\theta = (\mu, \sigma^2)$:

$$Y_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2).$$

(A mixture of Normals, with groups based on age and gender, would be more realistic but also more complicated.) The joint PDF of the data is

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y};\theta) &= \prod_{j=1}^{n} f_{Y_j}(y_j;\theta), \quad \text{since the } Y_j \text{ are i.i.d.} \\ &= \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_j-\mu)^2\right\}, \quad \text{since } Y_j \sim \mathcal{N}(\mu,\sigma^2) \\ &= \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\mu)^2\right\}. \end{aligned}$$

Example 2.1.10. Let Y_j be 1 if a certain baseball player gets on base in their jth inning in the playoffs and 0 otherwise. We know that Y_j is Bernoulli (since the Bernoulli is the *only* distribution with support $\{0,1\}$), but it is much less obvious whether it is reasonable to assume that the Y_j are i.i.d. Assume for this example that the Y_j are i.i.d., so we have the 1-dimensional parametric model

$$Y_j \overset{\text{i.i.d.}}{\sim} \text{Bern}(\theta).$$

We can write the PMF of Y_j compactly in two different ways:

$$P(Y_j = y_j;\theta) = y_j\theta + (1-y_j)(1-\theta), \text{ for } y_j \in \{0,1\},$$

and

$$P(Y_j = y_j;\theta) = \theta^{y_j}(1-\theta)^{1-y_j}, \text{ for } y_j \in \{0,1\}.$$

The latter is more useful here since it is in *product* form, which will mesh well with the fact that the joint CDF is the *product* of the marginal CDFs. We then have

$$P(\mathbf{Y} = \mathbf{y}; \theta) = \prod_{j=1}^{n} \theta^{y_j}(1-\theta)^{1-y_j} = \theta^{n\overline{y}}(1-\theta)^{n-n\overline{y}}, \quad \text{since } n\overline{y} = \sum_{j=1}^{n} y_j.$$

Note that the joint density depends on the data $(y_1, \ldots, y_n)$ only through their sample mean $\bar{y}$ (or, equivalently, only through their sum $\sum_{j=1}^{n} y_j$). This compression of the n-dimensional vector $(y_1, \ldots, y_n)$ down to the 1-dimensional scalar $\bar{y}$ happens without losing any information as far as the joint density is concerned. We will discuss this kind of compression in more detail in Chapter 7.

For example, in the 2013 post-season series between the Boston Red Sox and Tampa Bay Rays, the Red Sox player David Ortiz batted 13 times and got on base 5 times. So here

$$n = 13, \quad \sum_{j=1}^{n} y_j = 5, \quad \text{and} \quad \bar{y} = \frac{5}{13}.$$

As far as the joint density is concerned, it does not matter *which* 5 of his 13 at-bats were the ones where he got on base. The data for this example are from
`https://www.baseball-reference.com/players/o/ortizda01.shtml`.

The modeling assumption that the Y_j's are i.i.d. is a strong one. Perhaps David Ortiz hit in streaks. Statistical models do not require i.i.d. assumptions; instead, the organizing principle is to build models for the joint distribution of $Y_1, Y_2, \ldots, Y_n$.

2.1.4 When the $Y_1, \ldots, Y_n$ are a time series

So far we have mainly been focusing on *independent* $Y_1, \ldots, Y_n$. However, in many applications the subscript represents *time*. Then independence often doesn't make sense as an assumption, since so many real world situations involve something evolving over time. Going beyond independence to allow for random variables evolving over time is the realm of *time series* in statistics and *stochastic processes* in probability.

Recall that for any events $A_1, \ldots, A_n$,

$$P(A_1, A_2, \ldots, A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2)\cdots P(A_n|A_1, \ldots, A_{n-1}).$$

Using this fact from probability, we can obtain a decomposition for the joint density of $\mathbf{Y}$ even when independence does not hold.

Definition 2.1.11 (Prediction decomposition)**.** If each Y_j is discrete, then the *prediction decomposition* is the factorization of the joint PMF as

$$P(\mathbf{Y} = \mathbf{y}; \theta) = P(Y_1 = y_1; \theta) \prod_{j=2}^{n} P(Y_j = y_j | Y_1 = y_1, \ldots, Y_{j-1} = y_{j-1}; \theta).$$

Similarly, if each Y_j is continuous with PDF f_{Y_j}, then the prediction decomposition is the factorization of the joint PDF as

$$f_{\mathbf{Y}}(\mathbf{y}; \theta) = f_{Y_1}(y_1; \theta) \prod_{j=2}^{n} f_{Y_j}(y_j | Y_1 = y_1, \ldots, Y_{j-1} = y_{j-1}; \theta).$$

The prediction decomposition replaces a product based on *independence* with a product based on *conditioning* by forecasting.

Example 2.1.12 (Prediction decomposition for a Markov chain)**.** Bianca lives in Singapore, where (unlike Boston!) the weather varies only mildly with the months of the year. She would like to know the probability it will not rain at all over the next 3 days in order to plan when she should water her garden. But she knows the daily weather can be described by streaks — periods of dry weather followed by some rainy days. Standard weather forecasts only tell us, at best, the chance it will rain on a specific day or hour.

Let $Y_j \in \{0, 1\}$, taking the value 1 if it rains on the jth day in Singapore and 0 otherwise. Going beyond independence, the simplest model Bianca could use is a Markov chain. If we make the modeling assumption that $Y_1, \ldots, Y_n$ is a Markov chain started at its stationary distribution, then the distribution of Y_j given the past, $(Y_1, \ldots, Y_{j-1})$, only depends on the most recent information, Y_{j-1} (this is the Markov property), and the distribution of Y_j is the same for all j (by definition of stationary distribution). If it is deemed too unrealistic that tomorrow's weather should depend on the past only through today's weather, then a higher order Markov chain could be used; this would be more realistic but also more complicated (and thus harder to do computations with and to estimate the parameters for).

Under these assumptions, we have a Markov chain on the state space $\{0, 1\}$. Let the transition probabilities be

$$P(Y_j = k | Y_{j-1} = \ell) = p_{\ell k}, \quad \ell, k \in \{0, 1\}.$$

We have $p_{\ell 0} + p_{\ell 1} = 1$ for $\ell = 0$ and for $\ell = 1$, so there are really only 2 parameters rather than 4 that are needed to specify the transition matrix; once we know p_{01} and p_{11}, we also know p_{00} and p_{10}. Let

$$\psi = (p_{01}, p_{11}), \text{ with } p_{01}, p_{11} \in (0, 1).$$

Let

$$\pi = \mathrm{E}[Y_j; \psi] = P(Y_j = 1; \psi).$$

(In statistics, π sometimes does *not* denote the number $3.14159265358979323846\ldots$, because of statisticians' fondness for using Greek letters to denote parameters. Specifically, it is frequently used for a parameter that is a probability (though p is also often used for this purpose), since the Greek letter π is the analog of the English letter p.) Note that we are assuming that the chain starts out at its stationary distribution, so π does not change with j.

Bianca's estimand is

$$\theta = P(Y_j = 0, Y_{j+1} = 0, Y_{j+2} = 0 | Y_{j-1} = 0) = p_{00}^3.$$

Let us work out how π is related to ψ, and what the prediction decomposition looks like in this setting. We have

$$\begin{aligned} \mathrm{E}[Y_j; \psi] &= \mathrm{E}\left[\mathrm{E}[Y_j | Y_{j-1}; \psi); \theta]\right], \quad \text{by Adam's law,} \\ \mathrm{E}[Y_j | Y_{j-1}; \psi] &= p_{01}(1 - Y_{j-1}) + p_{11} Y_{j-1} = p_{01} + (p_{11} - p_{01})\, Y_{j-1}. \end{aligned}$$

Taking the expectation of both sides of the above equation and using the fact that

$$\pi = \mathrm{E}[Y_j; \psi] = \mathrm{E}[Y_{j-1}; \psi]$$

(by the assumption that the chain starts out at its stationary distribution), we have

$$\pi = p_{01} + (p_{11} - p_{01})\pi.$$

Thus,

$$\pi = \frac{p_{01}}{1 - p_{11} + p_{01}}, \quad 1 - \pi = \frac{1 - p_{11}}{1 - p_{11} + p_{01}}.$$

By the Markov property, the prediction decomposition here simplifies to

$$P(\mathbf{Y} = \mathbf{y}; \psi) = P(Y_1 = y_1; \psi) \prod_{j=2}^{n} P(Y_j = y_j | Y_{j-1} = y_{j-1}; \psi).$$

Since we are assuming that the chain starts out at its stationary distribution,

$$P(Y_1 = y_1; \psi) = \pi^{y_1} (1 - \pi)^{1 - y_1}.$$

So

$$\begin{aligned} P(\mathbf{Y} = \mathbf{y}; \psi) &= \pi^{y_1} (1 - \pi)^{1 - y_1} \prod_{j=2}^{n} \left\{ p_{11}^{y_j} (1 - p_{11})^{1 - y_j} \right\}^{y_{j-1}} \left\{ p_{01}^{y_j} (1 - p_{01})^{1 - y_j} \right\}^{1 - y_{j-1}} \\ &= \pi^{y_1} (1 - \pi)^{1 - y_1} p_{11}^{n_{11}} (1 - p_{11})^{n_{10}} p_{01}^{n_{01}} (1 - p_{01})^{n_{00}}, \end{aligned}$$

where

$$n_{\ell k} = \sum_{j=2}^{n} I(y_{j-1} = l) I(y_j = k), \quad \ell, k \in \{0, 1\}^2,$$

counts the number of moves from ℓ to k. So the joint PMF is determined by the following quantities: the two parameters p_{01} and p_{11} (since we know what p_{00}, p_{10}, and π are in terms of these), the initial data point y_1, and the four summary statistics $n_{00}, n_{01}, n_{10}, n_{11}$. Using techniques we develop later, we can make inferences about ψ, which in turn imply inferences about the estimand θ.

2.2 Likelihood

2.2.1 Definition and Intuition

Let $f_{\mathbf{Y}}(\mathbf{y}; \theta)$ be the joint density function of all the observations (perhaps conditioning on some predictor variables), under some model with parameter θ. Here we use "density" in a broad sense, including both the discrete case (where we have a probability mass function) and the continuous case (where we have a probability density function), so that we do not need to discuss the discrete and continuous cases separately.

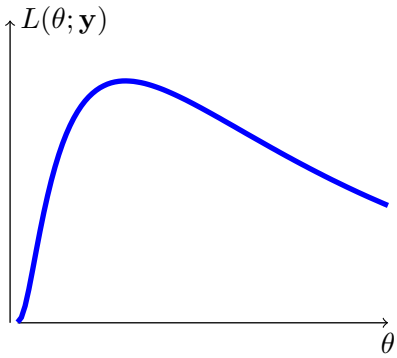


Figure 2.1: A likelihood function $L(\theta; \mathbf{y})$ plotted against θ, holding the data $\mathbf{y}$ fixed.

Definition 2.2.1 (Likelihood function)**.** Let $\mathbf{y}$ be the observed value of $\mathbf{Y}$. The function given by

$$L(\theta; \mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}; \theta),$$

regarded as a function of the parameter, with the data held constant, is called the *likelihood function*. That is, the likelihood function is the probability or probability density of the data given the parameters, as a function of the parameters. So the likelihood function is a function of θ, with $\mathbf{y}$ treated as fixed. A likelihood function is illustrated in Figure 2.1.

Notationally, it is conventional to separate θ and y by a semicolon: $L(\theta; \mathbf{y})$. Often we even write $L(\theta)$ for the likelihood function, leaving the $\mathbf{y}$ implicit, to simplify the notation further and to emphasize that the likelihood function is regarded as a function of θ.

In a density function, θ is fixed, while $\mathbf{y}$ varies. In a likelihood function, the roles are reversed: $\mathbf{y}$ is fixed, while θ varies. The idea is that we *know* $\mathbf{y}$ since we observed it, whereas we do not know θ, so we want to compare various possible candidate values of θ based on how consistent they are with the data. If $L(\theta_2; \mathbf{y})$ is much larger than $L(\theta_1; \mathbf{y})$, then the data seem more consistent with θ_2 than θ_1, in the sense that θ_2 makes the data that we actually observed much more likely than θ_1 does.

☣ **2.2.2.** Suppose that we are comparing two possible values for θ, say θ_1 and θ_2, and find that $L(\theta_2; \mathbf{y})$ is much larger than $L(\theta_1; \mathbf{y})$. Strictly speaking we cannot conclude that θ_2 is more likely than θ_1 to be the true value of θ (in the sense of higher probability); in fact, "more likely" is not even defined on the parameter space (at least, not yet), since we have not imposed a probability distribution on θ. In a frequentist approach, θ is regarded as fixed but unknown, and it does not have a distribution.

The likelihood function plays an essential role in statistical inference, for both Bayesian approaches and frequentist approaches. Let us quickly introduce why the likelihood function is important for both approaches, then later in the course we will go into much more depth about this.

Bayesian perspective: In a Bayesian approach, we have a prior density $\pi(\theta)$ for θ, and use Bayes' rule to obtain the posterior density:

$$\pi(\theta|\mathbf{y}) = \frac{\pi(\theta) f(\mathbf{y}|\theta)}{f(\mathbf{y})} \propto \pi(\theta) f(\mathbf{y}|\theta) = L(\theta; \mathbf{y}) \pi(\theta),$$

where the proportionality stems from the fact that we are treating $\mathbf{y}$ as fixed (so the denominator $f(\mathbf{y})$, which is the marginal density of $\mathbf{y}$, is viewed as a constant). That is, Bayes' rule says, in words:

The posterior is proportional to likelihood times prior.

So the two key ingredients for a Bayesian analysis are the prior distribution $\pi(\theta)$ and the likelihood function $L(\theta; \mathbf{y})$. Combining the likelihood and the prior, we obtain the posterior distribution, which we then base our inferences on.

Frequentist perspective: In a frequentist approach, θ does not have a posterior distribution, but we can use the likelihood function as a surrogate for assessing how plausible various possible values of θ are. One of the most widely used estimation techniques in statistics is *maximum likelihood estimation*, which says to estimate θ using

$$\hat{\theta} = \arg\max_{\theta} L(\theta; \mathbf{y}),$$

the parameter value that maximizes the likelihood function. This value is called the *maximum likelihood estimate* (MLE). The MLE is illustrated in Figure 2.2. We will study properties and examples of maximum likelihood estimation later. Additionally, *likelihood ratios* such as $L(\theta_2;\mathbf{y})/L(\theta_1;\mathbf{y})$ come up often in statistics, such as if we wish to decide between which of two hypotheses, $\theta = \theta_1$ or $\theta = \theta_2$, is better supported by the data.

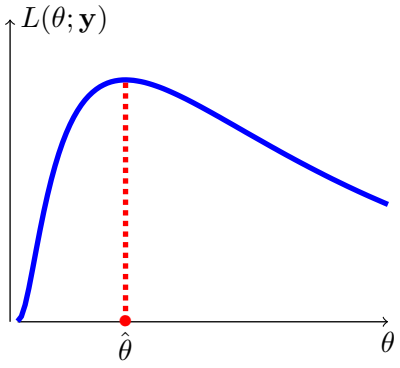


Figure 2.2: Plot of a likelihood function $L(\theta;\mathbf{y})$, with a dotted line at $\hat{\theta}$, the MLE.

Remark 1 (Equivalence of likelihood functions)**.** Two likelihood functions are viewed as equivalent if one is a positive constant times the other. In fact, the "constant" can even be a function of the data (it just cannot depend on the parameter)!

From a Bayesian point of view, the posterior is *proportional* to likelihood times prior; the normalization to make the posterior sum or integrate to 1 can be done *at the end*, so rescaling the likelihood function has no effect on the posterior. From a frequentist point of view, the primary uses of the likelihood function are for maximum likelihood estimation and likelihood ratios, neither of which are affected by rescaling the likelihood function. For example, if we use $7L(\theta;\mathbf{y})$ as our likelihood function instead of $L(\theta;\mathbf{y})$, then the MLE is the same since maximizing $L(\theta;\mathbf{y})$ is equivalent to maximizing $7L(\theta;\mathbf{y})$, and the likelihood ratios are the same since

$$\frac{7L(\theta_2;\mathbf{y})}{7L(\theta_1;\mathbf{y})} = \frac{L(\theta_2;\mathbf{y})}{L(\theta_1;\mathbf{y})}.$$

Example 2.2.3 (Poisson likelihood)**.** Recall that the PMF of a Poisson r.v. with parameter λ is

$$P(Y = y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!},$$

for $y = 0, 1, 2, \ldots$. So the likelihood function for n i.i.d. Poisson observations $y_1, \ldots, y_n$ is the product

$$L(\lambda;\mathbf{y}) = \prod_{j=1}^{n} e^{-\lambda}\lambda^j/y_j = \frac{e^{-n\lambda}\lambda^{\sum_{j=1}^n y_j}}{\prod_{j=1}^n y_j!}.$$

If desired, we can instead drop the expression involving the factorials (since it is a multiplicative factor that doesn't involve λ). This gives the simpler expression

$$L(\lambda; \mathbf{y}) = e^{-n\lambda}\lambda^{\sum_{j=1}^{n} y_j}$$

as our likelihood function (just make sure not to leave out terms involving λ).

Example 2.2.4 (Binomial likelihood)**.** Let $Y \sim \text{Bin}(n, p)$ with $n = 3$ and p unknown. Suppose that $Y = 2$ is observed. The likelihood function is

$$L(p) = \binom{3}{2} p^2(1-p).$$

As noted above, multiplying by a constant gives an equivalent likelihood function, so we are free to omit constant factors (for simplicity) or to introduce constant factors (to standardize or normalize $L(p)$). Three natural ways to write the same likelihood function are:

1. $L(p) = p^2(1-p)$ (this is the simplest, avoiding clutter);
2. $L(p) = \frac{27}{4}p^2(1-p)$ (this is standardized to have maximum 1);
3. $L(p) = 12p^2(1-p)$ (this is normalized so that it integrates to 1; note that this is the Beta(3,2) density function).

All three of these $L(p)$ are equally valid ways to express the likelihood function. For the last of the three, note that in a Bayesian approach with the prior $p \sim \text{Unif}(0, 1)$, the posterior is the likelihood and so we have a Beta(3,2) posterior, which is consistent with what we know about Beta-Binomial conjugacy from Chapter 8 of the Stat 110 book.

2.2.2 Log-likelihood

It is very common when working with likelihood to work with the *log-likelihood*

$$l(\theta; \mathbf{y}) = \log L(\theta; \mathbf{y})$$

instead. Two reasons that this is helpful are:

1. *Numerical stability.* Extremely small probabilities often come up in likelihood calculations. For example, if we observe 100 i.i.d. Bern(1/2) random variables, then the probability of the observed data $y_1, \ldots, y_n$ is
$$\prod_{j=1}^{100} (1/2)^{y_j}(1/2)^{1-y_j} = 2^{-100} \approx 7.89 \times 10^{-31}.$$

The log of this is

$$-100 \log 2 \approx -69.3,$$

which is safer and simpler to work with than an extremely small value like 7.89×10^{-31}.

2. *Sums rather than products.* As we have seen, the likelihood function is often expressed as a product. Taking the log converts the product to a sum, and often it's easier to work with sums than products. For example, in calculus the derivative of a sum is the sum of the derivatives, whereas the derivative of a product involves using the product rule. And in probability, for the sum of a lot of independent random variables we may be able to apply theorems such as the law of large numbers and central limit theorem, whereas it's less clear how to handle the product of a lot of independent random variables.

Since log is a continuous, strictly increasing function, maximizing the likelihood is equivalent to maximizing the log-likelihood, so when we study MLE later we are free to work with the log-likelihood.

If the Y_j are independent, then the likelihood function is

$$L(\theta; \mathbf{y}) = \prod_{j=1}^{n} f_{Y_j}(y_j; \theta)$$

and the log-likelihood function is

$$l(\theta; \mathbf{y}) = \sum_{j=1}^{n} \log f_{Y_j}(y_j; \theta).$$

Each data point contributes one term to the above sum.

Strategy: The concept of likelihood is very powerful. So often a productive problem-solving strategy in statistics is to *start by writing down the likelihood function.* In doing so, it often is helpful first to *consider the likelihood contribution of a single observation.* Once we have the likelihood function, it is often useful for building insight to *plot the likelihood function.*

☣ **2.2.5.** As explained earlier, we can drop a *multiplicative* constant (not depending on θ but possibly depending on the data) from a likelihood function $L(\theta)$. This corresponds to dropping an *additive* constant from the log-likelihood $l(\theta)$. But we can't drop a multiplicative constant from a log-likelihood or an additive constant from a likelihood! For example, with a single observation Y from the model $Y \sim \mathcal{N}(\theta, 1)$, the likelihood function is

$$L(\theta) = \frac{1}{\sqrt{2\pi}} e^{-(y-\theta)^2/2}.$$

We are free to drop the multiplicative constant and instead use

$$L(\theta) = e^{-(y-\theta)^2/2}.$$

In terms of log-likelihood, with the first version of the likelihood we have

$$l(\theta) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}(y-\theta)^2$$

while with the second version we have

$$l(\theta) = -\frac{1}{2}(y-\theta)^2.$$

Dropping the Normal normalizing constant in $L(\theta)$ corresponds to dropping the additive constant $-\frac{1}{2}\log(2\pi)$ in the log-likelihood. However, we *cannot* drop the factor of $\frac{1}{2}$ in either version of the log-likelihood. Note also how convenient the log-likelihood is to work with in this example: it is (with either version) simply a quadratic function of θ.

Example 2.2.6. Consider a tiny dataset measuring the number of cycles before failure for springs at stress levels of 950 N/mm^2 (Newtons per square millimeter). We saw this example first in Davison (2003). A very common model for time until failure is the Exponential distribution (see Section 5.5 of the Stat 110 book). We model the cycles before failure, measured in thousands of cycles, as

$$Y_j \overset{\text{i.i.d.}}{\sim} \text{Expo}(\theta), \text{ for } j = 1, 2, \ldots, n,$$

with density

$$f_{Y_1}(y;\theta) = \theta\exp(-y\theta), \quad y > 0.$$

Recall that $\text{E}(Y_1) = 1/\theta$, $\text{Var}(Y_1) = 1/\theta^2$. Assuming that the springs fail independently, all with the same Exponential distribution, the likelihood function is

$$\begin{aligned} L(\theta;\mathbf{y}) &= f_{\mathbf{Y}}(\mathbf{y};\theta) = \prod_{j=1}^{n} f_{Y_j}(y_j;\theta) \\ &= \theta^n \exp\left(-\theta\sum_{j=1}^{n} y_j\right). \end{aligned}$$

The similarity in form between this likelihood and the Poisson likelihood from Example (letting $\theta = \lambda$, both are of the form λ to a power times an exponential function of λ) is not a coincidence: the Poisson and Exponential distributions are connected via the notion of a Poisson process, as discussed in Section 5.6 of the Stat 110 book.

The log-likelihood function is

$$l(\theta;\mathbf{y}) = n\log\theta - \theta\sum_{j=1}^{n} y_j.$$

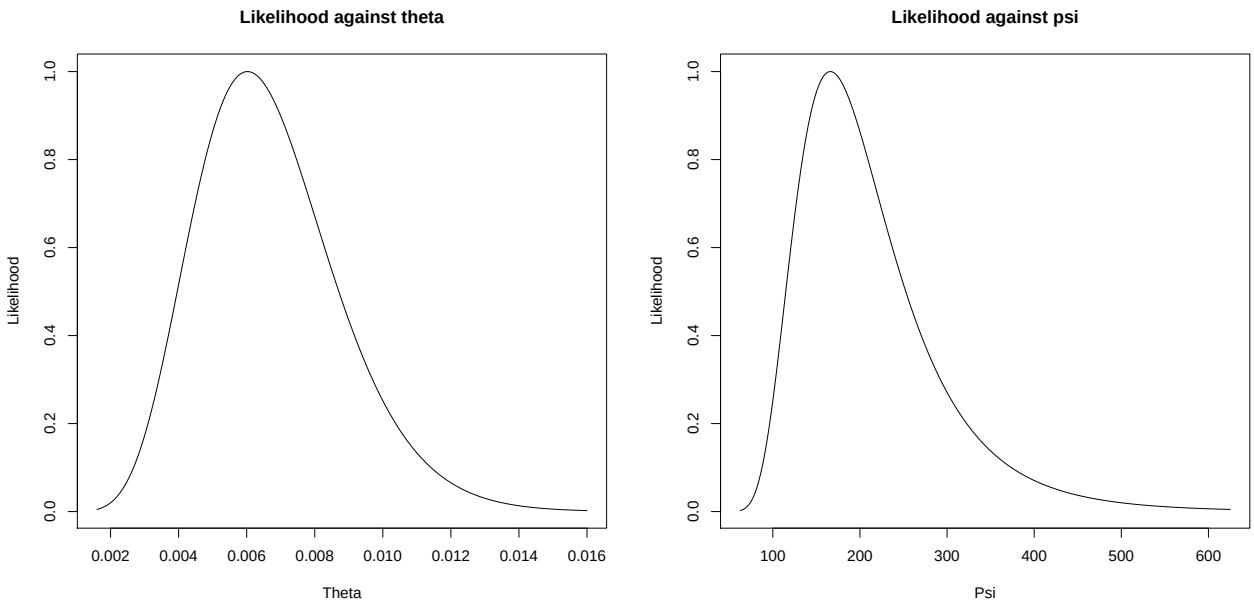


Figure 2.3: Likelihood function for the spring data, parameterized by θ (left) and $\psi = 1/\theta$ (right). The curves are different in shape but for any value θ_0, the height of the curve on the left at θ_0 equals the height of the curve on the right at the corresponding value $\psi_0 = 1/\theta_0$.

The data are

$$\mathbf{y} = (225, 171, 198, 189, 135, 162, 135, 117, 162)\,,$$

with sample size $n = 9$.

The left-hand side of Figure 2.3 shows the likelihood function (not the log-likelihood function), standardized so that the maximum is 1.

Example 2.2.7 (Censored and truncated data)**.** A widget-making company wants to study the reliability of their supposedly water-resistant widgets. The *survival time* of a widget that gets wet is defined as the length of time from when the widget gets wet until it stops working. Suppose that such survival times are i.i.d. Expo(λ) random variables, with time measured in days. We have a dataset $\mathbf{t} = (t_1, \ldots, t_n)$ of survival times (in days) of n widgets that got wet. Using the Exponential PDF, as in the previous example, the likelihood function is

$$L(\lambda; \mathbf{t}) = \prod_{j=1}^{n} \lambda e^{-\lambda t_j} = \lambda^n e^{-\lambda n \bar{t}},$$

where $\bar{t}$ is the sample mean of $t_1, \ldots, t_n$.

However, an important principle in statistics is to think about *how the data were collected* rather than take the data at face value. For example, are the data really i.i.d. draws from the distribution

of interest? Could there be correlations over time? Are there data we're *not* seeing that would have been very different from the data we are able to observe?

Truncated data: Suppose that it turns out that the experiment began by getting a set of widgets wet at noon on a Friday, and then the experimenter went home for the weekend, and came back at noon on the following Monday. Some widgets may have failed in that 3 day period, but no one was around to record when those failures occurred. The experimenter discarded data about such widgets, rather than taking a principled approach by recording all the relevant information. Then the data we have are implicitly conditioned on the survival time being more than 3 days.

Let $T \sim \text{Expo}(\lambda)$ be the survival time of a widget. By the memoryless property, the conditional distribution of $(T-3)|(T>3)$ is also $\text{Expo}(\lambda)$. We can also show this by computing the conditional PDF of T given $T > 3$:

$$f(t|T>3) = \frac{f(t)}{P(T>3)} = \frac{\lambda e^{-\lambda t}}{e^{-3\lambda}} = \lambda e^{-\lambda(t-3)}, \text{ for } t > 3.$$

This kind of conditional distribution is called a *truncated distribution*, and the resulting data are called *truncated data*. The likelihood is a product of terms following the above density:

$$L(\lambda; \boldsymbol{t}) = \lambda^n e^{-\lambda \sum_{j=1}^{j} (t_j - 3)}.$$

Here n is the number of widgets on which we have data. Initially there may have been more than n widgets, but we don't know how many there were since the experimenter did not keep track of how many widgets failed in the first 3 days.

Censored data: Now suppose instead that the truncation described above did not occur. There were n widgets initially. The experiment is run for 7 days, at which point we need to come up with some estimates. Suppose that at time 7, m of the n widgets have failed, with $m < n$. For the remaining widgets, we don't have the luxury of waiting around until they fail. The data here are called *censored data* since for widgets that have not yet failed, we do not know their failure times: we only know that they lasted more than 7 days.

Let $s_1, \ldots, s_m$ be the observed failure times, and $s_{m+1}, \ldots, s_n$ be the unobserved failure times (e.g., we know that $s_{m+1} \geq 7$ but we don't know the value of s_{m+1}).

A widget's contribution to the likelihood function for λ is the PDF at t if the widget was observed to have stopped working at time t, and is the probability of still being working after 7 days if its survival time was censored. Using the Exponential PDF and CDF, the likelihood function is

$$L(\lambda) = \left(\prod_{j=1}^{m} \lambda e^{-\lambda s_j} \right) \left(e^{-7\lambda} \right)^{n-m} = \lambda^m e^{-\lambda(m\bar{s} + 7(n-m))},$$

where $\bar{s}$ is the sample mean of $s_1, \ldots, s_m$. In forming the truncated data likelihood, we don't know how many data points we didn't observe. Here, in forming the censored data likelihood, we know that there are $n-m$ unobserved data points, and we take into account that we *did* observe that these data points all exceed 7 days.

2.2.3 Reparameterization

Some people like to parametrize the Exponential density in a different way, in terms of the *mean parameter* $\psi = 1/\theta$ instead of the *rate parameter* θ. (The parameter θ is called the rate parameter because of the close connection between an Expo(θ) random variable and a Poisson process of rate θ; the parameter ψ is called the mean parameter because Y_j has mean ψ.) In terms of ψ, the PDF of one data point is

$$f_{Y_1}(y;\psi) = \psi^{-1}\exp(-y\psi^{-1}), \quad y > 0.$$

The likelihood function is then

$$L(\psi;\mathbf{y}) = \prod_{j=1}^{n} \psi^{-1}\exp(-y_j\psi^{-1}) = \psi^{-n}\exp\left(-\psi^{-1}\sum_{j=1}^{n} y_j\right).$$

The change from θ to ψ is called a *reparameterization* of the model. One statistician may prefer working in terms of θ while another may prefer working in terms of ψ. It would be very sad if they obtained different inferences just because of using different parameterizations for the same model. Happily, it should not matter which parameterization was used. Note that $L(\psi;\mathbf{y})$ can be obtained from $L(\theta;\mathbf{y})$ simply by plugging in $\theta = \psi^{-1}$. For each point on the likelihood curve $L(\theta;\mathbf{y})$, there is a corresponding point on the likelihood curve $L(\psi;\mathbf{y})$.

The shape of the plot of $L(\psi;\mathbf{y})$ against ψ is different from that of the plot of $L(\theta;\mathbf{y})$ against θ, as seen in Figure 2.3, since the transformation from θ to ψ means that the horizontal axes in these plots are different, but the likelihood of any value of ψ equals the likelihood of the corresponding value of θ.

Generalization the discussion of reparameterization from the above example, we have the following comforting and convenient result.

Theorem 2.2.8 (Invariance under one-to-one transformation of the parameter)**.** *The likelihood function is unchanged under reparameterization, in the following sense. Consider a likelihood function $L(\theta;\mathbf{y})$ and let $\psi = g(\theta)$ be a reparameterization, where g is a known one-to-one function. Then*

$$L(\psi;\mathbf{y}) = L(\theta;\mathbf{y}).$$

Proof. The joint density of $\mathbf{Y}$ does not depend on whether we parameterize in terms of ψ or in terms of θ; either way we are getting a PMF or PDF for $\mathbf{Y}$, just with different notation in how to express the parameter. Therefore,

$$L(\psi; \mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}; \psi) = f_{\mathbf{Y}}(\mathbf{y}; \theta) = L(\theta; \mathbf{y}),$$

where for each possible value of θ on the right-hand side we are using the corresponding value $\psi = g(\theta)$ on the left-hand side, and for each possible value of ψ on the left-hand side we are using the corresponding value $\theta = g^{-1}(\psi)$ on the right-hand side. ∎

☣ **2.2.9.** The notation in Theorem 2.2.8 can sometimes seem confusing. In strict mathematical formalism, the L on the left-hand side is not the same function as the L on the right-hand side since the L on the left takes values of ψ as input, while the L on the right takes values of θ as input. We could instead introduce subscripts, like L_ψ on the left and L_θ on the right, but for brevity we are just using L to denote likelihood, where it is clear from the context which parameter is the input to that likelihood function.

People also sometimes wonder how it is possible that on the one hand the likelihood function is invariant, while on the other hand a plot of $L(\psi; \mathbf{y})$ as a function of ψ may look very different from a plot of $L(\theta; \mathbf{y})$ as a function of θ. To make an algebra analogy, let $f(x) = x^2$ and consider the change of variables $x = y^3$. Let $h(y) = y^6$. Then

$$f(x) = h(y).$$

But a plot of the function f looks very different from a plot of the function g. We are free though to choose whether to work with $f(x)$ or with $h(y)$. For example, if our goal is to find the value of x that maximizes $f(x)$, it would be equivalent to find the value y_0 of y that maximizes $h(y)$, and then use the corresponding value $x_0 = y_0^3$ for x.

In contrast, in a Bayesian approach we have to use change of variables techniques from probability (as in Chapter 8 of the Stat 110 book), e.g., Jacobians, if we want to convert between the posterior density of univariate θ and the posterior density of $\psi = g(\theta)$:

$$\begin{aligned}
\pi(\theta|\mathbf{y}) &\propto L(\theta; \mathbf{y})\pi(\theta) \\
&= L(\psi; \mathbf{y})\pi(\theta), \quad \text{by the invariance property of likelihood} \\
&= L(\psi; \mathbf{y})\pi(\psi)\left|\frac{\partial \psi}{\partial \theta}\right|, \quad \text{by change of variables} \\
&\propto \pi(\psi|\mathbf{y})\left|\frac{\partial \psi}{\partial \theta}\right|, \quad \text{by Bayes' rule.}
\end{aligned}$$

It is very convenient that for change of parameters with likelihoods, no Jacobians are required!

In Theorem 2.2.8 we considered what happens to the likelihood if we transform the *parameter*. It is also often helpful to transform the *data*.

Theorem 2.2.10 (Invariance under one-to-one transformation of the data)**.** *Let* $\mathbf{y} = (y_1, \ldots, y_n)$ *be the observed data, coming from a model with parameter* θ*. Let* h *be a known one-to-one function from* $\mathbb{R}^n$ *to* $\mathbb{R}^n$*. Use* h *to transform the data, letting* $\mathbf{x} = h(\mathbf{y})$*. Then taking the dataset to be* $\mathbf{x}$ *rather than* $\mathbf{y}$ *has no effect on the likelihood function:*

$$L(\theta; \mathbf{x}) = L(\theta; \mathbf{y}).$$

Proof. For simplicity, we will only write the proof in the case of a single observation y from a continuous distribution, with h a differentiable, strictly increasing function. Let Y be the r.v. that "crystallizes" to y, $x = h(y)$, and $X = h(Y)$. By the change of variables formula,

$$L(\theta; x) = f_X(x; \theta) = f_Y(y; \theta)\frac{1}{h'(y)}.$$

But $\frac{1}{h'(y)}$ is a multiplicative "constant" (not depending on θ), so it can be dropped. Then we can take the likelihood function for θ, based on the data x, to be

$$L(\theta; x) = f_Y(y; \theta) = L(\theta; y).$$

∎

Example 2.2.11 (Transforming a Log-Normal back to Normal)**.** Let

$$Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \text{Log-Normal}(\mu, \sigma^2),$$

and let $\theta = (\mu, \sigma^2)$ be the estimand. Recall from Chapter 6 of the Stat 110 book that the log of a Log-Normal is Normal. It may then be helpful to perform a log transformation, since Normal r.v.s. may be easier and more familiar to work with than Log-Normal r.v.s. So let

$$X_j = \log Y_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2).$$

Let y_j be the observed value of Y_j and $x_j = \log y_j$ be the observed value of X_j. Then the likelihood function $L(\theta)$ is the same regardless of whether we take our dataset to be $\mathbf{x}$ or $\mathbf{y}$. This makes sense intuitively since in transforming from $\mathbf{y}$ to $\mathbf{x}$, we neither created nor destroyed information about θ. We processed the data but we did not get new data, nor did we throw out data.

2.3 Statistics, estimators, and estimates

As before, let our data be the realization of the random variables

$$\mathbf{Y} = (Y_1, \ldots, Y_n).$$

Our goal is to learn about the estimand, which we will write here as θ.

Definition 2.3.1 (Statistic)**.** A *statistic* is a function of $Y_1, \ldots, Y_n$ (and possibly other known quantities). We can write a statistic as $T(\mathbf{Y})$, where computing the function T must *not* require knowledge of any unknown parameters.

Example 2.3.2 (Sample mean is a statistic)**.** The sample mean is a very widely used statistic:

$$T(\mathbf{Y}) = \overline{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j.$$

Some other statistics that we have encountered already are the sample variance, sample median, and sample p-quantile.

☣ **2.3.3.** A statistic is not allowed to depend on any unknown parameters, but its *distribution* often does depend on unknown parameters. For example, let $Y_1, \ldots, Y_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$, with μ and σ^2 unknown and $n \geq 2$. Since the sum of independent Normals is Normal,

$$\bar{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

We can *compute* $\bar{Y}$ just by knowing $Y_1, \ldots, Y_n$, but its *distribution* depends on μ, σ^2. For the sample variance

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2,$$

it is shown in Chapter 10 of the Stat 110 book that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

so S^2 has a scaled χ^2_{n-1} distribution where the scaling depends on σ^2 (though, interestingly, the distribution of S^2 does not depend on μ).

2.3.1 What is an estimator?

Definition 2.3.4 (Estimator)**.** Suppose that we use the random vector $\mathbf{Y} = (Y_1, ..., Y_n)$ to construct a statistic

$$\widehat{\theta} = T(\mathbf{Y}),$$

with the intention that this statistic should estimate an estimand θ. The statistic $\widehat{\theta}$ is called an *estimator*.

An estimator $\widehat{\theta}$ for θ is a random variable that can be computed based on the Y_j's, with the goal of being close to the estimand θ. All estimators are statistics, but a statistic is only an estimator if it has been charged with the task of estimating a particular estimand.

Next we define *bias*, which is a measure of how far off an estimator is from the estimand on average.

Definition 2.3.5 (Bias)**.** The *bias* of an estimator $\widehat{\theta}$ for θ is

$$\text{bias}(\widehat{\theta}) = \text{E}(\widehat{\theta}) - \theta.$$

We say that $\widehat{\theta}$ is *unbiased* for θ if its bias is 0, i.e., its expected value is θ.

To compute the bias, recall that by LOTUS,

$$\text{E}[\widehat{\theta}] = \int T(\mathbf{y}) f_{\mathbf{Y};\theta}(\mathbf{y}) d\mathbf{y},$$

if $\mathbf{Y}$ is continuous, and

$$\text{E}[\widehat{\theta}] = \sum_{\mathbf{y}} T(\mathbf{y}) P(\mathbf{Y} = \mathbf{y}; \theta),$$

if $\mathbf{Y}$ is discrete, where the integral and sum are over the support of $\mathbf{Y}$. Typically, the bias depends on θ (so we may be able to compute the bias theoretically but may not know the actual value of the bias due to θ being unknown).

The term *bias* is standard but unfortunate: it sounds terrible to use a biased estimator! Who wants to have to admit being biased? Who wants to have to defend to a general audience a statistical analysis involving a biased estimator? This definition is a *technical* meaning of "bias", within the estimation context. As we will see, it is extremely common in statistics that it is worthwhile to allow a little bit of bias in our estimator, if by doing so we can reduce the variance substantially.

☣ **2.3.6.** Just as a biased estimator is not necessarily bad, an unbiased estimator is not necessarily good. There are problems where no unbiased estimator exists, or where the only unbiased estimator that exists is ridiculous. It is desirable to have the bias be small, but other considerations such as the standard error also need to be taken into account.

Example 2.3.7 (An unbiased but absurd estimator)**.** Exercise 4.66 of the Stat 110 book is an example where an unbiased estimator exists but is ludicrous. Let $X \sim \text{Pois}(\lambda)$, and let the estimand be $\theta = e^{-3\lambda}$. The estimator $T = (-2)^X$ is unbiased for θ since

$$\text{E}[(-2)^X] - \theta = \sum_{k=0}^{\infty} (-2)^k \frac{\lambda^k}{k!} e^{-\lambda} - e^{-3\lambda} = e^{-\lambda} e^{-2\lambda} - e^{-3\lambda} = 0.$$

But it would be ridiculous to use this estimator: we have $0 < \theta \leq 1$, but T is *negative* if X is odd, and is at least 4 if X is even and positive. It turns out that T is the *only* unbiased estimator in this problem. Eliminating bias ends up forcing T to commit the category error of providing estimates that are, when $X \neq 0$, not even possible values of θ.

Definition 2.3.8 (Standard error)**.** The *standard error* of an estimator $\widehat{\theta}$ for θ is its standard deviation:

$$\mathrm{SE}(\widehat{\theta}) = \sqrt{\mathrm{Var}(\widehat{\theta})}.$$

This is a measure of how variable the estimator is.

Before resuming our statistical calculations, we state a handy mathematical fact.

Lemma 2.3.9 (Sum of squares identity)**.** For any random variables $Y_1, \ldots, Y_n$ and any constant c,

$$\sum_{j=1}^{n}(Y_j - c)^2 = \sum_{j=1}^{n}(Y_j - \bar{Y})^2 + n(\bar{Y} - c)^2.$$

This lemma is derived in the proof of Theorem 6.3.4 in the Stat 110 book, by writing

$$(Y_n - c)^2 = \{(Y_n - \bar{Y}_n) + (\bar{Y}_n - c)\}^2$$

and then expanding.

Example 2.3.10 (Bias of sample mean, sample variance, and sample standard deviation)**.** Suppose that $Y_1, \ldots, Y_n$ are uncorrelated with $\mathrm{E}[Y_j] = \mathrm{E}[Y_1]$ and $\mathrm{Var}[Y_j] = \mathrm{Var}[Y_1]$ for $j = 1, 2, \ldots, n$. Then

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$$

and

$$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2$$

are common estimators for $\mathrm{E}[Y_1]$ and $\mathrm{Var}(Y_1)$, respectively. The mean and variance of the sample mean are

$$\mathrm{E}[\bar{Y}] = \frac{1}{n}\sum_{j=1}^{n}\mathrm{E}[Y_j] = \mathrm{E}[Y_1], \qquad \mathrm{Var}[\bar{Y}] = \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{Var}[Y_j] = \frac{\mathrm{Var}[Y_1]}{n}.$$

In particular, the sample mean is unbiased, and its standard error is proportional to $1/\sqrt{n}$.

Next we will compute the bias of S^2. By Lemma 2.3.9,

$$\sum_{j=1}^{n}(Y_j - \bar{Y})^2 = \sum_{j=1}^{n} Y_j^2 - n(\bar{Y})^2,$$

Then

$$\begin{aligned}\mathrm{E}[S^2] &= \mathrm{E}\left[\frac{1}{n-1}\sum_{j=1}^{n} Y_j^2 - n(\bar{Y})^2\right] \\ &= \frac{1}{n-1}\left[n(\mathrm{Var}[Y_1] + \mathrm{E}[Y_1]^2) - n(\mathrm{Var}[\bar{Y}] + \mathrm{E}[\bar{Y}]^2)\right] = \mathrm{Var}[Y_1].\end{aligned}$$

So $\bar{Y}$ is unbiased for $\mathrm{E}[Y_1]$, with

$$\mathrm{SE}(\bar{Y}) = \sqrt{\mathrm{Var}[Y_1]/n},$$

and S^2 is unbiased for $\mathrm{Var}(Y_1)$. However, by Jensen's inequality (see Chapter 10 of the Stat 110 book)

$$\mathrm{E}[S] = \mathrm{E}[\sqrt{S^2}] \leq \sqrt{\mathrm{E}[S^2]} = \sqrt{\mathrm{Var}[Y_1]},$$

so $S = \sqrt{S^2}$ is a biased estimator of the standard deviation of Y_1.

We would like to construct an estimator $\widehat{\theta}$ for the estimand θ with low bias *and* low standard error. Often, however, there is a tradeoff between these, where we can decrease the standard error but only at the cost of increasing the bias, or vice versa. This is known as the *bias-variance tradeoff*, and is one of the most ubiquitous phenomena in statistics and machine learning. This tradeoff will be one of the main ideas discussed in the next chapter.

2.3.2 What is an estimate?

Definition 2.3.11 (Estimate)**.** An *estimate* is a realization of an estimator. So if our data $\mathbf{y}$ is a realization of $\mathbf{Y}$ and $T(\mathbf{Y})$ is an estimator of some estimand θ, then $T(\mathbf{y})$ is an estimate of θ.

Example 2.3.12 (Estimand vs. estimator vs. estimate)**.** Suppose that $Y_1, \ldots, Y_n$ are i.i.d. random variables with an unknown mean μ (the **estimand**). A simple, natural **estimator** for μ is the sample mean $\bar{Y}$. Once we have observed the data $y_1, \ldots, y_n$, we can compute the corresponding **estimate** $\bar{y}$. Before we observe the data, we have an estimator. After we observe the data, the Y_j's crystallize into the y_j's and the estimator crystallizes into an estimate.

☠ **2.3.13** (Pre-data vs. post-data points of view)**.** When scientists write $\widehat{\theta}$, sometimes they mean the estimator $\widehat{\theta} = T(\mathbf{Y})$ and sometimes they mean the estimate $\widehat{\theta} = T(\mathbf{y})$. This is rather unfortunate but it is very common. Hopefully, you will be able to tell from the context whether the author is discussing an estimator or an estimate. In your own writing, strive to use sufficiently precise language so that your readers do not get confused! And in thinking about a statistics problem, we often have to bounce back and forth between a pre-data view (thinking about the joint distribution of $Y_1, \ldots, Y_n$) and a post-data view (what to do with the observed data $y_1, \ldots, y_n$).

2.4 Sample moments and method of moments

2.4.1 Sample moments

A key probabilistic summary of the distribution of a random variable Y_1 is (if it exists) the kth *moment*, which by LOTUS is

$$\mathrm{E}[Y_1^k] = \int_{-\infty}^{\infty} y^k f_{Y_1}(y) dy, \quad k = 1, 2, \ldots.$$

See Chapter 6 of the Stat 110 book for various information about moments. For a statistical model where $Y_1, \ldots, Y_n$ are i.i.d., the kth *sample moment* is

$$M_k = \frac{1}{n} \sum_{j=1}^{n} Y_j^k.$$

The kth sample moment $\hat{\theta} = M_k$ is a natural estimator for the kth theoretical moment $\theta = \mathrm{E}[Y_1^k]$. The bias and variance of the kth sample moment are given by

$$\mathrm{bias}(M_k) = 0, \quad \text{and} \quad \mathrm{Var}(M_k) = \mathrm{Var}(Y_1^k)/n.$$

Also, the central limit theorem tells us the asymptotic distribution of M_k:

$$\sqrt{n}(M_k - \mathrm{E}[Y_1^k]) \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(Y_1^k)),$$

so long as $\mathrm{Var}(Y_1^k)$ exists.

The idea of estimating moments by sample moments generalizes to a technique called the *method of moments*.

2.4.2 Method of moments

One of the most widely used techniques for generating estimators is through the *method of moments*, which was introduced by Karl Pearson in 1894. Estimators generated in this way are not always the most accurate, but they remain useful to this day.

Definition 2.4.1 (Method of moments)**.** Consider the following strategy, the *method of moments principle*, for obtaining an estimator. Express the estimand in terms of moments, and then replace the estimand with the estimator and the theoretical moments with the sample moments. An estimator obtained in this way is called a *method of moments* estimator (MoM).

Example 2.4.2 (Poisson MoM)**.** We observe i.i.d. $Y_1, \ldots, Y_n \sim \mathrm{Pois}(\theta)$, with θ unknown. How should we estimate θ? Focus on the Poisson property that

$$\theta = \mathrm{E}_\theta[Y_1] = \sum_{y=0}^{\infty} y P(Y_1 = y; \theta), \tag{2.1}$$

where here we have opted to use a subscript on the expectation to explicitly remind us that it depends upon the value of θ. Such reminders are not necessary, but they are sometimes helpful so you will occasionally see these subscripts in our book.

The method of moments principle takes

$$\theta = \mathrm{E}_\theta[Y_1]$$

from (2.1), replacing moments by sample moments and estimands by estimators to get

$$\hat{\theta} = \frac{1}{n}\sum_{j=1}^{n} Y_j.$$

Suppose (for concreteness) then that the sample mean is $\bar{y} = 3.1$. Then, the method of moments estimate of θ is 3.1, which makes sense intuitively since the mean of a Pois(3.1) random variable is also 3.1.

Example 2.4.3 (Normal MoM)**.** We observe i.i.d. $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$, with μ and σ^2 unknown. Our estimand will be $\theta = (\mu, \sigma^2)$. If the data have sample mean $\bar{y} = 5.2$ and sample variance $s^2 = 2.3$, then it would be natural to estimate that the distribution that generated the data is $\mathcal{N}(5.2, 2.3)$. This turns out to be the method of moments estimator, if s^2 is defined to have n rather than $n-1$ in its denominator.

To see why this follows, first relate the estimand θ to moments

$$\mu = \mathrm{E}_\theta[Y_1], \quad \text{and} \quad \sigma^2 = \mathrm{E}_\theta[Y_1^2] - \mathrm{E}_\theta[Y_1]^2.$$

Then we replace estimands with the estimator notation and the moments by their sample moments, this delivers

$$\hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} Y_j, \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n} Y_j^2 - \left(\frac{1}{n}\sum_{j=1}^{n} Y_j\right)^2.$$

Now $\hat{\sigma}^2$ simplifies to

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n} (Y_j - \bar{Y})^2.$$

So the method of moments estimator of the mean is the sample mean and the estimator of the variance is the sample variance (with n rather than $n-1$ in the denominator). Notice that $\hat{\sigma}^2$ is biased.

☣ **2.4.4.** The sample moment is an unbiased estimator of the theoretical moment, but method of moments estimators are *not* unbiased in general.

More abstractly, consider a model for which the CDF is $F_{Y_1;\theta}$. Assume a decision has been made to focus on the kth moment of this distribution and that it is related to the estimand through the function

$$\alpha(\theta) = \mathrm{E}_\theta[Y_1^k].$$

Next, suppose $Y_j \overset{\text{i.i.d.}}{\sim} F_{Y_1;\theta}$, for $j = 1, \ldots, n$. The method of moments principle delivers

$$\alpha(\hat{\theta}) = \frac{1}{n}\sum_{j=1}^{n} Y_j^k.$$

This is only directly useful to the goal of estimating θ if the function α is invertible (analytically or numerically). Assuming that we can find the inverse function α^{-1}, we have

$$\theta = \alpha^{-1}\{\mathrm{E}[Y_1^k]\},$$

and the method of moments estimator

$$\hat{\theta} = \alpha^{-1}\left(\frac{1}{n}\sum_{j=1}^{n} Y_j^k\right).$$

Example 2.4.5 (Two MoM estimators for Exponential data)**.** In our springs example (Example 2.2.6), we modeled the time to failure of each spring with an Expo(θ) distribution. Suppose that we decide to estimate θ by the method of moments. Recall that

$$\mathrm{E}[Y_1] = \theta^{-1}$$

and

$$\mathrm{E}[Y_1^2] = \mathrm{Var}(Y_1) + \mathrm{E}[Y_1]^2 = 2\theta^{-2}.$$

Thus, there are two different ways of going from the estimands to moments:

$$\theta^{-1} = \mathrm{E}[Y_1], \quad \text{and} \quad 2\theta^{-2} = \mathrm{E}[Y_1^2].$$

These two routes then deliver two different methods of moment estimators of the same estimand θ:

$$\hat{\theta}^{-1} = \frac{1}{n}\sum_{j=1}^{n} Y_j, \quad \text{and} \quad 2\tilde{\theta}^{-2} = \frac{1}{n}\sum_{j=1}^{n} Y_j^2.$$

Rewriting, we have

$$\hat{\theta} = \frac{1}{\overline{Y}}, \quad \text{and} \quad \widetilde{\theta} = \sqrt{\frac{2}{\frac{1}{n}\sum_{j=1}^{n} Y_j^2}}.$$

So a method of moments estimator may not be unique. Which of these estimators is better? Later in this course, we will discuss principled approaches to deciding between estimators.

Example 2.4.6 (Pareto MoM)**.** The *Pareto distribution* is widely used as a model of extremes beyond some large threshold $t > 0$. For example, in 2017 the economist Sir Anthony B. Atkinson used the Pareto distribution to describe the upper tail of the income distribution in the U.K., in order to study income inequality and to estimate the socially optimal income tax rates. It is also used in the statistical analysis of athletics records and flooding. The Pareto(θ, t) density is

$$f_{Y_1}(y;\theta) = \frac{\theta t^\theta}{y^{\theta+1}}, \quad y > t,$$

where the parameter $\theta > 0$ is called the *tail index* and the parameter $t > 0$ is called the *threshold*. Let $Y_1, \ldots, Y_n$ be i.i.d. Pareto(θ, t). One interesting feature of this distribution is that

$$\mathrm{E}(Y_1^r) = \frac{\theta t^r}{\theta - r}$$

for $r < \theta$, but $\mathrm{E}(Y_1^r)$ is infinite for $r \geq \theta$. Thus, the tail index controls how many moments exist.

Suppose that the estimand is θ, which is unknown, and that t is known. At first it seems that a method of moments estimator is impossible to construct, since the first moment of Y_1 may not even exist (and whether it does depends on something that is unknown). However, using a change of variables calculation reveals that

$$\log(Y_1/t) \sim \mathrm{Expo}(\theta),$$

which implies that

$$\mathrm{E}[\log(Y_1/t)] = \theta^{-1}.$$

Replacing moments by sample moments and estimands by estimators delivers

$$\hat{\theta}^{-1} = \frac{1}{n}\sum_{j=1}^{n} \log(Y_j/t),$$

and, ultimately, the method of moments estimator for θ:

$$\widehat{\theta} = \frac{1}{\frac{1}{n}\sum_{j=1}^{n} \log(Y_j/t)}.$$

2.4.3 Multiple parameter version*

More abstractly, the case of general multidimensional estimands can be expressed compactly. If $\boldsymbol{\theta}$ is a K-dimensional vector and

$$\alpha(\boldsymbol{\theta}) = \mathrm{E}[h(Y_1)]$$

where h and α are functions whose inputs are K-dimensional vectors and whose outputs are K-dimensional vectors, then the K-dimensional $\hat{\boldsymbol{\theta}}$ which solves the K equations

$$\alpha(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{j=1}^{n} h(Y_j),$$

is the method of moment estimator. Again this is only immediately useful if the function α is invertible, in which case

$$\hat{\boldsymbol{\theta}} = \alpha^{-1}\left(\frac{1}{n}\sum_{j=1}^{n} h(Y_j)\right).$$

Example 2.4.7. For simplicity, assume the random variables X_1 and Y_1 each have a zero mean. Set the estimand as

$$\boldsymbol{\theta} = \begin{pmatrix} \frac{\mathrm{E}[X_1 Y_1]}{\mathrm{E}[X_1^2]} \\ \mathrm{E}[X_1^2] \end{pmatrix}.$$

Notice that the term $\mathrm{E}[X_1Y_1]/\mathrm{E}[X_1^2]$ is $\beta_{Y_1 \sim X_1}$, the regression of Y_1 on X_1 from Section 1.5, when the means are zero.

Assume the bivariate random vectors $\{(X_j, Y_j), j = 1, \ldots, n\}$ are i.i.d. pairs. Now build a bivariate statistic

$$T(\mathbf{X}, \mathbf{Y}) = \begin{pmatrix} \frac{1}{n}\sum_{j=1}^{n} X_j Y_j \\ \frac{1}{n}\sum_{j=1}^{n} X_j^2 \end{pmatrix}.$$

Then

$$\mathrm{E}[T(\mathbf{X}, \mathbf{Y})] = \begin{pmatrix} \mathrm{E}[X_1 Y_1] \\ \mathrm{E}[X_1^2] \end{pmatrix} = \begin{pmatrix} \frac{\mathrm{E}[X_1 Y_1]}{\mathrm{E}[X_1^2]} \times \mathrm{E}[X_1^2] \\ \mathrm{E}[X_1^2] \end{pmatrix} = \begin{pmatrix} \theta_1 \times \theta_2 \\ \theta_2 \end{pmatrix} = \alpha(\boldsymbol{\theta}).$$

Replacing expectations by sample quantities and the estimand by the method of moment estimator yields

$$\begin{pmatrix} \frac{1}{n}\sum_{j=1}^{n} X_j Y_j \\ \frac{1}{n}\sum_{j=1}^{n} X_j^2 \end{pmatrix} = \begin{pmatrix} \widehat{\theta}_1 \times \widehat{\theta}_2 \\ \widehat{\theta}_2 \end{pmatrix} = \alpha(\widehat{\boldsymbol{\theta}})$$

which solves out to

$$\widehat{\theta}_1 = \frac{\sum_{j=1}^{n} X_j Y_j}{\sum_{j=1}^{n} X_j^2} \quad \text{and} \quad \widehat{\theta}_2 = \frac{1}{n}\sum_{j=1}^{n} X_j^2.$$

Therefore, $\widehat{\theta}_1$ is a method of moments estimator. Chapter 6 will detail these types of *descriptive regressions*.

2.5 Recap

This chapter covered a great deal of ground. It introduced statistical models as the joint distribution of data, while we wish to learn the estimand. Parametric statistical models are joint distributions

indexed by a parameter. Typically in parametric models the estimand is the parameter. Using that structure, the parameter can be potentially inferred using the likelihood: directly through maximum likelihood estimation or in combination with a prior via Bayes' theorem.

The main concepts covered in this chapter are listed in Table 2.1. A clear distinction was drawn

Formula or idea	Description or name
$F_{\mathbf{Y};\theta}$	parametric model
θ	parameter
$\theta \in \Theta$	parameter space
$L(\theta;\mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y};\theta)$	likelihood
$\pi(\theta\|\mathbf{y}) \propto L(\theta;\mathbf{y})\pi(\theta)$	posterior
$\log L(\theta;\mathbf{y})$	log-likelihood
$L(\psi;\mathbf{y}) = L(\theta;\mathbf{y})$	reparameterization
θ	estimand
$\hat{\theta} = T(\mathbf{Y})$	estimator
$\hat{\theta} = T(\mathbf{y})$	estimate
$\text{bias}(\hat{\theta}) = \text{E}[\hat{\theta}] - \theta$	bias
$\text{SE}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$	standard error
replace expectations by averages, estimand by estimator	method of moments

Table 2.1: Main ideas and notation in Chapter 2.

between the estimand (what you want to learn), the estimator, and the estimate. This triple is absolutely essential in modern statistics. Finally, we introduced the method of moments, which replaces estimands with estimators and theoretical expectations with sample averages.

Bayesian methods, maximum likelihood estimation, and the method of moments are three of the most useful general-purpose estimation strategies used in statistics.

2.6 R, models, likelihood, estimation, and method of moments

Plotting data and results is enormously important in statistics, so you would expect that R has some sophisticated graphing functions. Section 2.6.1 provides a start to learning how to plot data in R, using simulated data to illustrate the main approaches. Section 2.6.2 shows how to read in data from a csv (comma seperated values) file and to save data to a csv file. The third subsection sets up a dataframe, which is a common data structure in R, used by many statistical functions within R. The

final subsection collects the code which produced the output used in various examples in this chapter.

2.6.1 Plotting of statistical data

The core elements of R provide strong ways of plotting statistical data, either via built-in commands or using a package such as ggplot2.

Table 2.2 provides a list of some of the most useful core plotting functions. Each function has many options which allow the user to specialize the plotting, e.g., adding axis labeling.

These functions are illustrated by again simulating data and then plotting a cross plot of data in x against y and then drawing a line through the data. We repeat this using a different plotting symbol and size.

X-Y plot and adding a straight line

```
set.seed(111)  # sets the seed of the random number generator
n = 400 # sample size
x = rt(n,4) # draw n values from t with 4 degrees of freedom
y = rnorm(n,0.2+0.9*x,0.3) # get draws from N(0.2+0.9x_j,0.3^2)

plot(x,y,xlab="x var",ylab="y var",main="Main")
abline(a=0.2,b=0.9,col="red")

# change plotting symbol, using small (cex=0.4) dots (pch=16)

plot(x,y,xlab="x var",ylab="y var",main="Main",pch=16,cex=0.4)
abline(a=0.2,b=0.9,col="red")
```

The plots will be displayed on the screen. You can use the R system to then save the plot. Importantly, the basic plot `plot()` is followed by `abline(a=0.2,b=0.9,col="red")` which generates a red line with an intercept of 0.2 and slope of 0.9. This line is placed on top of the original plot. Hence it is useful to think of this second command as decorating the `plot()` output. You can add other lines until you get bored. A new picture will appear when a new `plot()` command is issued.

If you want to save the plot through the R, declare a filename and then close it. The snippet below does this, in this case saving the picture to a pdf file.

Save a plot to a PDF file

```
pdf("SimplePlot.pdf")
  plot(x,y,xlab="x var",ylab="y var",main="Main",pch=16,cex=0.4)
  abline(a=0.2,b=0.9,col="red")
dev.off(); # device off, i.e., close the file
```

The right-hand side of Figure 2.4 shows the `SimplePlot.pdf`, while the left-hand side shows

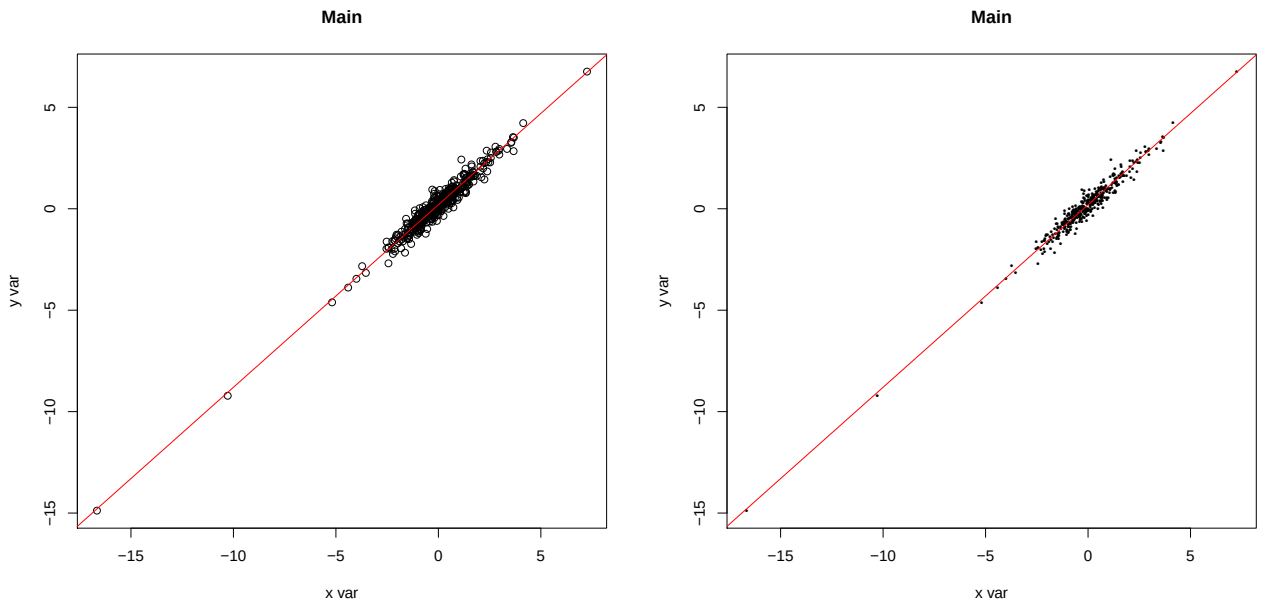


Figure 2.4: Figures produced by R using the `plot` function.

the same plot but without the use of the options `pch=16,cex=0.4` — so the plotting symbols are moderately sized circles (rather than small dots).

As discussed in Chapter 1, it is important to note that the file is saved to the working directory. The functions `getwd()` and `setwd()` are useful for checking and setting the working directory.

The function `curve` is a simple way of plotting likelihood functions against the parameter.

2.6.2 Reading and saving data

Reading and saving data is one of the most important and often frustrating aspects of applied statistics and data science. It feels trivial when it works and frustrating when it does not. In Windows, to read data from a file

```
"C:\Users\Neil\annual.csv"
```

into X, we can use

Reading in file under Windows

```
X = read.csv(file="C:\\Users\\Neil\\annual.csv", header=TRUE, sep=",");
```

and save data from X into a file

```
"C:\Users\Neil\MyData.csv"
```

using

Command	What it does
`plot(x)`	x plotted against index
`plot(x,type="l")`	x line plot against index
`points(x)`	add points to plot
`lines(x,y)`	add line to plot
`abline(h=3,col="red")`	add red horizontal line at height 3
`abline(v=1,col="blue")`	add blue vertical line at 1
`text(x,y,sNames,cex=0.7)`	add labels in "sNames", font size 0.7
`plot(x,y,xlab="x",ylab="y",main="Main")`	y against x plot
`curve(f,from=a,to=b)`	plots function $f(x)$ from $x = a$ to $x = b$
`pdf("out.pdf"); plot(x); dev.off();`	save plot as PDF in "out.pdf"
`jpeg("out.jpeg"); plot(x); dev.off();`	save plot as jpeg in "out.jpeg"
`png("out.png"); plot(x); dev.off();`	save plot as png in "out.png"
`arrows(x,y,z,w,length=0.1,angle=30)`	add arrows from x,y pairs to z,w pairs
`par(mfrow=c(2,3))`	figure is a 2 by 3 set of plots

Table 2.2: Basic plotting facilities in the core R language.

```
write.csv(X,"C:\\Users\\Neil\\MyData.csv", row.names = FALSE)
```

If no path is given, R will assume the file is in the "working directory". The strange \\ is needed for R to find the directory paths because of a collision with how R interprets backslashes.

On a Mac, paths are written using forward slashes rather than backslashes, which makes things simpler. For example, if the file is on the desktop, we can load it with

Reading in file on a Mac

```
X = read.csv(file="~/Desktop/annual.csv", header=TRUE, sep=",");
```

and we can save the data from X into a file on the desktop with

```
write.csv(X,"~/Desktop/MyData.csv", row.names = FALSE);
```

2.6.3 Data frames

For substantial applied projects it is helpful to be careful about labeling of variables and cases. R does this through a *data frame*. Many R packages assume that data is presented to their functions through a data frame. Roughly, data frames are "matrices" with rows and columns labeled by the names of variables and cases. But unlike in a matrix, the columns in a data frame can be of different

data types. For example, there can be columns where the entries are strings rather than numbers. The string aspect will appear in a moment, but first consider an example that looks like a standard matrix.

A data frame preinstalled in R is "mtcars", which gives the performance features of 32 cars that were measured in 1974.

Accessing existing dataframe

```
> head(mtcars)  # outputs first 5 cars
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
> nrow(mtcars)   # number of cars
[1] 32
> ncol(mtcars)    # number of features
[1] 11
> mtcars["Volvo 142E", "wt"] # weight of Volvo 142E, in 1000 lbs
[1] 2.78
```

One of the nicest data frame commands in this context is `help(mtcars)`, which documents the dataset.

How can we create a data frame? The example below is a data frame from imaginary students in Stat 111, allowing us to see how they performed in the course and how that relates to their performance in Stat 110. The notation "NA" is for "not available", i.e., the data is missing. In this case it could be because the student has not yet declared their concentration or does not have a secondary field.

Example of setting up a dataframe

```
>name = c("Josh","Jose","Joelle","Jane")
>state = c("CA","TX","MA","GA")
>concentration = c("Econ","Hist","Stat",NA)
>secondary = c("CS",NA,"Classics",NA)
>score111 = c(57,63,92,89)
>score110 = c(86,54,76,78)

>stat111.df = data.frame(name,state,concentration,secondary,
    score111,score110)
>stat111.df
    name state concentration secondary score111 score110
1   Josh    CA          Econ        CS       57       86
2   Jose    TX          Hist      <NA>       63       54
3 Joelle    MA          Stat   Classics       92       76
```

```
4    Jane     GA           <NA>       <NA>        89        78
>c(mean(stat111.df[,"score111"]),mean(stat111.df[,"score110"]))
[1] 75.25 73.50
```

2.6.4 Code for various examples in this chapter

Example 2.2.6 shows the impact of reparameterization for the likelihood for the springs data.

Code for Example 2.2.6

```
Y = c(225,171,198,189,135,162,135,117,162); n=9 # load data
nR = 800; nMin = 80 # for plotting
mAns = matrix(nrow=nR,ncol=3) # storage

for (i in 1:nR){  # compute logL for various parameters values
  mTheta = 0.00002*i # loop over different parameters
  mPsi = 1.0/mTheta
  logL = n*log(mTheta) - mTheta*sum(Y) #log L
  mAns[i,] = c(mTheta,logL,mPsi);
}

# mTheta version

#pdf("Lspring.pdf");
   plot(mAns[c(nMin:nR),1],
    exp(mAns[c(nMin:nR),2]-max(mAns[c(nMin:nR),2])),type="l",
        main="Likelihood against theta",
        xlab="Theta",ylab="Likelihood")
#dev.off()

#pdf("logLspring.pdf")
   plot(mAns[c(nMin:nR),1],mAns[c(nMin:nR),2],type="l",
    main="logL against theta", xlab="Theta",ylab="logL")
#dev.off()

# mPsi version

#pdf("LspringPsi.pdf")
   plot(mAns[c(nMin:nR),3],
    exp(mAns[c(nMin:nR),2]-max(mAns[c(nMin:nR),2])),type="l",
        main="Likelihood against psi",xlab="Psi",ylab="Likelihood")
#dev.off()

#pdf("logLspringPsi.pdf")
   plot(mAns[c(nMin:nR),3],mAns[c(nMin:nR),2],type="l",
    main="logL against psi", xlab="Psi",ylab="logL")
#dev.off()
```

Chapter 3

Loss Functions, Bias-Variance Tradeoff, and Asymptotics

3.1 Bias and variance of sample p-quantiles

In statistical inference, we hope that the estimator $\hat{\theta}$ will be close to the estimand θ. In Chapter 2, we introduced two important measures of the quality of an estimator:

$$\text{bias}(\hat{\theta}) = \text{E}[\hat{\theta}] - \theta \quad \text{and} \quad \text{SE}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}.$$

Recall from Chapter 1 that the sample p-quantile is a useful descriptive statistic. What are the bias and standard error of the sample p-quantile?

Focus on the case where the sample is i.i.d. random variables which are Uniform! Then the sample p-quantile will have a small bias, but otherwise is nicely behaved. Although Uniforms are cute, this might seem like an excessively specialized result. However, the result will later allow us to work out what happens for *all* continuous random variables, using the universality of the Uniform (see Section 5.3 of the Stat 110 book).

Example 3.1.1 (Sample quantiles of Uniform data)**.** Let $Y_1, \ldots, Y_n$ be i.i.d. Uniform random variables on $[0, 1]$. Then

$$Y_{(\lceil np \rceil)}$$

is the sample p-quantile estimator of the estimand $Q_{Y_1}(p)$, and $Q_{Y_1}(p) = p$ since the Unif$(0, 1)$ CDF is $F_{Y_1}(u) = u$ (recall that for a Uniform r.v., probability is proportional to length). As shown in Section 8.6 of the Stat 110 book,

$$Y_{(j)} \sim \text{Beta}(j, n - j + 1).$$

So

$$\text{E}[Y_{(j)}] = \frac{j}{n+1}, \quad \text{and} \quad \text{Var}[Y_{(j)}] = \frac{j\,(n-j+1)}{(n+1)^2\,(n+2)},$$

which implies, setting $j = \lceil np \rceil$, that

$$\text{bias}(Y_{(\lceil np \rceil)}) = \frac{\lceil np \rceil}{n+1} - p, \quad \text{and} \quad \text{Var}(Y_{(\lceil np \rceil)}) = \frac{\lceil np \rceil \left(n - \lceil np \rceil + 1\right)}{(n+1)^2 (n+2)}.$$

Therefore, the sample p-quantile is biased. Note that the ceiling function satisfies

$$x \leq \lceil x \rceil < x + 1.$$

Plugging in $x = np$, this implies that

$$-\frac{p}{n+1} \leq \frac{\lceil np \rceil}{n+1} - p < \frac{1-p}{n+1}.$$

So if n gets large, we have, for some constants a and b,

$$\text{bias}(Y_{(\lceil np \rceil)}) \approx \frac{b}{n}, \quad \text{and} \quad \text{Var}(Y_{(\lceil np \rceil)}) \approx \frac{p(1-p)}{n} = \frac{a}{n},$$

where we are using the fact that $\lceil np \rceil / n \to p$ as $n \to \infty$. Hence both the bias and the standard error go to 0 as n grows, but the bias goes to 0 at a faster rate (since the standard error is approximately on the order of $1/\sqrt{n}$ rather than $1/n$).

In the above example, we derived the mean and variance of the sample quantile of i.i.d. Unif$(0, 1)$ data. There is also a nice result for the asymptotic *distribution* of the sample quantile.

Theorem 3.1.2 (Asymptotic distribution of sample quantile)**.** *If $Y_1, \ldots, Y_n$ be i.i.d. Uniform random variables on $[0, 1]$ and $p \in (0, 1)$. Then as $n \to \infty$,*

$$\sqrt{n}\{Y_{(\lceil np \rceil)} - Q_{Y_1}(p)\} \xrightarrow{d} \mathcal{N}(0, p(1-p)). \tag{3.1}$$

The proof of this theorem uses the change of variables formula and Taylor approximations. The details of the proof get pretty technical, so the proof is in the starred Section 3.8.

3.2 Bias and variance are sometimes in conflict

We would like to construct estimators that have small bias *and* small variance. But in many areas of statistics and machine learning, bias and variance are in direct conflict. To resolve this conflict we will need some new ideas.

3.2.1 Nonparametric density estimator

We will illustrate the potential conflict between the bias and variance in the context of nonparametric estimation of the density function $f_Y(y)$.

Previously we introduced the empirical CDF

$$\hat{F}(y) = \frac{1}{n}\sum_{j=1}^{n} I\left(Y_j \leq y\right).$$

as a way to nonparametrically unbiasedly estimate a CDF $F(y) = \mathrm{E}[I(Y \leq y)]$, but the empirical CDF is always a *discrete* CDF: a step function, flat except that it jumps each time it reaches one of the data points, starting at 0 and ending at 1. This is illustrated in Figure 3.1.

Here though we are trying to estimate the *density*, not the CDF. We know for a continuous distribution that the derivative of the CDF is the PDF, so one idea would be to estimate the CDF and then take the derivative of the estimated CDF. But taking the derivative of the ECDF is utterly useless for estimating a density: the derivative of the ECDF is 0 everywhere, except for being undefined at the data points!

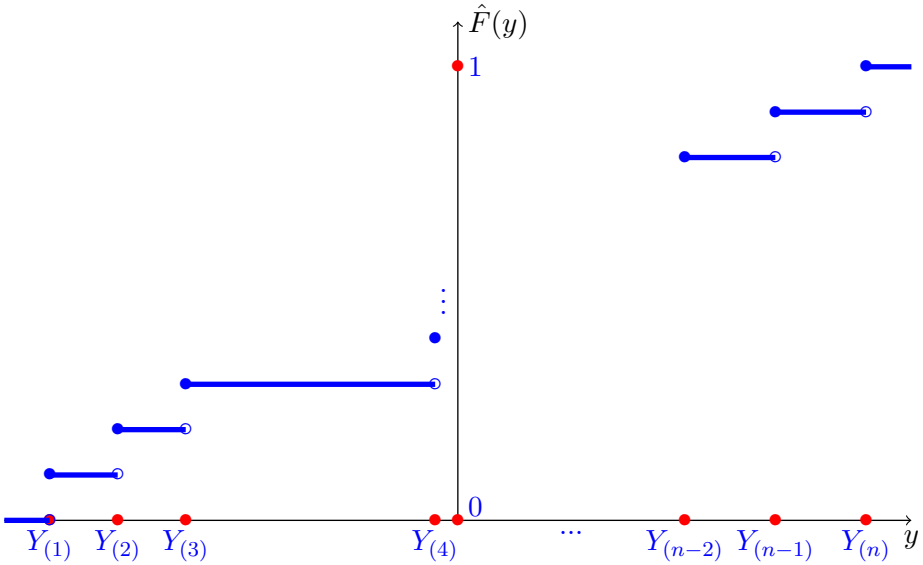


Figure 3.1: Empirical distribution function. On the x-axis we display the ordered data $Y_{(1)} \leq Y_{(2)} \leq \ldots \leq Y_{(n)}$.

For estimating the density of a *discrete* distribution, looking at a *histogram* is a natural approach to estimating the density (which is the PMF). For continuous distributions, histograms are still useful, but limited by being blocky rather than smooth. Instead, we will introduce *kernel density estimation*, which is a widely used method for density estimation in the continuous case. Unlike a histogram, kernel density estimation gives a smooth curve as an estimate for a smooth density.

Definition 3.2.1 (Kernel density estimator)**.** Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. with a CDF $F_{Y_1}(y)$ and PDF $f_{Y_1}(y)$. Let the estimand be $\theta = f_{Y_1}(y)$ for some value of y, and let $h > 0$. Let

$$\hat{\theta} = \frac{1}{hn} \sum_{j=1}^{n} I\left(Y_j \in (y - h/2, y + h/2]\right).$$

The estimator $\hat{\theta}$ is called the *kernel density estimator* (KDE) of $f_{Y_1}(y)$, with *rectangular kernel*. The number h is called a *bandwidth*. More generally, let K be a nonnegative function, called the *kernel function*. Then the kernel density estimator for this choice of K is

$$\hat{\theta} = \frac{1}{hn} \sum_{j=1}^{n} K\left(\frac{Y_j - y}{h}\right).$$

A widely used kernel aside from the rectangular kernel is to let K be the standard Normal PDF. In this book, by default when discussing KDE we will assume we are using the rectangular kernel.

In terms of the ECDF, we can write the KDE as

$$\hat{\theta} = \frac{\hat{F}(y + h/2) - \hat{F}(y - h/2)}{h}.$$

Taking the limit as $h \to 0$ on the right-hand side would yield the definition of the derivative of $\hat{F}$ at y. As noted earlier, the derivative of the ECDF is not useful, but here we fix h rather than letting $h \to 0$. Thus, the KDE can be thought of as a way to get an approximate notion of "slope" for the ECDF without actually taking the derivative.

As an alternative way to interpret $\hat{\theta}$, note that the KDE counts the number of data points in an interval of length h around y, dividing the count by h and n. Recall from Intuition 5.1.8 in the Stat 110 book that for small h we can think of $f_{Y_1}(y)h$ intuitively as being approximately the probability of Y_1 being in an interval of length h, centered at y. It then makes sense to estimate $f_{Y_1}(y)h$ by the *proportion* of data points in such an interval, which again gives the KDE.

A natural and important question for anyone who wants to use KDE is how to choose the bandwidth h to make the estimator as "good" as possible. It turns out that a conflict between bias and variance arises when choosing h.

Example 3.2.2 (Tradeoff in choice of bandwidth in KDE)**.** Let $Y_j \sim \mathcal{N}(11, 4^2)$ and $n = 100$, and simulate some i.i.d. data $y_1, \ldots, y_n$. The left hand side of Figure 3.2 plots $\hat{\theta}$ against y, as well as the true density function (black line). The `R` code for this is given in Section 3.11. To implement this we use $h = 5.0$ (blue line), which is 1.25 standard deviations of Y_1, and $h = 0.5$ (red line), 1/8 of a standard deviation. The small bandwidth estimator is very jagged (indicating high variance). Both

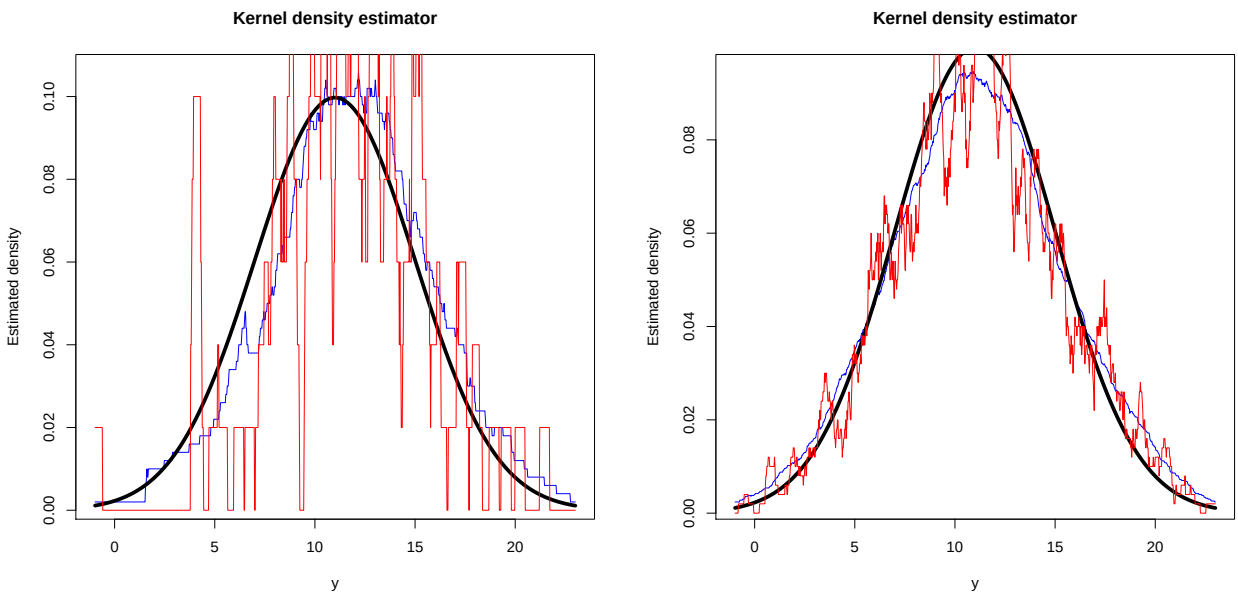


Figure 3.2: True $\mathcal{N}(11, 4^2)$ density (black line) and kernel density estimators with $h = 5.0$ (blue line) and $h = 0.5$ (red line). The left-hand side has $n = 100$ and the right-hand side has $n = 1000$.

estimators behave much better when the sample size increases to $n = 1000$, as shown on the right-hand side of Figure 3.2.

The following theorem gives information about the bias and standard error of the KDE, which in turn can help us to select a good value of h.

Theorem 3.2.3 (KDE bias and SE)**.** *Let the random variables $Y_1, Y_2, \ldots, Y_n$ be i.i.d. and $f_{Y_1}(y)$ be at least twice differentiable in y. Then for small h,*

$$\text{bias}[\hat{\theta}] \approx \frac{1}{24} h^2 f''_{Y_1}(y) \quad \textit{and} \quad \text{SE}(\hat{\theta}) \approx \sqrt{\frac{1}{nh} f_{Y_1}(y)}.$$

The proof, which (as with so many proofs in statistics) is based on Taylor approximation, is in Section 3.8. Note that the bias of $\hat{\theta}$ falls as h gets small, but the variance increases. This is an example of a *bias-variance tradeoff*. So how should we choose h? In the next section, we start to investigate the question of how we should handle this kind of tradeoff.

3.3 Loss functions, risk, and mean square error

In order to be more quantitatively precise at trading bias against variance, and in assessing the performance of estimators in general, we need ways to measure how well an estimator is doing. If one person says "I love the MLE!" but another says "I love being unbiased!", is this just a matter of

opinion, completely subjective, or are there principled ways to decide when to prefer one estimator to another?

One approach to answering this question is to say that before we can talk about one estimator being better than another, we need to think more broadly about *decision theory* or *decision analysis*. This sets up decisions using the following somewhat abstract language, which can be used here and in different contexts later:

- the *action* we are considering;
- the *loss* that will be incurred if the action is taken;

Then we can try to select the action that will minimize our expected loss. Expected loss will appear in many contexts in this book, e.g., where the actions are a prediction and the loss being the distance between the outcome and a prediction. But for now, we focus on decision theory in the context of selecting an estimator. For estimation, the "action" is selecting an estimator $\hat{\theta}$ of the estimand θ and the "loss" is the implications of $\hat{\theta}$ being distant from θ.

Definition 3.3.1 (Loss function)**.** A *loss* function is a function

$$\text{Loss}(\theta, \hat{\theta}),$$

interpreted as the loss or cost associated with using the estimate $\hat{\theta}$ when the estimand is θ. We require $\text{Loss}(\theta, \hat{\theta}) \geq 0$ and $\text{Loss}(\theta, \theta) = 0$ (the best case scenario, in which no loss is incurred, is when the estimate is exactly correct). An example of a loss function is drawn in Figure 3.3.

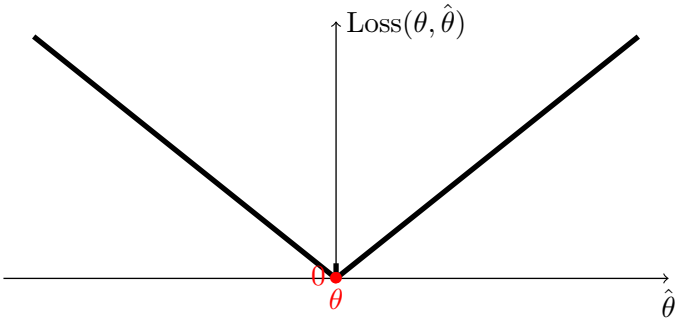


Figure 3.3: Example of a loss function, $\text{Loss}(\theta, \hat{\theta})$ plotted against $\hat{\theta}$. Here $\text{Loss}(\theta, \hat{\theta}) = |\hat{\theta} - \theta|$.

We take a frequentist approach for now, treating θ as an unknown constant (not having a distribution). The estimator $\hat{\theta} = T(\mathbf{Y})$ is, of course, a random variable. So $\text{Loss}(\theta, \hat{\theta})$ is also a random variable, as it is a function of $\hat{\theta}$ (for each fixed θ).

The *risk* of an estimator is its expected loss (averaged over all possible datasets). If we are able to compute the risks of various estimators we are considering, then we can try to select an estimator whose risk tends to be low.

Definition 3.3.2 (Risk function)**.** For the estimator $\hat{\theta} = T(\mathbf{Y})$, the *risk function* is the expected loss

$$\text{Risk}(\theta) = \text{E}_\theta[\text{Loss}(\theta, \hat{\theta})] = \int \text{Loss}(\theta, T(\mathbf{y})) f_{\mathbf{Y};\theta}(\mathbf{y}) d\mathbf{y},$$

where the integral is over the support of $\mathbf{Y}$.

Of course, when the true value of θ is unknown, the actual risk $\text{Risk}(\theta)$ is also typically unknown. So it may not be obvious how to choose between two estimators based on their risk functions if the risk functions, when plotted as a function of θ, crisscross. But the risk function may still be a useful tool for getting a sense of the performance of an estimator.

Before considering risk further, we need to specify the loss function $\text{Loss}(\theta, \hat{\theta})$. In some applications of decision theory, the loss is a literal cost, measurable in dollars or human lives. Many industrial and business decisions have this structure.

☣ **3.3.3.** In research problems it is often hard to foresee the potential benefits and losses from gaining more precise knowledge or overthrowing existing paradigms. For example, consider a problem involving measuring quantum effects in physics. It is far from clear what a suitable loss function for this problem would be.

In working out good general-purpose estimators more abstractly, though, it helps to have a more generic way to measure costs. We will look at a few standard ways to do so.

The most widely used loss function, at least when the parameter space is $\mathbb{R}$, is *squared error loss*. It is convenient mathematically, easy to write down, and dovetails well with the definition of variance.

Definition 3.3.4 (MSE)**.** The loss function

$$\text{Loss}(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$$

is called the *squared error loss*. Its expected value is the *mean square error* (MSE) of $\hat{\theta}$ is:

$$\text{MSE}(\hat{\theta}) = \text{E}_\theta[(\hat{\theta} - \theta)^2].$$

Sometimes the square root of the MSE is used instead, in order to have the units be the same as that of θ; this is called the *root mean square error* (RMSE).

Another natural, though less widely used, loss function is absolute error loss, which uses the *distance* between the estimate and the truth.

Definition 3.3.5 (MAE)**.** The loss function

$$\text{Loss}(\theta, \hat{\theta}) = |\hat{\theta} - \theta|$$

is called the *absolute error loss*. Its expected value is the *mean absolute error* (MAE) of $\hat{\theta}$ is

$$\text{MAE}(\hat{\theta}) = \text{E}_\theta[|\hat{\theta} - \theta|].$$

Absolute error loss punishes small errors more severely and large errors more gently than the RMSE. Which one of the two makes more sense in practice depends, of course, on the application. Mathematically, squared error loss is more pleasant to work with because the absolute value function is not differentiable at 0 and since there is so much geometry involving squared lengths, e.g., the Pythagorean theorem.

In some problems, the goal is simply to guess the correct value of θ, with no partial credit given for a wrong guess. Then we can consider a binary loss function that simply marks whether our guess is right or wrong.

Definition 3.3.6 (0–1 loss)**.** The loss function

$$\text{Loss}(\theta, \hat{\theta}) = I(\hat{\theta} \neq \theta)$$

is called *0–1 loss*. Note that its expected value is $\text{E}_\theta[\text{Loss}(\theta, \hat{\theta})] = P_\theta(\hat{\theta} \neq \theta)$.

All three of these special loss functions (squared error loss, absolute error loss, and 0–1 loss) handle overestimates and underestimates symmetrically: the cost of estimating θ as $\theta + c$ equals the cost of estimating θ as $\theta - c$. In the context of a specific problem, we may want an asymmetric loss function instead. Asymmetric losses are important in many real world situations (e.g., in designing a spam filter for email, putting a legitimate email in your spam folder may be more costly than putting a spam email in your inbox).

Definition 3.3.7 (Check loss)**.** The loss function

$$\text{Loss}(\theta, \hat{\theta}) = |\hat{\theta} - \theta|\{c_+ I(\hat{\theta} > \theta) + c_- I(\hat{\theta} < \theta)\}, \quad c_+, c_- \geq 0,$$

is called the *check loss*. It appears in Chapter 6 in connection with quantiles. If $c_+ > c_-$, then the check loss will give more weight to overestimation, while $c_+ < c_-$ racks up the losses if there is a great deal of underestimation.

We will discuss decision theory more later in the course; for now, we will focus on the most widely used general purpose loss function: squared error loss.

3.4 Bias-Variance tradeoff

We have already encountered examples where there is a tradeoff between bias and variance in estimation. This kind of situation is ubiquitous in statistics, machine learning, and data science. A quantitative version of the bias-variance tradeoff is given by the following identity.

3.4.1 Decomposing the mean square error

Theorem 3.4.1 (Bias-variance tradeoff)**.** *The mean square error of an estimator $\hat{\theta}$ for the estimand θ is the variance plus the square of the bias:*

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + \left[\mathrm{bias}(\hat{\theta})\right]^2.$$

Proof. Let $V = \hat{\theta} - \theta$. Then

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}) &= \mathrm{E}_\theta[V^2] \\
&= \mathrm{Var}_\theta(V) + (\mathrm{E}_\theta[V])^2 \\
&= \mathrm{Var}(\hat{\theta}) + \{\mathrm{E}(\hat{\theta}) - \theta\}^2 \\
&= \mathrm{Var}(\hat{\theta}) + \left[\mathrm{bias}(\hat{\theta})\right]^2.
\end{aligned}$$

∎

If you ever forget which term gets squared in the bias-variance decomposition, just think about the *units*. For example, if θ is an unknown distance measured in meters, then both variance and squared bias are measured in square meters, so it makes sense to add them.

Some observations about the bias-variance decomposition are below.

- For an unbiased estimator, the MSE equals the variance. For example, for i.i.d. data $Y_1, Y_2, \ldots, Y_n$ with mean μ and variance σ^2, $\bar{Y}$ is unbiased and $\mathrm{MSE}(\bar{Y}) = \sigma^2/n$.
- The necessary and sufficient condition for $\mathrm{MSE}(\hat{\theta})$ to go to zero as n goes to ∞ is that the variance and the absolute value of the bias go to 0 as $n \to \infty$.
- For many widely used estimators, such as the sample mean $\bar{Y}$, sample variance S^2, and sample p-quantiles, $\mathrm{Var}(\hat{\theta}) \approx a/n$ for n large and $\mathrm{bias}(\hat{\theta}) \approx b/n$ for n large, where n is the sample size

and a and b are constants. Since the bias gets squared in the MSE, the variance term matters more when n is large. Furthermore, for large n, the

$$\text{MSE}(\hat{\theta}) \approx a/n,$$

and as n gets large we have $\text{MSE}(\hat{\theta}) \to 0$.

Example 3.4.2 (Shrinking the sample mean towards 0)**.** Let $Y_1, \ldots, Y_n$ be i.i.d., and let the estimand be $\theta = \text{E}[Y_1]$. Consider the estimator

$$\hat{\theta} = c\bar{Y},$$

where $c \in [0, 1]$ is a known constant. The extreme case $c = 0$ gives the degenerate estimator that always estimates θ as 0, while the extreme case $c = 1$ gives the sample mean. Note that

$$|\hat{\theta}| = c|\bar{Y}| \leq |\bar{Y}|,$$

so $\hat{\theta}$ is closer to 0 than is $\bar{Y}$ (for $c < 1$ and $\bar{Y} \neq 0$). Statisticians call this kind of modification to an estimator *shrinkage towards zero*; it is a surprisingly powerful idea, as we will see in chapters 6 and 9. Now

$$\text{Var}(\hat{\theta}) = c^2 \frac{\text{Var}(Y_1)}{n},$$

and

$$\text{bias}(\hat{\theta}) = c\text{E}[\bar{Y}] - \text{E}[Y_1] = (c - 1)\theta.$$

Thus, the risk of $\hat{\theta}$ is

$$\text{E}_\theta[\text{Loss}(\theta, \hat{\theta})] = \text{MSE}(\hat{\theta}) = c^2 \frac{\text{Var}(Y_1)}{n} + (1 - c)^2\theta^2.$$

Figure 3.4 plots the the risk functions for $\bar{Y}$ (solid line) and $\hat{\theta}$ (dotted line). In terms of risk, $\hat{\theta}$ is better for small θ and $\bar{Y}$ is better for large θ. If you have a hunch that θ is small, but do not know its value, it might make sense to set $c < 1$ and deliberately induce bias, in order to suppress the variance. To formalize that, this risk can be found by noting that

$$\partial R(\theta)/\partial c = 2c\text{Var}(Y_1)/n - 2(1 - c)\theta^2.$$

Setting it to zero, the resulting "best" c is $\theta^2/\{\text{Var}(Y_1)/n + \theta^2\}$, which is close to 0 if θ is small.

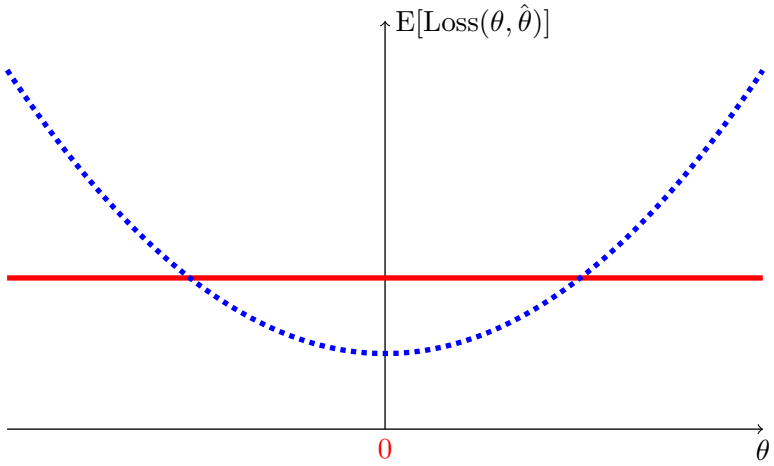


Figure 3.4: The risk function E[Loss$(\theta, \hat{\theta})$] plotted against θ, for two estimators: $\bar{Y}$ (solid line) and $\hat{\theta} = c\bar{Y}$ (dotted line), where $c \in [0, 1)$.

3.4.2 The KDE's MSE goes to 0, slowly

Recall that the kernel density estimator

$$\hat{\theta} = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h} I\left(Y_j \in (y - h/2, y + h/2]\right)$$

has mean and standard error of

$$\text{bias}[\hat{\theta}] \approx \frac{1}{24} h^2 f''_{Y_1}(y) \quad \text{and} \quad \text{SE}(\hat{\theta}) \approx \sqrt{\frac{1}{nh} f_{Y_1}(y)},$$

respectively. Thus the mean square error is

$$\text{MSE}(\hat{\theta}) \approx \frac{1}{24^2} h^4 \{f''_{Y_1}(y)\}^2 + \frac{1}{nh} f_{Y_1}(y).$$

It may seem alarming that n does not seem to appear in the first term: we hope that the MSE will go to 0 as $n \to \infty$, but why should the first term go to 0? Actually though, we can allow the choice of h to depend on n.

The right-hand side in the expression for MSE can be minimized by differentiating it with respect to h and setting $\tilde{h}$ as the value which solves

$$\frac{1}{144} \tilde{h}^3 \{f''_{Y_1}(y)\}^2 - \frac{1}{n\tilde{h}^2} f_{Y_1}(y) = 0.$$

This implies

$$n\tilde{h}^5 = 144 \frac{f_{Y_1}(y)}{\{f''_{Y_1}(y)\}^2},$$

so $\tilde{h} = cn^{-1/5}$, where c is a constant. So we should choose our bandwidth such that it goes to 0 very slowly as n grows. Feeding $\tilde{h}$ into the MSE, we get

$$\text{MSE}(\hat{\theta}) \approx \frac{1}{n\tilde{h}}\left[\frac{1}{24^2}n\tilde{h}^5\{f''_{Y_1}(y)\}^2 + f_{Y_1}(y)\right] = \frac{5}{4}\frac{f_{Y_1}(y)}{n\tilde{h}} = \frac{d}{n^{4/5}},$$

where d is a constant.

The bias and variance are in conflict, so optimally trading them off against one another leads to an overall MSE that does not fall to 0 with n at rate n^{-1} (as we see for typical estimators in parametric models such as like $\bar{Y}$, S^2, $\hat{Q}(p)$, $S_{X,Y}$, $r_{X,Y}$, and $b_{Y\sim X}$), but instead at the slower rate $n^{-4/5}$. Nonparametric methods have the major advantage of not requiring parametric assumptions but here we do pay a price, as seen in the slower decay rate for the MSE.

☣ **3.4.3.** If h varies with y, then the function $\widehat{f_n}(y)$ of y will typically not integrate to 1 over the range of y to one, even though the true density function must. Usually we want to estimate the entire density curve f so selecting a single h for all y is attractive. A natural measure of how good our estimate is is the *integrated* MSE, which essentially totals up the MSE across all points y:

$$\text{IMSE}(\widehat{f_n}) = \int_{-\infty}^{\infty} \text{MSE}(\widehat{f_n}(y))dy \approx \frac{1}{24^2}h^4\int_{-\infty}^{\infty}\{f''_{Y_1}(y)\}^2dy + \frac{1}{nh}.$$

Selecting $\tilde{h}$ over the whole range of y so that

$$n\tilde{h}^5 = 144\frac{1}{\int_{-\infty}^{\infty}\{f''_{Y_1}(y)\}^2dy},$$

we get

$$\text{IMSE}(\widehat{f_n}) \approx \frac{d_1}{n^{4/5}}$$

for some constant d_1. Again we see the $n^{-4/5}$ rate.

☣ **3.4.4.** Worse, this $n^{-4/5}$ rate is only the univariate case. If we nonparametrically estimate a d-dimensional density then you can show the best rate which MSE can go to zero is $n^{-4/(4+d)}$, which gets slower and slower as d increases! In statistics this is an example of the *curse of dimensionality*. Estimation gets materially harder for some procedures as the dimension of the problem increases.

This is a foundational point for modern statistics: using nonparametrics is not a free lunch. Some nonparametric methods are data hungry (although some are not, like the empirical CDF).

3.5 Consistency of estimators

Of course, we hope that our estimator will converge to the estimand as the sample size grows. An estimator with this property is called *consistent*.

Definition 3.5.1 (Consistency)**.** An estimator $\hat{\theta}$ is *consistent* for the estimand θ if $\hat{\theta}$ converges in probability to θ as the sample size $n \to \infty$, i.e., for every $\epsilon > 0$ we have

$$P(|\hat{\theta} - \theta| \geq \epsilon) \to 0$$

as $n \to \infty$. This is written in shorthand as

$$\hat{\theta} \xrightarrow{p} \theta.$$

Note that implicitly $\hat{\theta}$ depends on n. Sometimes it is clearer to write the dependence explicitly, with notation such as $\hat{\theta}_n$ that explicitly indicates the sample size n.

If we have a consistent estimator and a very large sample size, we can be confident that the estimate we compute using the data will be close to the truth. This is a rather minimal requirement to ask for when deciding on an estimator. An inconsistent estimator sometimes fails to get the right answer even with an *infinite* amount of data. In the real world, we always have a *finite* amount of data, but at the very least we should make sure we get the right answer as $n \to \infty$.

Example 3.5.2 (Some consistent estimators of the mean)**.** Let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. Let the estimand be $\mu = \mathrm{E}[Y_1]$, which we assume to be finite. Then the sample mean $\bar{Y}$ is a consistent estimator for μ by the weak law of large numbers (see Chapter 10 of the Stat 110 book).

But there are infinitely many other consistent estimators here too. For example, someone with a superstitious dislike of odd numbers could propose the estimator

$$\frac{1}{\lfloor n/2 \rfloor} \sum_{k=1}^{\lfloor n/2 \rfloor} Y_{2k},$$

which is the sample mean after discarding all the Y_i's with i odd. This is a very inefficient use of the data but it is still consistent, again by the law of large numbers. Another ridiculous but consistent estimator is

$$\bar{Y} + \frac{10^{100}}{\sqrt{n}}.$$

It is consistent since the first term converges in probability to μ and the second term converges to 0, but it has a massive bias for any sample size we are ever likely to be able to get in practice.

Example 3.5.3 (Consistency of ECDF)**.** Let $Y_1, \ldots, Y_n$ be i.i.d. r.v.s with CDF F and $\hat{F}$ be the empirical CDF. Then $\hat{F}(y)$ is a consistent estimator for $F(y)$ for each $y \in \mathbb{R}$. As shown in Chapter 1, the strong law of large numbers implies that, for each y,

$$\hat{F}(y) \to F(y) \text{ with probability 1.}$$

Better yet, an even stronger form of convergence holds. The *Glivenko-Cantelli theorem* says that the convergence of the ECDF to the CDF is *uniform* over all y, in the sense that

$$\sup_{y \in \mathbb{R}} |\hat{F}(y) - F(y)| \to 0 \text{ with probability 1.}$$

The "sup" here stands for "supremum", which is a generalization of the maximum (it is the maximum if the maximum exists, and in general it is the least upper bound).

A useful sufficient condition for consistency is that the MSE should go to 0 as the sample size $n \to \infty$.

Theorem 3.5.4 (Sufficient condition for consistency)**.** *If $\hat{\theta}$ is an estimator for the estimand θ and $\text{MSE}(\hat{\theta}) \to 0$ as $n \to \infty$, then $\hat{\theta}$ is consistent. In particular, since MSE is variance plus squared bias, to show that $\hat{\theta}$ is consistent it suffices to show that both the bias and the variance go to 0 as $n \to \infty$.*

Proof. Suppose that the MSE of $\hat{\theta}$ goes to 0. Then by Markov's inequality, for any $\epsilon > 0$ we have

$$P(|\hat{\theta} - \theta| \geq \epsilon) = P((\hat{\theta} - \theta)^2 \geq \epsilon^2) \leq \frac{\text{E}[(\hat{\theta} - \theta)^2]}{\epsilon^2} = \frac{\text{MSE}(\hat{\theta})}{\epsilon^2} \to 0.$$

∎

☣ **3.5.5.** It may be tempting to believe that $\hat{\theta} \xrightarrow{p} \theta$ implies that $\text{E}[\hat{\theta}] \to \theta$, since if two r.v.s are very likely to be very close to each other, then it may seem intuitively that their means should also be close. But this implication is false. For a counterexample, let the data be realizations of independent r.v.s $U, Y_1, \ldots, Y_n$ where $U \sim \text{Unif}(0,1)$ and $Y_j \sim \text{Bern}(\theta)$. Let

$$\hat{\theta} = \bar{Y} + nI(U \leq 1/n).$$

Then $\hat{\theta} \xrightarrow{p} \theta$, since $\bar{Y} \xrightarrow{p} \theta$ and for n large, the second term in $\hat{\theta}$ is very likely to be 0. But

$$\text{E}[\hat{\theta}] = \text{E}[\bar{Y}] + nP(U \leq 1/n) = \theta + 1.$$

As we mentioned earlier, many estimators have $\text{bias}(\hat{\theta}) \approx b/n$ and $\text{Var}(\hat{\theta}) \approx a/n$, which would mean that $\text{MSE}(\hat{\theta}) \approx a/n$. Other estimators have slower rates at which $\text{MSE}(\hat{\theta})$ converges to zero as n gets large. Table 3.1 gives some core examples of this for some descriptive statistics.

Recall from Stat 110 that a function of a random variable is a random variable, and that it is often useful to transform one random variable to another. In the statistical context, suppose that we have an estimator $\hat{\theta}$ which we know is consistent for estimating θ. But what if our estimand is $g(\theta)$ rather than θ? The obvious choice of estimator is then $g(\hat{\theta})$, but is this still consistent? The *continuous mapping theorem* (CMT) is a tool that helps answer such questions.

Statistics	Approximate MSE	Consistent?
sample mean	$\mathrm{Var}(Y_1)/n$	Yes
sample variance	$\mathrm{Var}[\{Y_1 - \mathrm{E}[Y_1]\}^2]/n$	Yes
sample covariance	$\mathrm{Var}[\{X_1 - \mathrm{E}[X_1]\}\{Y_1 - \mathrm{E}[Y_1]\}]/n$	Yes
sample p-quantile for uniform	$p(1-p)/n$	Yes
kernel density estimator	$d/n^{4/5}$	Yes
Y_1	$\mathrm{Var}(Y_1)$	No
$\frac{1}{2}(Y_1 + Y_2)$	$\mathrm{Var}(Y_1)/2$	No

Table 3.1: Consistency of some familiar summary statistics, shown by $\mathrm{MSE}(\hat{\theta}) \to 0$ as $n \to \infty$ and inconsistency for some silly estimators.

Theorem 3.5.6 (Continuous mapping theorem)**.** *Let $X, X_1, X_2, \ldots$ be a sequence of r.v.s and let g be a continuous function. If*

$$X_n \xrightarrow{p} X,$$

then

$$g(X_n) \xrightarrow{p} g(X).$$

Also, if

$$X_n \xrightarrow{d} X,$$

then

$$g(X_n) \xrightarrow{d} g(X).$$

In particular, it follows that if $\hat{\theta}$ is a consistent estimator for θ and g is a continuous function, then $g(\hat{\theta})$ is a consistent estimator for $g(\theta)$. Another useful property of convergence in probability is recorded below.

Theorem 3.5.7. *Let $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$. Then*

$$X_n + Y_n \xrightarrow{p} X + Y,$$

$$X_n - Y_n \xrightarrow{p} X - Y,$$

$$X_n Y_n \xrightarrow{p} XY,$$

and, if $P(Y_n = 0) = P(Y = 0) = 0$,

$$X_n / Y_n \xrightarrow{p} X/Y.$$

We now show consistency of many of the most widely used summary statistics: sample variance, sample covariance, sample standard deviation, regression, and sample p-quantile. Let $Y_1, Y_2, \ldots$ be i.i.d. r.v.s with a finite, positive variance. Extending this, let the pairs $(X_1, Y_1), (X_2, Y_2), \ldots$ be i.i.d. pairs of r.v.s, each with a finite, positive variance.

1. *Sample variance.* By the law of large numbers, $\bar{Y} \xrightarrow{p} \mathrm{E}[Y_1]$ and $\frac{1}{n}\sum_{j=1}^{n} Y_j^2 \xrightarrow{p} \mathrm{E}[Y_1^2]$. By the sum of squares identity,
$$S^2 = \frac{n}{n-1}\left(\frac{1}{n}\sum_{j=1}^{n} Y_j^2 - (\bar{Y})^2\right).$$
The $n/(n-1)$ factor goes to 1. Using the above results,
$$S^2 \xrightarrow{p} \mathrm{E}[Y_1^2] - (\mathrm{E}[Y_1])^2 = \mathrm{Var}[Y_1],$$
i.e., the sample variance consistently estimates the true variance.

2. *Sample covariance.* Again by the law of large numbers, $\bar{X} \xrightarrow{p} \mathrm{E}[X_1]$ and $\frac{1}{n}\sum_{j=1}^{n} X_j Y_j \xrightarrow{p} \mathrm{E}[X_1 Y_1]$. (It follows from the Cauchy-Schwarz inequality that $\mathrm{E}[X_1 Y_1]$ is finite.) Then
$$S_{X,Y} \xrightarrow{p} \mathrm{Cov}[X_1, Y_1],$$
i.e., the sample covariance consistently estimates the true covariance.

3. *Sample standard deviation.* We have $S^2 \xrightarrow{p} \mathrm{Var}(Y_1)$, so by the CMT,
$$S \xrightarrow{p} \sqrt{\mathrm{Var}(Y_1)}.$$

4. *Regression.* We have $S_X^2 \xrightarrow{p} \mathrm{Var}(X_1) > 0$ and $S_{X,Y} \xrightarrow{p} \mathrm{Cov}(X_1, Y_1)$, so
$$\frac{S_{X,Y}}{S_X^2} \xrightarrow{p} \frac{\mathrm{Cov}(X_1, Y_1)}{\mathrm{Var}(X_1)} = \beta_{Y \sim X}.$$

5. *Sample quantile.* Let the Y_j have a continuous, strictly increasing CDF F_{Y_1}. Setting $F_{Y_1}(y) = p$, we have
$$y = F_{Y_1}^{-1}(p) = Q_{Y_1}(p),$$
as illustrated in Figure 3.5. By universality of the Uniform,
$$U_j = F_{Y_1}(Y_j) \sim \mathrm{Unif}(0,1) \quad \text{and so} \quad Y_j = F_{Y_1}^{-1}(U_j) = Q_{Y_1}(U_j).$$
The ordering of the Uniforms and Ys are the same, as $Y_j = Q_{Y_1}(U_j)$, implying
$$Q_{Y_1}(U_{(\lceil np \rceil)}) = Y_{(\lceil np \rceil)}.$$

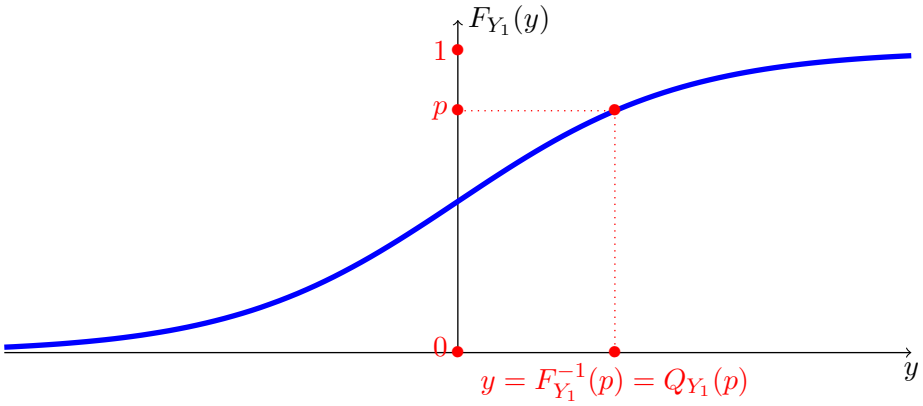


Figure 3.5: The cumulative distribution function $F_{Y_1}(y)$ drawn against y, showing the p-quantile $Q_{Y_1}(p)$.

The sample p-quantile, $U_{(\lceil np \rceil)}$, has a MSE which goes to 0 as n increases, so $U_{(\lceil np \rceil)} \xrightarrow{p} p$. As Q_{Y_1} is a continuous function we can use the continuous mapping theorem on $U_{(\lceil np \rceil)}$ to obtain the limit of $Y_{(\lceil np \rceil)}$. Hence

$$Y_{(\lceil np \rceil)} \xrightarrow{p} Q_{Y_1}(p),$$

the p-quantile of Y_1.

3.6 Large sample (asymptotic) approximations

Consistency tells us that the estimator $\hat{\theta}$ will eventually get close to the estimand θ when n is enormous. This is comforting, but it does not tell us the *rate* at which $\hat{\theta}$ gets close to θ, and it fails to satisfy what we said in Chapter 1:

> *A hallmark of statistics is that we aim not only to make statements based on data, but also to assess how confident we should be about those statements.*

To deliver on that promise, we need a more refined way to assess how close $\hat{\theta}$ is to θ when n is large, preferably a way that is also useful when n is moderate. Sometimes it is difficult to calculate the distribution or moments of $\hat{\theta}$ to understand how close it is to the estimand θ. In modern statistics, two forms of approximations to the distribution are commonly used:

- large sample (asymptotic) approximations;
- simulation.

A third, concentration inequalities, are also used, particularly in machine learning. We briefly discussed that topic in the starred section 3.9.

In this section we use large sample approximations, circling around the implications of the definition of convergence in distribution.

Definition 3.6.1. Let $X_1, X_2, \ldots$ be a sequence of random variables and F_{X_n} be the CDF of X_n. Let X be a random variable with CDF F_X. Then X_n *converges in distribution* to the random variable X if

$$F_{X_n}(x) \to F_X(x), \quad \text{for all } x \in \mathbb{R} \text{ such that } F_X \text{ is continuous at } x.$$

This is written in shorthand as $X_n \xrightarrow{d} X$.

In words, convergence in distribution says that if n is large then the distribution of X_n is roughly the same as X. In practice this is often used to get *approximations*. For example, we may want to find $P(a < X_n \leq b)$ but it may be far too difficult to do that analytically. But if $X_n \xrightarrow{d} X$ and n is large, then

$$P(a < X_n \leq b) = F_{X_n}(b) - F_{X_n}(a) \approx F_X(b) - F_X(a),$$

where hopefully working with the distribution of X is much more tractable than working with the distribution of X_n.

It is important to understand the distinction and relationship between convergence in probability and convergence in distribution. For example, in the weak law of large numbers we have convergence in probability, while in the central limit theorem we have convergence in distribution. Convergence in probability implies convergence in distribution, but not conversely except in the special case where X is a constant.

Theorem 3.6.2. *Let $X_n \xrightarrow{p} X$. Then $X_n \xrightarrow{d} X$. The converse is false in general. But if X is a constant c (i.e., X is a degenerate r.v. that always equals c), then $X_n \xrightarrow{p} X$ is equivalent to $X_n \xrightarrow{d} X$.*

We will focus on five main tools for deriving asymptotic distributions. Combining these powerful tools, we can obtain many useful asymptotic results. These tools are:

1. Law of large numbers (see Chapter 10 of the Stat 110 book);
2. Central limit theorem (see Chapter 10 of the Stat 110 book);
3. Continuous mapping theorem (see Theorem 3.5.6);
4. Slutsky's theorem (see Theorem 3.6.4);
5. Delta method (see Theorem 3.6.7).

For simple estimators, such as the sample average, we can sometimes directly establish the asymptotic distribution, typically using a central limit theorem (CLT) to obtain a result of the form

$$X_n = \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \omega^2).$$

For example, we know from the CLT that for i.i.d. Y_j with $\mathrm{Var}(Y_1) = \sigma^2 < \infty$, we have

$$\sqrt{n}(\bar{Y} - \mathrm{E}[Y_1]) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Often we want to obtain the asymptotic distribution of a more complicated estimator, built from simpler statistics.

3.6.1 Slutsky's Theorem

Suppose that our estimator is of the form $X_n + Y_n$, where we already have convergence in distribution results for X_n and for Y_n. A natural question then is whether we can combine these separate results into one for the sum.

☣ **3.6.3.** Suppose that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} Y$. It does *not* follow that $X_n + Y_n \xrightarrow{d} X + Y$. As a simple counterexample, let $X_n = Y_n \sim \mathcal{N}(0,1)$ and X, Y be i.i.d. $\mathcal{N}(0,1)$. Then $X_n + Y_n = 2X_n \sim \mathcal{N}(0,4)$, whereas $X + Y \sim \mathcal{N}(0,2)$. Clearly, the $\mathcal{N}(0,4)$ distribution does not converge to $\mathcal{N}(0,2)$. The problem is that the statement $X_n \xrightarrow{d} X$ is about the *marginal* distributions of X_n and X, and similarly for the statement $Y_n \xrightarrow{d} Y$, whereas the distribution of $X_n + Y_n$ depends heavily on the *joint* distribution of X_n and Y_n.

In the special case where Y_n converges to a *constant*, we do have a simple way to combine the convergence results for X_n and Y_n.

Theorem 3.6.4 (Slutsky's Theorem)**.** *If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is a constant, then*

1. $X_n + Y_n \xrightarrow{d} X + c$;
2. $X_n - Y_n \xrightarrow{d} X - c$;
3. $X_n Y_n \xrightarrow{d} cX$;
4. $X_n / Y_n \xrightarrow{d} X/c$, *if* $c \neq 0$.

Example 3.6.5 (Creating a parameter-free limiting distribution)**.** A core result in statistical inference can be established using Slutsky's Theorem. Suppose that $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \omega^2)$ and $\hat{\omega} \xrightarrow{p} \omega > 0$. Then we will show that

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{\hat{\omega}} \xrightarrow{d} \mathcal{N}(0,1).$$

Let $Z \sim \mathcal{N}(0,1)$, so that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \omega Z.$$

By Part 4 of Slutsky's theorem,

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{\hat{\omega}} \xrightarrow{d} \frac{\omega Z}{\omega} = Z,$$

as desired.

Note that the limiting distribution is *parameter-free*: we have the $\mathcal{N}(0,1)$ distribution not, say, the $\mathcal{N}(\theta, \theta^2)$ distribution. Statisticians call quantities which have parameter-free distributions *pivotal quantities*. Then, for example, if n is large then the random interval

$$\left[\hat{\theta} - 1.96\frac{\hat{\omega}}{\sqrt{n}}, \hat{\theta} + 1.96\frac{\hat{\omega}}{\sqrt{n}}\right],$$

will contain the estimand θ with probability approximately equal to $P(|Z| < 1.96) \approx 0.95$. This interval is an example of a *confidence interval*, a concept which we will discuss much more in Chapter 5. Having a confidence interval provides an answer to the promise we made above: not only do we estimate θ, but also we give an interval that helps quantify how uncertain we are about our estimate.

Example 3.6.6 (Asymptotic distributions of sample variance and sample covariance)**.** Slutsky's theorem is also useful for obtaining the asymptotic distributions of the sample variance S^2 and sample covariance $S_{X,Y}$. First use the sum of squares identity to decompose

$$\sqrt{n}\left[\frac{1}{n}\sum_{j=1}^{n}(Y_j - \bar{Y})^2 - \sigma^2\right] = \sqrt{n}\left[\frac{1}{n}\sum_{j=1}^{n}(Y_j - \mu)^2 - \sigma^2\right] - \sqrt{n}(\bar{Y} - \mu)^2.$$

But

$$\sqrt{n}(\bar{Y} - \mu) \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(Y_1))$$

by the CLT, so

$$\sqrt{n}(\bar{Y} - \mu)^2 \xrightarrow{p} 0.$$

So applying the Slutsky's theorem and a CLT on averages of squares, we have

$$\sqrt{n}\{S^2 - \mathrm{Var}[Y_1]\} \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(\{Y_1 - \mathrm{E}[Y_1]\}^2)).$$

Similar arguments give us that

$$\sqrt{n}\{S_{X,Y} - \mathrm{Cov}[X_1, Y_1]\} \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(\{X_1 - \mathrm{E}[X_1]\}\{Y_1 - \mathrm{E}[Y_1]\})).$$

3.6.2 Delta method

The continuous mapping theorem lets us establish a consistency result for $g(\hat{\theta})$ from a consistency result for $\hat{\theta}$. We now introduce the *delta method*, which gives us the asymptotic distribution of $g(\hat{\theta})$ from the asymptotic distribution of $\hat{\theta}$.

Theorem 3.6.7 (Delta method)**.** *Suppose that g is a differentiable function and*

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \omega^2).$$

Then

$$\sqrt{n}\{g(\hat{\theta}) - g(\theta)\} \xrightarrow{d} \mathcal{N}\bigg(0, (g'(\theta))^2\omega^2\bigg).$$

As an approximation*, this says that*

$$g(\hat{\theta}) \dot{\sim} \mathcal{N}\bigg(g(\theta), (g'(\theta))^2\frac{\omega^2}{n}\bigg),$$

for n large.

Proof. If n is large, then $\hat{\theta}$ is close to θ (with high probability). Taylor expand $g(\hat{\theta})$ about θ, yielding the approximation

$$g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta),$$

as the higher order terms should be smaller as they involve squares, cubes, etc. of $\{\hat{\theta} - \theta\}$ which is going to zero at rate $n^{-1/2}$. Rearranging,

$$\begin{aligned}\sqrt{n}\{g(\hat{\theta}) - g(\theta)\} &\approx g'(\theta)\sqrt{n}\{\hat{\theta} - \theta\} \\ &\xrightarrow{d} g'(\theta)\omega Z,\end{aligned}$$

where $Z \sim \mathcal{N}(0, 1)$. ∎

☣ **3.6.8.** The delta method may look absurd at first sight: if $\hat{\theta}$ is Normal, why should nonlinear transformations of it be approximately Normal? For example, if $Z \sim \mathcal{N}(0, 1)$ then Z^2 is χ^2_1 and e^Z is Log-Normal(0,1). The χ^2_1 and Log-Normal(0,1) distributions are *skewed* and the possible values are *nonnegative*; clearly they are not Normal, nor are they approximately Normal.

The key is that the delta method is an *asymptotic* result, which assumes that n is large. It may yield very poor approximations if n is not large. Assuming though that n is large, there is a simple intuitive explanation for why the delta method holds. Since g is differentiable, if we zoom in enough it looks *linear*. We see this in Figure 3.6. And we *are* zooming in a lot when n gets large, since $\hat{\theta}$ will then be close to θ. With g being approximately linear when n is large, $g(\hat{\theta})$ is approximately a linear function of $\hat{\theta}$, and we know that a linear function of a Normal is Normal.

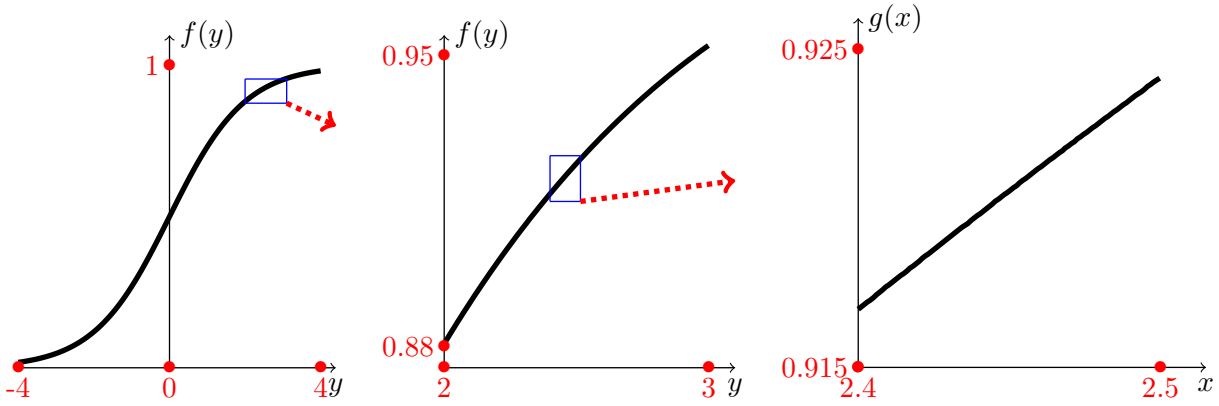


Figure 3.6: Zooming into the logistic function $g(x) = e^x/(1+e^x)$ against x, as we go from left to right (following the red dotted arrows, magnifying out the blue box), showing how it becomes approximately linear.

Example 3.6.9 (Estimating e^θ)**.** Let $Y_1, Y_2, \ldots$ be i.i.d. with mean θ and variance σ^2. Suppose that the estimand is $\psi = e^\theta$ rather than θ itself. The most obvious estimator to use for ψ is

$$\hat{\psi} = e^{\bar{Y}}.$$

This estimator is biased since by Jensen's inequality we have

$$\mathrm{E}[\hat{\psi}] = \mathrm{E}[e^{\bar{Y}}] > e^{\mathrm{E}[\bar{Y}]} = \psi.$$

However, we will see that *asymptotically* the bias goes away. By the CLT,

$$\sqrt{n}(\bar{Y} - \theta) \sim \mathcal{N}(0, \sigma^2).$$

Let $g(\theta) = e^\theta = \psi$, so $g'(\theta) = e^\theta = \psi$. Then by the delta method,

$$\sqrt{n}(\hat{\psi} - \psi) \xrightarrow{d} \mathcal{N}(0, \psi^2\sigma^2).$$

As an approximation, this says that for large n,

$$\hat{\psi} \dot{\sim} \mathcal{N}(\psi, \psi^2\sigma^2/n).$$

We also have another way to derive the asymptotic distribution of $\hat{\psi}$. Using the fact that

$$\bar{Y} \dot{\sim} \mathcal{N}(\theta, \sigma^2/n)$$

for large n, it follows (from a version of the continuous mapping theorem that works for convergence in distribution) that

$$\hat{\psi} \dot{\sim} \text{Log-Normal}(\theta, \sigma^2/n).$$

Estimator	Delta method				Asymptotic variance
	Parent	θ	$g(\theta)$	$g'(\theta)$	$\{g'(\theta)\}^2\omega^2$
$a+b\bar{Y}$	$\bar{Y}$	$\mathrm{E}[Y_1]$	$a+b\theta$	b	$b^2\mathrm{Var}(Y_1)$
S	S^2	$\mathrm{Var}(Y_1)$	$\sqrt{\theta}$	$\frac{1}{2\sqrt{\theta}}$	$\frac{\mathrm{Var}(\{Y_1-E[Y_1]\}^2)}{4\mathrm{Var}(Y_1)}$
$\log S^2$	S^2	$\mathrm{Var}(Y_1)$	$\log\theta$	$\frac{1}{\theta}$	$\frac{\mathrm{Var}(\{Y_1-E[Y_1]\}^2)}{\mathrm{Var}(Y_1)^2}$
$Y_{(\lceil np\rceil)}$	$U_{(\lceil np\rceil)}$	$Q_{U_1}(p)=p$	$Q_{Y_1}(p)$	$\frac{\partial Q_{Y_1}(p)}{\partial p}$	$p(1-p)\left(\frac{\partial Q_{Y_1}(p)}{\partial p}\right)^2$

Table 3.2: Terms in the delta method. If $Y_1 \sim \mathcal{N}(\mu,\sigma^2)$, then $\mathrm{Var}(\{Y_1-E[Y_1]\}^2)/\mathrm{Var}(Y_1)^2 = 2$, so the asymptotic distribution of $\log S^2$ is free of θ.

These two approximations seem incompatible, but they do yield similar results when n is large. For moderate n though, the Log-Normal approximation is preferable because it only rests on one layer of approximation (the CLT) whereas the delta method approximation rests on two (CLT plus Taylor approximation).

Table 3.2 gives some additional examples of the use of the delta method. Here the estimand is $g(\theta)$ and the asymptotic variance is $\{g'(\theta)\}^2\omega^2$, where $g'(\theta) = \partial g(\theta)/\partial\theta$.

In the p-quantile case, Table 3.2 lists that

$$\sqrt{n}\{Y_{(\lceil np\rceil)} - Q_{Y_1}(p)\} \xrightarrow{d} \mathcal{N}(0, p(1-p)\{\partial Q_{Y_1}(p)/\partial p\}^2).$$

It will be helpful here to recall that for a generic continuously differentiable invertible function q,

$$\frac{\partial q^{-1}(x)}{\partial x} = \frac{1}{\frac{\partial q(z)}{\partial z}}, \quad \text{where} \quad z = q^{-1}(x), \tag{3.2}$$

so, as Q_{Y_1} is the inverse of F_{Y_1} and f_{Y_1} is the derivative of F_{Y_1},

$$\sqrt{n}\{Y_{(\lceil np\rceil)} - Q_{Y_1}(p)\} \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{\{f_{Y_1}(Q_{Y_1}(p)\}^2}\right).$$

Thus the asymptotic variance will be large when the density of Y_1 is small at the estimand $Q_{Y_1}(p)$. The density will typically be small in the tails of the distribution of Y_1.

Example 3.6.10 (Asymptotic standard errors in births example)**.** Returning to the Ethel Burns example of durations of births. Recall that $\bar{y} \approx 7.72$ and $s \approx 3.57$, while sample p-quantiles $Y_{(\lceil np\rceil)}$ for $p = 0.025$, 0.5, and 0.975 are, respectively, approximately 2.57, 7.50, and 14.53. The asymptotic standard errors for the quantiles are approximately 0.40, 0.45, and 0.88, using $\sqrt{p(1-p)}/\{\sqrt{n}f_{Y_1}(Q_{Y_1}(p))\}$, where we have estimated f_{Y_1} using a kernel density estimator.

3.6.3 Asymptotic distribution of method of moments estimator

Using the central limit theorem and delta method, we can also derive the asymptotic distribution of a method of moments estimator. First, let us recall the setup. Suppose that $Y_j \overset{\text{i.i.d.}}{\sim} F_{Y;\theta}$, $j = 1, \ldots, n$ and that

$$\alpha(\theta) = \mathrm{E}_\theta[h(Y_1)].$$

Replacing moments by sample moments and estimands by estimators delivers

$$\alpha(\hat{\theta}) = \frac{1}{n}\sum_{j=1}^{n} h(Y_j).$$

Assume $\alpha(\theta)$ is invertible. Then $\hat{\theta} = \alpha^{-1}\{\frac{1}{n}\sum_{j=1}^{n} h(Y_j)\}$, and $\theta = \alpha^{-1}\{\mathrm{E}_{Y_1;\theta}[h(Y_1)]\}$.

If $\mathrm{Var}(h(Y_1)) < \infty$, then the CLT says that

$$\sqrt{n}\left[\frac{1}{n}\sum_{j=1}^{n} h(Y_j) - \mathrm{E}[h(Y_1)]\right] \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(h(Y_1))),$$

so the delta method implies that if α^{-1} is differentiable, then

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \left(\frac{\partial \alpha^{-1}(\mathrm{E}[h(Y_1)])}{\partial \mathrm{E}[h(Y_1)]}\right)^2 \mathrm{Var}(h(Y_1))\right).$$

Using the derivative of an inverse function as in equation (3.2), delivers the asymptotic distribution of the method of moment estimator:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \left[\frac{\partial \alpha(\theta)}{\partial \theta}\right]^{-2} \mathrm{Var}(h(Y_1))\right).$$

The estimator will be pretty imprecise if $\partial\alpha(\theta)/\partial\theta$ is close to zero, which is when $\alpha(\theta)$ is insensitive to θ. The asymptotic variance is quite simple to estimate, requiring us to differentiate $\alpha(\theta)$ and to estimate $\mathrm{Var}(h(Y_1))$.

3.7 Multivariate asymptotic approximations*

Asymptotic approximations are often useful. Here we briefly outline how the univariate approximations extend to the multivariate case. In statistics this is mostly used in the context of a (vector) estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, ..., \widehat{\theta}_K)^\intercal$ of the (vector) estimand $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\intercal$, as well as the test statistics developed in Chapter 8. We state the results in this subsection without proofs, except in one insightful case.

The first extension is purely probabilistic, extending the definitions of convergence in probability and convergence in distribution.

- Convergence in probability and distribution. If the sequence $\mathbf{X}_1, \mathbf{X}_2, ...$ of K-dimensional random vectors:
  - obeys, for every $\varepsilon > 0$,
    $$\lim_{n\to\infty} P(\|\mathbf{X}_n - \mathbf{X}\|_2 \geq \varepsilon) \to 0.$$
    then we write $\mathbf{X}_n \xrightarrow{p} \mathbf{X}$. Here, generically, the notation $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_K^2}$, denotes the Euclidean norm of $\mathbf{x} = (x_1, \ldots, x_K)^\mathtt{T}$.
  - obeys
    $$\lim_{n\to\infty} P(\mathbf{X}_n \leq \mathbf{x}) \to P(\mathbf{X} \leq \mathbf{x}), \quad \text{for all } \mathbf{x} \in \mathbb{R}^K,$$
    then we write $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$.

Applying this to statistics, the estimator $\widehat{\boldsymbol{\theta}}$ is *consistent* for the estimand $\boldsymbol{\theta}$ if $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$.

Most applications of asymptotics in statistics start with a law of large numbers and a central limit theorem. The ones for i.i.d. problems are:

- If $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are i.i.d. K-dimensional vectors with
  - a $K \times 1$ vector of means $\mathrm{E}[\mathbf{X}] = (\mathrm{E}[X_1], ..., \mathrm{E}[X_n])^\mathtt{T}$ exists, then $\overline{\mathbf{X}} \xrightarrow{p} \mathrm{E}[\mathbf{X}]$. This result is called the *multivariate law of large numbers.*
  - a $K \times K$ matrix of variances and covariances $\mathrm{Var}(\mathbf{X})$ exists, then
    $$\sqrt{n}(\overline{\mathbf{X}} - \mathrm{E}[\mathbf{X}]) \xrightarrow{d} \mathcal{N}_K(\mathbf{0}_K, \mathrm{Var}(\mathbf{X})).$$
    This result is called the *multivariate central limit theorem.* Here $\mathrm{Var}(\mathbf{X})$ has (i, j)th element $\mathrm{Cov}(X_i, X_j)$ and $\mathbf{0}_K$ is a K-dimensional vector of zeros.

To extend beyond sample averages, statisticians frequently use the continuous mapping theorem, Slutsky's Theorem, and the delta method in tandem. The first two of these are:

- Continuous mapping theorem. Assume g is a p-dimensional continuous function (i.e., it takes in K inputs and outputs a p-dimensional vector). Part 1: if $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$, then $g(\widehat{\boldsymbol{\theta}}) \xrightarrow{p} g(\boldsymbol{\theta})$. Part 2: if $\widehat{\boldsymbol{\theta}} \xrightarrow{d} \boldsymbol{\theta}$, then $g(\widehat{\boldsymbol{\theta}}) \xrightarrow{d} g(\boldsymbol{\theta})$.
- Slutsky's Theorem. Assume $\mathbf{X}_n \xrightarrow{d} \mathbf{X}$ and $\mathbf{Y}_n \xrightarrow{p} \mathbf{C}$. Then jointly
  $$\begin{pmatrix} \mathbf{X}_n \\ \mathbf{Y}_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \mathbf{X} \\ \mathbf{C} \end{pmatrix}.$$

The proof of the delta method is of a form which appears in many statistical calculations and so is instructive.

Theorem 3.7.1 (Multivariate delta method)**.** *Assume that* $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^{\mathtt{T}}$ *is* K*-dimensional,* $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_K)^{\mathtt{T}}$*,*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}_K\left(\mathbf{0}_K, \Sigma_{\boldsymbol{\theta}}\right)$$

as $n \to \infty$ *and* g *is a* p*-dimensional continuous differentiable function of all elements in* $\boldsymbol{\theta}$*. Then*

$$\sqrt{n}\left\{g(\widehat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})\right\} \xrightarrow{d} \mathcal{N}_p\left(\mathbf{0}_p, \left\{g'(\boldsymbol{\theta})\right\} \Sigma_{\boldsymbol{\theta}} \left\{g'(\boldsymbol{\theta})\right\}^{\mathtt{T}}\right),$$

where the $p \times K$ *matrix* $g'(\boldsymbol{\theta}) = \partial g(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^{\mathtt{T}}$*.*

Proof. Using a vector Taylor expansion

$$g(\widehat{\boldsymbol{\theta}}) \approx g(\boldsymbol{\theta}) + \sum_{j=1}^{K} \frac{\partial g(\theta)}{\partial \theta_j}\left(\widehat{\theta}_j - \theta_j\right) = g(\boldsymbol{\theta}) + g'(\boldsymbol{\theta})\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right),$$

then

$$\sqrt{n}\left\{g(\widehat{\boldsymbol{\theta}}) - g(\boldsymbol{\theta})\right\} \approx g'(\boldsymbol{\theta})\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right).$$

Recall that if $\mathbf{X} \sim \mathcal{N}_K(\boldsymbol{\mu}_{\mathbf{X}}, \boldsymbol{\Sigma}_{\mathbf{X}})$, a K-dimensional Multivariate Normal distribution, then

$$\mathbf{C}\mathbf{X} \sim \mathcal{N}_p(\mathbf{C}\boldsymbol{\mu}_{\mathbf{X}}, \mathbf{C}\boldsymbol{\Sigma}_{\mathbf{X}}\mathbf{C}^{\mathtt{T}}),$$

when $\mathbf{C}$ is a $p \times K$ matrix of constants. Applying that result here with $\mathbf{C} = g'(\boldsymbol{\theta})$ and $\mathbf{X} \sim \mathcal{N}_K\left(0, \Sigma_{\boldsymbol{\theta}}\right)$ yields the required result. ∎

Example 3.7.2. Suppose the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. Then the bivariate version of the CLT states

$$\sqrt{n}\begin{pmatrix} \overline{Y} - \mu_Y \\ \overline{X} - \mu_X \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_X^2 \end{pmatrix}\right),$$

as $n \to \infty$. Suppose $\mu_X \neq 0$, and we are interested in $\overline{Y}/\overline{X}$ as an estimator for μ_Y/μ_X. Many statistical problems are of this form. Define

$$U = Y - (\mu_Y/\mu_X)\, X.$$

Then the delta method implies that

$$\sqrt{n}\left(\frac{\overline{Y}}{\overline{X}} - \frac{\mu_Y}{\mu_X}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\mu_X^2}\sigma_U^2\right). \tag{3.3}$$

Why? Putting this into the delta method structure $g(\theta_1, \theta_2) = \theta_1/\theta_2$, so

$$g'(\theta) = \left(\frac{1}{\theta_2}, -\frac{\theta_1}{\theta_2^2}\right) = \frac{1}{\mu_X}\left(1, -\frac{\mu_Y}{\mu_X}\right), \quad \Sigma_{\boldsymbol{\theta}} = \begin{pmatrix} \sigma_Y^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_X^2 \end{pmatrix}.$$

Now

$$\begin{aligned} g'(\theta)\Sigma_{\boldsymbol{\theta}} g'(\theta)^{\mathsf{T}} &= \frac{1}{\mu_X^2}\left(1, -\frac{\mu_Y}{\mu_X}\right)\begin{pmatrix} \sigma_Y^2 & \sigma_{X,Y} \\ \sigma_{X,Y} & \sigma_X^2 \end{pmatrix}\begin{pmatrix} 1 \\ -\frac{\mu_Y}{\mu_X} \end{pmatrix} \\ &= \frac{1}{\mu_X^2}\left\{\sigma_Y^2 - 2\left(\frac{\mu_Y}{\mu_X}\right)\sigma_{X,Y} + \left(\frac{\mu_Y}{\mu_X}\right)^2 \sigma_X^2\right\} \\ &= \frac{1}{\mu_X^2}\sigma_U^2, \quad U = Y - \frac{\mu_Y}{\mu_X}X, \quad \mathrm{E}[U] = 0, \quad \sigma_U^2 = \mathrm{Var}(U), \end{aligned}$$

and then applying the delta method yields the result.

Example 3.7.3. (Ratio of sample means). In the special case of Example 3.7.2 where $\mu_Y = 0$, we have

$$\sqrt{n}\left(\overline{Y}/\overline{X}\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma_Y^2/\mu_X^2\right). \tag{3.4}$$

Example 3.7.4. (Regression etsimator). Specialize Example 3.7.2 to where $Y_i = Z_i W_i$ and $X_i = W_i^2$, then

$$\widehat{\theta} = \frac{\sum_{i=1}^n Z_i W_i}{\sum_{i=1}^n W_i^2}, \quad \theta = \frac{\mathrm{E}[ZW]}{\mathrm{E}[W^2]},$$

then,

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathrm{E}[W^2V^2]}{\mathrm{E}[W^2]^2}\right), \quad V = Z - \theta W,$$

as $U = VW$. The estimator $\widehat{\theta}$ appears in Chapter 6 on regression where Z is an outcome and W is a predictor.

Example 3.7.5. (Instrumental variables estimator). Specialize Example 3.7.2 to the case where $Y_i = Z_i A_i$ and $X_i = W_i A_i$, then

$$\widehat{\theta} = \frac{\sum_{i=1}^n Z_i A_i}{\sum_{i=1}^n W_i A_i}, \quad \theta = \frac{\mathrm{E}[ZA]}{\mathrm{E}[WA]},$$

so

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathrm{E}[A^2V^2]}{\mathrm{E}[WA]^2}\right), \quad V = Z - \theta W,$$

as $U = VA$. The estimator $\widehat{\theta}$ appears in the social and medical sciences, where A is called an *instrumental variable* for outcome Z and predictor W.

☣ **3.7.6.** In (3.3) it is crucial that $\mu_X \neq 0$. If μ_X is close to 0 but not 0, then this delta method is asymptotically valid but in practice it may well be useless as it may take an enormous n for the approximation to be reasonable for the distribution of $\overline{Y}/\overline{X}$. Exactly the same phenomenon is a concern for Example 3.7.5 when $\mathrm{E}[WA]$ is close to zero. In econometrics this is called a *weak instruments* problem.

Example 3.7.7. In the special case of Example 3.7.2 where $Y_i = Z_i W_i$ and $X_i = W_i$, which is binary, and

$$\widehat{\theta} = \frac{\sum_{i=1}^n Z_i W_i}{\sum_{i=1}^n W_i}, \quad \theta = \mathrm{E}[Z|W=1],$$

then

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \{\mathrm{E}[W]\}^{-1} \mathrm{Var}(Z|W=1)\right).$$

Why? The $\mu_Y = \mathrm{E}[ZW]$ and $\mu_X = \mathrm{E}[W]$, so $\mu_Y/\mu_X = \mathrm{E}[Z|W=1] = \theta$. Thus,

$$U = Y - (\mu_Y/\mu_X)\, X = \{Z - \mathrm{E}[Z|W=1]\}\, W_1.$$

By Eve's law,

$$\mathrm{Var}(U) = \mathrm{E}[W]\mathrm{Var}(Z_1|W=1),$$

as $W^2 = W$. Simplifying delivers the result. The estimator $\widehat{\theta}$ appears in Chapter 11. Often the result is written as

$$\sqrt{n}(\widehat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \mu_W^{-2}\sigma_U^2\right).$$

3.8 A couple of technical proofs *

First we prove Theorem 3.1.2.

Proof. Define

$$Z_{j,n} = \sqrt{n}(Y_{(j)} - p), \quad Y_{(j)} \sim U(j, n-j+1),$$

so $Y_{(j)} = p + Z_{j,n}/\sqrt{n}$ and $1 - Y_{(j)} = 1 - p - Z_{j,n}/\sqrt{n}$, while $\partial Z_{j,n}/\partial Y_{(j)} = \sqrt{n}$. We will study the distribution of $Z_{j,n}$ as n goes to infinity with $j = \lceil pn \rceil$.

By the change of variables formula (see Chapter 8 of the Stat 110 book), the form of the Beta density, and keeping track of only terms involving z, the density of $Z_{j,n}$ is

$$\begin{aligned} f_{Z_{j,n}}(z) &\propto (p + z/\sqrt{n})^{j-1}(1 - p - z/\sqrt{n})^{n-j}, \quad z \in \left[-\sqrt{n}p, \sqrt{n}(1-p)\right] \\ &= p^{j-1}(1-p)^{n-j}\left\{1 + z/\left(p\sqrt{n}\right)\right\}^{j-1}\left[1 - z/\left\{(1-p)\sqrt{n}\right\}\right])^{n-j} \\ &\propto \left\{1 + z/\left(p\sqrt{n}\right)\right\}^{j-1}\left[1 - z/\left\{(1-p)\sqrt{n}\right\}\right])^{n-j}, \end{aligned}$$

as the Jacobian does not depend upon z.

By Taylor approximation, for small x we have $\log(1+x) \approx x - x^2/2$. So setting $j = \lceil pn \rceil$, then if n is large $z/\sqrt{n}$ will be small so

$$\begin{aligned}\log f_{Z_{j,n}}(z) &= c + (\lceil pn \rceil - 1)\log\left\{1 + z/\left(p\sqrt{n}\right)\right\} + (n - \lceil pn \rceil)\log\left[1 - z/\left\{(1-p)\sqrt{n}\right\}\right] \\ &\approx c + \frac{z}{\sqrt{n}}\left\{(\lceil pn \rceil - 1)/p - (n - \lceil pn \rceil)/(1-p)\right\} \\ &\quad -\frac{1}{2}\left(\frac{z}{\sqrt{n}}\right)^2\left\{(\lceil pn \rceil - 1)/p^2 + (n - \lceil pn \rceil)/(1-p)^2\right\}.\end{aligned}$$

But

$$\begin{aligned}\left(\frac{\lceil pn \rceil - 1}{p}\right) - \left(\frac{n - \lceil pn \rceil}{1-p}\right) &= \left(\frac{np + (\lceil pn \rceil - np - 1)}{p}\right) - \left(\frac{n(1-p) - (\lceil pn \rceil - np)}{1-p}\right) \\ &= \left(\frac{\lceil pn \rceil - np - 1}{p}\right) - \left(\frac{\lceil pn \rceil - np}{1-p}\right) = O(1),\end{aligned}$$

and

$$\frac{\lceil pn \rceil - 1}{p^2} + \frac{n - \lceil pn \rceil}{(1-p)^2} = \frac{n}{p(1-p)} + O(1).$$

So

$$\log f_{Z_{j,n}}(z) \to c - \frac{z^2}{2p(1-p)},$$

the log-density of a $\mathcal{N}(0, p(1-p))$ variable, as claimed. ∎

Next, we prove Theorem 3.2.3.

Proof. For each Y_j, whether or not it falls in the interval $(y - h/2, y + h/2]$ is a Bernoulli trial:

$$I\left(Y_j \in (y - h/2, y + h/2]\right) \overset{\text{i.i.d.}}{\sim} \text{Bern}(F_{Y_1}(y + h/2) - F_{Y_1}(y - h/2)).$$

So

$$\mathrm{E}[\hat{\theta}] = \frac{1}{h}\mathrm{E}[I\left(Y_1 \in (y - h/2, y + h/2]\right)] = \frac{1}{h}[F_{Y_1}(y + h/2) - F_{Y_1}(y - h/2)]$$

and, as $\hat{\theta}$ is a scaled average,

$$\begin{aligned}\mathrm{Var}[\hat{\theta}] &= \frac{1}{nh^2}\mathrm{Var}[I\left(Y_1 \in (y - h/2, y + h/2]\right)] \\ &= \frac{1}{nh^2}\{F_{Y_1}(y + h/2) - F_{Y_1}(y - h/2)\}[1 - \{F_{Y_1}(y + h/2) - F_{Y_1}(y - h/2)\}].\end{aligned}$$

If F_{Y_1} is sufficiently smooth, then for $u \approx 0$ we can Taylor approximate

$$F_{Y_1}(y + u) \approx F_{Y_1}(y) + u f_{Y_1}(y) + \frac{1}{2}u^2 f'_{Y_1}(y) + \frac{1}{6}u^3 f''_{Y_1}(y),$$

where $f'_{Y_1}(y) = \partial f_{Y_1}(y)/\partial y$. Applying this expansion twice for small $h/2$,

$$F_{Y_1}(y + h/2) \approx F_{Y_1}(y) + (h/2)f_{Y_1}(y) + \frac{1}{2}(h/2)^2 f'_{Y_1}(y) + \frac{1}{6}(h/2)^3 f''_{Y_1}(y)$$

and

$$F_{Y_1}(y - h/2) \approx F_{Y_1}(y) - (h/2)f_{Y_1}(y) + \frac{1}{2}(h/2)^2 f'_{Y_1}(y) - \frac{1}{6}(h/2)^3 f''_{Y_1}(y).$$

The third term cancels when we take the difference, so

$$F_{Y_1}(y + h/2) - F_{Y_1}(y - h/2) \approx h f_{Y_1}(y) + \frac{1}{24}h^3 f''_{Y_1}(y),$$

which implies the desired results. ∎

3.9 Concentration inequalities*

Asymptotic approximations and simulations are the main way statisticians approximate the distribution of estimators and test statistics. Another form of approximation, called *concentration inequalities*, has also become common in recent years. The main theme is to give an upper bound on probabilities (which can be used to carry out conservative inference procedures), with no need for n to increase. The most well known versions of this are based on Markov's inequality, Chebyshev's inequality, and Chernoff's bound (see Chapter 10 of the Stat 110 book).

A widely used concentration inequality in statistics and machine learning is *Hoeffding's inequality.*

Theorem 3.9.1 (Hoeffding's inequality)**.** *Suppose $Y_1, ..., Y_n$ are independent (not necessarily i.i.d.), the $S_n = \sum_{j=1}^n Y_j$, with the j-th version bounded $a_j \leq Y_j \leq b_j$ for each $j = 1, 2, ..., n$. Then Hoeffding's inequality says that for all $t > 0$,*

$$P(S_n - \mathrm{E}[S_n] \geq t) \leq \exp\left(-\frac{2t^2}{\sum_{j=1}^n (b_j - a_j)^2}\right).$$

Proof. This is based on Hoeffding's lemma, which states that for any random variable $X \in [a, b]$ that

$$\mathrm{E}[e^{s(X - E[X])}] \leq \exp\left\{s^2(b-a)^2/8\right\}.$$

The proof of Hoeffding's lemma is beyond the scope of this book. But we will use the lemma to prove

Hoeffding's inequality. For all constants $s, t > 0$, the

$$\begin{aligned}
P(S_n - \mathrm{E}[S_n] &\geq t) \\
&= P(s\,(S_n - \mathrm{E}[S_n]) \geq st) \\
&= P(\exp\{s\,(S_n - \mathrm{E}[S_n])\}) \geq \exp(st)), \quad \text{as exp is an increasing function} \\
&\leq \exp(-st)\mathrm{E}[\exp\{s\,(S_n - \mathrm{E}[S_n])\}], \quad \text{by Markov's inequality} \\
&= \exp(-st)\prod_{j=1}^{n}\mathrm{E}[\exp\{s\,(Y_j - \mathrm{E}[Y_j])\}], \quad \text{by independence} \\
&\leq \exp(-st)\prod_{j=1}^{n}\exp(s^2(b_j-a_j)^2/8), \quad \text{by Hoeffding's lemma} \\
&= \exp\{S(s)\}, \quad \text{where} \quad S(s) = -st + \left(s^2/8\right)\sum_{j=1}^{n}(b_j-a_j)^2.
\end{aligned}$$

Now select $\widehat{s}$, the value of s to minimize $S(s)$ (yielding the tightest bound). But

$$\frac{\partial S(t)}{\partial t} = -t + (s/4)\sum_{j=1}^{n}(b_j-a_j)^2, \quad \text{so} \quad \widehat{s} = \frac{4t}{\sum_{j=1}^{n}(b_j-a_j)^2},$$

implying

$$\begin{aligned}
S(\widehat{s}) &= -\widehat{s}t + \left(\widehat{s}^2/8\right)\sum_{j=1}^{n}(b_j-a_j)^2 \\
&= -\frac{4t^2}{\sum_{j=1}^{n}(b_j-a_j)^2} + \frac{2t^2}{\sum_{j=1}^{n}(b_j-a_j)^2} \\
&= -\frac{2t^2}{\sum_{j=1}^{n}(b_j-a_j)^2}.
\end{aligned}$$

Plugging this in delivers the stated result.

∎

This bound is useful as it holds without knowledge of the distribution of the $Y_1, ..., Y_n$, nor does it need the random variables to be identically distributed.

Example 3.9.2. Assume that $Y_1, \ldots, Y_n$ are all Bernoulli variables or all Uniform random variables — but where $\mathrm{E}[Y_j]$ might change over j. Then, for $\varepsilon > 0$ and taking $t = n\varepsilon$, the

$$P(n^{-1}S_n - n^{-1}\mathrm{E}[S_n] \geq \varepsilon) = P(S_n - \mathrm{E}[S_n] \geq n\varepsilon) \leq \exp\left(-2\varepsilon^2 n\right),$$

as $b_j - a_j = 1$ for all j.

Formula or idea	Description or name
$\text{bias}(\hat{\theta}) = \text{E}[\hat{\theta}] - \theta, \quad \text{SE}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$	bias, standard error
$\hat{\theta} = \frac{1}{n}\sum_{j=1}^{n} \frac{1}{h} I(Y_j \in (y - h/2, y + h/2])$	kernel density estimator
$\text{Loss}(\theta, \hat{\theta})$	loss
$\text{E}[\text{Loss}(\theta, \hat{\theta})]$	risk
$\text{MSE}(\hat{\theta}) = \text{E}[(\hat{\theta} - \theta)^2] = \text{bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$	mean square error, bias-variance trade-off
$\text{MAE}(\hat{\theta}) = \text{E}[\lvert\hat{\theta} - \theta\rvert]$	mean absolute error
$\hat{\theta} \xrightarrow{p} \theta$	consistency of estimator $\hat{\theta}$ for estimand θ
if $X_n \xrightarrow{p} X$ then $g(X_n) \xrightarrow{p} g(X)$, if g continuous,	continuous mapping theorem part 1
if $X_n \xrightarrow{d} X$ then $g(X_n) \xrightarrow{d} g(X)$, if g continuous	continuous mapping theorem part 2
if $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{p} c$, then	Slutsky's theorem
(a) $X_n + Y_n \xrightarrow{d} X + c$, (b) $X_n - Y_n \xrightarrow{d} X - c$,	
(c) $X_n Y_n \xrightarrow{d} cX$, (d) $X_n / Y_n \xrightarrow{d} X/c$ if $c \neq 0$	
if $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \omega^2)$, and $g'(\theta)$ exists, then	delta method
$\sqrt{n}\{g(\hat{\theta}) - g(\theta)\} \xrightarrow{d} \mathcal{N}(0, \omega^2 g'(\theta)^2)$	
$\sqrt{n}(Y_{(\lceil np \rceil)} - Q_{Y_1}(p)) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{f_{Y_1}(Q_{Y_1}(p))^2}\right)$	sample quantile's asymptotic distribution
if $\text{Var}(h_1(Y_1)) < \infty$, α invertible,	MoM's asymp. dist. based on $\alpha(\theta) = \text{E}[h(Y_1)]$
then MoM $\hat{\theta}$ has $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \left[\frac{\partial \alpha(\theta)}{\partial \theta}\right]^{-2} \text{Var}(h_1(Y_1))\right)$	

Table 3.3: Main ideas and notation in Chapter 3. The middle section of this Table are a triple of results from probability theory which are extremely helpful in statistics.

3.10 Recap

Statistical estimators typically have a modest bias of roughly b/n and a variance of roughly a/n. Hence both decline at rate n^{-1}. We quantified how to think about overall measures of loss of estimators, so that we can choose between competing procedures. Most of our focus was on the mean square error, which is the variance plus the squared bias. In the above typically case, the mean square error is $a/n + b^2/n^2 \approx a/n$ when n is quite large. Thus the variance dominates the bias.

The main concepts covered in this chapter are listed in Table 3.3.

We saw a key example, the kernel density estimator of a density function, which has a more sophisticated bias and variance, depending upon a bandwidth parameter which is controlled by the

statistician. For the kernel the bias and variance were in conflict, with the bias being small for small bandwidth, but small bandwidths drive the variance higher. In that case the MSE allowed us to pool the bias and variance, providing a principled criteria for choosing the bandwidth.

The other development we had was to bring in additional tools from probability that allow us to study the behavior of estimators when n is large.

Two key forms of convergence are convergence in probability and convergence in distribution.

Five key tools for asymptotics are the law of large numbers, central limit theorem, continuous mapping theorem, Slutsky's theorem, and the delta method. Such tools allow us to find asymptotic approximations to the distributions of a wide variety of random variables. This in turn helps us to study a wide variety of estimators, when the sample size is large. Examples of the application of these asymptotic methods include in studying the limiting distribution of sample quantiles, sample covariances, and method of moments estimators.

Asymptotic approximations are not the only way the distribution of estimators and test statistics can be approximated in statistics. In Chapter 10 we will also develop sampling methods, such as the bootstrap, which are computationally more intensive but requite less use of mathematical derivations.

3.11 R, loss, bias-variance tradeoff, and asymptotics

Strings appear extensively in R, so we provide a short briefing on them here. There is also a subsection which illustrates how to run a simulation experiment in R. Finally, there is a subsection which contains the code for the examples in this Chapter.

3.11.1 Text and strings

In modern statistics there is enormous interest in the statistical analysis of *text data*: for speech recognition, translation, search, categorization, etc. The text will be held using *strings*. The core of R has many features which allow the use and manipulation of text. Here we will discuss some of the most basic ones.

Many of the common functions, such as `c()`, `cbind()`, and `replicate()`, in R also work for strings. Strings can also be combined with numerical quantities. Some basic features for working with strings can be seen in these examples:

Setting up and editing strings

```
> "First"
[1] "First"
> c("First","Second","Third")
```

```
[1] "First"  "Second" "Third"
> LETTERS[1:3]
[1] "A" "B" "C"
> letters[1:3];
[1] "a" "b" "c"
> replicate(2,LETTERS[1:3])
     [,1] [,2]
[1,] "A"  "A"
[2,] "B"  "B"
[3,] "C"  "C"
> cbind(c(1, 5, 3, 2, 5),
      c("shop", "zoo", "road", NA, "rail"))
     [,1] [,2]
[1,] "1"  "shop"
[2,] "5"  "zoo"
[3,] "3"  "road"
[4,] "2"  NA
[5,] "5"  "rail"
```

Strings can be directly manipulated within R, which is often helpful in generating large numbers of plots. To add together text the paste function is helpful:

Writing out graph titles

```
> sTitle="Graph title "
> tTitle="for QQ plot"
> paste(sTitle,tTitle,sep="")
[1] "Graph title for QQ plot"
```

If you like shouting while typing you can use convert text to all caps using:

Code converting to caps

```
> toupper("Conditioning is the soul of statistics")
[1] "CONDITIONING IS THE SOUL OF STATISTICS"
```

An alternative is to whisper:

Code converting to lower case

```
> tolower("PREDICTION and CauSALity are just DIfferent");
[1] "prediction and causality are just different"
```

3.11.2 Simulation experiments

Simulation plays an enormous role in modern statistics. One aspect of this is running simulation experiments to measure the performance of statistical procedures or statistical approximations. R is a very useful tool for carrying out simulation experiments. We will illustrate that here.

Return to Example 3.6.9, and assess the accuracy of the CLT for $\hat{\psi} = e^{\bar{Y}}$, where $Y_j \sim N(\theta, \sigma^2)$, for $j = 1, \ldots, n$. The estimand is $\psi = e^{\theta}$ and the delta method implies

$$\sqrt{n}(\hat{\psi} - \psi) \xrightarrow{d} \mathcal{N}(0, \psi^2\sigma^2).$$

A limitation of simulation experiments is that we have to select which particular values of n, θ, and σ^2 to study, rather than provide results which hold across all possible parameter values.

In this experiment we take $n = 4$, 10, 25, and 100 and $\theta = 2$ and $\sigma = 0.5$. The simulation size will be $B = 10,000$. We will measure the bias, the mean square error, and how close to Normal the simulations are using a Normal QQ plot — the Normal QQ plots the ordered data $y_{(1)}, \ldots, y_{(n)}$, the sample quantiles against the Normal theoretical quantiles $Q_{\mathcal{N}(0,1)}(1/(n+1)), \ldots, Q_{\mathcal{N}(0,1)}(n/(n+1))$. It should be roughly a straight line through 0 with unit slope if the Gaussian assumption holds.

Simulation experiment and QQ plot output

```
set.seed(111)
theta=2; sigma=0.5; psi = exp(theta) # parameters of experiment
B=10000; Hatpsi = rep(0,B); i=0  # experiment setup
mExper = matrix(0,4,3) # create matrix in which to store results
colnames(mExper)=c("n","bias","n*mse") # name stored results

for (n in c(4,10,25,100)){ i=i+1; # i is a counter
   for (b in (1:B)){
     y = rnorm(n,theta,sigma)
     Hatpsi[b] = exp(mean(y))
   }

   qqnorm(sqrt(n)*(Hatpsi -psi)/(psi*sigma),pch = 16,cex=0.4,
              main=paste("QQ plot sample size",n))
   abline(a=0,b=1,col="red")

   mExper[i,] = c(n,(mean(Hatpsi)-psi),n*mean((Hatpsi-psi)^2))

   pdf(paste("QQnormal",n,".pdf",sep =""))
     qqnorm(sqrt(n)*(Hatpsi -psi)/(psi*sigma),pch = 16,cex=0.4,
              main=paste("QQ plot sample size",n))
     abline(a=0,b=1,col="red")
   dev.off()
}

print(mExper) # print summaries

       n         bias    n*mse
[1,]   4 0.263632201 15.33893
[2,]  10 0.087814938 14.12222
```

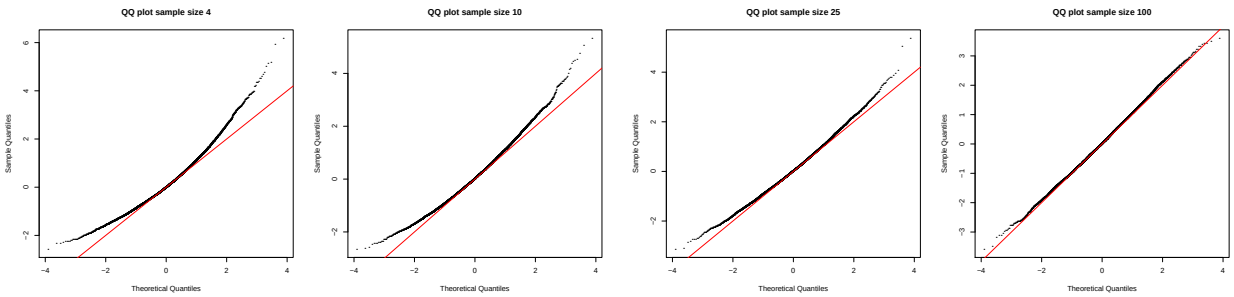


Figure 3.7: QQ plot for the $\sqrt{n}(\hat{\psi} - \psi)/(\psi\sigma)$ for various values of n.

```
[3,]   25 0.042718833 14.08224
[4,]  100 0.007681337 13.64232
```

The resulting QQ plots are collected in Figure 3.7. Overall the impression is that the CLT is not very useful even if $n = 25$, but by the time $n = 100$ it is a reasonable approximation. Notice how the bias falls with n, as expected. The $n \times$ MSE stabilizes as n increases, again, as we would expect.

In more detail, when $n = 4$ the sample quantiles in the left hand tail of the distribution are too small compared to a normal, while they are too large in the right hand tail. As n increases these features wane.

The code has a feature we have not seen before. To label the QQ plots with different values of n we set the main title of the picture using a function `paste` which allows us to add to a string of fixed text something which changes over the iteration.

3.11.3 Code for various examples in this chapter

The code produces the results for Example 3.2.2, which simulates data and computes the kernel density estimator, which are then compared to the true density function.

Code from Example 3.2.2

```
iPoints=1000;  mResults = matrix(0,iPoints,4) # storage

for (k in (1:2)){
  if (k==1) n = 100    # 2 sample size cases
  if (k==2) n = 1000

  Yobs = rnorm(n,mean=11.0,sd=4.0);  # simulated data

  for (j in (1:2)){
    if (j==1) h = 5.0  # 2 bandwidths
    if (j==2) h = 0.5
```

```
    for (i in (1:iPoints)){
      y = -1.0 + i*(24.0/iPoints)
      denEst = mean(abs(Yobs-y)<(0.5*h))/h

      if (j==1) mResults[i,1:3] = c(y,denEst,dnorm(y,11.0,4.0))
      if (j==2) mResults[i,4] = denEst
    }
  }

#  if (k==1) pdf("denEst1.pdf")  # plot results
#  if (k==2) pdf("denEst2.pdf")

  plot(mResults[,1],mResults[,2],type="l",
    main="Kernel density estimator", xlab="y",
    ylab="Estimated density",col="blue")
  points(mResults[,1],mResults[,3],type="l",col="black",lwd=4)

  points(mResults[,1],mResults[,4],type="l",col="red")

#  dev.off()
}
```

Chapter 4

Maximum Likelihood Estimation

4.1 Defining and finding the maximum likelihood estimate (MLE)

Maximum likelihood estimation is among the most widely used estimation techniques in all of statistics. As the name suggests, the idea is to choose the parameter value that maximizes the likelihood function from a parametric statistical model. In a sense, this is the parameter value that is the most consistent with the data, since it makes the observed data as probable as possible.

4.1.1 Introduction

Definition 4.1.1 (Maximum likelihood estimator)**.** The *maximum likelihood estimate* (MLE) of θ is the value $\hat{\theta}$ that maximizes the likelihood function $L(\theta; \mathbf{y})$. Mathematically, this is written as

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\max}\ L(\theta; \mathbf{y}).$$

Here "arg max" stands for "the argument that maximizes" out of all $\theta \in \Theta$, the permissible parameter space. The corresponding estimator is called the *maximum likelihood estimator*

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\max}\ L(\theta; \mathbf{Y}),$$

which depends on the random $\mathbf{Y}$. That is, if the maximum likelihood estimate is $T(\mathbf{y})$, then the maximum likelihood estimator is $T(\mathbf{Y})$. We say "the" value in this definition since typically in practice the MLE exists and is unique, but it is possible to come up with examples where the MLE does not exist or is not unique. An MLE is illustrated in Figure 4.1, which is repeated from Chapter 2.

☣ **4.1.2.** The notation $\hat{\theta}$ and the acronym MLE are often used both for the maximum likelihood *estimator* and for the maximum likelihood *estimate*; the context should make clear which is meant.

☣ **4.1.3.** The word "likelihood" has a technical meaning in statistics, referring to the likelihood function. In everyday English, "likelihood" is synonymous with "probability". Sometimes people say

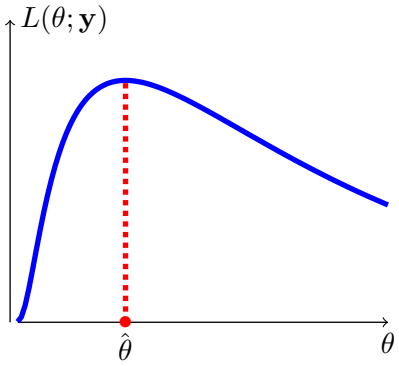


Figure 4.1: The likelihood $L(\theta; \mathbf{y})$ drawn against θ, showing $\hat{\theta}$, the maximum likelihood estimate.

the MLE is the "most likely" parameter value given the data. But in everyday English "most likely" means the same thing as "most probable", which does *not* give a valid interpretation of the MLE. In fact, we do not yet have a definition of what it even means for one θ value to be more probable than another, since we have not specified a distribution for θ.

Later when we discuss Bayesian statistics in more detail, we *will* specify a distribution for θ and then it makes sense to say one value of θ is more probable than another. Whether the MLE is the most probable value of θ given the data depends on the prior. If, for example, the parameter space is the interval $[a, b]$ and the prior is Uniform on $[a, b]$, then the MLE is the most probable value of θ given the data. But if the prior probability of a particular $\theta_0 \in [a, b]$ is much higher than the prior probability of a particular $\theta_1 \in [a, b]$, then even if the likelihood function favors θ_1 over θ_0 this may not be enough to offset the prior favoring θ_0 over θ_1.

Since log is a strictly increasing function, the MLE is also the value of θ that maximizes the log-likelihood function:

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\max} \ \log L(\theta; \mathbf{y}).$$

Usually it is easier to work with the log-likelihood than the likelihood when finding MLEs. One reason for this is that we often need to take derivatives to find the MLE, and it tends to be much easier to take the derivative of a sum than the derivative of a product.

Example 4.1.4 (Binomial)**.** Let $Y \sim \text{Bin}(n, p)$, with n known and p the estimand. The likelihood function is, dropping the binomial coefficient since it acts as a constant,

$$L(p; y) = p^y (1-p)^{n-y},$$

so the log-likelihood function is

$$l(p; y) = \log L(p; y) = y \log p + (n - y) \log(1 - p).$$

Our sample size is only 1 here, but note that if we had observed n i.i.d. Bern(p) random variables instead, with total y, then the likelihood function would be equivalent. So for likelihood purposes, we can compress n i.i.d. Bern(p) observations into one Binomial.

To find the MLE $\hat{p}$, we can set the derivative of $l(p; y)$,

$$l'(p; y) = \frac{\partial l(p; y)}{\partial p} = \frac{y}{p} - \frac{n-y}{1-p},$$

equal to 0:

$$\frac{y}{\hat{p}} - \frac{n-y}{1-\hat{p}} = 0,$$

which rearranges to

$$\hat{p} = \frac{y}{n}.$$

To check that we have found the maximum, note that the second derivative

$$l''(p; y) = \frac{\partial^2 l(p; y)}{\partial p^2} = -\frac{y}{p^2} - \frac{n-y}{(1-p)^2}$$

is negative for all p. So the log-likelihood is globally concave, which shows that $\hat{p}$ is the global maximum of $l(p)$. Therefore, we have found the unique MLE for this problem.

To compare the MLE with the MoM, note that to find a MoM $\tilde{p}$ we can simply equate the observed value of Y with the expected value of Y. Solving $y = n\tilde{p}$ yields the estimate $\tilde{p} = y/n$. So MLE agrees with MoM in this example. It also agrees with common sense since, e.g., if we obtain exactly 7 successes in 10 independent trials, then you do not need to know much statistics to come up with the guess 0.7 for the probability of success.

For a couple of quick metrics to assess the MLE, we can compute its bias and standard error. The bias is

$$\text{bias}(\hat{p}) = \text{E}[\hat{p}] - p = \frac{pn}{n} - p = 0,$$

which says that $\hat{p}$ is unbiased. The standard error is, using the result for the variance of a Binomial,

$$\text{SE}(\hat{p}) = \sqrt{\text{Var}(\hat{p})} = \sqrt{\frac{np(1-p)}{n^2}} = \sqrt{\frac{p(1-p)}{n}}.$$

Comfortingly, the standard error decreases as n increases, and goes to 0 as $n \to \infty$. Less comfortingly, the $\sqrt{n}$ in the denominator grows rather slowly. This situation is very common in statistics: many widely used estimators have standard errors that are approximately $a/\sqrt{n}$ for large n, where a is a constant and n is the sample size. This says, for example, that to reduce the standard error by a factor of ten we need a hundred times as much data.

Since p is unknown, the standard error of $\hat{p}$ is also unknown, but the function $p(1-p)$ is maximized at $p = 1/2$, so we can give an upper bound on the standard error that does not depend on p:

$$\mathrm{SE}(\hat{p}) \leq \sqrt{\frac{1/4}{n}} = \frac{1}{2\sqrt{n}}.$$

This says, e.g., that to guarantee a standard error of at most 0.01, we need to have a sample size of at least 2,500.

4.1.2 Normal distribution examples

Example 4.1.5 (Normal with known variance)**.** Let $Y_1, \ldots, Y_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$, with $\theta = \mu$ unknown but σ^2 known. This situation is contrived—why would we know the variance if we do not even known the mean?— but it does come up occasionally, and in any case it is a useful stepping stone toward the more complicated model where both parameters are unknown. The likelihood function is, dropping the Normal normalizing constant,

$$L(\mu; \mathbf{y}) = \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - \mu)^2\right\},$$

so the log-likelihood function is

$$l(\mu; \mathbf{y}) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_j - \mu)^2 = -\frac{1}{2\sigma^2}\left\{\sum_{j=1}^{n}(y_j - \bar{y})^2 + n(\bar{y} - \mu)^2\right\},$$

by the sum of squares identity. The log-likelihood function is a *quadratic* function of the unknown parameter, which turns out is an ideal situation as far as likelihood and maximum likelihood estimation are concerned.

It is easy to maximize $l(\mu; \mathbf{y})$: just set $\mu = \bar{y}$. Any other choice of μ would make $(\bar{y} - \mu)^2 > 0$, resulting in a strictly smaller value for the likelihood. So the MLE is

$$\hat{\mu} = \bar{y},$$

which makes sense intuitively and agrees with the MoM of a mean. Often it is difficult to find the exact distribution of the MLE, but here we already know from the fact that the sum of independent Normals is Normal, and properties of mean and variance that

$$\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

So $\hat{\mu}$ is unbiased, with standard error

$$\mathrm{SE}(\hat{\mu}) = \frac{\sigma}{\sqrt{n}}.$$

Once again the standard error goes to 0 as a constant over $\sqrt{n}$.

Example 4.1.6 (Normal with both parameters unknown)**.** More realistically, let $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$ be i.i.d., with both parameters unknown. We will parameterize the model in terms of the mean and standard deviation, $\theta = (\mu, \sigma)$, rather than mean and variance, but the results will be equivalent either way. Just make sure to be consistent and careful about which you are using; many careless errors in statistics stem from mixing up standard deviation and variance!

The likelihood function is the same as in the previous example, except that now it is a function of both μ and σ, and we need to include the σ from the Normal normalizing constant for each y_j:

$$L(\mu, \sigma; \mathbf{y}) = \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - \mu)^2\right\}.$$

So the log-likelihood function is

$$l(\mu, \sigma; \mathbf{y}) = -\frac{1}{2\sigma^2}\left\{\sum_{j=1}^{n}(y_j - \bar{y})^2 + n(\bar{y} - \mu)^2\right\} - n\log\sigma.$$

Typically some multivariable calculus would be needed at this point if finding an MLE mathematically: set the partial derivatives $\partial l(\mu, \sigma; \mathbf{y})/\partial\mu$ and $\partial l(\mu, \sigma; \mathbf{y})/\partial\sigma$ equal to 0, solve for μ and σ, and then apply the second derivative test to check whether we have found a maximum. We will go through this process at the end of this chapter (to illustrate techniques that can be used more generally), but first we will take a route that only needs single variable calculus.

First note that for any σ, the choice of μ that maximizes $l(\mu, \sigma; \mathbf{y})$ is $\mu = \bar{y}$. This is because μ only appears in the term $n(\bar{y} - \mu)^2$, and this term is 0 when $\mu = \bar{y}$ and positive when $\mu \neq \bar{y}$. Due to the minus sign at the very front of $l(\mu, \sigma; \mathbf{y})$, the optimal choice of μ is then $\mu = \bar{y}$.

Plugging in $\mu = \bar{y}$, we now want to find the value of σ that maximizes

$$-\frac{1}{2\sigma^2}t - n\log\sigma,$$

where

$$t = \sum_{j=1}^{n}(y_j - \bar{y})^2.$$

(Often it is helpful to introduce a little bit of notation to make expressions look simpler!) Setting the derivative with respect to σ equal to 0,

$$\frac{t}{\hat{\sigma}^3} - \frac{n}{\hat{\sigma}} = 0.$$

This rearranges to

$$\hat{\sigma}^2 = \frac{t}{n}.$$

We have found a maximum since the second derivative is

$$-\frac{3t}{\sigma^4}+\frac{n}{\sigma^2},$$

which, evaluated at $\hat{\sigma}^2 = t/n$, is

$$-\frac{3t}{t^2/n^2}+\frac{n}{t/n}=-\frac{2n^2}{t}<0.$$

Hence, the maximum likelihood estimator of θ is

$$\hat{\mu}=\bar{Y},\quad \hat{\sigma}^2=\frac{1}{n}\sum_{j=1}^{n}(Y_j-\bar{Y})^2.$$

We will show in the next section that the MLE of σ^2 is the square of the MLE of σ. This property is very convenient, as well as notationally pleasant since it makes the notation $\hat{\sigma}^2$ unambiguous (it does not matter whether the hat is applied first or the square is applied first).

The estimator $\hat{\sigma}^2$ is the sample variance, except that the sample variance is usually defined with $n-1$ rather than n in the denominator. For large n, it makes very little difference whether $n-1$ or n is used in the denominator. For small n, it can make a substantial difference.

Example 4.1.7 (MLE for Cauchy Model)**.** The Cauchy distribution, which is notorious for being challenging to work with, is the basis for a statistical model where the MLE is far less intuitive (and far less mathematically tractable) than in the Normal model. Recall from Chapter 7 of the Stat 110 book that the Cauchy distribution is the distribution of the ratio Z_1/Z_2, for Z_1, Z_2 i.i.d. $\mathcal{N}(0,1)$.

Consider the following Cauchy statistical model: let $Y_1, Y_2, \ldots, Y_n$ be i.i.d. with $Y_j = C_j + \theta$, where the C_j are i.i.d. Cauchy random variables. So the PDF of each Y_j is

$$f_{Y_1}(y;\theta)=\frac{1}{\pi\{1+(y-\theta)^2\}}.$$

We wish to estimate θ, which is called a *location parameter* since Y_j is obtained from C_j by shifting the location of the median, making the distribution symmetric about θ rather than 0.

Using the sample mean $\bar{Y}$ to estimate θ would be a disaster, since it turns out that $\bar{Y}$ has the same distribution as Y_1. For example, the sample mean based on a million Y_j's has the same distribution as Y_1, so if we used $\bar{Y}$ we would essentially be throwing away all but one of our million data points!

A much more prudent choice of estimator would be the sample median. Intuitively it makes sense to use the sample median rather than the sample mean here, because the Cauchy has very heavy tails and the sample median, $Y_{(\lceil 0.5n \rceil)}$, is much less sensitive than the sample mean to extreme observations. Specializing the asymptotic distribution of a quantile from Chapter 3 to the sample median for Cauchy data gives the asymptotic distribution

$$\sqrt{n}\{Y_{(\lceil 0.5n \rceil)}-\theta\} \xrightarrow{d} \mathcal{N}(0,\pi^2/4).$$

Still, it seems rather ad hoc to use the sample median without some principle to justify it. Instead, let us try to find the MLE of θ. The log-likelihood function is

$$l(\theta) = -n\log(\pi) - \sum_{j=1}^{n} \log\{1 + (y_j - \theta)^2\},$$

so

$$\frac{\partial l(\theta)}{\partial \theta} = 2\sum_{j=1}^{n} \frac{y_j - \theta}{1 + (y_j - \theta)^2}.$$

Unfortunately, the solution to the equation

$$\sum_{j=1}^{n} \frac{y_j - \hat{\theta}}{1 + (y_j - \hat{\theta})^2} = 0$$

is not available in closed form. Worse yet, the log-likelihood function turns out *not* to be concave and there may be multiple modes. On the bright side, iterative methods such as Newton's method (also known as Newton-Raphson) can often be applied to such problems. Newton's method is introduced in Section 4.5.

Using code given in Section 4.9.1, we simulated from the Cauchy location model, with a small sample size $n = 10$ and the true value of θ being 0. The log-likelihood function is plotted in Figure 4.2. The log-likelihood function looks quite irregular in the left hand side of Figure 4.2; recall that the ideal case scenario is for it to look *quadratic*.

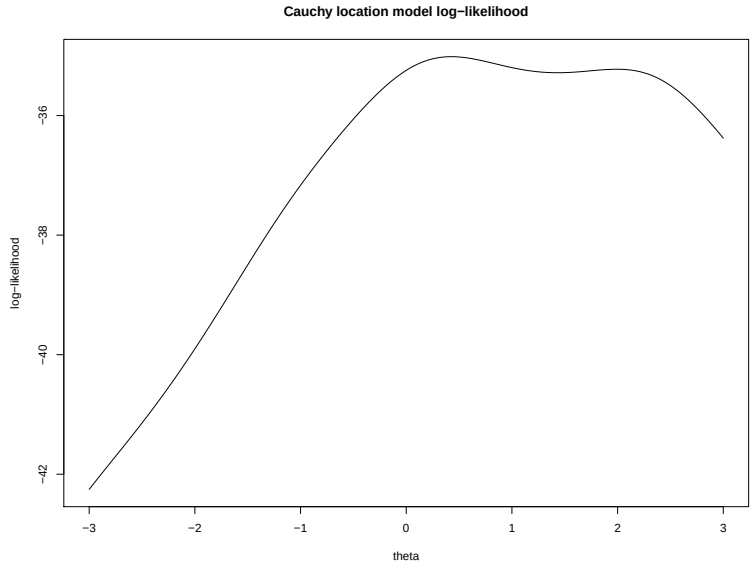


Figure 4.2: Log-likelihood function for one simulated example of 10 Cauchy data points with $\theta = 0$.

The sample mean in this simulation was -2.25 (very far from the truth), whereas the sample median was 0.17 (much better). The MLE was numerically found to be 0.43, a bit worse than the

sample median (though here we just ran this *once*, for illustrative purposes, rather than doing a comprehensive simulation study).

Re-running the above code with sample size $n = 100$ instead of $n = 10$, the log-likelihood function looks much more like a quadratic, as shown in Figure 4.3.

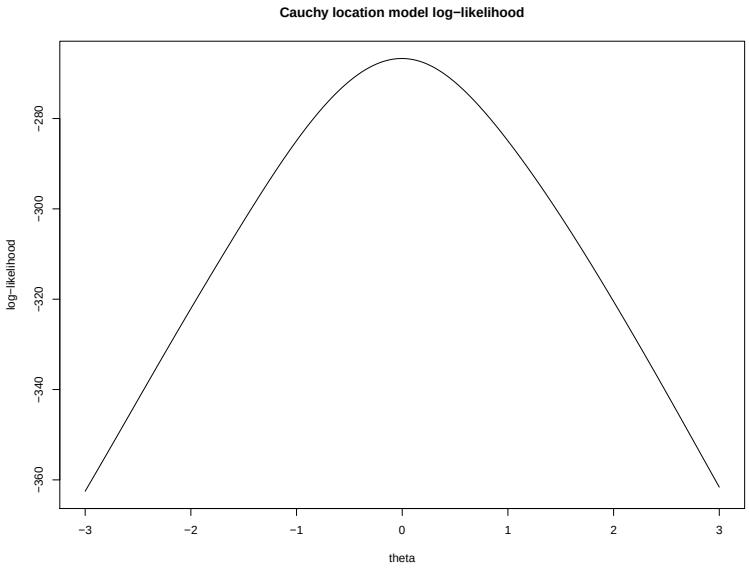


Figure 4.3: Log-likelihood function for one simulated example of 100 Cauchy data points with $\theta = 0$.

Now the sample mean is -0.52 (still bad, and only by coincidence is it better than when $n = 10$), the sample median is 0.018 (quite good), and the MLE is -0.0036 (extremely good).

4.2 Properties of the MLE

In addition to being intuitively appealing, the MLE has some excellent theoretical properties, under some assumptions. There are no panaceas in statistics; every method has limitations. We will see examples later in the course where another estimator is clearly preferable over the MLE. Overall though, the MLE tends to work well both in theory and in practice.

Some of the main properties that the MLE enjoys are as follows. All of these require some technical assumptions known as *regularity conditions*. Under these assumptions, which we will discuss more later, we have the following.

- The MLE is *invariant*, which means that if $\hat{\theta}$ is the MLE of θ, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.
- The MLE $\hat{\theta}$ is *consistent*, which means that it converges in probability to the true θ.

- The MLE is *asymptotically Normal* (so its distribution is approximately Normal if the sample size is large).
- The MLE is *asymptotically unbiased* (the bias approaches 0 as the sample size grows).
- The MLE is *asymptotically efficient* (no other asymptotically unbiased estimator will have a lower standard error asymptotically).

First let's study *invariance*, the first property on the above list.

Theorem 4.2.1 (Invariance of MLE)**.** *Let $\hat{\theta}$ be the MLE of θ, and let g be a one-to-one function. Then the MLE of $g(\theta)$ is $g(\hat{\theta})$.*

Proof. This result follows from the invariance property of likelihood. Let $\tau = g(\theta)$. Each point on the reparameterized likelihood curve $L(\tau)$ has a corresponding point on the original likelihood function $L(\theta)$, such that the likelihood value for τ is the same as the likelihood value for the corresponding θ. In particular, the value $\hat{\tau}$ that maximizes $L(\tau)$ is $\hat{\tau} = g(\hat{\theta})$. ∎

The invariance property of MLE is incredibly convenient. If we have calculated the MLE $\hat{\theta}$ for θ, and g is one-to-one, then we immediately know that the MLE of $g(\theta)$ is $g(\hat{\theta})$, with no further calculation required. In contrast, if we take a Bayesian approach, then a widely used estimator for a parameter is its *posterior mean*,

$$\hat{\theta}_{\text{Bayes}} = \text{E}(\theta|\mathbf{y}).$$

This procedure is *not* invariant though. Typically in probability $\text{E}[h(X)] \neq h(\text{E}[X])$ when h is a nonlinear function (moreover, we know which way the inequality goes by Jensen's inequality, if h is convex or concave). So in general

$$\widehat{g(\theta)}_{\text{Bayes}} = \text{E}[g(\theta)|\mathbf{y}] \neq g(\text{E}[\theta|\mathbf{y}]) = g(\hat{\theta}_{\text{Bayes}}).$$

Invariance of the MLE is so convenient, in fact, that we *define* it to be true even when g is not one-to-one.

Definition 4.2.2 (MLE under a parameter transformation that is not one-to-one)**.** If $\hat{\theta}$ is the MLE of θ and g is not a one-to-one function, then we define the MLE of $g(\theta)$ to be $g(\hat{\theta})$.

With this definition, invariance *always* applies, including when θ is a vector. The reason this separate definition is needed is that if g is not a one-to-one function, then $\tau = g(\theta)$ is *not* simply a reparameterization of the model: one value of τ can correspond to more than one distribution in the

model, making it unclear what $L(\tau)$ even means. For example, if the parameter space for θ is $(-\infty, \infty)$ we still say that the MLE of θ^2 is the square of the MLE of θ even though the squaring function is not one-to-one. And if $\theta = (\theta_1, \theta_2)$ is two-dimensional and the MLE is $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$, then it is natural to say that $\hat{\theta}_1$ is the MLE of θ_1, even though $(\theta_1, \theta_2) \mapsto \theta_1$ is not a one-to-one function (so it is unclear what the likelihood function of θ_1 viewed in isolation would even mean).

Example 4.2.3 (MLE of an unknown probability)**.** Let $Y_1, \ldots, Y_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$, with both parameters unknown. Let $(\hat{\mu}, \hat{\sigma})$ be the MLE of (μ, σ), which was obtained in Example 4.1.6. Invariance explains why the MLE of σ^2 is simply the square of $\hat{\sigma}$. Extending this example, suppose that Y_j is the height of the jth individual (in feet), our statistical model is that

$$Y_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2),$$

but our estimand θ is the probability of someone being more than six feet tall, which is

$$\theta = P(Y_1 > 6) = P\left(\frac{Y_1 - \mu}{\sigma} > \frac{6 - \mu}{\sigma}\right) = 1 - \Phi\left(\frac{6 - \mu}{\sigma}\right).$$

Then we just need to make the parameters don hats to get the MLE of θ:

$$\hat{\theta} = 1 - \Phi\left(\frac{6 - \hat{\mu}}{\hat{\sigma}}\right).$$

☣ **4.2.4.** Invariance of the MLE is *convenient*, but that does not mean that the resulting estimators are always *good*. For example, suppose that our estimand is $\theta = e^{\mu}$. By invariance, the MLE is $\hat{\theta} = e^{\hat{\mu}}$. If $\hat{\mu} = \bar{Y} \sim N(\mu, \sigma^2/n)$, then the MLE is Log-Normal:

$$\hat{\theta} \sim \mathcal{LN}(\mu, \sigma^2/n).$$

Using the result for the mean of a Log-Normal (see Chapter 6 of the Stat 110 book),

$$\mathrm{E}(\hat{\theta}) = \exp\left(\mu + \frac{\sigma^2}{2n}\right) = \theta e^{\sigma^2/(2n)}.$$

If n is large compared with σ^2, then the right-hand side is approximately θ. But if n is not large compared with σ^2, then $\hat{\theta}$ is severely biased. For example, if $n = \sigma^2$ then $\mathrm{E}(\hat{\theta}) \approx 1.65\,\theta$, i.e., on average the estimator is 65% larger than the truth.

4.3 Kullback-Leibler divergence

So far we have used the notation θ both for the estimand and for the argument in the likelihood $L(\theta; \mathbf{y})$. This has the virtue of simplicity, but in some situations could lead to confusion about when θ

denotes the fixed *true* value of the parameter and when it is a variable that can range freely over the parameter space. To study the properties of likelihood functions in more detail, it is helpful to have a separate notation for the estimand.

Notation 4.3.1. Let θ^* be the estimand, such that the random variables $Y_1, \ldots, Y_n$ are generated by the joint CDF $F_{\mathbf{Y};\theta^*}$. The hope, of course, is that our estimators will be close to θ^*. The distribution $F_{\mathbf{Y};\theta^*}$ is called the *data generating process*. To summarize our notation:

- $\hat{\theta} = \underset{\theta \in \Theta}{\arg\max}\ \log L(\theta; \mathbf{Y})$ is an estimator,
- θ is the argument in the likelihood function $L(\theta; \mathbf{Y})$, and
- θ^* is the true value or estimand, generating $\mathbf{Y}$ through $F_{\mathbf{Y};\theta^*}$.

Many of the useful properties of the MLE turn out to be closely related to the expected log-likelihood evaluated at θ,

$$\mathrm{E}[\log L(\theta; \mathbf{Y})] = \int \log L(\theta; \mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}; \theta^*) d\mathbf{y}.$$

This quantity averages the log-likelihood over the random $\mathbf{Y}$ produced by $F_{\mathbf{Y};\theta^*}$. In particular, it is a function of θ alone (for fixed θ^*), since we have averaged over all $\mathbf{Y}$. Note the distinction between the θ and the θ^* in the above equation!

In this section we will show that $\mathrm{E}[\log L(\theta; \mathbf{Y})]$, as θ varies, is globally maximized at θ^*. This is a fundamental property of likelihoods. It implies that high likelihoods tend to pop up for values of θ that are near θ^*. This result is *very* general, and does *not* assume the data are i.i.d. The result will follow for all models that satisfy some mild regularity conditions. In particular, it will apply to, e.g., various problems in time series, networks, and spatial statistics, where we have to account for dependence in the data.

To establish this result we introduce the *Kullback-Leibler divergence*, which is a key quantity in statistics, machine learning, and information theory. It measures how close a distribution is to the distribution which generates the data. We will give the definition for general PDFs and then specialize it to the context of a likelihood function.

Definition 4.3.2 (KL divergence)**.** The *Kullback-Leibler divergence* (KL divergence) from the CDF F to the CDF G is

$$D_{\mathrm{KL}}(F||G) = \mathrm{E}\left(\log \frac{f(\mathbf{Y})}{g(\mathbf{Y})}\right) = \mathrm{E}[\log f(\mathbf{Y}) - \log g(\mathbf{Y})] = \int \{\log f(\mathbf{y}) - \log g(\mathbf{y})\} f(\mathbf{y}) d\mathbf{y},$$

where f and g are the PDFs corresponding to F and G (in the case of discrete distributions, we replace the PDFs with PMFs). The expectations are computed with $\mathbf{Y}$ generated according to F, not G.

An important case of KL divergence is when $F = F_{\mathbf{Y};\theta^*}$ and $F = F_{\mathbf{Y};\theta}$. Then

$$D_{\mathrm{KL}}(F_{\mathbf{Y};\theta^*}||F_{\mathbf{Y};\theta}) = \mathrm{E}[\log L(\theta^*;\mathbf{Y}) - \log L(\theta;\mathbf{Y})].$$

Intuitively, $D_{\mathrm{KL}}(F_{\mathbf{Y};\theta^*}||F_{\mathbf{Y};\theta})$ measures how much the expected log-likelihood is higher at θ^* than at θ, computing the expectation under the true distribution $F_{\mathbf{Y};\theta^*}$.

☣ **4.3.3.** Kullback-Leibler divergence is sometimes called "Kullback-Leibler distance", but it is not a distance in the usual mathematical sense, since it is not symmetric and does not satisfy the triangle inequality. It is a *directional* measure: it measures the effect on the average log-likelihood if we start from the true distribution and then "move" to some other distribution.

Example 4.3.4 (KL divergence in a Normal model)**.** Suppose that

$$Y_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2),$$

and we have a sample of size n, where σ^2 is known. Then

$$\log \frac{L(\theta^*;\mathbf{Y})}{L(\theta;\mathbf{Y})} = -\frac{1}{2\sigma^2}\sum_{j=1}^{n}(Y_j-\theta^*)^2 + \frac{1}{2\sigma^2}\sum_{j=1}^{n}(Y_j-\theta)^2,$$

so

$$\begin{aligned} D_{\mathrm{KL}}(F_{\mathbf{Y};\theta^*}||F_{\mathbf{Y};\theta}) &= -\frac{n\sigma^2}{2\sigma^2} + \frac{n}{2\sigma^2}\mathrm{E}[(Y_1-\theta)^2] = -\frac{1}{2\sigma^2}n\sigma^2 + \frac{1}{2\sigma^2}n\{\mathrm{E}(Y_1-\theta^*)^2 + (\theta^*-\theta)^2\} \\ &= \frac{n(\theta^*-\theta)^2}{2\sigma^2}. \end{aligned}$$

The KL divergence is plotted in Figure 4.4, as a function of θ. Notice that $D_{\mathrm{KL}}(F_{\mathbf{Y};\theta^*}||F_{\mathbf{Y};\theta})$ is uniquely minimized where $\theta = \theta^*$, so in this case the expected log-likelihood is maximized at θ^*.

Theorem 4.3.5 (Additivity of KL divergence)**.** *If we observe independent $Y_1, \ldots, Y_n$, then*

$$D_{KL}(F_{\mathbf{Y};\theta^*}||F_{\mathbf{Y};\theta}) = \sum_{j=1}^{n} D_{KL}(F_{Y_j;\theta^*}||F_{Y_j;\theta}),$$

where $D_{KL}(F_{Y_j;\theta^*}||F_{Y_j;\theta})$ is the Kullback-Leibler divergence for the jth observation:

$$D_{KL}(F_{Y_j;\theta^*}||F_{Y_j;\theta}) = \mathrm{E}\left(\log \frac{L(\theta^*;Y_j)}{L(\theta;Y_j)}\right).$$

In the i.i.d. case,

$$D_{KL}(F_{\mathbf{Y};\theta^*}||F_{\mathbf{Y};\theta}) = nD_{KL}(F_{Y_1;\theta^*}||F_{Y_1;\theta}).$$

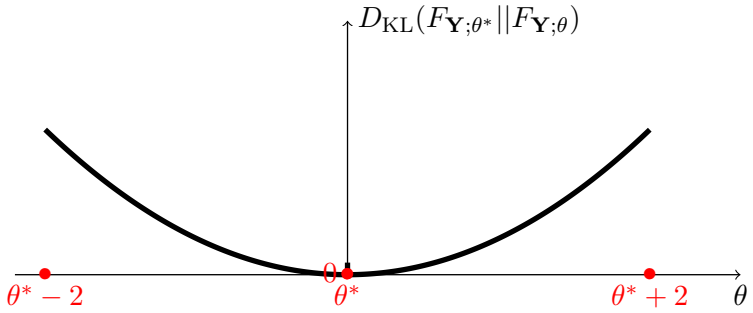


Figure 4.4: Example of the Kullback-Leibler divergence, $D_{\mathrm{KL}}(F_{\mathbf{Y};\theta^*}||F_{\mathbf{Y};\theta})$ plotted against θ.

Proof. By the independence assumption,

$$\log L(\theta; \mathbf{Y}) = \sum_{j=1}^{n} \log L(\theta; Y_j),$$

and likewise for $\log L(\theta^*; \mathbf{Y})$. The result then follows from linearity of expectation. ∎

We will now show that Kullback-Leibler divergence is always nonnegative. This simple property has far-reaching consequences.

Theorem 4.3.6 (Nonnegativity of KL divergence)**.** *For any CDFs F and G,*

$$D_{KL}(F||G) \geq 0.$$

This inequality is strict unless $F = G$, that is they are the same distribution functions.

Proof. The key idea is to use Jensen's inequality, which will be applicable since log is a concave function. For concreteness, we will write the proof in the continuous case; the discrete case can be handled analogously. First let us verify a fact which is useful in its own right: the expected value of the likelihood ratio $g(\mathbf{Y})/f(\mathbf{Y})$ is 1. By LOTUS,

$$\mathrm{E}\left(\frac{g(\mathbf{Y})}{f(\mathbf{Y})}\right) = \int \frac{g(\mathbf{y})}{f(\mathbf{y})} f(\mathbf{y}) \mathrm{d}\mathbf{y} = \int g(\mathbf{y}) \mathrm{d}\mathbf{y} = 1.$$

The function $-\log(x)$ is convex (since its second derivative is $1/x^2 > 0$), so Jensen's inequality says that for any positive random variable X,

$$-\mathrm{E}[\log(X)] \geq -\log \mathrm{E}[X].$$

Therefore,

$$D_{\mathrm{KL}}(F||G) = -\mathrm{E}\left\{\log \frac{g(\mathbf{Y})}{f(\mathbf{Y})}\right\} \geq -\log \mathrm{E}\left\{\frac{g(\mathbf{Y})}{f(\mathbf{Y})}\right\} = -\log 1 = 0,$$

as desired. If F and G are the same distribution then $D_{\mathrm{KL}}(F||G) = 0$; otherwise, $D_{\mathrm{KL}}(F||G) > 0$. ∎

In the context of likelihoods, again take $F = F_{\mathbf{Y};\boldsymbol{\theta}^*}$ and $G = F_{\mathbf{Y};\theta}$. Then this theorem means that if $F_{\mathbf{Y};\theta}$ and $F_{\mathbf{Y};\boldsymbol{\theta}^*}$ are the same CDF then $D_{\mathrm{KL}}(F_{\mathbf{Y};\theta^*}||F_{\mathbf{Y};\theta}) = 0$; otherwise, we will have that

$$D_{\mathrm{KL}}(F_{\mathbf{Y};\theta^*}||F_{\mathbf{Y};\theta}) > 0.$$

Thus θ^* is always the value which maximizes the expected log-likelihood function.

This is a remarkable result. It applies without specifying any particular form for the statistical model (the statistical model could be, for example, for network data); the sole assumption is that the mean of the log of the likelihood ratio exists. The result lends support for using maximum likelihood estimation: the MLE $\hat{\theta}$ is where the *observed* log-likelihood function has its peak, while θ^* is where the *expected* log-likelihood function has its peak. Using the MLE exploits the fact that the log-likelihood will tend to have its peak near θ^*.

We can also use the Kullback-Leibler divergence result to prove a consistency theorem for the MLE. It will be proved under an i.i.d. assumption, although it holds much more broadly.

Theorem 4.3.7 (Consistency of MLE)**.** *Suppose that the parameter space is finite and that the observations $Y_1, \ldots, Y_n$ are i.i.d. Also assume that for $\theta_1 \neq \theta_2$, the distribution function $F_{Y;\theta_1}$ is different from the distribution of $F_{Y;\theta_2}$ (this is known as* identifiability *of the model). Then the MLE $\hat{\theta}$ is consistent:*

$$\hat{\theta} \xrightarrow{p} \theta^*,$$

as the sample size $n \to \infty$.

Proof. For each θ, the strong law of large numbers implies that

$$\frac{1}{n}\left[\log L(\theta^*;\mathbf{Y}) - \log L(\theta;\mathbf{Y})\right] = \frac{1}{n}\sum_{j=1}^{n} \log \frac{f(Y_j;\theta^*)}{f(Y_j;\theta)} \to D_{\mathrm{KL}}(F_{Y_1;\theta^*}||F_{Y_1;\theta}),$$

with probability 1, recalling $D_{\mathrm{KL}}(F_{Y_1;\theta^*}||F_{Y_1;\theta})$ is the Kullback-Leibler divergence for a single random variable Y_1. Further, recall the probability 1 result means in particular that for all choices of $\epsilon > 0$ it is possible to find an N such that for all $n > N$,

$$\frac{1}{n}\left[\log L(\theta^*;\mathbf{Y}) - \log L(\theta;\mathbf{Y})\right] - D_{\mathrm{KL}}(F_{Y_1;\theta^*}||F_{Y_1;\theta}) > -\epsilon,$$

with probability one. Let

$$c = \min_{\theta \neq \theta^*} D_{\mathrm{KL}}(F_{Y_1;\theta^*}||F_{Y_1;\theta}) > 0,$$

and then select ϵ having the property that $0 < \epsilon < c$. Then, with probability 1 there is an N such that for all $n \geq N$ and all $\theta \neq \theta^*$,

$$\log L(\theta^*;\mathbf{Y}) - \log L(\theta;\mathbf{Y}) > n(c - \epsilon),$$

holds. This implies that $L(\theta^*; \mathbf{Y})$ is *vastly* larger than any other $L(\theta; \mathbf{Y})$ if n is very large, and the unique MLE is $\hat{\theta} = \theta^*$ for all $n \geq N$. Because of the finiteness of the parameter space, the MLE will *equal* θ^* with probability 1 for all sufficiently large n, which is actually a stronger conclusion than merely converging to θ^* in probability. ∎

Consistency of the MLE holds far more generally than under the above assumptions, but the proofs get much more technical when we move to more general parameter spaces or beyond i.i.d. data.

4.3.1 Score function

Definition 4.3.8 (Score function)**.** The *score function* is

$$s(\theta; \mathbf{y}) = \frac{\partial \log L(\theta; \mathbf{y})}{\partial \theta} = \frac{1}{L(\theta; \mathbf{y})} \frac{\partial L(\theta; \mathbf{y})}{\partial \theta}.$$

Figure 4.5 shows examples of the score functions drawn against θ for three different datasets: $s(\theta; \mathbf{y})$, $s(\theta; \mathbf{z})$ and $s(\theta; \mathbf{w})$.

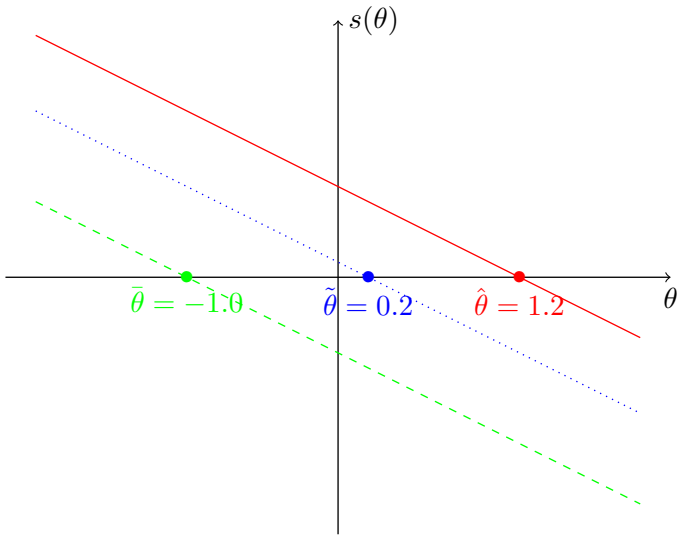


Figure 4.5: Example of the score function for 3 different datasets: $\mathbf{y}$, $\mathbf{z}$ and $\mathbf{w}$. Red line and MLE $\hat{\theta}$ corresponds to dataset $\mathbf{y}$, while blue line and MLE $\tilde{\theta}$ corresponds to $\mathbf{z}$ and green line and MLE $\bar{\theta}$ corresponds to $\mathbf{w}$.

Who cares what the score is? We have already seen one application of the score function: setting it equal to 0 and solving for θ is a standard approach to finding the MLE. We will soon see that the score function also helps us quantify the *uncertainty* of the MLE.

☣ **4.3.9.** Make sure to keep track of what is regarded as fixed and what is regarded as random (for the score function, but also more generally in statistics). As defined, the score function depends on

both θ and on $\mathbf{y}$. In some applications, such as when finding the MLE, we fix $\mathbf{y}$ and look at $s(\theta;\mathbf{y})$ as a function of θ. In other applications, it turns out to be useful to fix θ at its true value θ^* and look at the random variable $s(\theta^*;\mathbf{Y})$. In still other applications, we are interested in some specific hypothesized parameter θ_0 and we look at the statistic $s(\theta_0;\mathbf{Y})$. Note that $s(\theta^*;\mathbf{Y})$ is *not* a statistic since θ^* is unknown.

A fundamental result about the score function is that its mean is 0 and its variance takes on a nice form, when $\theta = \theta^*$. Again, notice this argument is for *general* statistical models of $\mathbf{Y}$, not requiring an i.i.d. assumption.

Theorem 4.3.10 (Information equality)**.** *Under some regularity conditions (mainly that the $f_{\mathbf{Y}}(\mathbf{y};\theta)$ is a smooth function in θ, the support of $\mathbf{Y}$ does not depend on θ, that the expected values needed below exist, and that we can DUThIS (differentiate under the integral sign) when needed below),*

$$\begin{aligned} \mathrm{E}[s(\theta^*;\mathbf{Y})] &= 0, \\ \mathrm{Var}\{s(\theta^*;\mathbf{Y})\} &= -\mathrm{E}[s'(\theta^*;\mathbf{Y})]. \end{aligned}$$

The prime in s' denotes taking the partial derivative with respect to θ, that is $s'(\theta^*;\mathbf{Y}) = \partial s(\theta^*;\mathbf{Y})/\partial\theta$.

Figure 4.6 illustrates the first equation, that the expected score is 0 at $\theta = \theta^*$. The second equation, that the variance of the score equals the expectation of the second derivative of the log-likelihood, is called the *information equality* in statistics. It will turn out to be a very useful result.

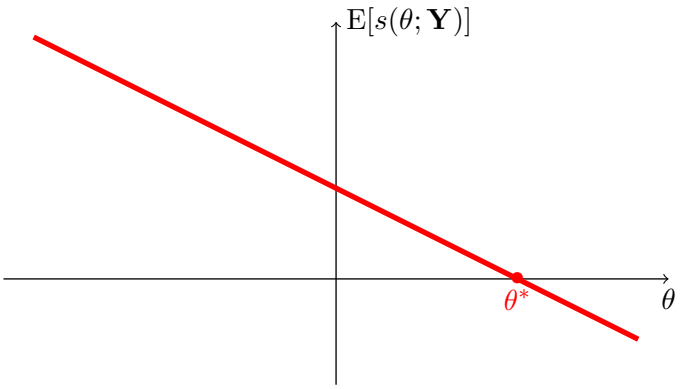


Figure 4.6: Expected value of the score function, $\mathrm{E}[s(\theta;\mathbf{Y})]$, which is zero at $\theta = \theta^*$.

Proof. To simplify notation, we will write the proof for the continuous case; the discrete case is

analogous. By LOTUS,

$$\begin{aligned}
\mathrm{E}[s(\theta;\mathbf{Y})] &= \int \frac{\partial \log L(\theta;\mathbf{y})}{\partial\theta} f(\mathbf{y};\theta^*)\mathrm{d}\mathbf{y}, \\
&= \int \frac{1}{L(\theta;\mathbf{y})}\frac{\partial L(\theta;\mathbf{y})}{\partial\theta} f(\mathbf{y};\theta^*)\mathrm{d}\mathbf{y}, \\
&= \int \frac{\partial f(\mathbf{y};\theta)}{\partial\theta}\frac{f(\mathbf{y};\theta^*)}{f(\mathbf{y};\theta)}\mathrm{d}\mathbf{y}, \quad \text{as } L(\theta;\mathbf{y}) = f(\mathbf{y};\theta).
\end{aligned}$$

In the special case where $\theta = \theta^*$, the above result reduces to

$$\mathrm{E}[s(\theta^*;\mathbf{Y})] = \int \left.\frac{\partial f(\mathbf{y};\theta)}{\partial\theta}\right|_{\theta=\theta^*}\mathrm{d}\mathbf{y}.$$

Assuming we can DUThIS (differentiate under the integral sign), we can interchange the order of integration and differentiation to get

$$\begin{aligned}
\mathrm{E}[s(\theta^*;\mathbf{Y})] &= \left(\frac{\partial}{\partial\theta}\int f(\mathbf{y};\theta)\mathrm{d}\mathbf{y}\right)\Big|_{\theta=\theta^*} \\
&= \left(\frac{\partial}{\partial\theta}1\right)\Big|_{\theta=\theta^*} \\
&= 0.
\end{aligned}$$

So the expected score is 0 at $\theta = \theta^*$. To find the variance of the score function at $\theta = \theta^*$, note that

$$\mathrm{E}[s'(\theta;\mathbf{Y})] = \int \frac{\partial^2 \log L(\theta;\mathbf{y})}{\partial\theta^2} f(\mathbf{y};\theta^*)\mathrm{d}\mathbf{y}.$$

Now

$$\begin{aligned}
\frac{\partial \log L(\theta;\mathbf{y})}{\partial\theta} &= \frac{1}{L(\theta;\mathbf{y})}\frac{\partial L(\theta;\mathbf{y})}{\partial\theta}, \quad \text{so} \\
\frac{\partial^2 \log L(\theta;\mathbf{y})}{\partial\theta^2} &= -\frac{1}{\{L(\theta;\mathbf{y})\}^2}\left\{\frac{\partial L(\theta;\mathbf{y})}{\partial\theta}\right\}^2 + \frac{1}{L(\theta;\mathbf{y})}\frac{\partial^2 L(\theta;\mathbf{y})}{\partial\theta^2} \\
&= -\left\{\frac{\partial \log L(\theta;\mathbf{y})}{\partial\theta}\right\}^2 + \frac{1}{L(\theta;\mathbf{y})}\frac{\partial^2 L(\theta;\mathbf{y})}{\partial\theta^2}.
\end{aligned}$$

Using the same argument as before,

$$\mathrm{E}\left\{\frac{1}{L(\theta^*;\mathbf{Y})}\frac{\partial^2 L(\theta^*;\mathbf{Y})}{\partial\theta^2}\right\} = 0,$$

so

$$\begin{aligned}
\mathrm{E}[s'(\theta^*;\mathbf{Y})] &= \int \frac{\partial^2 \log L(\theta^*;\mathbf{y})}{\partial\theta^2} f(\mathbf{y};\theta^*)\mathrm{d}\mathbf{y} \\
&= -\int \{s(\theta^*;\mathbf{y})\}^2 f(\mathbf{y};\theta^*)\mathrm{d}\mathbf{y}, \\
&= -\mathrm{Var}\{s(\theta^*;\mathbf{Y})\}.
\end{aligned}$$

∎

☣ **4.3.11.** One of the regularity conditions is that the support of the data should not depend on θ. Weird things can happen if this assumption does not hold, e.g., for the model $Y_j \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, \theta)$.

The result that the expected score at θ^* is zero is not surprising given that the Kullback-Leibler divergence $D_{\text{KL}}(F_{\mathbf{Y};\theta^*}||F_{\mathbf{Y};\theta})$ is minimized at $\theta = \theta^*$, but the direct proof is informative and leads onto the result for the variance of the score. This second result is fundamental in statistics.

4.3.2 Fisher information

Fisher information is a measure of how informative the data are about a estimand in a parametric statistical model. Fisher information plays a central role in statistical inference. For example, it turns out to be a very useful concept for understanding how the MLE behaves with large samples and for obtaining general bounds on how well we can estimate a parameter.

Definition 4.3.12 (Fisher information)**.** The *Fisher information* in the sample for a parameter θ in a parametric statistical model $F_{\mathbf{Y};\theta}$ is

$$\mathcal{I}_{\mathbf{Y}}(\theta^*) = \text{Var}\left\{s(\theta^*; \mathbf{Y})\right\} = \text{E}[s(\theta^*; \mathbf{Y})^2],$$

where we compute the variance under the assumption that the true parameter value is θ. Let

$$\mathcal{J}_{\mathbf{Y}}(\theta^*) = -\text{E}[s'(\theta^*; \mathbf{Y})].$$

Then

$$\mathcal{I}_{\mathbf{Y}}(\theta^*) = \mathcal{J}_{\mathbf{Y}}(\theta^*),$$

the *information equality*. In statistics it is traditional to suppress θ^* and write $\mathcal{I}_{\mathbf{Y}}(\theta)$ and $\mathcal{J}_{\mathbf{Y}}(\theta)$, without the stars, implicitly understanding the role of θ^*.

☣ **4.3.13.** Make sure not to mix up $\mathcal{I}_{Y_1}(\theta)$, which is the Fisher information in Y_1, a single observation, and $\mathcal{I}_{\mathbf{Y}}(\theta)$, which is the Fisher information *for the entire data*. If the data are i.i.d., then $\mathcal{I}_{\mathbf{Y}}(\theta) = n\mathcal{I}_{Y_1}(\theta)$.

It is not obvious from looking at the definition why Fisher information is a measure of information. Some intuition for this can be gleaned from thinking about the *curvature* of $\text{E}\log L(\theta; \mathbf{Y})$. If the expected log-likelihood function has a sharp peak at θ^*, the data can be very informative about θ. If the $\text{E}\log L(\theta; \mathbf{Y})$ is quite flat at θ^*, the data do not seem to be giving us much information that we can use for pinpointing the true parameter value.

To measure this curvature in the log-likelihood function, the natural approach is to take the second derivative of the log-likelihood function. For example, the function $-10(\theta-3)^2$ has a much sharper peak at $\theta = 3$ than the function $-(\theta-3)^2$. The Fisher information is the average value of the curvature (averaged over all possible datasets), so in a sense is the average amount of information that the data have to offer about the parameter.

Example 4.3.14 (Information equality for Poissons)**.** For $Y_1, \ldots, Y_n$ i.i.d. Poisson with mean θ, $\mathrm{E}(Y_1) = \mathrm{Var}(Y_1) = \theta$ so

$$\begin{aligned} \mathcal{I}_{\mathbf{Y}}(\theta) &= \mathrm{Var}\left\{s(\theta;\mathbf{Y})\right\} = \mathrm{Var}\left(\frac{n\overline{Y}}{\theta}\right) = \frac{n\theta}{\theta^2} = \frac{n}{\theta}, \\ \mathrm{E}\left\{s'(\theta;\mathbf{Y})\right\} &= -\mathrm{E}\left(\frac{n\overline{Y}}{\theta^2}\right) = -\frac{n}{\theta}, \end{aligned}$$

which verifies the information equality for this example.

Example 4.3.15 (Information equality for Exponentials)**.** For $Y_1, \ldots, Y_n$ i.i.d. Exponential r.v.s with mean ψ,

$$\begin{aligned} \mathcal{I}_{\mathbf{Y}}(\psi) &= \mathrm{Var}\left\{s(\psi;\mathbf{Y})\right\} = \mathrm{Var}\left(\frac{n\overline{Y}}{\psi^2}\right) = \frac{n^2}{\psi^4}\mathrm{Var}(\overline{Y}) = \frac{n}{\psi^2}, \\ \mathrm{E}\left\{s'(\psi;\mathbf{Y})\right\} &= \frac{n}{\psi^2} - \mathrm{E}\left(\frac{2n\overline{Y}}{\psi^3}\right) = -\frac{n}{\psi^2}, \end{aligned}$$

so again the information equality holds. If instead we are interested in the Fisher information about the rate parameter $\lambda = 1/\psi$, we can either redo the calculation or use the result about that we introduce next.

Fisher information is *not* invariant under reparameterization.

Theorem 4.3.16 (Fisher information when transforming the parameter)**.** *Let $\tau = g(\theta)$, where g is a differentiable function with $g'(\theta) \neq 0$. Then*

$$\mathcal{I}_{\mathbf{Y}}(\tau) = \frac{\mathcal{I}_{\mathbf{Y}}(\theta)}{\{g'(\theta)\}^2}.$$

Proof. Converting between the score function for τ and the score function for θ,

$$s(\tau;\mathbf{Y}) = \frac{\partial}{\partial\tau}\log L(\theta;\mathbf{Y}) = \left(\frac{\partial}{\partial\theta}\log L(\theta;\mathbf{Y})\right)\frac{\partial\theta}{d\tau} = \frac{s(\theta;\mathbf{Y})}{g'(\theta)}.$$

Taking the variance of both sides gives the desired formula. ∎

Example 4.3.17 (Fisher information for Cauchy model)**.** For the Cauchy model, the score for a single observation is

$$s(\theta; Y_1) = \frac{2(Y_1 - \theta)}{1 + (Y_1 - \theta)^2}.$$

Thus,

$$\mathcal{I}_{Y_1}(\theta) = 4\mathrm{E}\left[\frac{(Y_1 - \theta)^2}{(1 + (Y_1 - \theta)^2)^2}\right] = \frac{4}{\pi}\int_{-\infty}^{\infty} \frac{x^2}{(1 + x^2)^3}\mathrm{d}x = \frac{1}{2},$$

where the integral is tricky but tractable.

Recall, switching back to using θ^* explicitly for clarity, $\mathcal{J}_{\mathbf{Y}}(\theta^*) = -\mathrm{E}[s'(\theta^*; \mathbf{Y})]$ equals $\mathcal{I}_{\mathbf{Y}}(\theta^*)$ by the information equality. The

$$\mathcal{J}_{\mathbf{y}}(\theta^*) = -s'(\theta^*; \mathbf{y}),$$

for the data we have in our hand, is called the *observed information* in the sample. The observed information varies over different datasets, and averages out to the Fisher information. The observed information is often estimated as

$$\hat{\mathcal{J}}_{\mathbf{y}} = -s'(\hat{\theta}; \mathbf{y}),$$

where $\hat{\theta}$ is the MLE.

4.3.3 Cramér-Rao lower bound

The Cramér-Rao lower bound (CRLB) provides a fundamental limit to how well we can estimate a parameter unbiasedly, based on the Fisher information. Intuitively, the less information we have about θ, the higher the variance will be for estimating θ unbiasedly. The Cramér-Rao lower bound make this intuition precise through a strikingly simple-looking inequality.

Theorem 4.3.18 (CRLB)**.** *Let $\hat{\theta}$ be an unbiased estimator of θ in a parametric statistical model $F_{\mathbf{Y};\theta}$. Under regularity conditions,*

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}_{\mathbf{Y}}(\theta)}.$$

Proof. The CRLB follows from the fact that correlation is between -1 and 1. Consider the covariance of the score function (which we will write as $S = S(\theta^*; Y)$ for short, where we again explicitly use θ^* for clarity) and the estimator (which we will write as $T(\mathbf{Y})$, or T for short). Since correlation is between -1 and 1 or by the Cauchy-Schwarz inequality, we have

$$\{\mathrm{Cov}(S, T)\}^2 \leq \mathrm{Var}(s)\mathrm{Var}(T).$$

This is already starting to look like the CRLB since $\mathrm{Var}(S) = \mathcal{I}(\theta^*)$, and $\mathrm{Var}(T)$ is what we are trying to find a lower bound on.

The only other thing we need to do is to compute the left-hand side. We can do this with DUThIS. To simplify notation, we will write the proof for the continuous case, but the same argument works regardless of whether the data are discrete or continuous.

$$\begin{aligned}
\mathrm{Cov}(S,T) &= \mathrm{E}[ST] - \mathrm{E}[S]\mathrm{E}[T] \\
&= \mathrm{E}[ST] \\
&= \int \frac{\partial f_{\mathbf{Y}}(\mathbf{y};\theta)/\partial\theta}{f_{\mathbf{Y}}(\mathbf{y};\theta)} T(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y};\theta) d\mathbf{y} \\
&= \frac{\partial}{\partial\theta} \int T(\mathbf{y}) f_{\mathbf{Y}}(\mathbf{y};\theta) d\mathbf{y} \\
&= \frac{\partial}{\partial\theta} \mathrm{E}_\theta[T(\mathbf{Y})] \\
&= \frac{\partial}{\partial\theta} \theta \\
&= 1,
\end{aligned}$$

where all the quantities get evaluated at θ^*. Thus,

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(T) \geq \frac{1}{\mathrm{Var}(S)} = \frac{1}{\mathcal{I}_{\mathbf{Y}}(\theta^*)}.$$

∎

Since for an unbiased estimator MSE equals variance, the CRLB tells us what the best MSE we can hope for if we are using an unbiased estimator. In some problems, no unbiased estimator achieves the CRLB (in fact, as we have seen, in some problems no unbiased estimator exists). In other problems, we *can* find an unbiased estimator that achieves the CRLB. And in many other problems, we can find an estimator that is *asymptotically* unbiased and which *asymptotically* achieves the CRLB. This will be a major theme later in this book.

Example 4.3.19 (CRLB in Normal model)**.** Consider a statistical model where $Y_1, \ldots, Y_n$ are i.i.d. from the $\mathcal{N}(\mu, \sigma^2)$ distribution with $\theta = \mu$ unknown but σ^2 known. Then

$$\mathrm{Var}(\bar{Y}) = \frac{\sigma^2}{n} = \frac{1}{\mathcal{I}_{\mathbf{Y}}(\mu)}.$$

So the sample mean, which is also the MLE, achieves the CRLB.

Example 4.3.20 (CRLB in Poisson model)**.** Consider n i.i.d. observations $Y_1, \ldots, Y_n$ from the Pois(λ) model, with λ unknown. Then

$$\mathrm{Var}(\bar{Y}) = \frac{\lambda}{n} = \frac{1}{\mathcal{I}_{\mathbf{Y}}(\lambda)}.$$

So again the sample mean, which is also the MLE of λ, achieves the CRLB. Now suppose that our estimand is

$$\theta = P(Y_1 = 0) = e^{-\lambda},$$

instead of λ. By invariance, the MLE of θ is

$$\hat{\theta} = e^{-\bar{Y}}.$$

Then $\hat{\theta}$ is biased, but we will show later that *asymptotically* it is unbiased and *asymptotically* it achieves the CRLB. Another natural estimator for θ is the proportion of 0's in the data:

$$\tilde{\theta} = \frac{1}{n}\sum_{j=1}^{n} I(Y_j = 0).$$

Note that $\tilde{\theta}$ is a method of moments estimator (without even needing the Poisson assumption). By linearity, $\tilde{\theta}$ is unbiased. Its variance is

$$\mathrm{Var}(\tilde{\theta}) = \frac{n}{n^2}\mathrm{Var}[I(Y_1 = 0)] = \frac{\theta(1-\theta)}{n}.$$

To see whether $\tilde{\theta}$ achieves the CRLB, let $g(\lambda) = e^{-\lambda}$, and note that

$$\mathcal{I}_{Y_1}(\theta) = \frac{\mathcal{I}_{Y_1}(\lambda)}{(g'(\lambda))^2} = \frac{1}{\lambda e^{-2\lambda}} = \frac{1}{(\theta)^2 \log(1/\theta)}.$$

Then

$$\mathrm{Var}(\tilde{\theta}) = \frac{\theta(1-\theta)}{n} > \frac{\theta^2 \log(1/\theta)}{n} = \frac{1}{\mathcal{I}_{\mathbf{Y}}(\theta)},$$

using the fact that $\log x \leq x - 1$ for all $x > 0$ (with strict inequality except at $x = 1$). So $\tilde{\theta}$ does not achieve the CRLB.

As we have discussed, in and of itself unbiasedness is not a good criterion to focus on. Happily, there is an extended version of the CRLB, one which also applies to biased estimators.

Theorem 4.3.21 (Extended CRLB)**.** *Let $\hat{\theta}$ be an estimator of θ in a parametric statistical model $F_{\mathbf{Y};\theta}$, such that $\mathrm{E}[\hat{\theta}] = g(\theta)$. Under regularity conditions,*

$$\mathrm{Var}(\hat{\theta}) \geq \frac{\{g'(\theta)\}^2}{\mathcal{I}_{\mathbf{Y}}(\theta)}.$$

The proof is the same as the proof for the unbiased case, except that when we evaluate $\mathrm{E}[T(\mathbf{Y})]$ we now have $g(\theta)$ instead of θ, so $\mathrm{Cov}(S, T)$ is $\partial g(\theta)/\partial\theta$ rather than $\partial\theta/\partial\theta = 1$.

4.3.4 Asymptotic distribution of the MLE

In addition to the good properties we have already seen, such as invariance and consistency, the MLE has excellent asymptotic properties: for large sample size, it is *approximately* the case that the MLE is Normal, unbiased, and achieves the CRLB. Again we use the θ^* notation at the start, to make the exposition clear.

Theorem 4.3.22 (Asymptotic distribution of the MLE)**.** *Let $\hat{\theta}$ be the MLE of a scalar parameter θ, based on i.i.d. observations $Y_1, \ldots, Y_n$ from $F_{\mathbf{Y};\theta^*}$. Under regularity conditions, the asymptotic distribution of $\hat{\theta}$ is given by the following:*

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}\left(0, \mathcal{I}_{Y_1}^{-1}(\theta^*)\right), \tag{4.1}$$

(that is converges in distribution) as the sample size $n \to \infty$.

As an *approximation*, this result says that for large n,

$$\hat{\theta} \dot{\sim} \mathcal{N}\left(\theta^*, \frac{1}{n\mathcal{I}_{Y_1}(\theta^*)}\right). \tag{4.2}$$

This theorem is fundamental and striking. The big picture is:

1. The bias of the MLE is small when n is large: even when we scale up $(\hat{\theta} - \theta^*)$ by $\sqrt{n}$ the result is centered at 0. Since $\text{bias}(\hat{\theta}) \to 0$, we say that the MLE is *asymptotically unbiased*.

2. The variance of the MLE is roughly $1/\{n\mathcal{I}_{Y_1}(\theta^*)\}$ which, by the CRLB, means that the MLE is asymptotically efficient.

3. The random interval
$$\left[\hat{\theta} - \frac{1.96}{\sqrt{n\mathcal{I}_{Y_1}(\hat{\theta})}}, \hat{\theta} + \frac{1.96}{\sqrt{n\mathcal{I}_{Y_1}(\hat{\theta})}}\right],$$
covers the estimand roughly 95% of the time across many replications of the experiment. The number 1.96, which is approximately the 0.975 quantile of the standard Normal distribution, is ubiquitous in statistics. We will discuss interval estimation in much more detail in Chapter 5.

☣ **4.3.23.** Note that in the *limit* statement in equation (4.1), n does not appear on the right-hand side. (How could it? It was sent to ∞.) In the *approximation* statement in equation (4.2), n does appear in the variance, which makes sense since the variance of $\hat{\theta}$ should be small when the sample size n is large. We could also have written $n\mathcal{I}_{Y_1}(\theta^*)$ as $\mathcal{I}_{\mathbf{Y}}(\theta^*)$; again, make sure not to mix up $\mathcal{I}_{Y_1}(\theta^*)$ and $\mathcal{I}_{\mathbf{Y}}(\theta^*)$.

Proof. The key idea of the proof is to Taylor-expand the score function about θ^*. We will write the Taylor expansion as an *approximation*; for a rigorous proof, we would need Taylor's theorem with remainder (i.e., include an error term), and then we could use some analysis and the regularity conditions to keep the remainder term under control. By Taylor approximation about θ^*,

$$0 = s(\hat{\theta}; \mathbf{Y}) \approx s(\theta^*; \mathbf{Y}) + (\hat{\theta} - \theta^*) s'(\theta^*; \mathbf{Y}).$$

This is a linear approximation of $s(\hat{\theta}; \mathbf{Y})$, and will be good if $\hat{\theta}$ is close to θ^*. And since the MLE is consistent, $\hat{\theta}$ *will* be close to θ^* (with high probability) for n large. But we also know that $s(\hat{\theta}; \mathbf{Y}) = 0$; that is exactly the equation we typically use to find the MLE! Rearranging the approximation and multiplying both sides by $\sqrt{n}$,

$$\sqrt{n}(\hat{\theta} - \theta^*) \approx \frac{\sqrt{n}\, s(\theta^*; \mathbf{Y})}{-s'(\theta^*; \mathbf{Y})} = \frac{\sqrt{n} \frac{1}{n} s(\theta^*; \mathbf{Y})}{-\frac{1}{n} s'(\theta^*; \mathbf{Y})}.$$

Next, apply the CLT to the numerator and the LLN to the denominator (we multiplied by $1/n$ in the numerator and denominator to set up for this strategy!). The numerator

$$\sqrt{n} \frac{1}{n} s(\theta^*; \mathbf{Y}) = \sqrt{n} \left[\frac{1}{n} \sum_{j=1}^{n} s(\theta^*; Y_j) \right]$$

is a scaled sample average of the zero mean scores, so converges in distribution to $\mathcal{N}(0, \mathcal{I}_{Y_1}(\theta^*))$ by the CLT. By the LLN, the denominator converges with probability 1 to the constant

$$-\mathrm{E}[s'(\theta^*; Y_1)] = \mathcal{I}_{Y_1}(\theta^*)$$

by the information equality. Letting $Z \sim \mathcal{N}(0, 1)$, the limiting distribution is

$$\frac{\sqrt{\mathcal{I}_{Y_1}(\theta^*)}\, Z}{\mathcal{I}_{Y_1}(\theta^*)} = \frac{Z}{\sqrt{\mathcal{I}_{Y_1}(\theta^*)}} \sim \mathcal{N}\left(0, \frac{1}{\mathcal{I}_{Y_1}(\theta^*)}\right).$$

∎

Having used notation to distinguish between θ and θ^* in developing the theory, in examples we will sometimes drop explicit use of the stars, to remove clutter.

Example 4.3.24 (Asymptotic distribution of MLE in Cauchy model)**.** In the Cauchy model,

$$\mathcal{I}_{Y_1}(\theta^*) = 1/2,$$

by Example 4.3.17. So the MLE has the asymptotic distribution

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, 2).$$

Compare this to the sample median, where

$$\sqrt{n}\{Y_{(\lceil 0.5n \rceil)} - \theta\} \xrightarrow{d} \mathcal{N}(0, 0.25\pi^2).$$

The ratio of the asymptotic variance of the MLE to the corresponding result for the sample median is $8/\pi^2 \approx 0.81$. Intuitively, this can be interpreted as saying that using the sample median rather than the MLE is like throwing away around 20% of the data before computing the estimate. On the other hand, the median is very easy to compute and would still be a sensible estimator for the median even if the assumption that the distribution is Cauchy were wrong, whereas the MLE of the Cauchy is not available in closed form and is strongly tied to the Cauchy assumption.

In the $\mathcal{N}(\mu, \sigma^2)$ case where $\theta = \mu$, then the MLE is $\bar{Y}$ and $\sqrt{n}(\bar{Y} - \theta) \sim \mathcal{N}(0, \sigma^2)$, whereas

$$\sqrt{n}\{Y_{(\lceil 0.5n \rceil)} - \theta\} \xrightarrow{d} \mathcal{N}(0, 0.5\pi\sigma^2).$$

So in this case the ratio of asymptotic variances is $2/\pi \approx 0.64$. Intuitively, this says that using the sample median is like throwing out 36% of the data, in return for gaining some robustness against deviations from Normality.

There are different morals you can draw from this example. If you build good statistical models and use a MLE, you typically get pretty good results. Some intuitive non-MLE estimators are likely to have poorer properties if the models are roughly right. However, the sample median provides non-stupid answers when the data can contain very wild observations, which is attractive if we want to automatically apply a statistical procedure and do not care a great deal about losing some statistical efficiency, or when we are especially concerned about model misspecification. We could carefully, laboriously derive the MLE for a bad model when it would have been simpler and more robust to have used the sample median.

4.4 Likelihoods based on conditional distributions

The core to building a statistical model is writing down the joint density (or PMF) of the data, typically written $f_{\mathbf{Y}}(\mathbf{y})$. The likelihood is then the joint density, varying the parameters.

Here we extend this much further, using a single new idea whose scope in statistical inference is vast. No fresh mathematical ideas are introduced in this section. But from a statistical perspective, the idea developed here allows us to build a much wider set of useful models, while still allowing us to continue to use the appealing likelihood framework. The idea is subtle and very important for applications. It yet again reinforces one of the big ideas from Stat 110:

Conditioning is the soul of statistics.

Suppose the data segments into pieces called $\mathbf{x}$ and $\mathbf{y}$ (e.g., $\mathbf{y}$ are outcomes and $\mathbf{x}$ are predictors), then the statistical model is the joint distribution $f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})$. Of course, the joint density equals the marginal density times the conditional density:

$$f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) f_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y}).$$

Statisticians often decide to study solely the conditional distribution, either out of convenience (we will see an example of this at the end of this section) or because their scientific focus is on the conditional distribution (our next example). In parametric models the statistical model becomes:

$$f_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y};\theta).$$

Then we can define the likelihood for this conditional density

$$L(\theta;\mathbf{y}|\mathbf{x}) = f_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y};\theta),$$

so the MLE takes the form

$$\hat{\theta} = \underset{\theta\in\Theta}{\arg\max}\ \log L(\theta;\mathbf{y}|\mathbf{x}).$$

Example 4.4.1 (Predictive regression)**.** Suppose that the data arrives in pairs, $(x_1,y_1),\ldots,(x_n,y_n)$. Then an important statistical model is when the $(X_1,Y_1),\ldots,(X_n,Y_n)$ are i.i.d. pairs from $F_{X,Y}$. To boil the problem down to its essence, in this example suppose $\mathrm{E}[X_1] = \mathrm{E}[Y_1] = 0$ throughout. For statistical prediction problems, the linear regression (when both means are zero) where

$$\mathrm{E}[Y|X=x,\theta] = \theta x$$

is popular (if you're already familiar with regression: the intercept is missing as we have set all the means to zero). This says if I have seen $X \approx x$ then on average the $Y \approx \theta x$. A Gaussian version of this is the *Gaussian regression model*

$$Y|(X=x,\theta) \sim \mathcal{N}(\theta x, \sigma^2).$$

Again we simplify by assuming σ^2 is known. The resulting density

$$f_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\mathbf{y};\theta) = \prod_{j=1}^{n} f_{Y_j|X_j=x_j}(y_j;\theta) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\theta x_j)^2\right],$$

delivers the following log-likelihood, score, and MLE:

$$l(\theta) = -\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\theta x_j)^2, \quad s(\theta;\mathbf{Y}|\mathbf{X}=\mathbf{x}) = \frac{1}{\sigma^2}\sum_{j=1}^{n} x_j(Y_j-\theta x_j), \quad \hat{\theta} = \frac{\sum_{j=1}^{n} x_j Y_j}{\sum_{j=1}^{n} x_j^2}.$$

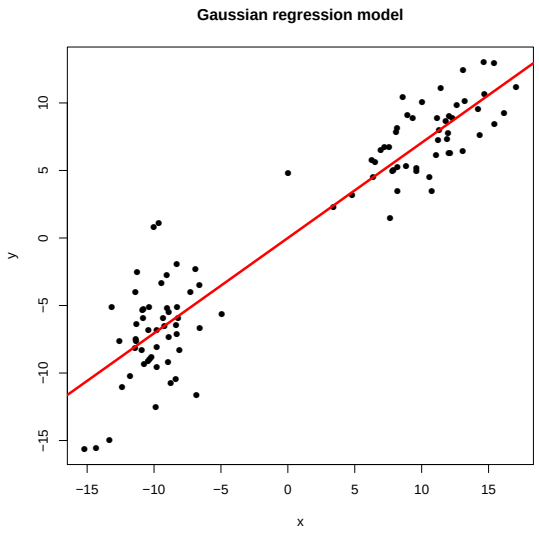


Figure 4.7: Gaussian regression simulated data, with highly bimodal, zero mean X. The red line is the line $y = \hat{\theta}x$, where $\hat{\theta} \approx 0.705$ is the MLE. Here $\text{SE}(\hat{\theta}) \approx 0.029$.

For scientists focusing on prediction, the attractive aspect of this approach is that this likelihood was established without any modeling of $f_{\mathbf{X}}(\mathbf{x})$, the marginal distribution of $\mathbf{X}$, which they might not care about. Figure 4.7 illustrates this using simulation with $n = 100$, the pairs being i.i.d., $\sigma = 3$, $\theta = 0.7$ and a highly bimodal density for

$$f_{X_1}(x) = 0.5\{\varphi(x; 10, 3^2) + \varphi(x; -10, 2^2)\},$$

where $\varphi(x; \mu, \omega^2)$ is the Normal density with mean μ and variance ω^2, evaluated at x. So here we have a relatively simple analysis of prediction which sidesteps entirely the fact that the distribution of $\mathbf{X}$ is pretty tricky.

Building likelihoods from conditional distributions is enticing, but do the attractive likelihood properties carry over? It turns out the answer is yes, but all the results condition on $\mathbf{X} = \mathbf{x}$!

- We have

  $$D_{KL}(F_{Y|X=x;\theta^*} || F_{Y|X=x;\theta}) \geq 0,$$

  for every value of x. That is, the conditional expectation of the log-likelihood is maximized at the true value θ^*, whatever predictors we see.

- The expected score is still 0:

  $$\text{E}[s(\theta^*; \mathbf{Y}, \mathbf{x})|\mathbf{X} = \mathbf{x}] = 0.$$

- The information equality still holds:

$$\mathcal{I}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\theta^*) = \mathrm{Var}[s(\theta^*; \mathbf{Y}, \mathbf{x})|\mathbf{X} = \mathbf{x}] = -\mathrm{E}[s'(\theta^*; \mathbf{Y}, \mathbf{x})|\mathbf{X} = \mathbf{x}].$$

- The Cramér-Rao lower bound still holds. For any conditionally unbiased $\hat{\theta}$, i.e., $\hat{\theta}$ such that $\mathrm{E}[\hat{\theta}|\mathbf{X} = \mathbf{x}] = \theta$, we have

$$\mathrm{Var}(\hat{\theta}|\mathbf{X} = \mathbf{x}) \geq \frac{1}{\mathcal{I}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\theta)}.$$

- For a wide variety of problems, the MLE is still approximately Normal for n large:

$$\hat{\theta}|(\mathbf{X} = \mathbf{x}) \dot{\sim} \mathcal{N}\left(\theta, \frac{1}{\mathcal{I}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\theta)}\right).$$

Why? It looks like we need to repeat all of our previous work, investigating the properties of scores and Fisher information under conditional densities. However, just recall from Chapter 2 of the Stat 110 book that conditional probabilities *are* probabilities! Conditional densities for $\mathbf{Y}|(\mathbf{X} = \mathbf{x})$ are just density functions but with a different probability function than for $\mathbf{Y}$ alone. Hence likelihoods based on conditional distributions, are just likelihoods and all of the main properties we expect from likelihood based quantities carry over here, but where every property is *conditional* on $\mathbf{X} = \mathbf{x}$.

Example 4.4.2 (Continuing Example 4.4.1)**.** The Fisher information in the sample is

$$\mathcal{I}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\theta^*) = \mathrm{Var}\left(\frac{1}{\sigma^2}\sum_{j=1}^{n} x_j(Y_j - \theta x_j)\right) = \frac{1}{\sigma^4}\sum_{j=1}^{n} x_j^2 \mathrm{Var}(Y_j|X = x_j) = \frac{1}{\sigma^2}\sum_{j=1}^{n} x_j^2.$$

The conditional mean and conditional variance of the MLE are

$$\begin{aligned}
\mathrm{E}[\hat{\theta}|X = x] &= \left(\sum_{j=1}^{n} x_j^2\right)^{-1} \sum_{j=1}^{n} x_j \mathrm{E}[Y_j|X_j = x_j] = \theta, \\
\mathrm{Var}[\hat{\theta}|X = x] &= \left(\sum_{j=1}^{n} x_j^2\right)^{-2} \sum_{j=1}^{n} x_j^2 \mathrm{Var}[Y_j|X_j = x_j] = \sigma^2\left(\sum_{j=1}^{n} x_j^2\right)^{-1} = \mathcal{I}^{-1}_{\mathbf{Y}|\mathbf{X}=\mathbf{x}}(\theta^*).
\end{aligned}$$

Hence the MLE is conditionally unbiased and achieves the conditional CRLB, so is CRLB efficient.

Importantly, the Fisher information depends upon $\sum_{j=1}^{n} x_j^2$, so if the $x_1, ..., x_n$ are wildly variable the Fisher information will be very high and very precise estimation of θ is possible. If the $x_1, ..., x_n$ are dull and largely close to zero, the Fisher information will be tiny and we can learn little about θ however clever our estimation strategy is.

Example 4.4.3 (Markov chain)**.** Recall the two-state Markov chain example from Chapter 2. Let $\theta = (p_{01}, p_{11})$, where $P(Y_j = k|Y_{j-1} = l, \theta) = p_{lk}$. The log-likelihood is

$$\begin{aligned} l(\theta; \mathbf{y}) &= \log P(\mathbf{Y} = \mathbf{y}; \theta) \\ &= y_1 \log \pi + (1 - y_1) \log(1 - \pi) \\ &+ n_{11} \log p_{11} + n_{10} \log(1 - p_{11}) + n_{01} \log p_{01} + n_{00} \log(1 - p_{01}), \end{aligned}$$

where $\pi = \mathrm{E}(Y_j; \theta) = p_{01}/(1 - p_{11} + p_{01})$ and $n_{lk} = \sum_{j=2}^{n} I(y_{j-1} = l)I(y_j = k)$. In the statistical analysis of time series y_1 is called the *initial value* of the process.

This log-likelihood has its own beauty, but the presence of the initial value in the likelihood means that the MLE has to be found by numerical optimization (which is fine, but a small pain). This can be cleaned up by deciding to use a conditional likelihood, conditioning on the initial value y_1

$$\begin{aligned} l(\theta; \mathbf{y}) &= \log P(Y_2 = y_2, ..., Y_n = y_n|Y_1 = y_1, \theta) \\ &= c + n_{11} \log p_{11} + n_{10} \log(1 - p_{11}) + n_{01} \log p_{01} + n_{00} \log(1 - p_{01}), \end{aligned}$$

The resulting MLE can be found analytically as

$$\frac{\partial l(\theta; \mathbf{y})}{\partial p_{01}} = \frac{n_{01}}{p_{01}} - \frac{n_{00}}{1 - p_{01}}, \quad \frac{\partial l(\theta; \mathbf{y})}{\partial p_{11}} = \frac{n_{11}}{p_{11}} - \frac{n_{10}}{1 - p_{11}}$$

so

$$\hat{p}_{01} = \frac{n_{01}}{n_{01} + n_{00}} \quad \text{and} \quad \hat{p}_{11} = \frac{n_{11}}{n_{11} + n_{10}}.$$

This is an example where we choose to use the conditional likelihood not because of the scientific question, but simply out of convenience.

4.5 Numerical optimization of the likelihood*

The MLE can sometimes be found analytically, but more often it is found by numerical optimization of the log-likelihood function. How is this carried out reliably and rapidly?

For many statistical models the log-likelihood function can be shown to be concave in θ, so statisticians often use the theory of concave optimization to numerically determine the MLE. Figure 4.8 provides two examples of concave log-likelihood functions based on 4 i.i.d. data points: one from a Gaussian model where

$$f_{Y_1}(y; \theta) = \exp(-(y - \theta)^2/8)/\sqrt{4 \times 2\pi},$$

the other from a Laplace model where

$$f_{Y_1}(y; \theta) = \exp(-|y - \theta|/2)/4.$$

For these two simple models it is possible to see where the maximum is (in the Laplace case there is not a unique global maximizer). But for more sophisticated problems we need numerical methods to find the MLE.

At its core, like asymptotic approximations from probability theory, optimization theory introduces few new statistical ideas — rather it vastly broadens the areas where statistical methods can be applied with confidence.

Definition 4.5.1 (Concave function)**.** A function g whose domain is I is *concave* if

$$g(px_1 + (1-p)x_0) \geq pg(x_1) + (1-p)g(x_0), \tag{4.3}$$

for all $x_0, x_1 \in I$ and $p \in [0,1]$. Strict concavity is where the $\geq$ is replaced by a strict inequality.

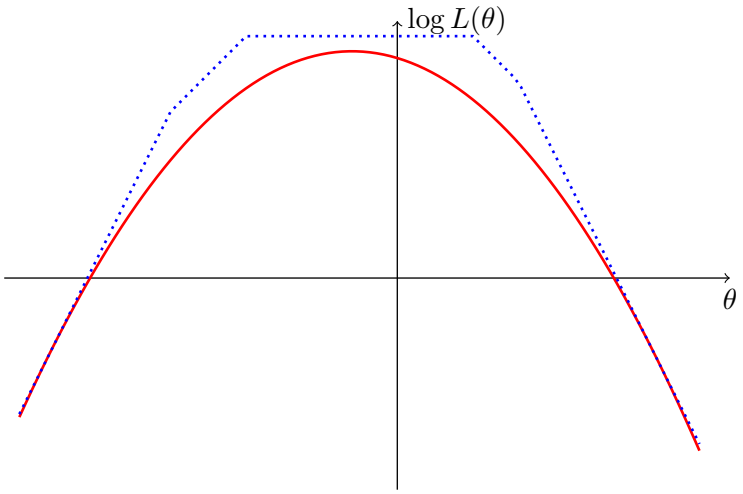


Figure 4.8: Two concave log-likelihood functions for data $\mathbf{y} = (0.5, 0.8, -1, -1.5)$: Gaussian model (red line) $\log L(\theta) = -\sum_{j=1}^{4}(y_j - \theta)^2/8$ and Laplace model (dotted blue line) $\log L(\theta) = -\sum_{j=1}^{4}|y_j - \theta|/2$. The Gaussian model has a strictly concave log-likelihood. The Laplace model has a concave log-likelihood.

Log-likelihoods for the sample are the sum of log-likelihoods for individual data points if the data are i.i.d. or the sum of the log-likelihoods for sequences of predictions for time series. How do we tell if the log-likelihood for the sample is concave? The following theorem can be used, which allows us just to check concavity for one data point or prediction at a time.

Theorem 4.5.2. *Suppose $g(x) = \sum_{j=1}^{n} w_j g_j(x)$, where the weights $w_1, \ldots, w_n$ are nonnegative and $x \in I$. If $g_1, \ldots, g_n$ are each concave, then g is concave. Strict concavity follows if one of the functions is strictly concave and the weight is not zero.*

Proof. The general case follows by induction if it true for $n = 2$, and $x_0, x_1 \in I$ and $p \in [0, 1]$. Then

$$\begin{aligned} g(px_1 + (1-p)x_0) &= w_1 g_1(px_1 + (1-p)x_0) + w_2 g_2(px_1 + (1-p)x_0) \\ &\geq w_1\{pg_1(x_1) + (1-p)g_1(x_0)\} + w_2\{pg_2(x_1) + (1-p)g_2(x_0)\} \\ &= p\{w_1 g_1(x_1) + w_2 g_2(x_1)\} + (1-p)\{w_1 g_1(x_0) + w_2 g_2(x_0)\} \\ &= pg(x_1) + (1-p)g(x_0). \end{aligned}$$

The strict concavity result follows immediately replacing the inequality sign by a strict inequality sign. ∎

Example 4.5.3 (Laplace model)**.** For a Laplace model the log-likelihood for the first data point is $\log L(\theta; y_1) = -|y_1 - \theta|$. Is this concave? Focusing on the core issue, this is the same as asking if $g(x) = -|a + x|$, is concave. Then

$$\begin{aligned} g(px_1 + (1-p)x_0) &= -|p(a + x_1) + (1-p)(a + x_0)| \\ &\geq -|p(a + x_1)| - |(1-p)(a + x_0)| \\ &= -p|a + x_1| - (1-p)|a + x_0| = pg(x_1) + (1-p)g(x_0), \end{aligned}$$

as required. So the Laplace log-likelihood for one data point is concave, and then Theorem 4.5.2 implies that $\log L(\theta; \mathbf{y}) = c - \sum_{j=1}^{n} |y_j - \theta|$ is concave.

If a numerical maximization routine finds a maximum, it is helpful to know if it will be a global maximum (rather than just a local maximum) and if that global maximum is unique? If the function being maximized is concave then the next theorem guarantees that any local maximum is a global maximum. If it is strictly concave then the global maximum is unique (without strictness, the function can be flat at the top).

Theorem 4.5.4. *Any local maximum of a concave function g on domain I is also a global maximum. If the function is strictly concave, then the function will have at most one global maximum on I.*

Proof. We prove the first result. Assume $x_0, x_1 \in I$, while p is tiny and x_1 is a local maximum. The local maximum property implies

$$g(x_1) \geq g(x_1 + p(x_0 - x_1)).$$

Since g is concave,

$$g(x_1 + p(x_0 - x_1)) = g(px_0 + (1-p)x_1) \geq pg(x_0) + (1-p)g(x_1). \tag{4.4}$$

Combining the two inequalities,

$$g(x_1) \geq pg(x_0) + (1-p)g(x_1).$$

Canceling terms yields the desired result whatever the choice of x_0:

$$g(x_1) \geq g(x_0).$$

The second result follows as equation (4.4) becomes a strict inequality under strict concavity, which the last 2 inequalities then inherit. ∎

The second derivative of the log-likelihood function controls Fisher information and the large sample behavior of the MLE. It also relates to concavity through the following theorem.

Theorem 4.5.5. *If g is twice continuously differentiable on I then g is concave if and only if for all $x \in I$,*

$$g''(x) = \frac{\partial^2 g(x)}{\partial x^2} \leq 0.$$

Proof. We will prove that if $x_0, x_1 \in I$ then

$$(x_1 - x_0)g'(x_0) \geq g(x_1) - g(x_0) \geq (x_1 - x_0)g'(x_1), \tag{4.5}$$

then the stated result follows by the differentiability assumption. Start with equation (4.3). Then, rearranging terms yields

$$(x_1 - x_0)\frac{g(x_0 + p(x_1 - x_0)) - g(x_0)}{p(x_1 - x_0)} \geq g(x_1) - g(x_0).$$

Taking $p \to 0$, then we produce the left hand side of (4.5). We produce the right hand side using the same technique, but writing $q = 1 - p$, so equation (4.3) becomes

$$g((1-q)x_1 + qx_0) \geq (1-q)g(x_1) + qg(x_0).$$

Then proceed as before, yielding

$$(x_1 - x_0)\frac{g(x_1 + q(x_0 - x_1)) - g(x_1)}{q(x_0 - x_1)} \leq g(x_1) - g(x_0).$$

Taking $q \to 0$ yields the left hand side of (4.5). ∎

Many models used in applied statistics have concave log-likelihoods.

Example 4.5.6 (Count regression model)**.** Follow Example 4.4.1, but switch from the Gaussian regression model to the count regression model which has

$$Y|X = x \sim \text{Pois}(\mu(x;\theta)),$$

where $\mu(x;\theta) = \exp(\theta x)$. Thus

$$P(Y = y|X = x;\theta) = \frac{\exp\{-\mu(x;\theta)\}\mu(x;\theta)^y}{y!},$$

which implies that, for conditionally independent pairs of data,

$$\log L(\theta) = c + \sum_{j=1}^{n} -\exp(\theta x_j) + y_j \theta x_j.$$

Thus

$$s(\theta;\mathbf{Y}|\mathbf{X} = \mathbf{x}) = \sum_{j=1}^{n} x_j\{y_j - \exp(\theta x_j)\}, \quad s'(\theta;\mathbf{Y}|\mathbf{X} = \mathbf{x}) = -\sum_{j=1}^{n} x_j^2 \exp(\theta x_j).$$

Hence the log-likelihood is concave. It is strictly concave under the conditions that θ is bounded and at least one of the $x_1, \ldots, x_n$ is not zero. So in practice the MLE will be unique, even though we cannot find the MLE analytically — it has to be determined using an algorithm.

Assume $g(x)$ is strictly concave in the univariate x and twice continuously differentiable on I. Now find

$$\hat{x} = \arg\max_{x \in I} g(x),$$

using numerical methods. A fundamental method for numerically approximating $\hat{x}$ is the Newton-Raphson method.

Definition 4.5.7 (Newton-Raphson)**.** The *Newton-Raphson method* or *Newton's method*, has the form

$$x_i = x_{i-1} + \left\{-\frac{\partial^2 g(x_{i-1})}{\partial x^2}\right\}^{-1} \frac{\partial g(x_{i-1})}{\partial x},$$

for $i = 1, 2, \ldots$, which is iterated until numerical convergence where $|x_i - x_{i-1}| < \epsilon$ where ϵ is some tolerance level. It is known to converge quickly once x_i is in the neighborhood of $\hat{x}$. The term $x_0 \in I$ is called the "initial value" of the optimizer.

Example 4.5.8 (Continued from Example 4.5.6)**.** To find the MLE for the count regression we iterate

$$\hat{\theta}_i = \hat{\theta}_{i-1} + \left(\sum_{j=1}^{n} x_j^2 \exp(\hat{\theta}_{i-1} x_j)\right)^{-1} \sum_{j=1}^{n} x_j(y_j - \exp(\hat{\theta}_{i-1} x_j)),$$

until convergence. In practice this algorithm usually converges in a handful of iterations.

Why does Newton-Raphson work?

Assume that $\hat{x}$ is an interior point, then it is given by the solution to

$$\frac{\partial g(x)}{\partial x}\Big|_{\hat{x}} = \frac{\partial g(\hat{x})}{\partial x} = 0.$$

Start with an "initial value" $x_0 \in I$ and carry out a Taylor expansion

$$\frac{\partial g(x)}{\partial x} \approx \frac{\partial g(x_0)}{\partial x} + \frac{\partial^2 g(x_0)}{\partial x^2}(x - x_0).$$

Now define x_1 which is the value of x which makes the right-hand side 0. This is given by

$$x_1 = x_0 + \left\{ -\frac{\partial^2 g(x_0)}{\partial x^2} \right\}^{-1} \frac{\partial g(x_0)}{\partial x}.$$

The term in the curly brackets must be positive, so the sign of $x_1 - x_0$ is the same as the sign of $\partial g(x_0)/\partial x$. Iterating this algorithm is the *Newton-Raphson method*.

In the last couple of decades a great deal of statistical research has focused on problems where concave optimization is needed for functions which are continuous but not everywhere differentiable, e.g., the Laplace log-likelihood in Figure 4.8. Newton-Raphson type methods cannot be directly applied to those problems. Statistical examples of this include quantile regression and LASSO. See Boyd and Vandenberghe (2004) for further reading about fast optimization methods for those and other important problems.

4.6 Multiple parameter version*

This book is about the main ideas in statistical inference. In most applications statistical models have multiple parameters, so the application of the above results needs some extension. These extensions do not really introduce any fundamentally new statistical ideas unless the dimensional of the parameters is quite large — which is a topic beyond the scope of this book. Hence our focus has been on single parameter cases.

For a moderate number of parameters, the extension of likelihood methods really only involves the use of some results on vectors and matrices from linear algebra. Here we outline the results without proofs.

Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^{\mathsf{T}}$ be a $K \times 1$ vector. Calculating the MLE now involves maximizing the likelihood

$$L(\boldsymbol{\theta}; \mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}; \boldsymbol{\theta})$$

over K variables. The arguments about Kullback-Leibler divergence do not change with K, which means

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\max} \log L(\boldsymbol{\theta}; \mathbf{y})$$

is still consistent for the estimand or true value $\boldsymbol{\theta}^*$.

Again a key quantity for likelihood analysis is the score, but this is now a $K \times 1$ vector

$$s(\boldsymbol{\theta}; \mathbf{y}) = \begin{pmatrix} \frac{\partial \log L(\boldsymbol{\theta}; \mathbf{y})}{\partial \theta_1} \\ \vdots \\ \frac{\partial \log L(\boldsymbol{\theta}; \mathbf{y})}{\partial \theta_K} \end{pmatrix} = \frac{\partial \log L(\boldsymbol{\theta}; \mathbf{y})}{\partial \boldsymbol{\theta}}.$$

Under the usual conditions, the expected score vector

$$\mathrm{E}[s(\boldsymbol{\theta}^*; \mathbf{Y})] = 0,$$

a $K \times 1$ vector of zeros.

We can define the *Fisher information matrix* for the sample as

$$\mathcal{I}_{\mathbf{Y}}(\boldsymbol{\theta}^*) = \mathrm{Var}(s(\boldsymbol{\theta}^*; \mathbf{Y})),$$

the $K \times K$-dimensional variance-covariance matrix of the score. The information equality for vector parameters is that

$$\mathcal{I}_{\mathbf{Y}}(\boldsymbol{\theta}^*) = \mathcal{J}_{\mathbf{Y}}(\boldsymbol{\theta}^*)$$

where

$$\mathcal{J}_{\mathbf{Y}}(\boldsymbol{\theta}^*) = -\mathrm{E}\left[\frac{\partial^2 \log L(\boldsymbol{\theta}^*; \mathbf{Y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}}\right],$$

the expectation of the Hessian matrix.

Here we will assume that $\mathcal{I}_{\mathbf{Y}}(\boldsymbol{\theta}^*)$ is an invertible matrix, that is the matrix inverse $\mathcal{I}_{\mathbf{Y}}(\boldsymbol{\theta}^*)^{-1}$ exists. Then the vector version of the Cramér-Rao lower bound, is that if $\hat{\boldsymbol{\theta}}$ is a vector of unbiased estimators then for any $K \times 1$ vector of constants $\mathbf{a}$, then

$$\mathrm{Var}(\mathbf{a}^{\mathsf{T}}\hat{\boldsymbol{\theta}}) = \mathbf{a}^{\mathsf{T}}\mathrm{Var}(\hat{\boldsymbol{\theta}})\mathbf{a} \geq \mathbf{a}^{\mathsf{T}}\mathcal{I}_{\mathbf{Y}}(\boldsymbol{\theta}^*)^{-1}\mathbf{a}.$$

In the language of matrix algebra, we say that $\mathrm{Var}(\hat{\boldsymbol{\theta}}) - \mathcal{I}_{\mathbf{Y}}(\boldsymbol{\theta}^*)^{-1}$, is *nonnegative definite* or *positive semidefinite*.

In the i.i.d. case multivariate case, we again have $\mathcal{I}_{\mathbf{Y}}(\boldsymbol{\theta}^*) = n\mathcal{I}_{Y_1}(\boldsymbol{\theta}^*)$, and asymptotic distribution of the MLE of $\boldsymbol{\theta}$ will then be Multivariate Normal

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{Y_1}(\boldsymbol{\theta}^*)^{-1}).$$

Taking these results together, the vector MLE is asymptotically unbiased, asymptotically efficient in the Cramér-Rao lower bound sense, and asymptotically Gaussian.

Example 4.6.1 (Mean and variance in Normal example)**.** Return to the statistical model where $Y_1, \ldots, Y_n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, with $\boldsymbol{\theta} = (\mu, \sigma^2)^{\mathtt{T}}$, so here $K = 2$. Then

$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}; \mathbf{y}) = -\frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{j=1}^{n} (y_j - \mu)^2.$$

The score is a 2×1 vector $s(\boldsymbol{\theta}) = (s(\mu; \mathbf{y}), s(\sigma^2; \mathbf{y}))^{\mathtt{T}}$ where

$$s(\mu; \mathbf{y}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \mu} = \frac{1}{\sigma^2} \sum_{j=1}^{n} (y_j - \mu), \quad s(\sigma^2; \mathbf{y}) = \frac{\partial l(\boldsymbol{\theta})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{j=1}^{n} (y_j - \mu)^2.$$

Setting $s(\hat{\boldsymbol{\theta}}; \mathbf{y}) = (0, 0)^{\mathtt{T}}$ and we have the equations

$$\frac{1}{\hat{\sigma}^2} \sum_{j=1}^{n} (y_j - \hat{\mu}) = 0, \quad -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4} \sum_{j=1}^{n} (y_j - \hat{\mu})^2 = 0.$$

Solving out

$$\hat{\mu} = \frac{1}{n} \sum_{j=1}^{n} y_j = \bar{y}, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{j=1}^{n} (y_j - \bar{y})^2.$$

We check this is a maximum in the likelihood by calculating the Hessian

$$\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathtt{T}}} = \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4} \sum_{j=1}^{n} (y_j - \mu) \\ -\frac{1}{\sigma^4} \sum_{j=1}^{n} (y_j - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{j=1}^{n} (y_j - \mu)^2 \end{pmatrix}.$$

Thus

$$\frac{\partial^2 l(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathtt{T}}} = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2\hat{\sigma}^4} \end{pmatrix},$$

which is negative definite so long as $\hat{\sigma} > 0$, completing successfully the check.

The Fisher information in a single observation $\mathcal{I}_{Y_1}(\boldsymbol{\theta})$ is

$$\begin{aligned} \mathrm{Var}(s(\boldsymbol{\theta}; Y_1)) &= \mathrm{Var} \begin{pmatrix} \frac{1}{\sigma^2}(Y_1 - \mu) \\ \frac{1}{2\sigma^4}(Y_1 - \mu)^2 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\sigma^4} \mathrm{Var}(Y_1) & \frac{1}{2\sigma^6} \mathrm{E}[(Y_1 - \mu)^3] \\ \frac{1}{2\sigma^6} \mathrm{E}[(Y_1 - \mu)^3] & \frac{1}{4\sigma^8} \mathrm{Var}\{(Y_1 - \mu)^2\} \end{pmatrix} \end{aligned}$$

As $Y_1 \sim \mathcal{N}(\mu, \sigma^2)$ then $\mathrm{E}[(Y_1 - \mu)^3] = 0$ by symmetry of the normal and $\mathrm{Var}\{(Y_1 - \mu)^2\} = 2\sigma^4$. Simplifying, delivers

$$\mathcal{I}_{Y_1}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

A different way of computing the Fisher information is via the expected Hessian:

$$\mathrm{E}\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathtt{T}}}\right] = \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{n}{\sigma^4} \mathrm{E}[Y_1 - \mu] \\ -\frac{n}{\sigma^4} \mathrm{E}[Y_1 - \mu] & \frac{n}{2\sigma^4} - \frac{n}{\sigma^6} \mathrm{E}[(Y_1 - \mu)^2] \end{pmatrix} = \begin{pmatrix} -\frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} - \frac{n}{\sigma^4} \end{pmatrix} = -n\mathcal{I}_{Y_1}(\boldsymbol{\theta}),$$

as expected from the information equality.

The form of the Fisher information implies the CRLB is

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}) = n^{-1}\begin{pmatrix}\sigma^2 & 0\\ 0 & 2\sigma^4\end{pmatrix}.$$

so the unbiased $\bar{Y}$ achieves the CRLB. Now $\hat{\sigma}^2$ is biased so the CRLB does not apply.

Asymptotically,

$$\sqrt{n}\left[\begin{pmatrix}\hat{\mu}\\ \hat{\sigma}^2\end{pmatrix} - \begin{pmatrix}\mu\\ \sigma^2\end{pmatrix}\right] \xrightarrow{d} \mathcal{N}\left(0, \begin{pmatrix}\sigma^2 & 0\\ 0 & 2\sigma^4\end{pmatrix}\right),$$

so $\hat{\sigma}^2$ is asymptotically unbiased, asymptotically efficient, and asymptotically Gaussian.

4.7 Estimation when model approximates the truth*

Suppose a researcher builds a parameteric model $F_{\mathbf{Y};\theta}$. She estimates θ by the maximum likelihood estimator

$$\widehat{\theta} = \underset{\theta\in\Theta}{\arg\max} \log L(\theta; \mathbf{Y}).$$

However, the actual data is from $F_{\mathbf{Y}}$. What does her MLE estimate in practice?

To frame this question quantitatively, it is helpful to recall the Kullback-Leibler divergence between the true F_{Y} and the model is

$$D_{\mathrm{KL}}(F_{\mathrm{Y}}||F_{\mathrm{Y};\theta}) = \mathrm{E}\left[\log\frac{f(\mathbf{Y})}{f(\mathbf{Y};\theta)}\right],$$

where the expectation is with respect to $F_{\mathbf{Y}}$. The value of θ in Θ which makes the model $F_{\mathbf{Y};\theta}$ closest to $F_{\mathbf{Y}}$, measured using the Kullback-Leibler divergence, is denoted by θ^*. It is

$$\theta^* = \underset{\theta\in\Theta}{\arg\min}\, D_{\mathrm{KL}}(F_{\mathbf{Y}}||F_{\mathbf{Y};\theta}) = \underset{\theta\in\Theta}{\arg\max}\, \mathrm{E}\left[\log f(\mathbf{Y};\theta)\right],$$

as $\mathrm{E}\left[\log f(\mathbf{Y})\right]$ is parameter free. In statistics θ^* is often called the *pseudo-true* value of the model $F_{\mathbf{Y};\theta}$. If $F_{\mathbf{Y}} = F_{\mathbf{Y};\theta_0}$, a special case of the model, then $\theta^* = \theta_0$ and the corresponding $D_{\mathrm{KL}}(F_{\mathbf{Y}}||F_{\mathbf{Y};\theta_0}) = 0$. When this is not the case then

$$D_{\mathrm{KL}}(F_{\mathbf{Y}}||F_{\mathbf{Y};\theta^*}) > 0,$$

for the best parametric model of type $F_{\mathbf{Y};\theta}$ cannot match $F_{\mathbf{Y}}$.

☣ **4.7.1.** In general it is not clear if θ^* is unique. A sufficient condition for uniqueness is that the expected log-likelihood is strict concave, which follows if the log-likelihood is strictly concave (as Theorem 4.5.2 says that sums of non-negatively weighted concave functions are concave and an expectation is a weighted sum).

Example 4.7.2. Assume that $Y_j \overset{\text{i.i.d.}}{\sim} F_{Y_1}$ and the researcher models $Y_j \overset{iid}{\sim} Po(\theta)$. We saw in Section 4.3 that the i.i.d. assumptions imply $D_{\mathrm{KL}}(F_{\mathrm{Y}}||F_{\mathrm{Y};\theta}) = nD_{\mathrm{KL}}(F_{Y_1}||F_{Y_1;\theta})$. Then

$$D_{\mathrm{KL}}(F_{Y_1}||F_{Y_1;\theta}) = \mathrm{E}\left[\log f(Y_1)\right] + \theta - \mathrm{E}[Y_1]\log\theta + \mathrm{E}\left[\log Y_1!\right].$$

Then $\partial D_{\mathrm{KL}}(F_{Y_1}||F_{Y_1;\theta})/\partial\theta = \left(1 - \theta^{-1}\mathrm{E}[Y_1]\right)$, while $\partial^2 D_{\mathrm{KL}}(F_{Y_1}||F_{Y_1;\theta})/\partial\theta^2 = \theta^{-2}\mathrm{E}[Y_1]$, so $\theta^* = \mathrm{E}[Y_1]$.

Let us now dig into some of the implications of

$$\theta^* = \underset{\theta\in\Theta}{\arg\max}\,\mathrm{E}\left[\log f(\mathbf{Y};\theta)\right].$$

If (a) $\mathrm{E}\left[\log f(\mathbf{Y};\theta)\right]$ is continuously differentiable and (b) θ^* is an interior point of Θ, then θ^* has the property that

$$\left.\frac{\partial \mathrm{E}\left[\log f(\mathbf{Y};\theta)\right]}{\partial\theta}\right|_{\theta=\theta^*} = 0. \tag{4.6}$$

Further, if additionally (c) the range of support of Y does not depend upon θ, then by DUThIS on (4.6),

$$\mathrm{E}\left[s(\theta^*;\mathbf{Y})\right] = 0, \quad \text{recalling} \quad s(\theta;\mathbf{Y}) = \frac{\partial \log f(\mathbf{Y};\theta)}{\partial\theta}.$$

Theorem 4.7.3 (Consistency of MLE)**.** *Suppose that the parameter space Θ is finite and that the observations $Y_1,\ldots,Y_n$ are i.i.d. F_{Y_1} and the model is $F_{Y_1;\theta}$. Then the MLE, with probability 1, converges to the pseudo-true value:*

$$\widehat{\theta} \to \theta^*.$$

This means that the MLE is consistent for the pseudo-true value, fitting the parametric model as close as possible to the truth (where close is defined in terms of the Kullback-Leibler divergence).

Proof. This follows using the same line of argument as in the proof of Theorem 4.3.6, but now by the SLLN

$$\frac{1}{n}\sum_{j=1}^{n}\left\{\log f(Y_j) - \log f(Y_j;\theta)\right\} \to D_{\mathrm{KL}}(F_{Y_1}||F_{Y_1;\theta}),$$

with probability one and we use that $D_{\mathrm{KL}}(F_{Y_1}||F_{Y_1;\theta})$ is minimized at θ^*. ∎

How accurate is $\widehat{\theta}$ as an estimator of θ^*? Theorem 4.7.4 provides a guide. It goes back to Sir David R. Cox (1961) and Peter Huber (1967).

Theorem 4.7.4. *Let $\widehat{\theta}$ be the MLE of a scalar parameter θ from the model $F_{Y_1;\theta}$, based on observations $Y_1,...,Y_n$ which are i.i.d. from F_{Y_1}. Under regularity conditions,*

$$\sqrt{n}\left(\widehat{\theta} - \theta^*\right) \xrightarrow{d} N\left(0, \frac{\mathcal{I}_{Y_1}(\theta^*)}{\mathcal{J}_{Y_1}(\theta^*)^2}\right)$$

where $\mathcal{I}_{Y_1}(\theta^*) = \mathrm{Var}(s(\theta^*; Y_1))$ *and* $\mathcal{J}_{Y_1}(\theta^*) = -\mathrm{E}\left[s'(\theta^*; Y_1)\right]$. *Both the expectation and variance are computed under* F_{Y_1}.

Crucially, under F_{Y_1} the $\mathcal{I}_{Y_1}(\theta^*)$ does not need to equal $\mathcal{J}_{Y_1}(\theta^*)$ (that is the information equality cannot be rolled out). This makes the asymptotic distribution of the MLE slightly more complicated. The term $\mathcal{I}_{Y_1}(\theta^*)/\mathcal{J}_{Y_1}(\theta^*)^2$ is typically written in so-called *sandwich form*

$$\mathcal{J}_{Y_1}(\theta^*)^{-1}\mathcal{I}_{Y_1}(\theta^*)\mathcal{J}_{Y_1}(\theta^*)^{-1},$$

which works in the multiparameter case.

Proof. Now

$$0 = s(\widehat{\theta}; \mathbf{Y}) \approx s(\theta^*; \mathbf{Y}) + (\widehat{\theta} - \theta^*)s'(\theta^*; \mathbf{Y}),$$

so, rearanging

$$\sqrt{n}\left(\widehat{\theta} - \theta^*\right) \approx \frac{\sqrt{n}\frac{1}{n}s(\theta^*; \mathbf{Y})}{-\frac{1}{n}s'(\theta^*; \mathbf{Y})}.$$

The

$$\frac{1}{n}s(\theta^*; \mathbf{Y}) = \frac{1}{n}\sum_{j=1}^{n} s(\theta^*; Y_j), \quad \frac{1}{n}s'(\theta^*; \mathbf{Y}) = \frac{1}{n}\sum_{j=1}^{n} s'(\theta^*; Y_j)$$

are averages of i.i.d. terms. The former has terms with zero means, so by the CLT,

$$\sqrt{n}\frac{1}{n}s(\theta^*; \mathbf{Y}) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{Y_1}(\theta^*)),$$

while by the SSLN,

$$-\frac{1}{n}s'(\theta^*; \mathbf{Y}) = -\mathrm{E}\left[s'(\theta^*; Y_1)\right] = \mathcal{J}_{Y_1}(\theta^*).$$

So by Slutsky's Theorem, the stated result for the CLT of $\widehat{\theta}$ follows by the properties of the normal distribution. ∎

The presence of the terms $\mathcal{I}_{Y_1}(\theta^*)$ and $\mathcal{J}_{Y_1}(\theta^*)$ in the asymptotic variance, which are computed under $F_{\mathbf{Y}}$, looks practically difficult as we do not know $F_{\mathbf{Y}}$. But for i.i.d. data they can be consistently estimated by

$$\widehat{\mathcal{I}}_{Y_1}(\theta^*) = \frac{1}{n}\sum_{j=1}^{n}\left\{\frac{\partial \log g(Y_j; \widehat{\theta})}{\partial \theta}\right\}^2 \quad \text{and} \quad \widehat{\mathcal{J}}_{Y_1}(\theta^*) = -\frac{1}{n}s'(\widehat{\theta}; \mathbf{Y}).$$

The term $\sqrt{\widehat{\mathcal{I}}_{Y_1}(\theta^*)/n\widehat{\mathcal{J}}_{Y_1}(\theta^*)^2}$ is often called the *robust standard error* of the MLE. They are commonly used in modern applied statistics.

Example 4.7.5. Continuing Example 4.7.2, then $\theta^* = \mathrm{E}[Y_1]$, so the MLE estimates $\mathrm{E}[Y_1]$. Further,

$$\mathcal{I}_{Y_1}(\theta^*) = \mathrm{Var}(Y_1)/\theta^{*2}, \quad \mathcal{J}_{Y_1}(\theta^*) = \mathrm{E}[Y_1]/\theta^{*2} = 1/\theta^*,$$

so $\mathcal{I}_{Y_1}(\theta^*)\mathcal{J}_{Y_1}(\theta^*)^{-2} = \mathrm{Var}(Y_1)$, and thus

$$\sqrt{n}\left(\widehat{\theta} - \theta^*\right) \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(Y_1)).$$

If the data is Poisson, as the modeler specifies, then the simpler result $\mathcal{I}_{Y_1}(\theta^*) = \mathcal{J}_{Y_1}(\theta^*) = \theta^* = \mathrm{E}[Y_1]$ holds, yielding the classic

$$\sqrt{n}\left(\widehat{\theta} - \theta^*\right) \xrightarrow{d} \mathcal{N}(0, \mathrm{E}[Y_1]).$$

4.8 Recap

The maximum likelihood estimator is the value of the parameter of a parametric statistical model $F_{\mathbf{Y};\theta}$ which maximizes the likelihood. Using Kullback-Leibler divergence we can see that the expected log-likelihood is maximized at the true parameter value — the estimand. This means that as $n \to \infty$, the MLE converges in probability to the estimand θ^*, where the data comes from $F_{\mathbf{Y};\theta^*}$, so the MLE is consistent.

The main points of the chapter are summarized in Table 4.1.

Beyond the interesting features of MLE, the likelihood itself has some foundational properties. The variance of the score is called the *Fisher information*, whose inverse provides a lower bound on the variance of any unbiased estimator. This result is called the *Cramér-Rao lower bound*. The asymptotic distribution of the MLE shows that the MLE is approximately unbiased, efficient, and Normal.

From an applied perspective, noticing all of the likelihood theory can be exported to deal with models build conditionally, out of $F_{\mathbf{Y}|(\mathbf{X}=\mathbf{x})}(\mathbf{y};\theta)$, where X can be selected be make computations or inference more meaningful (e.g., the X are predictors, so $Y|X$ corresponds to prediction problems) or convenient (e.g., initial conditions for a Markov chain, to make the likelihood particularly simple to manipulate).

4.9 `R` and maximum likelihood estimation

4.9.1 Functions

The core of `R` has many useful built-in functions, such as `mean()`, `quantile()` and `lm()`. Sometimes it is necessary to build your own function, so they can be repeatedly called by your own code. Here we will describe the basics of functions in `R`, and illustrate this by providing code for the extended example of MLE in the Cauchy case given in Section 4.1.7.

Formula or idea	Description or name
$L(\theta;\mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y};\theta)$	likelihood
$\hat{\theta} = \underset{\theta\in\Theta}{\arg\max} \log L(\theta;\mathbf{y})$	maximum likelihood estimate
$\hat{\theta} = \underset{\theta\in\Theta}{\arg\max} \log L(\theta;\mathbf{Y})$	maximum likelihood estimator
$s(\theta;\mathbf{y}) = \partial \log L(\theta;\mathbf{y})/\partial\theta$	score
$\psi = g(\theta)$ then $\hat{\psi} = g(\hat{\theta})$	invariance of MLE
θ^*	estimand. Data generated by $F_{\mathbf{Y};\theta^*}$
$D_{\mathrm{KL}}(F\|\|G) = E\left(\log \frac{f(Y)}{g(Y)}\right)$	Kullback-Leibler divergence
$D_{\mathrm{KL}}(F_{Y;\theta^*}\|\|F_{Y;\theta}) \geq 0$	expected log-likelihood maximized at θ^*
$\hat{\theta} \xrightarrow{p} \theta^*$	MLE consistent
$\mathrm{E}[s(\theta^*;\mathbf{Y})] = 0$, and $\mathrm{Var}(s(\theta^*;\mathbf{Y})) = -\mathrm{E}[s(\theta^*;\mathbf{Y})]$	properties of score, derived using DUThIS
$\mathcal{I}_{\mathbf{Y}} = \mathrm{Var}(s(\theta^*;\mathbf{Y}))$	Fisher information
$\mathrm{Var}(\hat{\theta}) \geq \mathcal{I}_{\mathbf{Y}}^{-1}$	Cramér-Rao inequality for unbiased estimator
$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}_{Y_1}^{-1})$	asymptotic distribution of MLE
$\hat{\theta} \dot{\sim} \mathcal{N}(\theta^*, \mathcal{I}_{\mathbf{Y}}^{-1})$	approximate distribution of MLE
$L(\theta;\mathbf{y}) = f_{\mathbf{Y}\|(\mathbf{X}=\mathbf{x})}(\mathbf{y};\theta)$	likelihoods based on $\mathbf{Y}\|(\mathbf{X}=\mathbf{x});\theta$

Table 4.1: Main ideas and notation in Chapter 4. DUThIS is differentiation under the integral sign.

General structure of a function in R

```
function_name <- function(arguments) {
   do something

   return something
}
```

Example 4.9.1 (Reproducing `rnorm` function)**.** To illustrate the use of the `function` structure in R, we reproduce the core function `rnorm()`.

Setup a random number generator

```
our_random_numbers <- function(n, mu=0, sigma=1) {
   rnorm(n,mu,sigma)
}
```

We will discuss why we have not made any return statement in a moment. Likewise, we will explain in a moment why we wrote `mu=0, sigma=1` rather than just `mu,sigma` as arguments.

Then we will run this three times:

Run function multiple times

```
> set.seed(111)
> our_random_numbers(3,2.3,1.2)
[1] 2.582265 1.903117 1.926051
> our_random_numbers(5,2.3,1.2)
[1] -0.4628148  2.0949487  2.4683339  0.5030880  1.0877739
> x = our_random_numbers(5,2.3,1.2)
> x
[1] 1.161829 1.707245 2.091591 1.812081 4.514764
```

The function behaves the same way as `rnorm()`: each time we call it, it produces some Normal random numbers and we control how many of them to make, their mean, and their standard deviation. We can direct the output from the function into `x` and then display it.

What about the odd-looking `mu=0, sigma=1` in the function? This provides *default values* of `mu` and `sigma` for the function. So if you do not specify their values you generate $\mathcal{N}(0,1)$ variables.

Role of default values

```
> set.seed(111)
> our_random_numbers(5)
[1] 0.2352207 -0.3307359 -0.3116238 -2.3023457 -0.1708760
```

It is crucial to understand what these functions return and what they throw away. They give back the answer to the last command in the function, and if we're not careful everything else can get thrown away! Suppose that for some reason we had an extra line `x=3` at the end of our code:

Getting the wrong thing

```
> our_random_numbers <- function(n,mu=0,sigma=1) {
+     rnorm(n,mu,sigma)
+     x = 3
+ }
> set.seed(111)
> y = our_random_numbers(6)
> y
[1] 3
```

Then, as illustrated above, the output would just be 3; the random numbers we generated got discarded. This can be frustrating!

If we just want to *print* results from along the way we can use `print` on intermediate results of interest within our function, e.g., use `print(rnorm(n,mu,sigma)` instead of just `rnorm(n,mu,sigma)`. But if we want to *save* multiple items, we can do so by putting the results we want to keep in a suitable storage container. In some cases, a vector, matrix, or data frame will work nicely for that. In

other cases, a *list* is more flexible. For now, all we need to know about a list is that it is a collection of objects. Below is a quick example of outputting more than one vector, using a list. Lists will be more extensively in more detail in Section 7.4.1.

Outputting more than one thing

```
> our_random_numbers <- function(n,mu=0,sigma=1) {
+     x = rnorm(n,mu,sigma)
+     y = rt(n,3)
+
+     list(nRVs=x,tRVs = y)
+ }
> set.seed(111)
> our_random_numbers(5)
$nRVs
[1]  0.2352207 -0.3307359 -0.3116238 -2.3023457 -0.1708760

$tRVs
[1] 0.1302441 0.4972099 1.8883730 0.4498688 1.3538764
```

4.9.2 Code for various examples in this chapter

Now let's return to the extended example of MLE in the Cauchy case given in Section 4.1.7. Recall in that problem the log-likelihood is

$$l(\theta) = -n\log(\pi) - \sum_{j=1}^{n} \log\{1 + (y_j - \theta)^2\}.$$

The code below plots the log-likelihood against θ and then finds the MLE. The code is shown when $n = 10$; the same code is used when $n = 100$ by changing line 2.

Code from Section 4.1.7

```
set.seed(111)
n <- 10  # sample size
y <- rcauchy(n) # simulate data
loglik <- function(theta) sum(-log(pi)-log(1+(y-theta)^2))  # logL fn

theta <- seq(-3,3,0.01); # range of theta to evaluate logL

plot(theta,sapply(theta,loglik),type="l",
    ylab="log-likelihood",
    main="Cauchy location model log-likelihood")
mle <- optimize(loglik,lower=-3,upper=3,maximum =TRUE) # find MLE
c(mean(y),median(y),mle)  # print sample mean, sample median and MLE
```

The plot appears as Figure 4.2 above.

The simulations for the regression Example 4.4.1 are generated by the following code. The plot appears as Figure 4.7.

Code for Example 4.4.1

```
set.seed(111)
n = 100
w = 1.0*(runif(n)>0.5)
x = w*rnorm(n,10,3) + (1.0-w)*rnorm(n,-10,2)
y = 0.7*x + rnorm(n,0.0,3.0)

plot(x,y,pch=16,main="Gaussian regression model")
abline(a=0,b=mean(y*x)/mean(x^2),col="red",lwd=3)

FisherInfo = (1.0/(3.0^2))*sum(x^2)
seMLE = sqrt(1.0/FisherInfo)
```

Chapter 5

Confidence Intervals

5.1 Introduction

So far we have mainly been focusing on *point estimation*, where we deliver a single estimate for our estimand. But we have also emphasized the importance of quantifying the *uncertainty* in estimators, and we have used the remarkable fact that the *same data* that provides an estimator can also be used to assess the accuracy of the estimator. In *interval estimation*, the goal of our inference is to provide a range of plausible values for our estimand. In this chapter we will start to introduce concepts and techniques for thinking about and constructing such intervals.

Definition 5.1.1 (Interval estimation)**.** An *interval estimate* $C(\mathbf{y})$ of a scalar estimand θ based on data $\mathbf{y}$ is an interval $[L(\mathbf{y}), U(\mathbf{y})]$, where the lower bound $L(\mathbf{y})$ and upper bound $U(\mathbf{y})$ are functions of the data, such that $L(\mathbf{y}) \leq U(\mathbf{y})$ for all $\mathbf{y}$. The corresponding random interval $[L(\mathbf{Y}), U(\mathbf{Y})]$, where $\mathbf{Y}$ are the random vectors that give rise to the data, is called an *interval estimator*, written $C(\mathbf{Y})$. Intuitively, the goal is that the probability should be high that $C(\mathbf{Y})$ contains the estimand.

To simplify notation we have written $[L(\mathbf{y}), U(\mathbf{y})]$ as a closed interval, but we can also use an open interval or half-open interval, and we allow infinite-length intervals such as $[L(\mathbf{y}), \infty)$ and $(-\infty, U(\mathbf{y})]$ as possibilities.

This is illustrated in Figure 5.1, which displays ten realizations of the random interval. Nine cover the fixed estimand θ and one (the one drawn in dotted red) does not.

Definition 5.1.2 (Coverage)**.** Let $C(\mathbf{Y})$ be an interval estimator for θ and $C(\mathbf{y})$ be the corresponding interval estimate. If $\theta \in C(\mathbf{y})$, we say that the interval *covers* θ. The probability of the interval estimator covering θ if θ is the true estimand, $P_{\mathbf{Y};\theta}(\theta \in C(\mathbf{Y}))$, is called the *coverage probability* of the interval estimator. Note that the coverage probability is a function of θ.

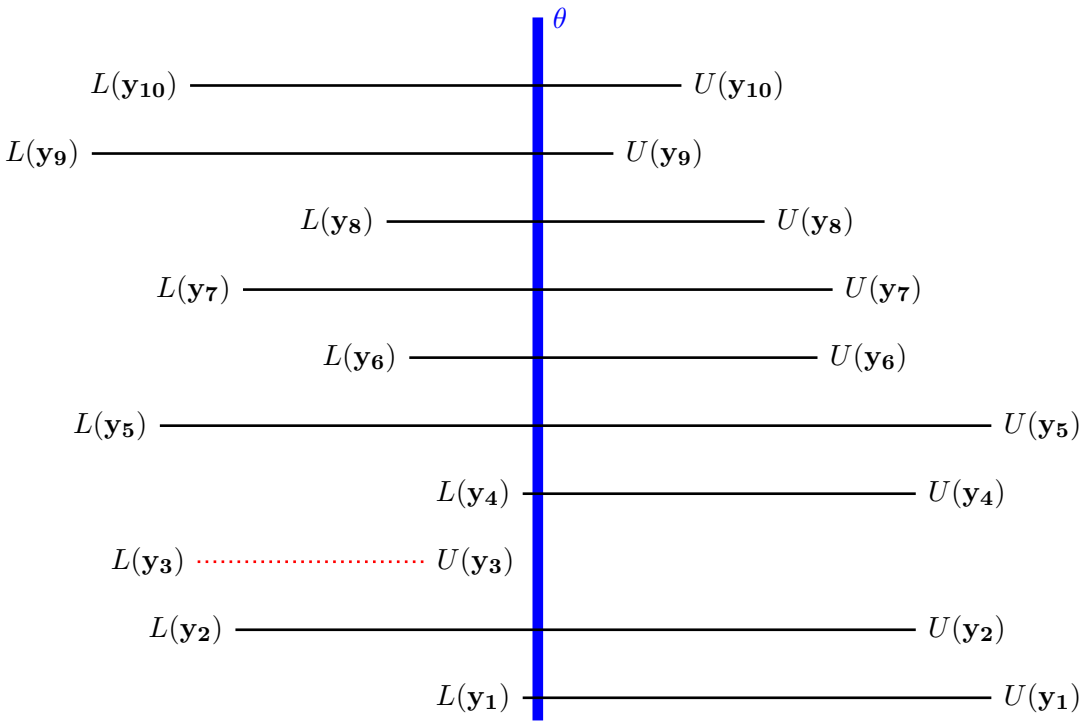


Figure 5.1: The random vector $\mathbf{Y}$ yields the random interval $[L(\mathbf{Y}), U(\mathbf{Y})]$. Shown are ten intervals, $[L(\mathbf{y}_1), U(\mathbf{y}_1)], \ldots, [L(\mathbf{y}_{10}), U(\mathbf{y}_{10})]$, with each corresponding to one replication of the experiment and represented by a horizontal line. The vertical bar is the true value θ. In this example, nine of the intervals (the solid black lines) cover θ, while one (the red dotted line) does not.

Intuitively, we would like to use an interval estimator that has high coverage probability. However, it does not make sense to look *only* at coverage probability.

☣ **5.1.3.** Construct an interval for θ which is $(-\infty, \infty)$ with probability 19/20 and the empty set otherwise. We call this the DD interval estimator, as it is inspired by Joe's 20-sided dice (from playing Dungeons and Dragons as a kid). This magical strategy has 95% coverage for all values of θ — a universal interval estimator which can be rolled out for all applied problems! But the DD interval estimator is, of course, utterly useless. This example illustrates the importance of considering the *lengths* of our interval estimators, not just the coverage probabilities.

There is no standard loss function for interval estimators, unlike for point estimation, where MSE is the most widely used choice. But there is a standard *goal*: fix a desired coverage probability, typically 0.95, and then create an interval estimator with the prescribed coverage probability for all θ, such that the resulting interval estimates tend to be reasonably short in length.

Definition 5.1.4 (Confidence Interval)**.** Fix a number α with $0 < \alpha < 1$. (In practice, the most common choice is $\alpha = 0.05$.) The interval estimator $C(\mathbf{Y}) = [L(\mathbf{Y}), U(\mathbf{Y})]$ is a $(1 - \alpha)$ *confidence*

interval (CI) if it has coverage probability $1-\alpha$ for all possible values of θ:

$$P_{\mathbf{Y};\theta}(\theta \in C(\mathbf{Y})) = 1-\alpha.$$

The constant $1-\alpha$ is called the *level* of the confidence interval. The half-width $0.5\{U(\mathbf{Y}) - L(\mathbf{Y})\}$ is called the *margin of error*.

In more than one dimension, we can define *confidence regions* analogously, e.g., if θ is a two-dimensional estimand we can consider confidence rectangles (or other regions in the plane) instead of confidence intervals.

☣ **5.1.5.** Confidence intervals are widely misinterpreted. For example, if the 95% CI $[0.1, 0.4]$ for θ is calculated from the data, a common mistake is to say that we can be 95% confident that θ is between 0.1 and 0.4, or that the probability is 0.95 that θ is between 0.1 and 0.4. This is a category error since the statement "θ is between 0.1 and 0.4" is deterministic, either true or false: we are currently working in a frequentist setting, so θ is *fixed*. It is the *interval estimator* that is random here, not the estimand.

Despite being nonsensical, this misinterpretation of CIs is unfortunately common. For example, Hoekstra et al. `http://www.ejwagenmakers.com/inpress/HoekstraEtAlPBR.pdf` reports on a study of 442 psychology students and 120 psychology researchers, asking them 6 true/false questions about how to interpret a confidence interval for a mean estimand where the CI was $[0.1, 0.4]$, such as whether it is true or false that "There is a 95% probability that the true mean lies between 0.1 and 0.4". The answer was false for all 6 questions, but both students and researchers marked an average of at least 3 of the 6 false interpretations as true.

The fact that this misinterpretation of CIs is so common is a sign that people *want* to be able to make statements like "the probability is 0.95 that θ is between 0.1 and 0.4, given the data". We *are* able to make such statements if we take a Bayesian approach, where probability statements about θ are made based on the posterior distribution of θ. Then we have what is called a *credible interval* rather than a confidence interval, e.g., a 95% credible interval $[L(\mathbf{y}), U(\mathbf{y})]$ for θ satisfies

$$P(\theta \in [L(\mathbf{y}), U(\mathbf{y})] | \mathbf{Y} = \mathbf{y}) = 0.95,$$

where now θ is regarded as a random variable (to capture our uncertainty about it) and $\mathbf{Y} = \mathbf{y}$ is conditioned on (since we observed it). We will discuss the Bayesian approach in much more detail in Chapter 9.

5.2 Constructing confidence intervals

We already have one natural approach for trying to construct a confidence interval for an estimand θ: start with the ML estimator $\hat{\theta}$, and then look at $\hat{\theta}$ plus or minus some number of standard errors, i.e., use the interval

$$C(\mathbf{Y}) = \left[\hat{\theta} - c \times \mathrm{SE}(\hat{\theta}), \hat{\theta} + c \times \mathrm{SE}(\hat{\theta})\right],$$

where c is a known positive constant. Very often in practice, c is chosen to be 1.96, for reasons that we will explain shortly. The true standard error $\mathrm{SE}(\hat{\theta})$ of $\hat{\theta}$ is also typically unknown, in which case it gets replaced by an estimated standard error (of course, the consequences of this replacement on the properties of the interval should then be studied).

Let us start with an example that illustrates an ideal case scenario, and then discuss more general strategies for constructing CIs.

Example 5.2.1 (Ideal case scenario: a Normal estimator)**.** We wish to create a $1-\alpha$ confidence interval for θ. Suppose that we are in the happy situation where $\hat{\theta}$ is Normal. Specifically, let

$$\hat{\theta} \sim \mathcal{N}(\theta, \sigma^2),$$

with σ^2 known. A simple but powerful approach is then to standardize $\hat{\theta}$ to get a standard Normal random variable:

$$\frac{\hat{\theta} - \theta}{\sigma} \sim \mathcal{N}(0,1).$$

The key here is that $\mathcal{N}(0,1)$ is a *known* distribution, not involving any unknown parameters, so we are in a good position to make probability statements about $\frac{\hat{\theta}-\theta}{\sigma}$. Specifically,

$$P\left(a \leq \frac{\hat{\theta} - \theta}{\sigma} \leq b\right) = F_{\mathcal{N}(0,1)}(b) - F_{\mathcal{N}(0,1)}(a),$$

where $F_{\mathcal{N}(0,1)} = \Phi$ is the $\mathcal{N}(0,1)$ CDF. We can choose constants a and b to obtain whatever probability we want on the right-hand side. For a $1-\alpha$ CI, we want the right-hand side to equal $1-\alpha$, and we can easily achieve this by setting

$$a = Q_{\mathcal{N}(0,1)}\left(\frac{\alpha}{2}\right), \quad b = Q_{\mathcal{N}(0,1)}\left(1 - \frac{\alpha}{2}\right),$$

where $Q_{\mathcal{N}(0,1)}$ is the quantile function of a $\mathcal{N}(0,1)$ random variable. This choice of a, b is not unique but it is simple, standard, and has the nice property that the resulting interval will be symmetric about $\hat{\theta}$. To simplify notation, let

$$c_p = Q_{\mathcal{N}(0,1)}(1-p)$$

for any $p \in (0,1)$. Note that $c_{1-p} = -c_p$ by symmetry of the Normal distribution (see Section 5.4 of the Stat 110 book). Thus setting

$$a = -c_{\alpha/2}, \quad b = c_{\alpha/2},$$

delivers

$$P\left(Q_{\mathcal{N}(0,1)}\left(\frac{\alpha}{2}\right) \leq \frac{\hat{\theta}-\theta}{\sigma} \leq Q_{\mathcal{N}(0,1)}\left(1-\frac{\alpha}{2}\right)\right) = P\left(-c_{\alpha/2} \leq \frac{\hat{\theta}-\theta}{\sigma} \leq c_{\alpha/2}\right) = 1-\alpha.$$

Rearranging,

$$P\left(\hat{\theta} - c_{\alpha/2}\,\sigma \leq \theta \leq \hat{\theta} + c_{\alpha/2}\,\sigma\right) = 1-\alpha.$$

Thus,

$$C(\mathbf{Y}) = \left[\hat{\theta} - c_{\alpha/2}\,\sigma, \hat{\theta} + c_{\alpha/2}\,\sigma\right]$$

is a $1-\alpha$ CI for θ, with margin of error $c_{\alpha/2}\,\sigma$. As shorthand, we sometimes write this interval as

$$\hat{\theta} \pm c_{\alpha/2}\,\sigma;$$

it is a CI centered at the MLE, with $c_{\alpha/2}$ standard errors of slack in each direction.

In practice, the most common choice of α is 0.05, and we obtain our CI by cutting off 2.5% at each tail. Then

$$c_{\alpha/2} = Q_{\mathcal{N}(0,1)}(0.975) \approx 1.96.$$

This helps explain why the number 1.96 is ubiquitous in statistics.

At the heart of why Example 5.2.1 works is that the quantity $\frac{\hat{\theta}-\theta}{\sigma}$ has a *known* distribution, which happened to be $\mathcal{N}(0,1)$. In statistical terms, we say that we created a *pivotal quantity*.

Definition 5.2.2 (Pivot)**.** A *pivotal quantity* or *pivot* is a random variable whose distribution is known.

☣ **5.2.3.** It is essential not to confuse a pivot with a statistic.

- A pivot typically depends on unknown parameters but its distribution cannot depend on unknown parameters.
- A statistic cannot depend on unknown parameters but its distribution typically depends on unknown parameters.

For example, suppose that we observe $Y_1, \ldots, Y_n$ which are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, with μ and σ both unknown. Then $\bar{Y}$ is a statistic (we can compute it from the data, without having to know μ or σ). But it is not a pivot, since

$$\bar{Y} \sim \mathcal{N}(\mu, \sigma^2/n),$$

so the distribution of $\bar{Y}$ depends on unknown parameters. In contrast, the expression on the left-hand side of

$$\sqrt{n}\left(\frac{\bar{Y} - \mu}{\sigma}\right) \sim \mathcal{N}(0, 1)$$

is a pivot since its distribution is known. It is not a statistic though, since to compute it would require knowing μ and σ in addition to knowing the data.

Example 5.2.1 is idealized but suggests much more general lines of thinking. We will examine three major ways to go beyond Example 5.2.1.

1. The estimator $\hat{\theta}$ is often *approximately* Normal even if it is not exactly Normal, and then we can obtain an approximate CI using the method of Example 5.2.1. Also, there is no rule that says we must base our CIs on the MLE $\hat{\theta}$; we are free to use any estimator in place of $\hat{\theta}$ as a starting point and then use a method similar to that of Example 5.2.1. We often use *asymptotics* to obtain approximate distributions.

2. We may be able to create a different pivot, with a known distribution other than Normal, and then follow a similar approach except, e.g., with Gamma quantiles rather than Normal quantiles.

3. We may use an approach called *resampling* to study the distribution of $\hat{\theta}$ computationally rather than mathematically, and then based on this there are several methods for obtaining approximate CIs. We will study resampling in Chapter 10, focusing on a powerful technique called the *bootstrap*.

5.3 Asymptotic approximations

In practice we often need to make approximations. It is often true that for fixed sample size n, the distribution of an estimator $\hat{\theta}$ is very complicated and very hard to derive mathematically, whereas for large n we can say that $\hat{\theta}$ is approximately Normal. We have already seen three major reasons why an amazing variety of estimators are asymptotically Normal:

- The *central limit theorem*. For example, if $\hat{\theta} = \bar{Y}$ where the Y_j are i.i.d. then the CLT tells us that $\hat{\theta}$ is asymptotically Normal.

- The *delta method*. For example, if $\hat{\theta} = \bar{Y}^3$ where we already know from the CLT that $\bar{Y}$ is asymptotically Normal, then $\hat{\theta}$ is also asymptotically Normal.

- Two of the most widely used estimators, the MLE and the MoM, are (under some regularity conditions) asymptotically Normal.

Definition 5.3.1 (Approximate Pivot)**.** Suppose that n is large and, based on asymptotics, we know that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

Then

$$\frac{\hat{\theta} - \theta}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1),$$

and the expression on the left-hand side is called an *approximate pivot*

When we have an approximate pivot, we can obtain an approximate CI by proceeding as in Example 5.2.1 (with approximate equalities in place of equalities). It follows that if σ is known then the interval

$$C(\mathbf{Y}) = \left[\hat{\theta} - c_{\alpha/2}\,\sigma/\sqrt{n}, \hat{\theta} + c_{\alpha/2}\,\sigma/\sqrt{n}\right]$$

is an approximate $1 - \alpha$ CI. Usually in practice σ is unknown, but can be estimated with some consistent estimator $\hat{\sigma}$. Then it is natural to use the interval

$$C(\mathbf{Y}) = \left[\hat{\theta} - c_{\alpha/2}\,\hat{\sigma}/\sqrt{n}, \hat{\theta} + c_{\alpha/2}\,\hat{\sigma}/\sqrt{n}\right],$$

though it is not obvious what effect plugging in $\hat{\sigma}$ for σ has on the coverage probability for fixed n; this may need to be studied via simulation. Asymptotically this substitution is fine though, since if

$$\sqrt{n}\left(\frac{\hat{\theta} - \theta}{\sigma}\right) \xrightarrow{d} \mathcal{N}(0, 1)$$

and $\hat{\sigma} \xrightarrow{p} \sigma$, then by the continuous mapping theorem and Slutsky's theorem,

$$\sqrt{n}\left(\frac{\hat{\theta} - \theta}{\hat{\sigma}}\right) = \sqrt{n}\left(\frac{\hat{\theta} - \theta}{\sigma}\right)\frac{\sigma}{\hat{\sigma}} \xrightarrow{d} \mathcal{N}(0, 1).$$

☣ **5.3.2.** The asymptotic approach only gives *approximate* confidence intervals. The mathematical statement is that, if $C(\mathbf{Y})$ is the interval estimator, then

$$\lim_{n\to\infty} P_{\mathbf{Y};\theta}(\theta \in C(\mathbf{Y})) = 1 - \alpha$$

as the sample size $n \to \infty$. This does not say, for a fixed n, how close the coverage probabilities are to $1 - \alpha$.

Confusingly, people often say "confidence interval" when they mean "approximate confidence interval". Some so-called $1 - \alpha$ confidence intervals are just aspirational, and in reality the coverage probability is far from $1 - \alpha$ for at least some possible values of θ. In such situations it is clearer to call $1 - \alpha$ the *nominal* level of the interval estimator. The hope is that the coverage probabilities will be close to $1 - \alpha$ for all θ but this may not be true, or it may be true but not yet demonstrated.

A further terminological issue is that some people define an $1-\alpha$ confidence interval with $\geq$ rather than $=$ in the definition, i.e., the coverage probability is *at least* $1 - \alpha$ for all θ. An advantage of this is that it may be easier to ensure that the coverage probability is at least $1 - \alpha$ for all θ than to get it to equal $1 - \alpha$ for all θ. This is especially true if the data are discrete, in which case it may be impossible to directly construct an interval estimator whose coverage probability is exactly $1 - \alpha$ for all θ. Using $\geq$ rather than $=$ is a smaller difference than it might seem. For example, if our goal was 95% coverage and we somehow ended up with 99% coverage, we might think "that's even better!", but we might also be able to shorten our intervals.

Example 5.3.3 (Election and Binomial CI)**.** An election is being held with two candidates, Candidate A and Candidate B. Each voter will vote either for Candidate A or for Candidate B. Suppose that we want a 95% confidence interval for p, the proportion of voters who will vote for Candidate A.

A survey of n voters is conducted, asking each voter whom they will vote for. Let Y be the number who say they will vote for A, and assume that $Y \sim \text{Bin}(n, p)$. (In practice, there are many complications such as people refusing to respond to the question, people not showing up to vote, people who change their minds or who are undecided about whom to vote for, and how to obtain a sample that looks like the population of voters.) The ML estimator is

$$\hat{p} = \frac{Y}{n},$$

and we know a lot about the distribution of $\hat{p}$ since we know a lot about the Binomial distribution. For example, we know that

$$\text{E}[\hat{p}] = p, \quad \text{Var}(\hat{p}) = \frac{p(1-p)}{n}.$$

It is not possible to find a CI with exactly 95% coverage probability for all p, because of the discreteness of the Binomial. Also, the needed calculations are much easier using a Normal than a Binomial, so we will use a Normal approximation for $\hat{p}$. Interpreting Y as the sum of n i.i.d. Bern(p)

random variables (imagine an indicator random variable for each respondent, equal to 1 if they will vote for Candidate A and 0 otherwise), we have a CLT for $\hat{p}$:

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \xrightarrow{d} \mathcal{N}(0,1).$$

This asymptotic statement is valid for all $p \in (0,1)$, but the convergence will be fast if p is close to $1/2$ and slow if p is close to 0 or 1.

This suggests using the interval

$$C(\mathbf{Y}) = \left[\hat{p} - 1.96\sqrt{p(1-p)/n}, \hat{p} + 1.96\sqrt{p(1-p)/n}\right],$$

but we cannot do that since p is unknown. If we knew p so that we could compute $\sqrt{p(1-p)/n}$, we would not need a CI in the first place! One approach to this difficulty is to note that $p(1-p)$ is maximized at $p = 1/2$ (as is easily seen by plotting the function $p(1-p)$ for p from 0 to 1), so

$$p(1-p) \leq \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

Widening the interval by rounding 1.96 up to 2 and replacing the unknown $p(1-p)$ by the known constant $1/4$ gives the simple and conservative interval

$$\left[\hat{p} - \frac{1}{\sqrt{n}}, \hat{p} + \frac{1}{\sqrt{n}}\right].$$

In widening the interval the coverage probability can only increase. Still, the conservative interval may be wider than necessary.

Another approach to is to replace the unknown $\sqrt{p(1-p)/n}$ by an estimator. By invariance, the MLE of $\sqrt{p(1-p)/n}$ is $\sqrt{\hat{p}(1-\hat{p})/n}$. This yields the interval

$$C(\mathbf{Y}) = [\hat{p} - 1.96\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + 1.96\sqrt{\hat{p}(1-\hat{p})/n}].$$

Let's check that asymptotically this is a valid 95% CI. By the LLN and the continuous mapping theorem,

$$\sqrt{\hat{p}(1-\hat{p})} \xrightarrow{p} \sqrt{p(1-p)}.$$

So by Slutsky's theorem,

$$\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \cdot \frac{\sqrt{p(1-p)/n}}{\sqrt{\hat{p}(1-\hat{p})/n}} \xrightarrow{d} \mathcal{N}(0,1).$$

Note that the slow growth of $\sqrt{n}$ is rather alarming if we want a small margin of error. Assume for simplicity that we use the conservative interval $\hat{p} \pm \frac{1}{\sqrt{n}}$. A typical sample size for a political poll is $n = 1,000$, but then the margin of error is about 0.03, which is problematic in a close election. If we want to get the margin of error down to 0.01, then we need $n = 10,000$, ten times as large a sample.

☣ **5.3.4.** Suppose that $\hat{p} = 0.52$, with a margin of error of 0.03. It is often reported in such a situation that the election is a "statistical tie" or that both candidates are equally likely to win, since the interval $[0.49, 0.55]$ contains values where A would win and values where B would win. There are several problems with this.

- The term "statistical tie" is not a well-defined statistical concept. (It can be formalized as a statistical hypothesis test, with a more precise statement being that the null hypothesis that the candidates have equal support would not be rejected based on the data. Hypothesis testing of this type will be discussed in Chapter 8.)
- The margin of error depends on the confidence level (e.g., 0.95), but the choice of confidence level is somewhat arbitrary. It is important when making statements to be clear about the confidence level!
- In a frequentist framework, $p > 0.5$ is a deterministic statement. It is either true or false! We need a Bayesian framework to be able to make probability statements about statements such as the chance that $p > 0.5$.
- There is no reason to think that all values in the interval $[0.49, 0.55]$ are equally plausible. The ML estimate 0.52 is better supported by the data than the endpoint 0.49.

To conclude this example, let us assess the coverage probability of the interval

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$$

via simulation. Asymptotically the coverage probability is 95%, but what is it for fixed n? A particular concern is what happens if p is close to 0 or 1, since the Binomial convergence to Normal via the CLT is slow in that case. Even without a computer we can see a potential disaster: if p is very close to 0 and n is not huge, it would not be surprising if every trial was a failure, yielding $Y = 0$, $\hat{p} = 0$, and the absurd interval $[0, 0]$ for p. Similarly, if p is very close to 1 then it may be unsurprising to observe $Y = n$, $\hat{p} = 1$, resulting in the absurd interval $[1, 1]$ for p.

Finding the coverage probability mathematically is difficult, but finding it via simulation is easy. For each p for which we want the coverage probability, we just need to generate a large number of i.i.d. $\text{Bin}(n, p)$ random variables, compute the interval for each one, and count what proportion of them contain p. Let $n = 30$ (the traditional rule of thumb for how large a sample size is needed for the CLT to give a good approximation). In Figure 5.2, we plot coverage probability as a function of p, based on a simulation with a million replications for each p in the grid of values $\{0.001, 0.002, \ldots, 0.998, 0.999\}$.

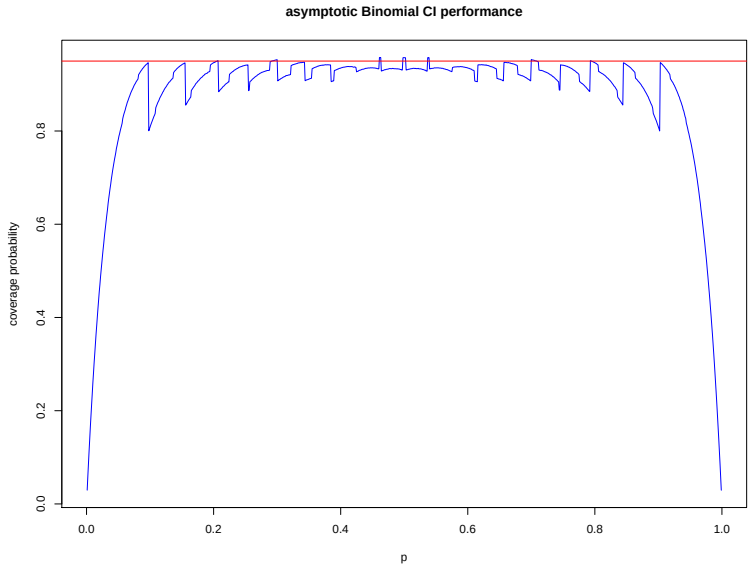


Figure 5.2: Coverage probability of the interval $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$ as a function of p, based on the sample size $n = 30$. A red horizontal line is shown at 0.95, the nominal confidence level. The coverage probability is almost always less than the nominal level, and is far below the nominal level if p is close to 0 or 1.

The coverage probability is abysmal for p close to 0 or 1. For example, the coverage probability is only 3% if $p = 0.001$ and 26% if $p = 0.01$. For $p = 0.1$ the coverage probability is a more respectable 81% but is still far below the desired 95%. The coverage probability is below 95% almost across the board, though it is approximately 95% when p is close to 0.5.

Example 5.3.5. Suppose that the estimand is a p-quantile $\theta = Q(p)$, and the estimator is the sample p-quantile $\hat{\theta} = \hat{Q}(p)$. Again we want an interval estimator for θ, not just the point estimator $\hat{\theta}$. Here p is known since it is chosen (e.g., we choose $p = 1/2$ if we are interested in the median) but θ is unknown since the true quantile function Q is unknown. Let the desired confidence level be 0.95.

As stated in Chapter 3, the asymptotic distribution of $\hat{\theta}$ is

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{(f_{Y_1}(\theta))^2}\right).$$

This suggests using the interval

$$C(\mathbf{Y}) = \left[\hat{\theta} - \frac{1.96}{f_{Y_1}(\theta)}\sqrt{\frac{p(1-p)}{n}}, \hat{\theta} + \frac{1.96}{f_{Y_1}(\theta)}\sqrt{\frac{p(1-p)}{n}}\right].$$

The $f_{Y_1}(\theta)$ appearing in the denominators is unknown, but we can estimate it using kernel density estimation, thus obtaining an approximate 95% CI for θ.

Example 5.3.6 (Duration of birth)**.** In the duration of birth example from Chapter 1, we reported the five estimates

$$\overline{y} \approx 7.72, \quad s_n \approx 3.57, \quad \hat{Q}(0.05) \approx 2.57, \quad \hat{Q}(0.5) \approx 7.50, \quad \hat{Q}(0.95) \approx 14.53,$$

and the empirical CDF $\widehat{F}(y)$ for $y \in [0, 20]$. These are estimates for the estimands

$$\mathrm{E}(Y_1), \quad \mathrm{SD}(Y_1), \quad Q(0.05), \quad Q(0.5), \quad Q(0.95),$$

and the CDF $F(y)$ for $y \in [0, 20]$, respectively.

We now provide a confidence interval for each estimand. We do not know the distributions of the estimators, but all of them are asymptotically Normal. So we will develop asymptotic confidence intervals. Let

$$\widehat{\mu}_4 = \frac{1}{n-1} \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^4,$$

and $\widehat{f}_n$ be the kernel density estimator (using a standard method known as the *Silverman rule of thumb* to choose the bandwidth).

The confidence intervals for the duration of birth example are reported in Table 5.1, with $\alpha = 0.05$. The large sample distributions we used for these calculations are:

$$\frac{\sqrt{n}\left\{\overline{Y} - \mathrm{E}(Y_1)\right\}}{\sqrt{S_n^2}} \xrightarrow{d} \mathcal{N}(0,1) \text{ by the CLT}, \quad \frac{\sqrt{n}\left(S_n^2 - \sigma^2\right)}{\sqrt{\widehat{\mu}_4 - \left(S_n^2\right)^2}} \xrightarrow{d} \mathcal{N}(0,1) \text{ by Chapter 3},$$

$$\frac{\sqrt{n}\left\{\hat{Q}(p) - Q(p)\right\}}{\sqrt{\frac{p(1-p)}{\left\{\widehat{f}_n(\hat{Q}(p))\right\}^2}}} \xrightarrow{d} \mathcal{N}(0,1) \text{ by Chapter 3}, \quad \frac{\sqrt{n}\left\{\widehat{F}(y) - F(y)\right\}}{\sqrt{\widehat{F}(y)(1-\widehat{F}(y))}} \xrightarrow{d} \mathcal{N}(0,1) \text{ by the CLT}.$$

To obtain the interval for the standard deviation, we compute the confidence interval for the variance and then take the square roots of the endpoints. This transformation makes sense since, in general, for any positive numbers σ, a, b with $a < b$, we have $\sigma^2 \in [a, b]$ if and only if $\sigma \in [\sqrt{a}, \sqrt{b}]$. The resulting interval for the standard deviation is not centered at the point estimate.

The right endpoint of the interval for the $\widehat{F}(15)$ is 1.003, which is absurd since the estimand must be between 0 and 1. It can be truncated at 1 (this truncation has no effect on the coverage probability). Alternatively, we could use the delta method to approximate the distribution of the transformed estimator $\log\left\{\widehat{F}(15)/(1-\widehat{F}(15))\right\}$, use that to form an interval for $\log\{F(15)/(1-F(15))\}$, and then transform back to get an interval for $F(15)$.

Estimate	value	Nominal 95% CI	
	$\hat{\theta}$	Lower	Upper
$\overline{y}$	7.72	7.00	8.44
s_n	3.57	2.98	4.07
$\hat{Q}(0.05)$	2.57	1.69	3.44
$\hat{Q}(0.5)$	7.50	6.51	8.48
$\hat{Q}(0.95)$	14.5	12.0	16.9
$\widehat{F}(5)$	0.263	0.174	0.352
$\widehat{F}(10)$	0.768	0.684	0.853
$\widehat{F}(15)$	0.968	0.933	1.003

Table 5.1: Nominal 95% confidence intervals, based on asymptotic approximations, for summary statistics for the duration of birth example from Chapter 1. The table also contains the summary statistics for the data.

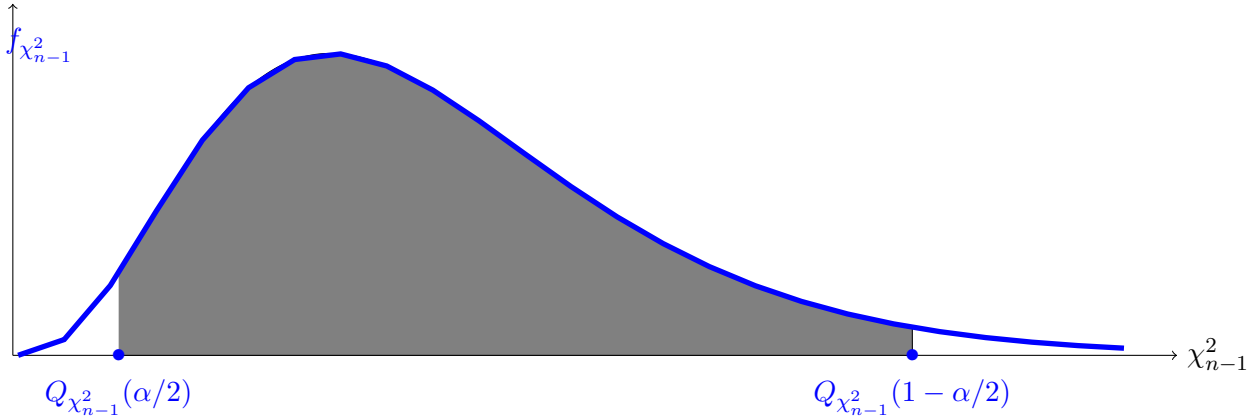


Figure 5.3: $1-\alpha$ confidence interval for $V \sim \chi^2_{n-1}$, placing $\alpha/2$ in each tail so shaded area contains $1-\alpha$ of the probability. Drawn here for $n = 7$.

5.4 Pivots with non-Gaussian distributions

A pivot must have a *known* distribution; there is no requirement that this distribution be Normal.

Example 5.4.1 (χ^2 distribution)**.** Let the data be i.i.d. $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$, with both parameters unknown. Suppose that we want a $1-\alpha$ CI for σ. The sample standard deviation is

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2}.$$

Theorem 10.4.3 in the Stat 110 book says that

$$V = (n-1)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-1},$$

so we have found a pivot. Hence, cutting off area $\alpha/2$ from each tail of the PDF of the χ^2_{n-1} distribution, we have

$$1-\alpha = P\left(Q_{\chi^2_{n-1}}(\alpha/2) \leq \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \leq Q_{\chi^2_{n-1}}(1-\alpha/2)\right).$$

This idea is illustrated in Figure 5.3 when $n = 7$. Inverting V flips the inequality signs, so

$$1-\alpha = P((n-1)\hat{\sigma}^2/Q_{\chi^2_{n-1}}(\alpha/2) \geq \sigma^2 \geq (n-1)\hat{\sigma}^2/Q_{\chi^2_{n-1}}(1-\alpha/2)).$$

Then the interval

$$C(\mathbf{Y}) = \left[\hat{\sigma}\sqrt{\frac{(n-1)}{Q_{\chi^2_{n-1}}(1-\alpha/2)}}, \hat{\sigma}\sqrt{\frac{(n-1)}{Q_{\chi^2_{n-1}}(\alpha/2)}}\right]$$

is a $1-\alpha$ confidence interval for σ.

Example 5.4.2 (t distribution)**.** Again, let the data be i.i.d. $Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$, with both parameters unknown. Suppose that we want a $1-\alpha$ CI for μ. If σ were known then, as shown earlier,

$$C(\mathbf{Y}) = \left[\bar{Y} - c_{\alpha/2}\sigma n^{-1/2}, \bar{Y} + c_{\alpha/2}\sigma n^{-1/2}\right]$$

would work, where $c_{\alpha/2} = Q_{\mathcal{N}(0,1)}(1-\alpha/2)$. But σ is unknown. We could replace σ by the sample standard deviation $\hat{\sigma}$, but then we would only have an approximate CI. Instead, let us construct a pivot, the *t-statistic*

$$T = \frac{\bar{Y}-\mu}{\hat{\sigma}/\sqrt{n}} = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \times \frac{\sigma}{\hat{\sigma}}.$$

The intuition for where T comes from is that we are trying to standardize the estimator $\bar{Y}$: first subtract the mean μ, then divide by the standard deviation, which is $\sigma/\sqrt{n}$. But it will be awkward trying to get a CI for μ when there are σ's floating around, so we replace σ by $\hat{\sigma}$.

We then have

$$T \sim t_{n-1},$$

a Student-t distribution with $n-1$ degrees of freedom. (For information about the Student-t distribution, see Section 10.4 of the Stat 110 book.)

Proof. Recall that if $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_{n-1}$, with Z and V independent, then

$$\frac{Z}{\sqrt{V/(n-1)}} \sim t_{n-1}.$$

Writing

$$Z = \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0,1), \quad V = \frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Moreover, the two ratios are independent by Example 7.5.9 in the Stat 110 book (the fact that the sample mean is independent of the sample variance is a very special property of the Normal distribution). Therefore, $T \sim t_{n-1}$. ∎

Thus, T is a pivotal quantity. It is *not* a statistic since μ is unknown. If μ is replaced by a hypothesized value μ_0 then we obtain a famous statistic known as a t *statistic*, which forms the basis of a widely used hypothesis testing procedure known as a t *test.* (We will discuss hypothesis testing in detail in Chapter 8.)

We can now obtain a CI based on T, using quantiles of the Student-t distribution

$$P\left(Q_{t_{n-1}}(\alpha/2) \leq T \leq Q_{t_{n-1}}(1-\alpha/2)\right) = (1-\alpha/2) - \alpha/2 = 1-\alpha.$$

Rearranging,

$$P\left(\bar{Y} - Q_{t_{n-1}}(1-\alpha/2)\hat{\sigma}/\sqrt{n} \leq \mu \leq \bar{Y} - Q_{t_{n-1}}(\alpha/2)\hat{\sigma}/\sqrt{n}\right) = 1-\alpha.$$

By symmetry of the t distribution about 0, $Q_{t_{n-1}}(\alpha/2) = -Q_{t_{n-1}}(1-\alpha/2)$. So a $1-\alpha$ CI for μ is

$$C(\mathbf{Y}) = \left[\bar{Y} - Q_{t_{n-1}}(1-\alpha/2)\hat{\sigma}/\sqrt{n}, \bar{Y} + Q_{t_{n-1}}(1-\alpha/2)\hat{\sigma}/\sqrt{n}\right].$$

Note that as n grows, $\bar{Y}$ gets closer and closer to μ (by the LLN) and the interval gets narrower and narrower, which makes sense intuitively. If one forgets to include the n, or multiplies instead of divides by $\sqrt{n}$, it should be easy to detect the mistake by looking at what happens to the length of the interval as n grows.

For large n, the t_{n-1} distribution is very close to $\mathcal{N}(0,1)$, whereas for small n the much heavier tails of the t_{n-1} distribution are evident. For example,

$$Q_{t_4}(0.975) \approx 2.78, \quad Q_{t_9}(0.975) \approx 2.26, \quad Q_{t_{49}}(0.975) \approx 2.01, \quad Q_{t_{499}}(0.975) \approx 1.96.$$

So for $\alpha = 0.05$, when n is at least moderately large we are essentially using the Normal quantile

$$Q_{\mathcal{N}(0,1)}(0.975) \approx 1.96.$$

We next consider an example that allows us to compare several different approaches, each of which is applicable to a wide variety of other problems.

Example 5.4.3 (Interval estimation for the Exponential rate parameter)**.** Let the data be i.i.d. r.v.s $Y_1, \ldots, Y_n \sim \text{Expo}(\lambda)$, with the rate parameter λ unknown. Let $\mu = 1/\lambda$ be the mean parameter. Suppose that we want a 95% CI for λ. We will consider three methods. The first two rely on asymptotics, while the third only uses exact probability calculations.

Method 1: Transformation. First we will construct an asymptotic CI for μ, and then we will transform it to get an asymptotic CI for λ. The most natural estimator for μ is $\bar{Y}$, noting $\bar{Y}$ is both the MLE and the MoM for μ. Now let us develop an interval estimator based on $\bar{Y}$. By the CLT,

$$\sqrt{n}\left(\bar{Y} - \mu\right) \xrightarrow{d} \mathcal{N}(0, \mu^2),$$

since $\mathrm{E}[Y_j] = \mu, \mathrm{Var}(Y_j) = \mu^2$. Estimating

$$\mathrm{SE}(\bar{Y}) = \sqrt{\mu^2/n} = \mu/\sqrt{n}$$

by its MLE,

$$\hat{\mathrm{SE}}(\bar{Y}) = \bar{Y}/\sqrt{n},$$

we have that

$$C(\mathbf{Y}) = \left[\bar{Y} - 1.96\,\bar{Y}/\sqrt{n}, \bar{Y} + 1.96\,\bar{Y}/\sqrt{n}\right]$$

is an approximate 95% CI for μ when n is large. This interval may perform poorly if n is small, e.g., it is nonsensical if $n < 4$ since then it contains negative values. Assume for this method and the next method that $n \geq 4$ (of course, we would want n to be considerably larger than this anyway for it to be reasonable to apply an asymptotic approach).

It is easy to transform the interval for μ to an interval for λ. Note that for any positive numbers a, b with $a \leq b$, we have

$$a \leq \mu \leq b \text{ if and only if } \frac{1}{b} \leq \lambda \leq \frac{1}{a}.$$

Therefore,

$$C(\mathbf{Y}) = \left[\frac{1}{\bar{Y} + 1.96\,\bar{Y}/\sqrt{n}}, \frac{1}{\bar{Y} - 1.96\,\bar{Y}/\sqrt{n}}\right]$$

is an approximate 95% CI for λ when n is large. The coverage probability of the interval for μ is the same as that of the interval for λ.

Method 2: Delta Method. The MLE of μ is $\bar{Y}$, so by invariance the MLE of $\lambda = 1/\mu$ is

$$\hat{\lambda} = \frac{1}{\bar{Y}}.$$

We will now construct an asymptotic CI for λ based on $\hat{\lambda}$. As noted in the previous method, the asymptotic distribution of $\bar{Y}$ is

$$\sqrt{n}\left(\bar{Y} - \mu\right) \xrightarrow{d} \mathcal{N}(0, \mu^2).$$

The delta method gives us a way to obtain the asymptotic distribution for $\hat{\lambda}$ from that of $\bar{Y}$. Letting $g(\mu) = 1/\mu$, we have $g'(\mu) = -1/\mu^2$, so the delta method gives

$$\sqrt{n}\left[g(\bar{Y}) - g(\mu)\right] \xrightarrow{d} \mathcal{N}(0, g'(\mu)^2\mu^2).$$

That is,

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{d} \mathcal{N}(0, \lambda^2).$$

Alternatively, we can obtain the same result by computing the Fisher information for λ and using the result from Chapter 4 about the asymptotic distribution of the MLE.

So

$$\text{SE}(\hat{\lambda}) \approx \sqrt{\frac{\lambda^2}{n}} = \frac{\lambda}{\sqrt{n}}$$

for large n, and we estimate this using $\hat{\text{SE}}(\hat{\lambda}) \approx \hat{\lambda} n^{-1/2}$. We then have the following asymptotic CI for λ:

$$C(\mathbf{Y}) = \left[\hat{\lambda} - 1.96\hat{\lambda}n^{-1/2}, \hat{\lambda} + 1.96\hat{\lambda}n^{-1/2}\right] = \left[\frac{1 - 1.96/\sqrt{n}}{\bar{Y}}, \frac{1 + 1.96/\sqrt{n}}{\bar{Y}}\right].$$

Method 3: Pivot. If we remember properties of the Exponential, we can construct a pivot without having to rely on any asymptotics. Specifically, recall that the sum of i.i.d. Exponentials has a Gamma distribution. Also recall from Section 8.4 of the Stat 110 book that if $G \sim \text{Gamma}(n, \lambda)$, then $\lambda G \sim \text{Gamma}(n, 1)$. We have $n\bar{Y} \sim \text{Gamma}(n, \lambda)$, so $\lambda n\bar{Y} \sim \text{Gamma}(n, 1)$. Therefore, $\lambda n\bar{Y}$ is a pivot and we can create a 95% CI based on Gamma quantiles:

$$P(Q_{\text{Gamma}(n,1)}(0.025) \leq \lambda n\bar{Y} \leq Q_{\text{Gamma}(n,1)}(0.975)) = 0.975 - 0.025 = 0.95.$$

Rearranging,

$$P\left(\frac{Q_{\text{Gamma}(n,1)}(0.025)}{n\bar{Y}} \leq \lambda \leq \frac{Q_{\text{Gamma}(n,1)}(0.975)}{n\bar{Y}}\right) = 0.95,$$

so

$$C(\mathbf{Y}) = \left[\frac{Q_{\text{Gamma}(n,1)}(0.025)}{n\bar{Y}}, \frac{Q_{\text{Gamma}(n,1)}(0.975)}{n\bar{Y}}\right]$$

is a 95% CI for λ.

5.5 Recap

As the name suggests, an interval estimate is an interval of plausible values for a parameter rather than a single guess

$$C(\mathbf{y}) = [L(\mathbf{y}), U(\mathbf{y})]$$

From a frequentist standpoint, the main task of interval estimation is to construct a confidence interval with some prescribed confidence level $1 - \alpha$ so that

$$P_{\mathbf{Y};\theta}(\theta \in C(\mathbf{Y})) = 1 - \alpha.$$

The main ideas and notation for confidence intervals are listed in Table 5.2.

Three main strategies for constructing confidence intervals are *pivots*, *asymptotic pivots*, and *resampling*. For asymptotics, we typically use the CLT, the delta method, and the theorems from Chapters

Formula or idea	Description or name
$C(\mathbf{y}) = [L(\mathbf{y}), U(\mathbf{y})]$	interval estimate
$C(\mathbf{Y}) = [L(\mathbf{Y}), U(\mathbf{Y})]$	interval estimator
$P(\theta \in C(\mathbf{Y}))$	coverage probability of interval $C(\mathbf{Y})$
If $P(\theta \in C(\mathbf{Y})) = 1 - \alpha$	then $C(\mathbf{Y})$ is $1 - \alpha$ Confidence Interval (CI) ($\mathbf{Y}$ random, θ fixed)
If $\lim_{n\to\infty} P(\theta \in C(\mathbf{Y})) = 1 - \alpha$	then $C(\mathbf{Y})$ is approximate $1 - \alpha$ Confidence Interval
How to build $C(\mathbf{Y})$?	pivot, asymptotic pivot or resampling
What is a pivot?	quantity with known distribution (parameter free)
What is an asymptotic pivot?	quantity with known asymptotic distribution (parameter free)
$\hat{\theta} \pm Q_{\mathcal{N}(0,1)}(1 - \alpha/2) \times \hat{SE}(\hat{\theta})$	common approximate $1 - \alpha$ Confidence Interval
$\hat{\theta} \pm 1.96 \times \hat{SE}(\hat{\theta})$	common approximate 95% Confidence Interval
$P(\theta \in C(\mathbf{y})\|\mathbf{Y} = \mathbf{y})$	Bayesian credible interval ($\mathbf{y}$ fixed, θ random)

Table 5.2: Main ideas and notation in Chapter 5.

3 and 4 about the asymptotic distributions of the MoM estimator and the MLE. For pivots we use a lot of results about probability distributions and their connections with each other. Resampling will be discussed in detail later in the course.

From a Bayesian standpoint, the main task of interval estimation is to construct a *credible interval*

$$P(\theta \in C(\mathbf{Y})|\mathbf{Y} = \mathbf{y})$$

with some prescribed posterior probability $1 - \alpha$; Bayesian methods will be considered in detail later in the course.

Confidence intervals are often misinterpreted and often confused with credible intervals. From a frequentist perspective, the interval is random and the estimand is fixed, and the correct interpretation is in terms of imagining that the experiment is repeated many times, with a new interval computed for each replication. Great care is needed both with the concepts and with the terminology.

5.6 R

We have already introduced enough R for basic computations and simulations involving confidence intervals. So in this section we take the opportunity to discuss an influential graphics package in R, ggplot2, as well as to provide code for the examples given above.

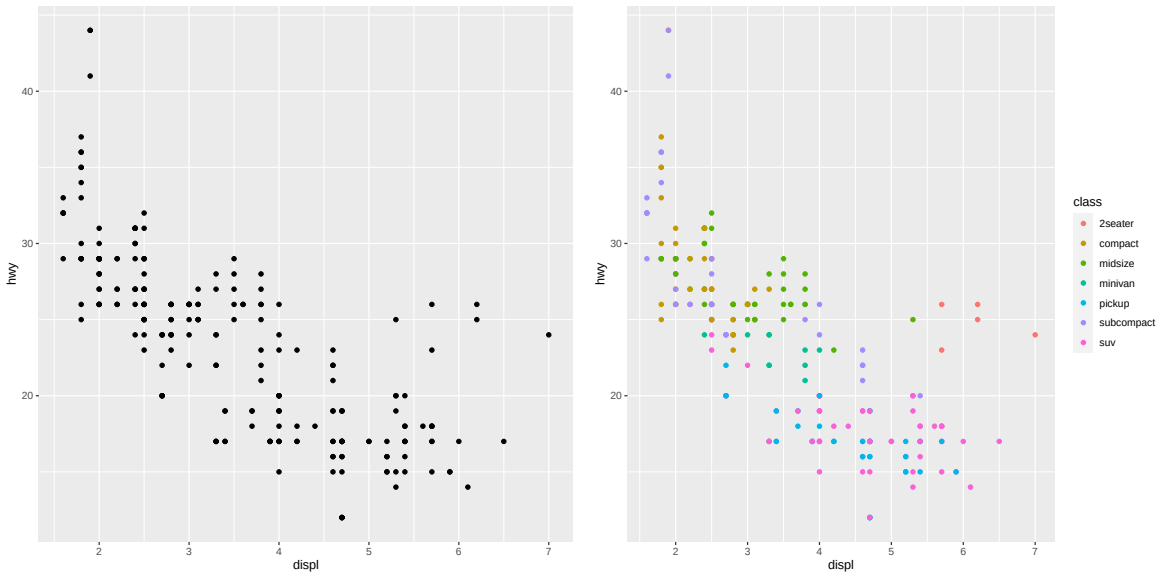


Figure 5.4: Left-hand side: plot of miles per gallon on a highway (`hwy`) against engine size (`displ`). Right-hand side: same as left-hand side, except using color to indicate class of car.

5.6.1 ggplot2: flexible plotting

`ggplot2` is a R graphics package with various appealing features that are not provided by base `R`. It was authored by Hadley Wickham in the 2000s, and is an implementation of Leland Wilkinson (1999)'s *The Grammar of Graphics*. There is an enormous amount of online documentation for `ggplot2`. A nice introduction is available at `https://r4ds.had.co.nz/data-visualisation.html`.

There are, of course, also many built-in graphics capabilities in `R`, but `ggplot2` gives an entirely different way of conceptualizing and implementing graphics, and facilitates the creation of many useful statistical plots.

We can install and load `ggplot2` as follows:

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

The `ggplot2` package has a data frame called `mpg`, which has data on the performance of various car models.

```
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy))
```

`help("mpg")` documents the `mpg` data. The variable `displ` measures the engine displacement (size), in liters, and `hwy` is miles per gallon on a highway. Again the plots can be saved to a PDF file by using `pdf("filename.pdf"); dev.off()` as the start and stop of the graphing.

The `ggplot(data = mpg)` says we are using the data frame `mpg` and want to do some plotting —

alone it produces nothing interesting to see. The `+` says there is more. `geom_point` says layer on top points. `aes` stands for *aesthetic*, and maps substantive variables `displ` to `x` and `hwy` to `y`. The result is on the left hand side of Figure 5.4.

More fun can be found by adding color to the aesthetics of the picture:

```
ggplot(data = mpg) +
    geom_point(mapping = aes(x = displ, y = hwy, color=class))
```

where in the `mpg` dataframe `class` is a type of car. This is shown on the right-hand side of Figure 5.4 The added color allows us to see a third dimension in the picture. Other possible aesthetics include the plotting symbol (shape), size, and opacity (alpha). If you want to set all the points to have a single color, then the color is set outside the aesthetic

```
ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy),
    color="blue"))
```

The same holds for size, symbol, etc.

To finish this extremely short primer for `ggplot2`, let us see some of the payoff of using this approach. Figure 5.5 shows the results from

```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy)) +
  facet_wrap(~ class, nrow = 2)
```

which carries out the same plot as before but adds *facets* — plots of subsets of the data, here subsetting by the car `class` category. We can see the SUV, minivan, and pickup cars have a lower `hwy` for the same engine size. This is quite an abstract way of writing the graphs, which takes some investment into the `ggplot2` language, but once there, there is great simplicity in it.

Finally, if like us you are not a fan of the shaded backgrounds in the plots, you can add `theme_bw()` to remove them:

```
  ggplot(data = mpg) + geom_point(mapping = aes(x = displ, y = hwy)) +
    theme_bw()
```

5.6.2 Code for various examples in this chapter

The first piece of code computes the coverage probability, which can be far from the nominal coverage probability, for a case where the data is binary.

Code which produces Figure 5.2

```
set.seed(111)
n <- 30
```

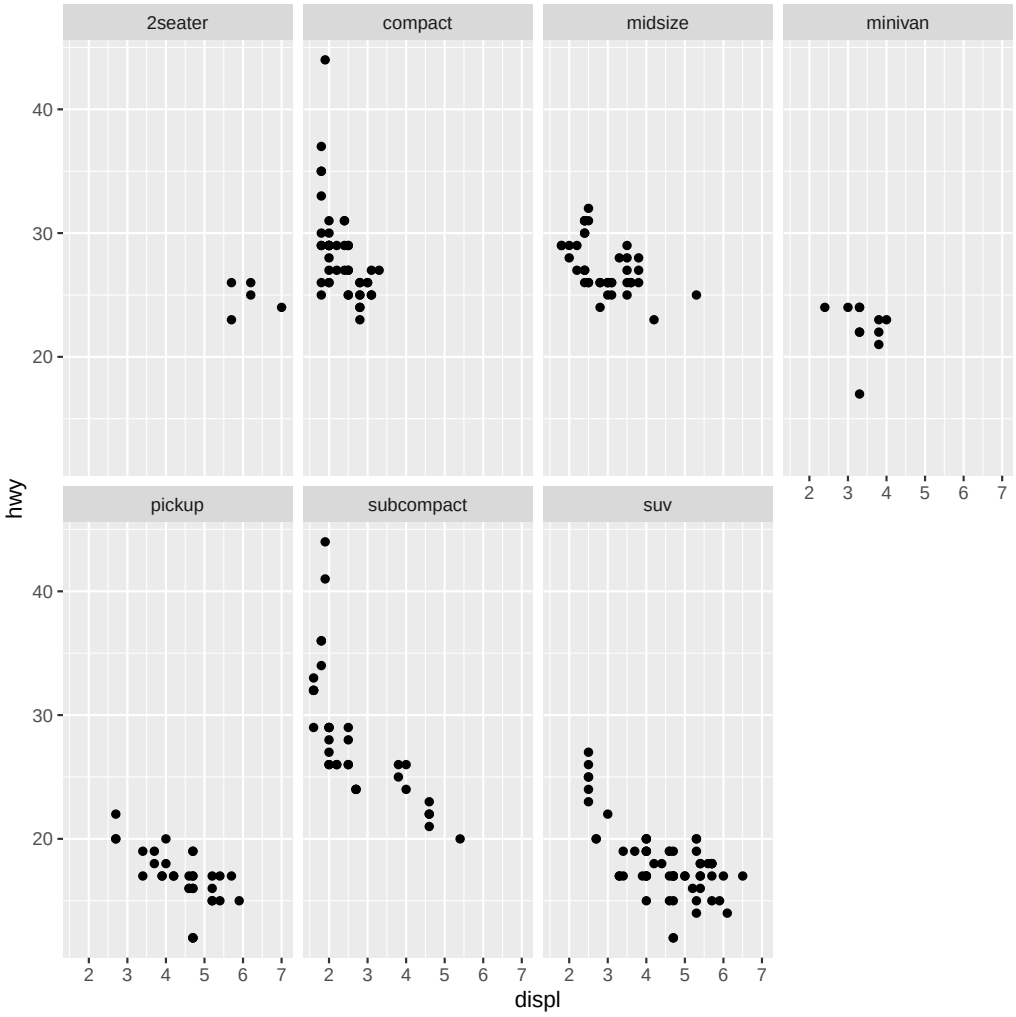


Figure 5.5: Plot of miles per gallon on a highway (`hwy`) against engine size (`displ`) for different classes of car.

```
p <- 0.001*(1:999)
nreps <- 10^6
coverage.prob <- rep(NA,999)
for(i in 1:999) {
  y <- rbinom(nreps, n, p[i])
  p.hat <- y/n
  cover <- (p.hat - 1.96*sqrt(p.hat*(1-p.hat)/n) < p[i]) &
         (p[i] < p.hat + 1.96*sqrt(p.hat*(1-p.hat)/n))
  coverage.prob[i] <- sum(cover)/nreps
}
plot(p,coverage.prob, col = "blue", type = "l",ylab = "coverage
     probability", main = "asymptotic Binomial CI performance")
abline(h = 0.95, col = "red")
```

The second piece of code, computes the terms in Example 5.3.6, which reports confidence intervals for descriptive statistics for the duration of birth data. The results are given in Table 5.1.

Code which produces Table 5.1

```
set.seed (111)
load("births.rda") # load data using R format

Mresults = matrix(nrow=8,ncol=3) # build storage for results
colnames(Mresults) = c("est","2.5% CI","97.5% CI")
rownames(Mresults) = c("mean","sd","0.05Q","0.5Q","0.95Q","F(5)",
            "F(10)","F(15)")

n = length(births$time); # estimates
mean(births$time); sd(births$time)
quantile(births$time,probs = c(0.05,0.5,0.95))
mean(births$time <= 5.0) #ECDF at y=5
mean(births$time <= 10.0) #ECDF at y=10
mean(births$time <= 15.0) #ECDF at y=15

# CI for sample mean
Mresults[1,] = c(mean(births$time),
    mean(births$time)-1.96*(sd(births$time)/sqrt(n)),
    mean(births$time)+1.96*(sd(births$time)/sqrt(n)))

# CI for sample sd
K = mean((births$time-mean(births$time))^4)
var1 = var(births$time)
Mresults[2,] = sqrt(c(var1,var1-1.96*sqrt((K-(var1^2))/n),
            var1+1.96*sqrt((K-(var1^2))/n)))

# CI for sample quantiles
pvec = c(0.05,0.5,0.95)
```

```
for (i in 1:3){
  p = pvec[i]
  x = quantile(births$time,probs=p)
  h = 1.06*(n^-0.2)*sd(births$time); # Silverman's bandwidth rule
  var1 = p*(1.0-p)/((density(births$time,bw=h,kernel="rectangular",
    from=x,to=x,n=1)$y)^2)
  Mresults[2+i,] = c(x,x-1.96*sqrt(var1/n),x+1.96*sqrt(var1/n))
}

# CI for ECDF
xvec =c(5.0,10.0,15.0)

for (i in 1:3){
  xc = xvec[i]
  yInd = births$time <= xc
  Mresults[5+i,] = c(mean(yInd),
    mean(yInd)-1.96*sqrt(mean(yInd)*(1.0-mean(yInd)))/sqrt(n),
        mean(yInd)+1.96*sqrt(mean(yInd)*(1.0-mean(yInd)))/sqrt(n))
}

print(Mresults)
```

Chapter 6

Regression

6.1 Regression

Regression is a widely used term in statistics and applied science. Here we introduce it using two different meanings, which unfortunately often get conflated elsewhere, resulting in various confusions. For both meanings, the data for each individual comes as a pair $(\mathbf{X}, Y)$, where $\mathbf{X}$ may be a scalar or a vector of variables and Y is a scalar variable. We will then study the following two frameworks.

- *Conditional* learning of
$$\mu(\mathbf{x}) = \mathrm{E}[Y|\mathbf{X} = \mathbf{x}].$$
This is called *predictive regression*, as we are using $\mathbf{X}$ to predict Y. In this framework, we focus on the conditional distribution of Y given $\mathbf{X}$.

- *Unconditional* learning of a *linear* function $\mu_L(\mathbf{x})$, so that the magnitude of $Y - \mu_L(\mathbf{X})$ is small on average over the random pair $(\mathbf{X}, Y)$. This is called *linear projection* or *descriptive regression*. In this framework, we focus on summarizing the joint behavior of the pairs $(\mathbf{X}, Y)$.

Associated with both frameworks are various models and estimators. The most famous models in the regression setting, both of which we will delve into in this chapter, are *linear regression* and *logistic regression*. Estimators for regression can be derived using the MLE, MoM, and Bayesian principles, as we have been doing in earlier chapters, and also through a new principle called *least squares*.

6.2 Predictive regression

A central goal in statistics and machine learning is *prediction*: what value of an *outcome variable* Y is likely, given some *predictor variables* $\mathbf{X} = (X_1, \ldots, X_K)$ that have been observed to equal $\mathbf{x} = (x_1, \ldots, x_K)$?

Example 6.2.1. Let $\mathbf{X}$ be the quantifiable characteristics or record of an application to Harvard College, e.g., grades, alumni interviewer assessment, statement, interests, gender, race, etc. Let Y be binary, taking the value 1 if the applicant gets an offer from Harvard College and 0 otherwise. A *predictive regression* gives the probability that $Y = 1$ given the individual's record $\mathbf{x}$. Such modeling formed part of the legal case Students for Fair Admissions, Inc. v. President and Fellows of Harvard College (Harvard Corporation) in 2019.

Many statistical models and methods focus on $\mathrm{E}[Y|\mathbf{X} = \mathbf{x}]$, the expected outcome given the predictors. Conditional expectation is also a central concept in probability theory (see Chapter 9 of the Stat 110 book). In statistics, the goal will be to learn this function from data.

Definition 6.2.2 (Predictive regression)**.** The task of estimating the conditional expectation

$$\mu(\mathbf{x}) = \mathrm{E}[Y|\mathbf{X} = \mathbf{x}]$$

is called *predictive regression*. The variable Y is called the *outcome variable*, while the $\mathbf{X}$ variables are called *predictors*, *covariates*, *regressors*, or *features*. Running a statistical method in this setting is sometimes called *regressing* Y on $\mathbf{X}$.

☣ **6.2.3.** Sometimes predictors are called "independent variables" and the outcomes is called the "dependent variable", but this is confusing terminology since independence versus dependence is already a central concept in statistics and, worse yet, we also have the notion of linear independence versus linear dependence from linear algebra. Sometimes the outcome variable is called the "response", though this could sound like it is saying that Y responds to a change in $\mathbf{X}$ (as in cause and effect), whereas we are currently discussing *prediction*, not *causation* (the latter is discussed in depth in Chapter 11).

In predictive regression models, it is essential to be clear not only about what is being assumed about the conditional expectation, but also about the conditional variance of Y given $\mathbf{X}$.

Definition 6.2.4 (Homoskedasticity and heteroskedasticity)**.** Assume that

$$\sigma^2(\mathbf{x}) = \mathrm{Var}(Y|\mathbf{X} = \mathbf{x})$$

exists. If $\sigma^2(\mathbf{x})$ does not vary with $\mathbf{x}$, then the predictive regression is called *homoskedastic*. Otherwise it is *heteroskedastic*.

Naturally, when we make predictions we would like notation and terminology for how close the prediction comes to the actual outcome.

Definition 6.2.5 (Regression error)**.** For predictive regression, the *regression error* is the random variable

$$U(\mathbf{x}) = Y - \mathrm{E}[Y|\mathbf{X} = \mathbf{x}].$$

So the regression error is the difference between the random outcome and the predicted outcome, given that the predictors are $\mathbf{x}$.

Broadly speaking, a lot of statistical work is about trying to separate signal from noise. As an immediate consequence of the above definitions, we have the following result, which decomposes Y into *signal* (the predicted part $\mu(\mathbf{x})$) and *noise* (the random error $U(\mathbf{x})$).

Theorem 6.2.6 (Signal-noise decomposition)**.** *With notation as above,*

$$Y = \mu(\mathbf{x}) + U(\mathbf{x}).$$

Next, we will show that the regression error has mean 0 (both conditionally and unconditionally) and is uncorrelated with the predictors.

Theorem 6.2.7 (Regression error: mean 0, uncorrelated with predictors)**.** *For a random pair* $(\mathbf{X}, Y)$*, write the regression error (for* $\mathbf{X}$ *random) as*

$$U(\mathbf{X}) = Y - \mathrm{E}[Y|\mathbf{X}].$$

Then

$$\mathrm{E}[U(\mathbf{X})|\mathbf{X} = \mathbf{x}] = 0,$$

$$\mathrm{E}[U(\mathbf{X})] = 0,$$

and for each predictor variable X_j*,*

$$\mathrm{Cov}(U(\mathbf{X}), X_j) = 0.$$

Proof. By construction,

$$\mathrm{E}[U(\mathbf{X})|\mathbf{X} = \mathbf{x}] = \mathrm{E}[U(\mathbf{x})|\mathbf{X} = \mathbf{x}] = \mathrm{E}[Y|\mathbf{X} = \mathbf{x}] - \mathrm{E}[Y|\mathbf{X} = \mathbf{x}] = 0,$$

for all $\mathbf{x}$. By Adam's law, we also have $\mathrm{E}[U(\mathbf{X})] = \mathbf{0}$ unconditionally. Again by Adam's law, as long as the covariance exists, for each predictor variable X_j we have

$$\mathrm{Cov}(U(\mathbf{X}), X_j) = \mathrm{E}[X_j U(\mathbf{X})] = \mathrm{E}[\mathrm{E}[X_j U(\mathbf{X})|\mathbf{X}]] = \mathrm{E}[X_j \mathrm{E}[U(\mathbf{X})|\mathbf{X}]] = \mathrm{E}[0 X_j] = 0.$$

∎

It is also important to consider the variance of Y, both conditionally and unconditionally. Recall $\sigma^2(\mathbf{x}) = \mathrm{Var}(Y|\mathbf{X} = \mathbf{x})$. Then $\sigma^2(\mathbf{x}) = \mathrm{Var}(U|\mathbf{X} = \mathbf{x})$, so by Eve's law,

$$\mathrm{Var}(U) = \mathrm{E}[\mathrm{Var}(U|\mathbf{X} = \mathbf{x})] + \mathrm{Var}(\mathrm{E}[U|\mathbf{X} = \mathbf{x}]) = \mathrm{E}[\sigma^2(\mathbf{X})],$$

if this variance exists. Likewise, Eve's law says

$$\mathrm{Var}(Y) = \mathrm{E}[\sigma^2(\mathbf{X})] + \mathrm{Var}(\mu(\mathbf{X})),$$

so a summary measure of the unconditional effectiveness of the prediction is

$$R^2 = \frac{\mathrm{Var}(\mu(\mathbf{X}))}{\mathrm{Var}(Y)} = 1 - \frac{\mathrm{Var}(U)}{\mathrm{Var}(Y)},$$

the share of the variation of Y contributed by the variation in the prediction.

☣ **6.2.8.** The measure $\mathrm{Var}(\mu(\mathbf{X}))$ is also called "explained variance", but there may not be much of an explanation; it may be clearer to say that R^2 is the share of the variation of Y that is *accounted for* by the variation in the prediction.

6.2.1 Linear regression

The most widely used version of predictive regression is the following model, which is linear in the parameters.

Definition 6.2.9 (Linear regression model)**.**

$$\mathrm{E}(Y|\mathbf{X} = \mathbf{x}, \boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \cdots + \theta_K x_K,$$

where $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_K)^\intercal$. This is called a *linear regression* model and the elements of $\boldsymbol{\theta}$ are called the *regression coefficients*. Typically, θ_0 is called the *intercept* and $\theta_1, \ldots, \theta_K$ are called *slopes*.

This interpretation of the regression coefficients is illustrated on the left-hand side of Figure 6.1.

☣ **6.2.10.** If, for example,

$$\mathrm{E}(Y|\mathbf{X} = \mathbf{x}, \boldsymbol{\theta}) = \theta_0 + \theta_1 x I_{x \geq 0} + \theta_2 x I_{x < 0},$$

then this is still called a linear regression, as $\mathrm{E}[Y|\mathbf{X} = \mathbf{x}, \boldsymbol{\theta}]$ is linear in the *parameters*, although it is non-linear in predictors — with different coefficients on the predictor according to the sign of the predictor. This is shown on the right-hand side of Figure 6.1. A widely used version of this comes up when we have more than one predictor. The model

$$\mathrm{E}(Y|\mathbf{X} = \mathbf{x}, \boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1 x_2,$$

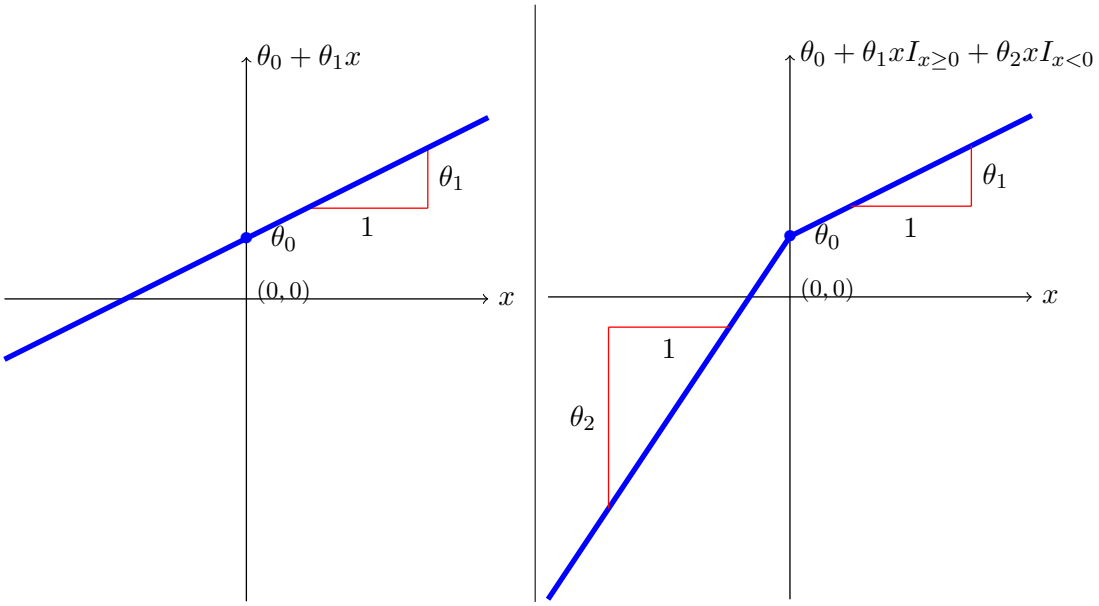


Figure 6.1: Two examples of linear in parameters predictive regression models for $\mathrm{E}[Y|\mathbf{X} = \mathbf{x}, \boldsymbol{\theta}]$. The red triangles show the gradients with respect to x.

is linear in parameters, so is still a linear regression. This model allows the two predictors x_1 and x_2 to *interact*, by including the *interaction term* $\theta_3 x_1 x_2$. Likewise, if x is a predictor in a linear regression, we can also include nonlinear functions of x (e.g., x^2) as predictors without having to leave the linear regression framework. So linear regression models are much more general than they would seem if one is not paying careful attention to the fact that they are required to be linear in the parameters, not in the predictors.

☣ **6.2.11** (Extrapolation)**.** Linear models are assumed to hold for all $\mathbf{x}$, which greatly simplifies modeling, analysis, and prediction. However, this means linear models can make predictions based on values of $\mathbf{x}$ far away from any we have seen in the data. Reaching outside the range of the observed predictors is called *extrapolation* in statistics. This is potentially useful, but is also dangerous, as such predictions rely almost entirely on the assumption of linearity, not so much on the data themselves. As a rule of thumb, it is much safer to *interpolate* between existing data points than to *extrapolate* to regions beyond where we have much data. Figure 6.2 shows a comical illustration of this by Randall Munroe of `xkcd.com`.

6.2.2 Logistic regression

Enormous numbers of datasets we see in research, public policy, and industry focus on a *binary* (0 or 1) outcome Y. Often we think of 1 as "success" and 0 as "failure", though we can define these

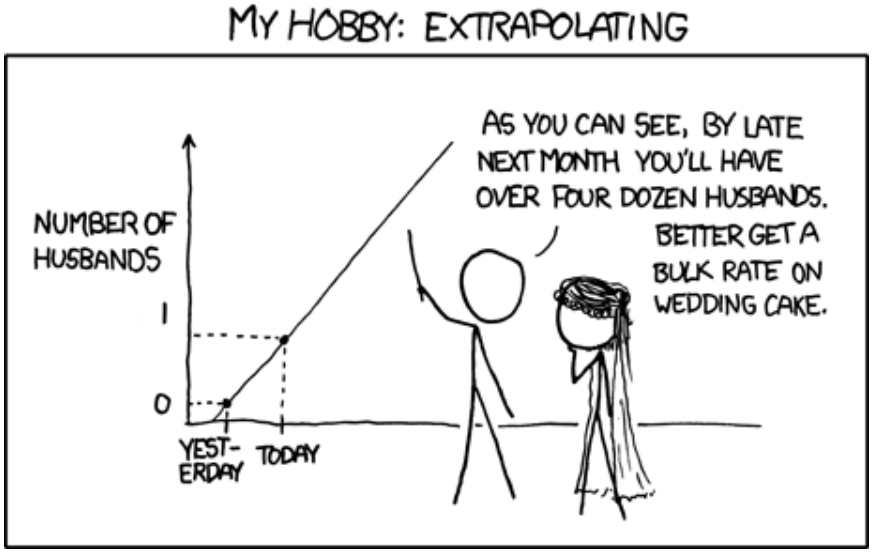


Figure 6.2: This comic by Randall Munroe, entitled "Extrapolating", is from `xkcd.com`, produced here under "Creative Commons Attribution - NonCommercial 2.5 License", detailed at `https://xkcd.com/license.html`.

however we want (as long as each outcome is success or failure, but not both). So "success" is not necessarily good and "failure" is not necessarily bad.

Example 6.2.12. For the technology company "iBiting", Y is 1 if an individual on one of iBiting's major webpage clicks through a sponsored link from xShop and 0 otherwise. For a Pharma company "aDrug" Y is 1 if a patient has 6 months of remissions from a particular cancer and 0 if not. For the Governor of Massachusetts, Y is 1 if a resident of Cambridge does not use a car to go to work and 0 if that person does. For a development economist at J-PAL, Y is 1 if a kid in a very poor country keeps on going to school after the age of 12 and 0 if not.

If our statistical focus is predictive, then again look at the regression

$$\mathrm{E}\left(Y|\mathbf{X}=\mathbf{x}\right)=P(Y=1|\mathbf{X}=\mathbf{x})=\mu(\mathbf{x}).$$

Hence the regression tells us that

$$P(Y=y|\mathbf{X}=\mathbf{x})=\mu(\mathbf{x})^{y}\left\{1-\mu(\mathbf{x})\right\}^{1-y}=\begin{cases}\mu(\mathbf{x}), & y=1,\\ 1-\mu(\mathbf{x}), & y=0.\end{cases}$$

The *odds*,

$$\frac{\mu(\mathbf{x})}{1-\mu(\mathbf{x})},$$

are monotonically increasing in $\mu(\mathbf{x})$. The odds are the probability of a success divided by the probability of a failure. If the odds are 3, for example, then the probability of a success is 3 times higher

than the probability of a failure: i.e., $0.75/(1-0.75) = 3$. Likewise, odds of 4 correspond to probability of 0.8.

Remark 2. The term "odds" comes from betting. For example, 3 to 1 odds in favor (or 1 to 3 odds against) is the same as a probability of 3 in 4, which corresponds to $\mu(x) = 0.75$.

Often in statistics the *log-odds*

$$\lambda(\mathbf{x}) = \log\left\{\frac{\mu(\mathbf{x})}{1-\mu(\mathbf{x})}\right\} = \log\mu(\mathbf{x}) - \log\left\{1-\mu(\mathbf{x})\right\},$$

appears, e.g.,

$$\log P(Y = y|\mathbf{X} = \mathbf{x}) = \log\{1-\mu(\mathbf{x})\} + y\log\left(\frac{\mu(\mathbf{x})}{1-\mu(\mathbf{x})}\right) \tag{6.1}$$

$$= \log\{1+e^{\lambda(\mathbf{x})}\} + y\lambda(\mathbf{x}). \tag{6.2}$$

The log-odds is the log-probability of a success minus the log-probability of a failure.

The log-odds function comes up so often in statistics that it has its own name: the *logit* function.

Definition 6.2.13 (Logit function)**.** The *logit* function is defined by

$$\text{logit}(p) = \log(p/(1-p)),$$

for $0 < p < 1$. The *inverse logit* function, which is also called the *logistic* function, *sigmoid* function, or *expit* function, is the inverse of the logit function:

$$\text{logit}^{-1}(x) = \frac{e^x}{1+e^x},$$

for all $x \in \mathbb{R}$.

Therefore, the probability of success is the logistic transform of the log-odds:

$$\mu(\mathbf{x}) = \frac{e^{\lambda(\mathbf{x})}}{\{1+e^{\lambda(\mathbf{x})}\}},$$

which is shown on the left hand side of Figure 6.3. For each value of $\mathbf{x}$,

$$\frac{\partial\mu(\mathbf{x})}{\partial\lambda(\mathbf{x})} = \mu(\mathbf{x})\left\{1-\mu(\mathbf{x})\right\} = \text{Var}(Y_1|\mathbf{X}_1 = \mathbf{x}).$$

The result on the variance comes from $Y|\mathbf{X} = \mathbf{x}$ being a Bernoulli trial — the conditional mean and conditional variance are locked together. This is shown on the right hand side of Figure 6.3.

The most common binary predictive regression is logistic regression, which statisticians mostly associate with the formalization by Sir David R. Cox in 1958.

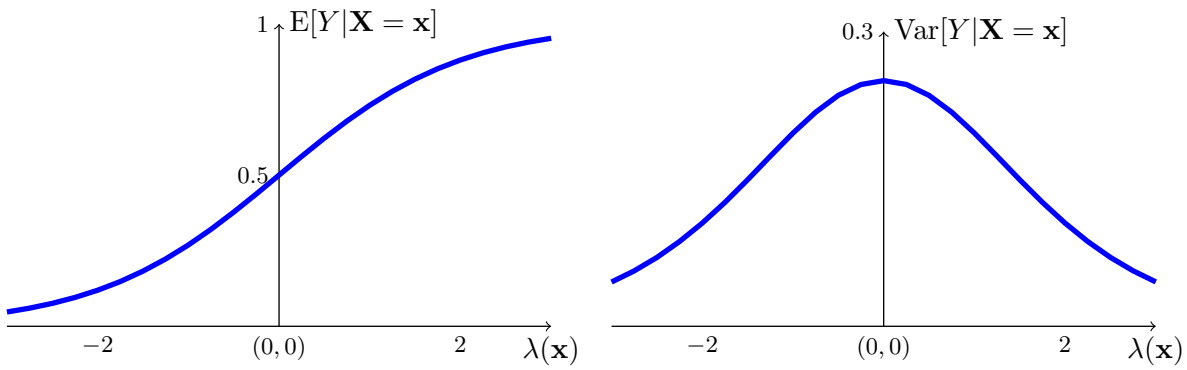


Figure 6.3: LHS draws $\mathrm{E}[Y|\mathbf{X}=\mathbf{x}]$ as a logistic function of the log-odds $\lambda(\mathbf{x})$. RHS draws $\mathrm{Var}[Y|\mathbf{X}=\mathbf{x}]$ as a function of $\lambda(\mathbf{x})$.

Definition 6.2.14 (Logistic regression)**.** The *logistic regression* model assumes that the probability of success, given the predictor variables, is

$$P(Y=1|\mathbf{X}=\mathbf{x},\boldsymbol{\theta})=\mu(\mathbf{x}|\boldsymbol{\theta})=\text{logit}^{-1}\left(\theta_0+\theta_1x_1+...+\theta_Kx_K\right)=\frac{\exp\left(\theta_0+\theta_1x_1+...+\theta_Kx_K\right)}{1+\exp\left(\theta_0+\theta_1x_1+...+\theta_Kx_K\right)}.$$

Logistic regression is one of the most useful methods in modern statistics and machine learning, providing interpretable results while being computationally attractive. It is often competitive with much more complicated methods so serves as a good point of comparison when deciding whether to use alternatives.

Example 6.2.15. As is so often the case, simulating the data helps with understanding the model. Consider the bimodal density

$$f_{X_1}(x)=0.5\varphi(x;3,3^2)+0.5\varphi(x;-3,2^2),$$

where $\varphi(x,\mu,\sigma^2)$ is the $\mathcal{N}(\mu,\sigma^2)$ PDF. So X_1 follows a *mixture* of Normal distributions. Take $n=100$, $K=1$, $\theta_0=0$, and $\theta_1=0.4$. The left-hand side of Figure 6.4 shows a plot of the simulated $(x_1,y_1),...,(x_n,y_n)$, drawn using black dots. The $x_1,...,x_n$ have two clusters, centered at -3 and 3, but having some mass spread between those two points. The $y_1,...,y_n$ are hard to see, as they are only binary and smear together graphically. But they cluster as mostly 0 when x is low and mostly 1 when x is high. The red dots are the corresponding $P(Y_j=1|X=x_j,\theta)$, following the logistic function upwards as x increases.

Statisticians like randomness! They like it so much they often add randomness to the data to learn more; Chapters 10 and 11 will mostly be about this theme. This may sound absurd: how can *adding*

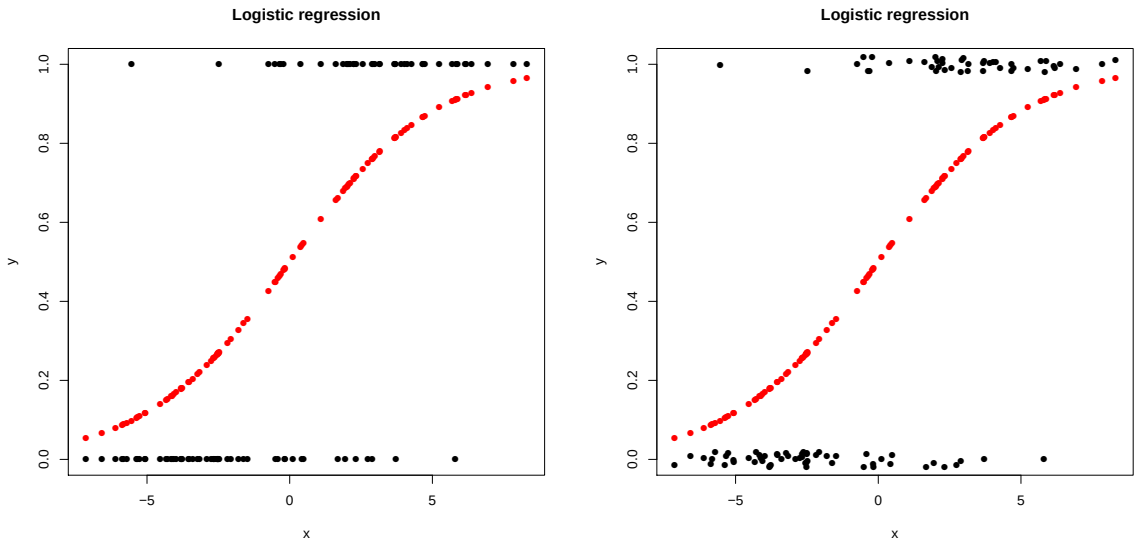


Figure 6.4: Left-hand side: Black dots are data, x_j is the jth predictor, and y_j is the binary outcome. The red dots are the corresponding $P(Y_j = 1|X = x_j, \theta)$, which follows the logistic function. Right-hand side: same as LHS except we jittered the plotted y_j to make them more visible, by adding Unif$(-0.02, 0.02)$ simulated noise to each y_j.

noise to something help? The right-hand side of Figure 6.4 is an example of this. In the left-hand side plot, it is hard to see the individual binary outcomes — they smear. A clearer impression can be gleaned by adding a tiny amount of randomness to each of the $y_1, \ldots, y_n$ in the plot. This is called *jittering* in statistics. This is carried out on the right-hand side. The outcome data are now much clearer. The code for this example is in Section 6.9.

The odds for logistic regression are

$$\begin{aligned}\frac{\mu(\mathbf{x}|\boldsymbol{\theta})}{1-\mu(\mathbf{x}|\boldsymbol{\theta})} &= \frac{\exp\left(\theta_0+\theta_1x_1+\ldots+\theta_Kx_K\right)}{1+\exp\left(\theta_0+\theta_1x_1+\cdots+\theta_Kx_K\right)} \Big/ \frac{1}{1+\exp\left(\theta_0+\theta_1x_1+\cdots+\theta_Kx_K\right)} \\ &= \exp\left(\theta_0+\theta_1x_1+\cdots+\theta_Kx_K\right),\end{aligned}$$

which means the log-odds are linear in parameters

$$\log\left\{\frac{\mu(\mathbf{x}|\boldsymbol{\theta})}{1-\mu(\mathbf{x}|\boldsymbol{\theta})}\right\} = \theta_0+\theta_1x_1+\cdots+\theta_Kx_K = \lambda(\mathbf{x}|\boldsymbol{\theta}),$$

where $\theta_{\mathbf{0}} = \lambda(\mathbf{0}|\boldsymbol{\theta})$, and

$$\theta_j = \frac{\partial\lambda(\mathbf{x}|\boldsymbol{\theta})}{\partial x_j} = \frac{1}{\mu(\mathbf{x}|\theta)\{1-\mu(\mathbf{x}|\theta)\}}\frac{\partial\mu(\mathbf{x}|\boldsymbol{\theta})}{\partial x_j}.$$

Hence logistic regression coefficients are the rate of change not of $\mu(\mathbf{x}|\boldsymbol{\theta}) = \mathrm{E}(Y|X = \mathbf{x})$ but of the log-odds $\lambda(\mathbf{x}|\boldsymbol{\theta})$.

6.3 Statistical models of predictive regression

A common statistical model for predictive regressions based on the pairs of data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ is the parametric predictive distribution of

$$Y_1, \ldots, Y_n | \mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_n = \mathbf{x}_n; \boldsymbol{\theta}.$$

Here we can focus on the conditional distribution of outcome given predictors, without worrying about how the predictions are generated. Let's assume that the outcomes given the predictors are independent and that for the jth outcome only the jth predictors are relevant, which for continuous data means that

$$f(y_1, \ldots, y_n | \mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_n = \mathbf{x}_n; \boldsymbol{\theta}) = \prod_{j=1}^{n} f(y_j | \mathbf{X}_j = \mathbf{x}_j; \boldsymbol{\theta}).$$

☣ **6.3.1.** If the *pairs* $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ are independent, then the above conditional independence statement holds. Note that we do not require an independence of assumption *within* the predictors $\mathbf{X}_j$ for individual j. For example, $\mathbf{X}_j$ may have been generated by a time series.

The log-likelihood for this conditional density is

$$\begin{aligned} \log L(\boldsymbol{\theta}) &= \log f(y_1, ..., y_n | \mathbf{X}_1 = \mathbf{x}_1, ..., \mathbf{X}_n = \mathbf{x}_n, \boldsymbol{\theta}) \\ &= \sum_{j=1}^{n} \log f(y_j | \mathbf{X}_j = \mathbf{x}_j, \boldsymbol{\theta}), \end{aligned}$$

which is convenient since each term in the sum represents the contribution of each pair. So the score and Fisher information in the sample are also sums:

$$s(\boldsymbol{\theta}) = \frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{j=1}^{n} s_j(\boldsymbol{\theta}), \quad s_j(\boldsymbol{\theta}) = \frac{\partial \log f(y_j | \mathbf{X}_j = \mathbf{x}_j, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$

and

$$\begin{aligned} \mathcal{I}_{\mathbf{Y}|(\mathbf{X}=\mathbf{x})}(\boldsymbol{\theta}^*) &= \mathrm{Var}\left\{ s(\boldsymbol{\theta}^*, \mathbf{Y}) | \mathbf{X}_1 = \mathbf{x}_1, ..., \mathbf{X}_n = \mathbf{x}_n \right\} \\ &= \sum_{j=1}^{n} \mathcal{I}_{Y_j|\mathbf{X}_j=\mathbf{x}_j}(\boldsymbol{\theta}^*), \quad \mathcal{I}_{Y_j|\mathbf{X}_j=\mathbf{x}_j}(\boldsymbol{\theta}^*) = \mathrm{Var}\left\{ s_j(\boldsymbol{\theta}^*; Y_j) | \mathbf{X}_j = \mathbf{x}_j \right\}, \end{aligned}$$

where $\mathcal{I}_{Y_j|\mathbf{X}_j=\mathbf{x}_j}(\boldsymbol{\theta}^*)$ is the Fisher information for the jth data pair.

The maximum likelihood estimator of $\boldsymbol{\theta}$ for predictive regression is

$$\widehat{\boldsymbol{\theta}} = \underset{\theta \in \Theta}{\arg\max} \sum_{j=1}^{n} \log f(Y_j | \mathbf{X}_j = \mathbf{x}_j, \boldsymbol{\theta}).$$

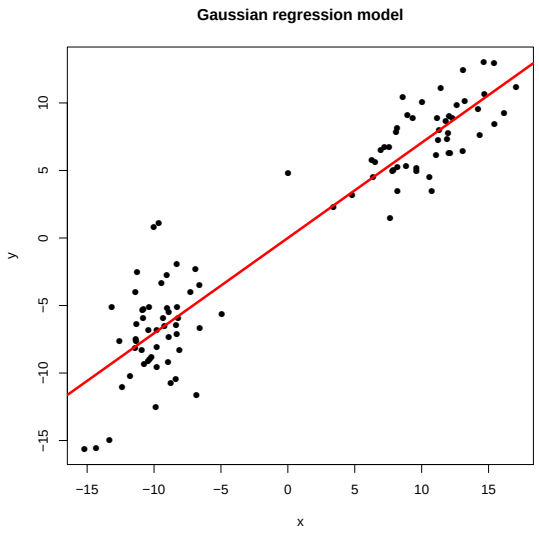


Figure 6.5: Gaussian regression simulated data, with highly bimodal, zero mean X. Red line is $\hat{\theta}x$, where $\hat{\theta} \approx 0.705$ is the MLE. The MLE has $\text{SE}(\hat{\theta}) \approx 0.029$.

6.3.1 Gaussian linear regression without intercept

The *Gaussian linear regression* model assumes the scatter around $\text{E}(Y|\mathbf{X} = \mathbf{x}, \boldsymbol{\theta})$ is Gaussian (Normal). To focus on statistical ideas rather than algebraic clutter, for now we will work with a single predictor ($K = 1$) and no intercept. Then θ will be a scalar regression coefficient, representing a slope. This type of model might make sense if we knew that $\text{E}(Y_1) = \text{E}(X_1) = 0$; we will discuss some more general cases later. Then the Gaussian regression model becomes

$$Y_j|(X_1 = x_1, ..., X_n = x_n), \theta \overset{\text{ind.}}{\sim} \mathcal{N}(\theta x_j, \sigma^2).$$

Again to focus on the core ideas, let's assume that σ^2 is known and constant (so we are assuming homoskedasticity).

We have already seen in Example 4.4.1 a detailed example of a Gaussian linear regression model, where we focused on statistical model building opportunities offered by the ability to choose to condition. The cross-plot of the outcomes and predictors from this example is repeated here in Figure 6.5. The attractive feature of the conditional approach is that even though the predictors are highly bimodal, this complication is not particularly relevant for the predictive regression.

The corresponding likelihood, score, and Fisher information in the sample are

$$\log L(\theta) = -\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - \theta x_j)^2, \quad s(\theta) = \frac{1}{\sigma^2}\sum_{j=1}^{n} x_j(y_j - \theta x_j), \quad \mathcal{I}(\theta) = \frac{1}{\sigma^2}\sum_{j=1}^{n} x_j^2.$$

The maximum likelihood estimate is

$$\hat{\theta} = \Big(\sum_{j=1}^{n} x_j^2\Big)^{-1}\Big(\sum_{j=1}^{n} x_j y_j\Big).$$

The MLE $\hat{\theta}$ is a celebrated estimator in statistics. Here we have derived it as a MLE for the Gaussian linear regression model. We will see shortly that $\hat{\theta}$ is also a method of moments estimator, even without the Gaussian assumption. Most famously it is the least squares estimator, based on the least squares principle — again this will be discussed shortly. We call $\hat{\theta}$ the *least squares* estimator.

Definition 6.3.2 (Residuals)**.** Once we have estimated θ with $\hat{\theta}$, it is natural to predict Y_j with $\hat{\theta}x_j$. The value $\hat{\theta}x_j$ is called the *fitted value* or *predicted value* of Y_j. The difference between the actual value of y_j and the predicted value is called the *residual*, and denoted by

$$\hat{U}_j = y_j - x_j\hat{\theta}.$$

☣ **6.3.3.** Often people mix up residuals with errors. The residuals are *observable*: they are statistics that can be computed from the data. In contrast, the errors are *unobservable* if, as is almost always the case in practice, the regression coefficients are unknown. Note that the residual and the error are related by

$$\hat{U}_j = U_j - x_j(\hat{\theta} - \theta) = U_j - x_j\frac{\sum_{i=1}^{n} x_i U_i}{\sum_{i=1}^{n} x_i^2},$$

A useful geometric property of the residuals is that they are orthogonal to the predictor variables.

Theorem 6.3.4 (Residuals are orthogonal to predictors)**.** *With notation as above, we have*

$$\sum_{j=1}^{n} x_j\hat{U}_j = 0.$$

Proof. Note that

$$\sum_{j=1}^{n} x_j^2\hat{\theta} = \hat{\theta}\sum_{j=1}^{n} x_j^2 = \sum_{j=1}^{n} x_j y_j,$$

so

$$\sum_{j=1}^{n} x_j\hat{U}_j = \sum_{j=1}^{n}(x_j y_j - x_j^2\hat{\theta}) = \sum_{j=1}^{n} x_j y_j - \sum_{j=1}^{n} x_j^2\hat{\theta} = 0.$$

∎

For prediction problems it is good practice to plot $\hat{U}_j$ against x_j for $j = 1, \ldots, n$, to help spot anomalies, violations of assumptions, and missing elements in the prediction, e.g., non-linear terms in x_j that should have been included. Such *residual plots* become particularly helpful with multiple predictors.

Properties of $\hat{\theta}$

Some main properties of $\hat{\theta}$ are given in Theorem 6.3.5. So that the properties can be used later for the MoM estimator and for least squares, they are stated under general conditions, not assuming that the Gaussian linear regression model is true. The properties under the Gaussian model will appear as part (c) of Lemma 6.3.6, stated soon.

Theorem 6.3.5 (Properties of the least squares estimator)**.** *Assume that the pairs* $(X_1, Y_1), ..., (X_n, Y_n)$ *have conditionally independent outcomes. Inference will condition the random variables* $\mathbf{X}$ *at the observed values* $\mathbf{x} = (x_1, ..., x_n)$*. Write* $\mu_j = \mathrm{E}[Y_j|X_j = x_j]$ *and* $\sigma_j^2 = \mathrm{Var}(Y_j|X_j = x_j)$*. Then*

$$\mathrm{E}[\hat{\theta}|\mathbf{X} = \mathbf{x}] = \Big(\sum_{j=1}^{n} x_j^2\Big)^{-1}\Big(\sum_{j=1}^{n} x_j\mu_j\Big), \quad \mathrm{Var}(\hat{\theta}|\mathbf{X} = \mathbf{x}) = \Big(\sum_{j=1}^{n} x_j^2\Big)^{-2}\Big(\sum_{j=1}^{n} x_j^2\sigma_j^2\Big).$$

Proof. Conditioning on the predictors, $\hat{\theta}$ is linear in the outcomes, so (as expectations of sums are sums of expectations)

$$\mathrm{E}[\hat{\theta}|\mathbf{X} = \mathbf{x}] = \Big(\sum_{j=1}^{n} x_j^2\Big)^{-1}\sum_{j=1}^{n} x_j\mathrm{E}[Y_j|\mathbf{X} = \mathbf{x}] = \Big(\sum_{j=1}^{n} x_j^2\Big)^{-1}\Big(\sum_{j=1}^{n} x_j\mu_j\Big).$$

Conditioning on the predictors $\hat{\theta}$ is linear in the conditionally independent outcomes, so

$$\mathrm{Var}[\hat{\theta}|\mathbf{X} = \mathbf{x}] = \Big(\sum_{j=1}^{n} x_j^2\Big)^{-2}\sum_{j=1}^{n} x_j^2\mathrm{Var}[Y_j|\mathbf{X} = \mathbf{x}] = \Big(\sum_{j=1}^{n} x_j^2\Big)^{-2}\Big(\sum_{j=1}^{n} x_j^2\sigma_j^2\Big).$$

∎

The following lemma gives the properties under successively stronger conditions, finishing with the Gaussian linear regression model.

Lemma 6.3.6. (a) If $\mu_j = \theta x_j$, then

$$\mathrm{E}[\hat{\theta}|\mathbf{X} = \mathbf{x}] = \theta,$$

so then the estimator is conditionally unbiased for θ.

(b) Under homoskedasticity, i.e., $\sigma_j^2 = \sigma^2$,

$$\mathrm{Var}[\hat{\theta}|\mathbf{X} = \mathbf{x}] = \sigma^2\Big(\sum_{j=1}^{n} x_j^2\Big)^{-1}.$$

(c) If $Y_j|\mathbf{X} = \mathbf{x} \sim \mathcal{N}(\theta X_j, \sigma^2)$, then $\hat{\theta}$ is conditionally unbiased for θ and conditionally achieves the Cramér-Rao lower bound

$$\mathrm{Var}(\hat{\theta}|\mathbf{X} = \mathbf{x}) = \mathcal{I}(\theta)^{-1}.$$

Furthermore,

$$\hat{\theta}|\mathbf{X} = \mathbf{x} \sim \mathcal{N}\left(\theta, \sigma^2\Big(\sum_{j=1}^{n} x_j^2\Big)^{-1}\right).$$

Proof. Special case of Theorem 6.3.5. ∎

As with the sample mean, even if Gaussianity does not hold then we typically have *approximate* Gaussianity if the sample size is reasonably large:

$$\hat{\theta}|(\mathbf{X} = \mathbf{x}) \dot{\sim} \mathcal{N}(\theta, \mathrm{Var}[\hat{\theta}|(\mathbf{X} = \mathbf{x})]),$$

if n is moderately large. A more difficult problem is that the variance of the MLE is delicate, as it can be inflated by heteroskedasticity.

☣ **6.3.7.** Part (a) of Lemma 6.3.6 implies that conditional unbiasedness of $\hat{\theta}$ only needs $\mathrm{E}[Y_j|(\mathbf{X} = \mathbf{x})] = \theta x_j$, not homoskedasticity. Many researchers are uncomfortable assuming homoskedasticity in computing the standard error of $\hat{\theta}$, but still want to use $\hat{\theta}$. Recall that

$$\mathrm{Var}(\sqrt{n}\hat{\theta}|(\mathbf{X} = \mathbf{x})) = \Big(n^{-1}\sum_{j=1}^{n} x_j^2\Big)^{-2}\Big(n^{-1}\sum_{j=1}^{n} x_j^2\sigma_j^2\Big), \quad \sigma_j^2 = \mathrm{Var}(Y_j|(\mathbf{X} = \mathbf{x})),$$

which can be estimated, for moderately large n, by

$$\Big(n^{-1}\sum_{j=1}^{n} x_j^2\Big)^{-2}\Big(n^{-1}\sum_{j=1}^{n} x_j^2\hat{U}_j^2\Big). \tag{6.3}$$

Why? From (6.3.3), then

$$\hat{U}_j^2 - U_j^2 = -2x_jU_j\frac{\sum_{i=1}^{n} x_iU_i}{\sum_{i=1}^{n} x_i^2} + x_j^2\left(\frac{\sum_{i=1}^{n} x_iU_i}{\sum_{i=1}^{n} x_i^2}\right)^2.$$

Thus, $n^{-1}\sum_{j=1}^{n} x_j^2(\hat{U}_j^2 - \sigma_j^2)$ is

$$\frac{\sum_{j=1}^{n} x_j^2(U_j^2 - \sigma_j^2)}{n} - 2\left(\frac{n^{-1}\sum_{j=1}^{n} x_jU_j}{n^{-1}\sum_{j=1}^{n} x_i^2}\right)\left(\frac{\sum_{j=1}^{n} x_j^3U_j}{n}\right) + \left(\frac{n^{-1}\sum_{j=1}^{n} x_jU_j}{n^{-1}\sum_{j=1}^{n} x_j^2}\right)^2\left(\frac{\sum_{j=1}^{n} x_j^4}{n}\right).$$

This difference will typically converge to zero in probability. Why? Each item in the first average has a zero conditional mean and is conditionally independent. The second and third terms get small as

$$n^{-1}\sum_{j=1}^{n} x_jU_j \xrightarrow{p} 0,$$

$$n^{-1}\sum_{j=1}^{n} x_j^3U_j \xrightarrow{p} 0.$$

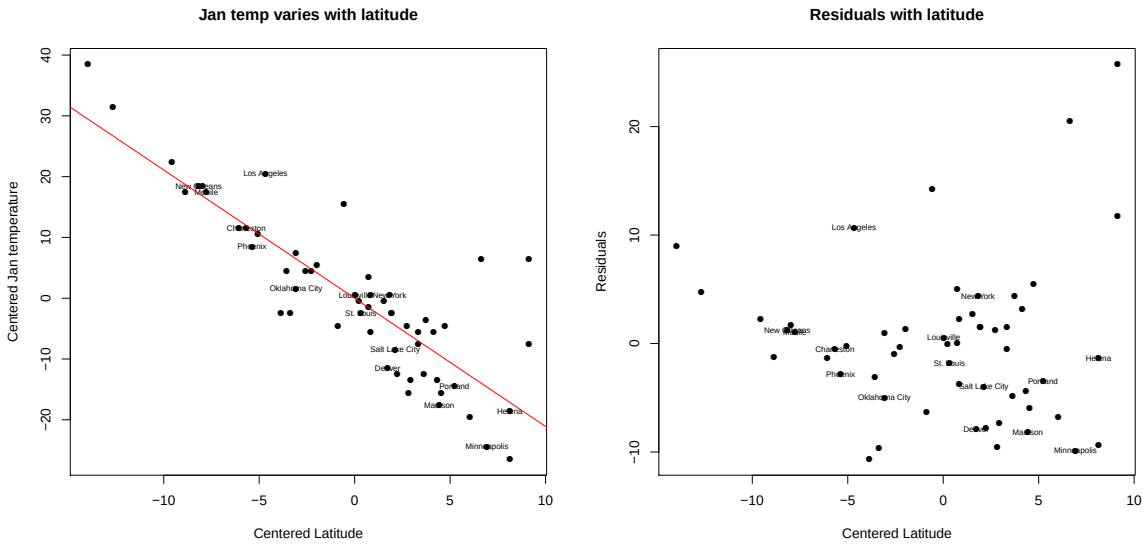


Figure 6.6: Left: Plot of centered January temperature plotted against centered latitude. Also given is the fitted regression line $\hat{\theta}x$, where the ML estimate $\hat{\theta} \approx -2.10$. So the prediction of the January temperature falls by about two degrees with every increase in latitude (moving north). For the MLE $\text{SE}(\hat{\theta}) \approx 0.177$ while the robust version is around 0.226. Right: Regression residuals $\hat{u}_j$ plotted against x_j.

(Proving these convergence results formally would require extensions of the law of large numbers beyond the i.i.d. case, which is beyond the scope of this book.)

The *heteroskedastically robust* variance estimator (6.3) appears quite commonly in applications of regression, particularly in the social and health sciences. The use of the word "robust" is confusing, almost requiring a biohazard within a biohazard. This estimator of the variance is not robust in the same statistical sense that a sample median is robust. For the sample median, robustness refers to being able to handle outliers or heavy-tailed distributions well, whereas the sample mean can be drastically altered by a single outlier. In contrast, robustness for this variance estimator refers to robustness to violations of the homoskedasticity assumption. The heteroskedastically robust variance estimator is *not* robust in the sense of the sample median, since it is based on two sums and sums are very sensitive to outliers.

Other researchers will look at residual plots to assess if heteroskedasticity is a problem for their application and make a decision about this before changing the model to improve fit or deploying a heteroskedastically robust variance to patch up the inference for the MLE.

Example 6.3.8 (How cold is it up North?)**.** Assume that you are in the Northern Hemisphere. If you tell me the latitude of a town (recall as you go north your latitude increases), how cold will it

be in January in that town? That is a prediction problem. The data are from Peixoto (1990), which defines temperature as average daily minimum temperature in Fahrenheit over all the January days from 1931 to 1960, and reports results for 58 U.S. towns. Before we start we *center* all temperatures and latitudes, with the sample mean temperature being about 26.5 (that is cold!) and the sample mean latitude being about 38.9 (around Louisville, Kentucky). In general, for a variable for which we have data $z_1, \ldots, z_n$, the *centered* version of the variable is obtained by subtracting the sample mean: $z_1 - \bar{z}, \ldots, z_n - \bar{z}$. This centers the variable to have a new sample mean of 0, since

$$\frac{1}{n}\sum_{j=1}^{n}(z_j - \bar{z}) = \frac{1}{n}\sum_{j=1}^{n} z_j - \frac{n}{n}\bar{z} = \bar{z} - \bar{z} = 0.$$

The left-hand side of Figure 6.6 shows the raw centered data plus a maximum likelihood estimate fit of the linear regression model. The slope is around 2.1, which suggests if we see two towns one degree latitude apart (to the North) the temperature typically is around 2.1 degrees cooler in the more northern town (1 degree latitude is around 69 miles). The code for this is in Section 6.9.

The right-hand side of Figure 6.6 is a plot of the residuals $\hat{u}_j$ against the predictor x_j. There is no great pattern in this plot, although perhaps the predictive errors are more positive for large values of $|x_j|$. This suggests it might be worthwhile exploring adding $|x_j|$ as another predictor to the linear regression, although the effect of doing this is likely to be very modest.

6.3.2 Gaussian linear regression with intercept

The ideas from the previous subsection can readily be extended to the case where there is an intercept. Let's work out what the maximum likelihood estimate is in this context. In a starred section we further generalize to the case where there are additional predictors. For now, let's consider an intercept θ_0 and a slope θ_1, so the Gaussian linear model becomes

$$Y|(X = x, \theta_0, \theta_1) \sim \mathcal{N}(\theta_0 + \theta_1 x, \sigma^2).$$

The log-likelihood is then

$$l(\theta_0, \theta_1) = -\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j - \theta_0 - \theta_1 x_j)^2.$$

Note that *maximizing* the log-likelihood is equivalent to *minimizing* the following sum of squares:

$$S(\theta_0, \theta_1) = \sum_{j=1}^{n}(y_j - \theta_0 - \theta_1 x_j)^2.$$

In Section 6.4.2 and Section 6.5 we will discuss minimizing sums of squares much more. Meanwhile, let's find $(\hat{\theta}_0, \hat{\theta}_1)$ that minimizes $S(\theta_0, \theta_1)$. Setting the partial derivative of $S(\theta_0, \theta_1)$ with respect to

θ_0 equal to 0, we have

$$-2\sum_{j=1}^{n}(y_j - \hat{\theta}_0 - \hat{\theta}_1 x_j) = 0,$$

which simplifies to

$$\bar{y} = \hat{\theta}_0 + \hat{\theta}_1 \bar{x}.$$

This equation neatly parallels the fact that, applying Adam's law to

$$\mathrm{E}[Y|X] = \theta_0 + \theta_1 X,$$

we have

$$\mathrm{E}[Y] = \theta_0 + \theta_1 \mathrm{E}[X].$$

Next, we set the partial derivative of $S(\theta_0, \theta_1)$ with respect to θ_1 equal to 0, which gives

$$-2\sum_{j=1}^{n}(y_j - \hat{\theta}_0 - \hat{\theta}_1 x_j)x_j = 0.$$

So

$$\sum_{j=1}^{n} x_j y_j = \hat{\theta}_0 n\bar{x} + \hat{\theta}_1 \sum_{j=1}^{n} x_j^2 = \left(\bar{y} - \hat{\theta}_1 \bar{x}\right) n\bar{x} + \hat{\theta}_1 \sum_{j=1}^{n} x_j^2.$$

Solving for $\hat{\theta}_1$ then gives

$$\hat{\theta}_1 = \frac{\sum_{j=1}^{n} x_j y_j - n\bar{x}\bar{y}}{\sum_{j=1}^{n} x_j^2 - n\bar{x}^2}.$$

For the denominator,

$$\frac{1}{n}\sum_{j=1}^{n} x_j^2 - \bar{x}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

by the sum of squares identity, mirroring the fact that

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

For the numerator, with some algebra it can be shown that

$$\frac{1}{n}\sum_{j=1}^{n} x_j y_j - \bar{x}\bar{y} = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y}),$$

mirroring the fact that

$$\mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])] = \mathrm{Cov}(X, Y).$$

So $\hat{\theta}_1$ is the sample covariance of $\mathbf{x}$ and $\mathbf{y}$, divided by the sample variance of $\mathbf{x}$.

Thus, the MLE is given by

$$\hat{\theta}_1 = \frac{\frac{1}{n}\sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})}{\frac{1}{n}\sum_{j=1}^n (x_j - \bar{x})^2}, \quad \hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}.$$

Note that $\hat{\theta}_1$ looks like $\hat{\theta}$ from the no-intercept case, with $x_j - \bar{x}$ in place of x_j and $y_j - \bar{y}$ in place of y_j. This makes sense intuitively since transforming x_j to $x_j - \bar{x}$ and y_j to $y_j - \bar{y}$ centers these variables to have sample mean 0.

6.3.3 Logistic regression

Recall the logistic regression for binary outcomes is setup in terms of the log-odds

$$\lambda(\mathbf{x}|\boldsymbol{\theta}) = \log\left\{\frac{\mu(\mathbf{x}|\boldsymbol{\theta})}{1-\mu(\mathbf{x}|\boldsymbol{\theta})}\right\} = \theta_0 + \theta_1 x_1 + \cdots + \theta_K x_K,$$

where $\mu(\mathbf{x}|\boldsymbol{\theta}) = P(Y = 1|\mathbf{x};\boldsymbol{\theta})$. The log-likelihood is, as ever, the log of the joint (conditional, as we are dealing with predictive regression) probability. Using equation (6.2) the

$$\begin{aligned}\log L(\theta) &= \log P(Y_1 = y_1, ..., Y_n = y_n | \mathbf{X}_1 = \mathbf{x}_1, ..., \mathbf{X}_n = \mathbf{x}_n, \theta) \\ &= -\sum_{j=1}^n \log\left\{1 + e^{\lambda(\mathbf{x}_j|\theta)}\right\} + \sum_{j=1}^n y_j \lambda(\mathbf{x}_j|\theta).\end{aligned}$$

As with linear regression, to focus on statistical ideas without algebraic clutter we will set $\theta_0 = 0$ and $K = 1$. Then the score function is

$$\begin{aligned}\frac{\partial \log L(\theta)}{\partial\theta} &= \sum_{j=1}^n \frac{\partial\lambda_j}{\partial\theta}\frac{\partial\left\{y_j\lambda_j - \log\left(1+e^{\lambda_j}\right)\right\}}{\partial\lambda_j}, \quad \lambda_j = \lambda(x_j|\theta) = x_j\theta \\ &= \sum_{j=1}^n \frac{\partial\lambda_j}{\partial\theta}(y_j - \mu_j), \quad \mu_j = e^{\lambda_j}/(1+e^{\lambda_j}) \\ &= \sum_{j=1}^n x_j(y_j - \mu_j).\end{aligned}$$

Thus the Fisher information in the sample, conditional on the predictors, is

$$\mathcal{I}_{\mathbf{Y}|(\mathbf{X}=\mathbf{x})}(\theta) = \sum_{j=1}^n x_j^2 \mu_j(1-\mu_j),$$

recalling that $\mathrm{Var}(Y_j|X = x, \theta) = \mu_j(1-\mu_j)$.

The Hessian for the log-likelihood is

$$H(\theta) = \frac{\partial^2 l(\theta)}{\partial\theta^2} = -\sum_{j=1}^n x_j \frac{\partial\mu(x_j|\theta)}{\partial\theta} = -\sum_{j=1}^n x_j \frac{\partial\lambda_j}{\partial\theta}\frac{\partial\mu_j}{\partial\lambda_j} = -\sum_{j=1}^n x_j^2\mu_j(1-\mu_j) = -\mathcal{I}_{\mathbf{Y}|(\mathbf{X}=\mathbf{x})}(\boldsymbol{\theta}).$$

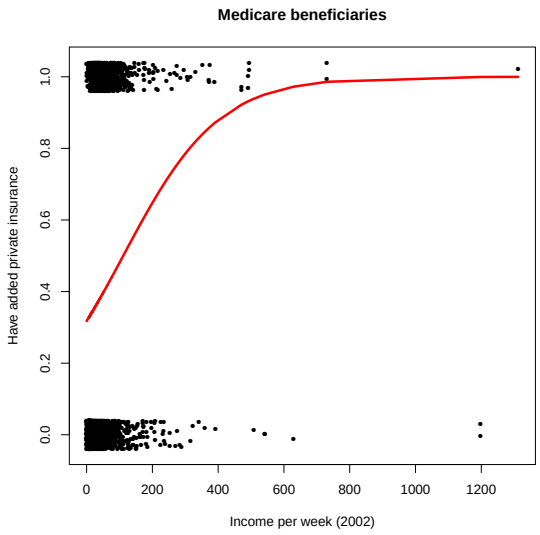


Figure 6.7: Black dots are binary take up of insurance among Medicare beneficiaries, plotted against weekly income. Estimated probability of take up is roughly $\exp(-0.764+0.00686x)/\{1+\exp(-0.764+0.00686x)\}$ and is plotted in red against income.

As $H(\theta)$ is always non-positive, the log-likelihood is globally concave in θ. This means the likelihood can be maximized quickly and reliably, even if $\boldsymbol{\theta}$ is in quite high dimensions. This is a major reason ML estimation of logistic regression is so commonly used. It is implemented in all statistical packages. For large n, we have

$$\hat{\theta}|(\mathbf{X}=\mathbf{x}) \dot{\sim} \mathcal{N}(\theta, \mathcal{I}_{\mathbf{Y}|(\mathbf{X}=\mathbf{x})}(\boldsymbol{\theta})^{-1}).$$

Example 6.3.9. Medicare beneficiaries have a choice to buy additional insurance to cover non-funded items of medical costs. Sometimes this insurance is provided by employers, other times it is paid for by the individual. Here we look at some old data from 2002 which illustrates this, taken from the Health and Retirement Survey (this data set comes from Cameron and Trivedi (2005)). Around 38% of those in the survey took out insurance. How is the take up rate effected by other factors? Here we look at how it relates to weekly income.

To do this we fit a slightly more complicated model than discussed above, fitting

$$P(Y=1|X=x) = \text{logit}^{-1}(\theta_0+\theta_1 x) = \exp(\theta_0+\theta_1 x)/\{1+\exp(\theta_0+\theta_1 x)\}.$$

Figure 6.7 plots insurance take up against weekly income. As the outcome is binary it is hard to see any important relationships. The results from the ML fit are $\hat{\theta}_0 \approx -0.764$ and $\hat{\theta}_1 \approx 0.0068$, while $SE(\hat{\theta}_1) \approx 0.00086$. The code for this example is given in Section 6.9.

Notice there is no call for robust standard errors here, for the binary regression's $\mu(x)$ determines the conditional mean and conditional variance as the outcomes are binary.

6.4 Linear regression, method of moments, and least squares

So far we have estimated a linear predictive regression using MLE. Of course, other estimation strategies can be used. Here we focus on the method of moments and a new one, called *least squares*.

6.4.1 Method of moments

In predictive linear regression, $\mathrm{E}[Y|X = x, \theta] = \theta x$ is a conditional moment. As it specifies a moment it is tempting to develop a method of moments procedure to estimate θ. If this conditional moment holds then, by Adam's law, unconditionally

$$\mathrm{E}[XY] = \mathrm{E}[X\mathrm{E}[Y|X]] = \mathrm{E}[\theta X^2] = \theta\mathrm{E}[X^2].$$

Hence if we make the assumption that the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d., then

$$\theta = \frac{\mathrm{E}[XY]}{\mathrm{E}[X^2]}, \quad \text{so} \quad \hat{\theta}_{\mathrm{MoM}} = \frac{\sum_{j=1}^n X_j Y_j}{\sum_{j=1}^n X_j^2}.$$

Thus this MoM estimator is the MLE for the Gaussian linear regression — but now we regard both the predictors and outcomes as random (recall that for the MLE, we conditioned on the predictors). Notice that Gaussianity plays no role at all in the MoM calculations.

We can again study its properties. The MoM approach has only assumed that $\mathrm{E}[Y|X = x, \theta] = \theta x$ holds. So conditional on the predictors:

$$\mathrm{E}[\hat{\theta}_{\mathrm{MoM}}|\mathbf{X}] = \theta, \quad \mathrm{Var}(\hat{\theta}_{\mathrm{MoM}}|\mathbf{X}) = \left(\sum_{j=1}^n x_j^2\right)^{-2} \sum_{j=1}^n x_j^2 \sigma_j^2,$$

by Lemma 6.3.6. Under $\mathrm{Var}(Y_j|X_j = x_j) = \sigma^2$, homoskedasticity, the conditional variance becomes the same as the variance of the MLE for the Gaussian linear regression model.

☣ **6.4.1.** Notice the vital assumption that $\mathrm{E}(X_j^2)$ does not change with j has sneaked in, which follows from the assumption that the pairs (X_j, Y_j) are i.i.d. Previously, we just specified that the outcomes, conditioned on the predictors, were independent, which is a vastly weaker assumption.

The idea of converting the conditional moment into an unconditional moment can be taken too far: $\mathrm{E}[X^2Y] = \theta\mathrm{E}[X^3]$, which delivers another MoM estimator $\sum_{j=1}^n X_j^2 Y_j / \sum_{j=1}^n X_j^3$, which has terrible properties. Worse yet, $\mathrm{E}[X^3Y] = \theta\mathrm{E}[X^4]$, so $\sum_{j=1}^n X_j^3 Y_j / \sum_{j=1}^n X_j^4$ is another terrible MoM estimator.

More interestingly,

$$\mathrm{E}[\mathrm{sign}(X)Y] = \theta\mathrm{E}[|X|],$$

using the fact that $X\text{sign}(X) = |X|$, produces the MoM estimator

$$\hat{\theta}_C = \frac{\sum_{j=1}^n \text{sign}(X_j)Y_j}{\sum_{j=1}^n |X_j|}.$$

The estimator $\hat{\theta}_C$ is sometimes called the *Cauchy estimator* after a 1836 paper by Augustin-Louis Cauchy. It has some attractive robustness properties that are detailed in an exercise.

This is the familiar challenge with the method of moments strategy: it produces estimator after estimator, only limited by one's patience, delivering the good (one case here), the bad (two cases here), and the interesting (one case).

6.4.2 Least squares

There is a long history of *fitting procedures* in statistics, which start with a numerical procedure which implies the estimator and so, in turn, the estimand. Perhaps the most storied case of this is least squares, which was developed by Carl Friedrich Gauss and Adrien-Marie Legendre who worked independently around 1800.

In the context of predictive regression, the *least squares* (LS) estimator is selected by minimizing the sum of the squared prediction errors. For our linear predictive regression problem the jth prediction error is $Y_j - \theta x_j$, and the sum of squared prediction errors is

$$S(\theta) = \sum_{j=1}^n (Y_j - \theta x_j)^2,$$

so

$$\hat{\theta}_{\text{LS}} = \underset{\theta \in \mathbb{R}}{\arg\min}\ S(\theta) = \frac{\sum_{j=1}^n x_j Y_j}{\sum_{j=1}^n x_j^2},$$

which is the maximum likelihood estimator for the Gaussian linear regression model and the (sensible) MoM estimator. Hence the LS estimator will certainly behave well under the Gaussian regression model, and more broadly if $\text{E}[Y_j | X_j = x_j] = \theta x_j$, then

$$\text{E}[\hat{\theta}_{\text{LS}} | (\mathbf{X} = \mathbf{x})] = \theta, \quad \text{Var}(\hat{\theta}_{\text{LS}} | (\mathbf{X} = \mathbf{x})) = \left(\sum_{j=1}^n x_j^2 \right)^{-2} \sum_{j=1}^n x_j^2 \text{Var}(Y_j | X_j = x_j).$$

We will talk much more about least squares in the linear projection section later.

Remark 3 (Minimizing loss)**.** The least squares criterion is a special case of selecting θ to minimize the sum of losses

$$S(\theta) = \sum_{j=1}^n \text{Loss}(Y_j, \theta x_j),$$

where $\text{Loss}(a, b) \geq 0$ is a loss function of predicting a using b, such that $\text{Loss}(a, a) = 0$. Minimizing the sum of losses is one of the great themes of machine learning, where the loss function is crafted to suit the scientific problem at hand. Aside from least squares, the most celebrated special case of this is minimizing the sum of absolute errors

$$S(\theta) = \sum_{j=1}^{n} |Y_j - \theta x_j|,$$

which is called *least absolute deviation*. It was first studied by Rudjer Boscovich in 1757.

☣ **6.4.2.** It may look innocuous moving from squares to absolute deviations, but it is a fundamental change as it adjusts the *estimand* (not much is bigger in statistics than changing the estimand: perhaps going from describing to predicting, or predicting to causality!). Recall from Section 6.1 of the Stat 110 book that the mean is the best guess for a random variable in a mean square error sense:

$$\text{E}[Y] = \arg\min_{a \in \mathbb{R}} \text{E}[(Y - a)^2].$$

A very quick proof of this fact is to write

$$\text{E}[(Y - a)^2] = \text{Var}(Y) + (a - \text{E}[Y])^2,$$

where the $\text{Var}(Y)$ term has nothing to do with a and the $(a - \text{E}[Y])^2$ clearly is minimized by taking $a = \text{E}[Y]$. The same section of the Stat 110 book also proves that the *median* is the best guess for a random variable in a mean absolute error sense:

$$Q_Y(0.5) = \arg\min_{a \in \mathbb{R}} \text{E}[|Y - a|].$$

Changing the loss function typically changes the estimand! Select your loss function with care.

6.5 Linear projection and descriptive regression

So far, regression has been *predictive*: we see $\mathbf{X}$, then try to predict Y, by modeling $\text{E}[Y|\mathbf{X} = \mathbf{x}]$. This section looks at using regression in a *descriptive* or summary manner, looking at $(\mathbf{X}, \mathbf{Y})$ *jointly* rather than focusing on the conditional distribution of Y given $\mathbf{X}$. This approach connects to an approach to regression called *linear projection*.

Suppose that X, Y have a joint distribution, where again we assume that X is scalar to focus on statistical concepts. Some familiar summary measures of the joint distribution are, for example,

$$\text{E}[X], \quad \text{E}[Y], \quad \text{Var}(X), \quad \text{Var}(Y), \quad \text{Cov}(X, Y), \quad \text{Cor}(X, Y).$$

Another summary is the *descriptive regression*

$$\beta_{Y\sim X} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}.$$

This is just a rewrite of various centered moments: not a parameterized statistical model. Is $\beta_{Y\sim X}$ an interesting summary measure?

The answer is yes. To think about it clearly, allow $\text{E}[X]$ and $\text{E}[Y]$ to have non-zero means, and allow an intercept.

One way of thinking about $\beta_{Y\sim X}$ is to find θ where

$$(\alpha,\theta) = \underset{(a,b)\in\mathbb{R}^2}{\arg\min}\, S(a,b), \quad \text{where} \quad S(a,b) = \text{E}[(Y-a-bX)^2].$$

Then, by DUThIS,

$$\frac{\partial S(a,b)}{\partial a} = -2\text{E}[(Y-a-bX)], \quad \frac{\partial S(a,b)}{\partial b} = -2\text{E}[X(Y-a-bX)] \tag{6.4}$$

which implies that

$$\alpha = \text{E}[Y] - \theta\text{E}[X], \quad \theta = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} = \beta_{Y\sim X}.$$

So $\alpha+\theta X$ best mimics Y in the sense that any other linear mimic $a+bX$ will have (by construction) a larger expected square error $\text{E}[(Y-a-bX)^2]$. Note that this result is analogous to the result for the MLE in a Gaussian linear regression with an intercept (from Section 6.3.2), with population quantities instead of sample quantities, but here we are *not* making any parametric assumptions.

It follows that

$$\alpha+\theta X = \text{E}[Y] + \beta_{Y\sim X}(X-\text{E}[X]).$$

The right-hand side is important and has its own nomenclature.

Definition 6.5.1. (Linear projection). Assume that the random variables X, Y each have a finite variance. Then the *linear projection* of Y on X at $X=x$ is defined as

$$\mu_L(x) = \text{E}[Y] + \beta_{Y\sim X}(x-\text{E}[X]).$$

The linear projection $\mu_L(x)$ is *not* the conditional expectation $\text{E}[Y|X=x]$ in general. The conditional expectation $\text{E}[Y|X=x]$ is the function of x that best approximates Y, in the sense of minimizing the expected square error; the linear projection $\mu_L(X)$ is the best *linear* function of x for approximating Y.

Writing the linear error as

$$U_L = Y - \mu_L(X),$$

then by construction $\text{E}[U_L]=0$ and $\text{E}[XU_L]=0$, due to the derivatives (6.4).

☣ **6.5.2.** Often U_L is used to write $Y = \mu_L(X) + U_L$. Mathematically this is, of course, correct. But it could give the false impression that X causes Y, by tempting us to think that if X is moved then Y must move. One can quickly overcome this causal urge by thinking: we can project X on Y just as easily as we can project Y on X, so the same logic would say Y causes X. So it is nonsense — not even deserving to be called casual inference, let alone causal inference — to conclude from $Y = \mu_L(X) + U_L$ that X causes Y.

Remark 4. How does linear projection relate to predictive regression? Recall that $Y = \mu(X) + U$, where $\mu(x) = \mathrm{E}[Y|X = x]$ and $U = Y - \mu(X)$, implying that $\mathrm{E}[U\mu(X)] = 0$. Then

$$\begin{aligned} Y - \mu_L(X) &= \{\mu(X) - \mu_L(X)\} + \{Y - \mu(X)\} \\ &= \{\mu(X) - \mu_L(X)\} + U \\ &= \text{linear approximation error} + \text{prediction error}. \end{aligned}$$

After some algebra it can be shown that

$$\beta_{Y\sim X} = \mathrm{Cov}(\mu(X), X)/\mathrm{Var}(X) = \beta_{\mu(X)\sim X}.$$

Now turn to statistical inference for this problem. First, the statistical model will be that the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. pairs from the joint distribution $F_{X,Y}$. For simplicity we will set $\mathrm{E}[Y] = \mathrm{E}[X] = 0$ and have the estimand

$$\theta = \frac{\mathrm{E}[XY]}{\mathrm{E}[X^2]}.$$

Then a method of moments estimator of this is

$$\hat{\theta} = \frac{n^{-1}\sum_{j=1}^n X_j Y_j}{n^{-1}\sum_{j=1}^n X_j^2},$$

which is also the least squares estimator

$$\hat{\theta} = \underset{a\in\mathbb{R}}{\arg\min} \sum_{j=1}^n (Y_j - aX_j)^2.$$

Example 6.5.3. (A stock's beta) Let the pairs $(x_1, y_1), ..., (x_n, y_n)$ be the daily returns over the last 5 years on an exchanged traded fund called "SPDR" which approximates the S&P500 (U.S.) market index and Apple stock, be modeled as i.i.d. draws from a joint distribution X, Y. For simplicity model these both as having zero means (in reality the means are both very close to zero, but importantly positive, while stock returns have some mild predictability, so the i.i.d. assumption is not entirely right either). We will estimate $\beta_{Y\sim X} = \{\mathrm{Var}(X)\}^{-1}\mathrm{Cov}(X, Y)$, using $\hat{\theta} = \sum_{j=1}^n x_j y_j / \sum_{j=1}^n x_j^2 \approx 1.256$.

Figure 6.8: Cross plots of daily geometric returns for Apple and S&P500 index from 1 January 2015 to 13 February 2020. Also shown is the OLS (MoM and MLE) estimated line $\hat{\theta}x$. The estimated slope is $\hat{\theta} \approx 1.256$ for both MoM and MLE, but the standard errors are different due to the different underlying assumptions behind the methods. The MLE has $\text{SE}(\hat{\theta}) \approx 0.038$; the MoM's SE is approximately 0.049.

Figure 6.8 shows a scatterplot of the data, together with $\hat{\theta}x$. The caption gives the estimated standard error of $\hat{\theta}$, computed two different ways. The code for all this is featured in Section 6.9.

Why is this economically interesting? Informally, $\beta_{Y\sim X}$, which is called the "market beta" in finance, is a summary measure of how Apple and the S&P500 index move together, but an alternative for that is the correlation. Why is $\beta_{Y\sim X}$ popular in this context? The $Y - \beta_{Y\sim X}X$ is the return on a portfolio: holding 1 Apple share and owing (in finance this is called shorting) $\beta_{Y\sim X}$ shares in the S&P500. It is the portfolio with the smallest expected squared return of any portfolio of the form $Y - bX$. Hence it is the one with the narrowest risk (measured through expected squared returns) of any portfolio which is holding 1 Apple share in that class. This conclusion does not need a predictive model, just a descriptive model of how X, Y vary together.

☣ **6.5.4.** Linear projection is powerful, allowing regression to be used without commiting to the linear model being true. This looks very appealing in Example 6.5.3 where we are making decisions which are linear: building returns on an investment by scaling up returns X on the S&P500. There are some dangers lurking here though. The clearest version is where Y is binary. Then $\mu(x) = \text{E}[Y|X = x]$ must be between 0 and 1 for any x. But the linear projection $\mu_L(x)$ could be, for example, 1.3 or -0.2 for some x — which is a category error if you think of $\mu_L(x)$ being a good approximation to $\mu(x)$. Guess how quickly you would get fired if you suggest to a TV host on election night that the data

implies the probability a particular candidate will win is 130%? Or, how quickly you would go bust betting on sports outcomes if your linear projection spits out the chance a particular team will win is -20%? Unthinking use of linear projection can lead to disaster.

What are the properties of $\hat{\theta}$? This is not easy. If we condition on the predictors it does not help (under linear projection, we no nothing about the conditional distribution of Y given X). We need to obtain unconditional results and so use asymptotic approximations as the estimator is non-linear in the random pairs $(X_1, Y_1), ..., (X_n, Y_n)$. Write $U_{L,j} = Y_j - \theta X_j$, and note that $\mathrm{E}[XU_L] = 0$. Then

$$\hat{\theta} = \theta + \frac{n^{-1}\sum_{j=1}^{n} X_j U_{L,j}}{n^{-1}\sum_{j=1}^{n} X_j^2}.$$

As the numerator is random, we do not know how to compute $\mathrm{E}[\hat{\theta}]$ for this estimator. By the CLT

$$n^{1/2}\left(\frac{1}{n}\sum_{j=1}^{n} X_j U_{L,j}\right) \xrightarrow{d} \mathcal{N}\Big(0, \mathrm{Var}(XU_L)\Big),$$

the LLN on $n^{-1}\sum_{j=1}^{n} X_j^2$, and Slutsky's theorem, we have

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathrm{Var}(XU_L)}{(\mathrm{E}[X^2])^2}\right),$$

so long as $\mathrm{Var}(XY) < \infty$ and $\mathrm{Var}(X^2) < \infty$. In practice we can estimate $\mathrm{Var}(XU)$ by the method of moments estimator $n^{-1}\sum_{j=1}^{n} X_j^2(Y_j - \hat{\theta}X_j)^2$ and $\mathrm{E}[X^2]$ by the sample average of the squared regressors. This takes us back to Biohazard 6.3.7 and robust standard errors.

Remark 5. Consider the predictive regression model $\mathrm{E}(Y|X = x) = \theta x$, recalling the notation

$$\sigma^2(x) = \mathrm{Var}(Y|X = x) = \mathrm{Var}(U|X = x),$$

where $U = Y - \theta X$. We showed earlier that $\mathrm{E}[U|X] = 0$. By Eve's Law

$$\mathrm{Var}(XU) = \mathrm{E}[X^2\mathrm{Var}(U|X)] + \mathrm{Var}(X\mathrm{E}[U|X]) = \mathrm{E}[X^2\sigma^2(X)],$$

resulting in the same unconditional limit result:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, \frac{\mathrm{E}[X^2\sigma^2(X)]}{(\mathrm{E}[X^2])^2}\right).$$

This says that starting with predictive regression and then converting to an unconditional result produces the same limit distribution as if we had started with descriptive regression.

6.6 Multiparameter regression*

Most of the statistical ideas in predictive and linear projection can be expressed by thinking about the $K = 1$, single parameter case. However, there are some elegant multiparameter results which are worthwhile highlighting when expressed using matrix algebra.

☣ **6.6.1.** One of the challenges in using linear algebra is statistics is the clash of notational conventions. In statistics and probability, capital letters, like Y, are typically used to denote random variables, while lowercase letters, like y, are typically used to denote data. In linear algebra, capital letters are typically matrices and lowercase letters are typically vectors. This can be confusing. Here we will do our best to be clear, by using bold to denote linear algebra terms. Even then sometimes it is hard to be completely consistent, but the meaning should be clear from the context.

6.6.1 Linear predictive regression

Think of the data appearing in a sequence of pairs, $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$, where y_j are scalars but $\mathbf{x}_j = (1, x_{j1}, ..., x_{jK})^\intercal$. (It will be convenient to carry around the 1's. Obviously, conditioning on a constant has no impact.) Focus on the predictive linear regression problem. Write

$$\mathrm{E}[Y_j|\mathbf{X}_j = \mathbf{x}_j, \boldsymbol{\theta}] = \theta_0 + \theta_1 x_{j1} + ... + \theta_K x_{jK} = \mathbf{x}_j^\intercal \boldsymbol{\theta},$$

where $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_K)^\intercal$.

The Gaussian prediction regression model has, for $j = 1, ..., n$,

$$Y_j|\mathbf{x}_j, \boldsymbol{\theta} \overset{\text{ind}}{\sim} \mathcal{N}(\mathbf{x}_j^\intercal \boldsymbol{\theta}, \sigma^2).$$

This can be written compactly as

$$\mathbf{y}|\mathbf{X}, \boldsymbol{\theta} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\theta}, \sigma^2 I_n),$$

a multivariate Gaussian distribution, where I_n is the $n \times n$ identity matrix, $\mathbf{y} = (Y_1, ..., Y_n)^\intercal$, and $\mathbf{X}$ is the $n \times (K + 1)$ matrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & ... & x_{1K} \\ 1 & x_{21} & ... & x_{2K} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & ... & x_{nK} \end{pmatrix}.$$

Note some deviations from our usual notation: the bold $\mathbf{y}$ is a vector of random variables, while we have conditioned on a matrix bold $\mathbf{X}$ of random variables.

The log-likelihood for this statistical (conditional) model is

$$\log L(\theta; \mathbf{y}) = -\frac{1}{2\sigma^2} \sum_{j=1}^{n} (y_j - \mathbf{x}_j^\intercal \boldsymbol{\theta})^2 = -\frac{1}{2\sigma^2} (\mathbf{y} - \mathbf{X}\theta)^\intercal (\mathbf{y} - \mathbf{X}\theta),$$

while, for $a, b = 0, 1, ..., K$, then

$$\frac{\partial \log L}{\partial \theta_a} = \frac{1}{\sigma^2}\sum_{j=1}^{n} x_{ja}(y_j - \mathbf{x}_j^{\mathsf{T}}\boldsymbol{\theta}) \quad \text{and} \quad H_{a,b} = \frac{\partial^2 \log L}{\partial \theta_a \partial \theta_b} = -\frac{1}{\sigma^2}\sum_{j=1}^{n} x_{ja}x_{jb}.$$

Then

$$s(\boldsymbol{\theta}; \mathbf{y}) = \begin{pmatrix} \frac{\partial \log L(\boldsymbol{\theta};\mathbf{y})}{\partial \theta_0} \\ \vdots \\ \frac{\partial \log L(\boldsymbol{\theta};\mathbf{y})}{\partial \theta_K} \end{pmatrix} = \frac{1}{\sigma^2}\sum_{j=1}^{n} \mathbf{x}_j(y_j - \mathbf{x}_j^{\mathsf{T}}\boldsymbol{\theta}) = \frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\theta),$$

and the Hessian is

$$\begin{aligned} H &= \begin{pmatrix} H_{0,0} & H_{0,1} & ... & H_{0,K} \\ H_{1,0} & H_{1,1} & ... & H_{1,K} \\ \vdots & \vdots & \ddots & \vdots \\ H_{K,0} & H_{K,1} & ... & H_{K,K} \end{pmatrix} = -\frac{1}{\sigma^2}\sum_{j=1}^{n} \begin{pmatrix} 1 & x_{j1} & ... & x_{jK} \\ x_{j1} & x_{j1}^2 & ... & x_{j1}x_{jK} \\ \vdots & \vdots & \ddots & \vdots \\ x_{jK} & x_{jK}x_{j1} & ... & x_{jK}^2 \end{pmatrix} \\ &= -\frac{1}{\sigma^2}\sum_{j=1}^{n} \mathbf{x}_j\mathbf{x}_j^{\mathsf{T}} \\ &= -\frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}\mathbf{X}. \end{aligned}$$

This implies that the MLE and the Fisher information in the sample are, respectively,

$$\hat{\theta} = \left(\sum_{j=1}^{n} \mathbf{x}_j\mathbf{x}_j^{\mathsf{T}}\right)^{-1} \sum_{j=1}^{n} \mathbf{x}_j y_j = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}, \quad \mathcal{I}_{\mathbf{y}|\mathbf{X}}(\theta) = \frac{1}{\sigma^2}\sum_{j=1}^{n} \mathbf{x}_j\mathbf{x}_j^{\mathsf{T}} = \frac{1}{\sigma^2}\mathbf{X}^{\mathsf{T}}\mathbf{X},$$

with the former assuming $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is invertible. Again, $\hat{\theta}$ is usually called the least squares estimator in the literature.

Matrix algebra enables the properties of $\hat{\theta}$ to be found elegantly. In particular,

$$\mathrm{E}[\hat{\theta}|\mathbf{X}] = \mathrm{E}[(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}|\mathbf{X}] = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathrm{E}[\mathbf{y}|\mathbf{X}] = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}\theta = \theta,$$

as $\mathrm{E}[\mathbf{y}|\mathbf{X}] = \mathbf{X}\theta$ and

$$\mathrm{Var}[\hat{\theta}|\mathbf{X}] = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathrm{Var}[\mathbf{y}|\mathbf{X}]\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1},$$

if $\mathrm{Var}[\mathbf{y}|\mathbf{X}] = \sigma^2 I_n$ (which will hold under the Gaussian model). Hence $\hat{\theta}$ is unbiased and achieves the Cramér-Rao lower bound. Under the Gaussian prediction model,

$$\hat{\theta}|\mathbf{X} \sim \mathcal{N}(\theta, \sigma^2(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}),$$

as $\hat{\theta}|\mathbf{X}$ is linear in $\mathbf{y}$ and $\mathbf{y}|\mathbf{X}$ is Gaussian. The multiparameter Bayes version of this result is discussed in Chapter 9, which focuses on Bayesian inference.

The vector of *regression errors* is $\mathbf{u} = \mathbf{y} - \mathbf{X}\theta$, and the vector of *regression residuals* is

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\theta} = (I_n - \mathbf{P}_X)\mathbf{y}, \quad \mathbf{P}_X = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}.$$

The $n \times n$ square matrix $\mathbf{P}_X$ is a *projection matrix* from linear algebra, since $\mathbf{P}_X\mathbf{P}_X = \mathbf{P}_X$ and $\mathbf{P}_X$ is a symmetric matrix. Geometrically, $\mathbf{P}_X$ projects onto the column space of X, i.e., $\mathbf{P}_X\mathbf{y}$ is the linear combination of the columns of X that is closest to $\mathbf{y}$. Now

$$\mathbf{X}^\mathsf{T}\mathbf{P}_X = \mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T} = \mathbf{X}^\mathsf{T},$$

so $\mathbf{X}^\mathsf{T}(I_n - \mathbf{P}_X) = \mathbf{0}$. Therefore, $\mathbf{X}^\mathsf{T}\hat{\mathbf{u}} = \mathbf{0}$, which says that the vector of regression residuals is orthogonal to each predictor variable. Furthermore,

$$(I_n - \mathbf{P}_X)(I_n - \mathbf{P}_X) = I_n^2 - I_n\mathbf{P}_X - \mathbf{P}_X I_n - \mathbf{P}_X^2 = I_n - \mathbf{P}_X - \mathbf{P}_X + \mathbf{P}_X = I_n - \mathbf{P}_X.$$

So

$$\mathrm{Var}(\hat{\mathbf{u}}|\mathbf{X}) = (I - \mathbf{P}_X)(\sigma^2 I_n)(I - \mathbf{P}_X) = \sigma^2(I_n - \mathbf{P}_X).$$

☣ **6.6.2.** Extending the results in Biohazard 6.3.7 on the effect of dropping the Gaussianity assumption, the property $\mathrm{E}[\hat{\theta}|\mathbf{X}] = \theta$ holds so long as $\mathrm{E}[\mathbf{y}|\mathbf{X}] = \mathbf{X}\theta$. Now suppose there is heteroskedasticity,

$$\mathrm{Var}(Y_j|\mathbf{X}_j] = \mathbf{x}_j) = \sigma_j^2,$$

which we write using matrix algebra as

$$\mathrm{Var}(Y|\mathbf{X}, \boldsymbol{\theta}) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2),$$

a $n \times n$ diagonal matrix. Then

$$\mathrm{Var}[\hat{\theta}|\mathbf{X}] = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathrm{Var}[\mathbf{y}|\mathbf{X}]\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\left[\sum_{j=1}^n \mathbf{x}_j\mathbf{x}_j^\mathsf{T}\sigma_j^2\right](\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}.$$

Thus we might "robustly" estimate $\mathrm{Var}(\sqrt{n}\hat{\theta}|\mathbf{X})$ by

$$\left[n^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\right]^{-1}\left[n^{-1}\sum_{j=1}^n \mathbf{x}_j\mathbf{x}_j^\mathsf{T}\hat{u}_j^2\right]\left[n^{-1}\mathbf{X}^\mathsf{T}\mathbf{X}\right]^{-1}.$$

Example 6.6.3 (Continuing Example 6.3.8)**.** Without centering anything this time, regress the temperature (y_j) on an intercept (1), latitude (x_{j1}), and longitude (x_{j2}). The resulting parameter estimates are $\hat{\theta}_0 \approx 98.6$, $\hat{\theta}_1 \approx -2.16$, and $\hat{\theta}_2 \approx 0.13$. Their standard errors are approximately 8.33, 0.17, and 0.06, respectively. The latitude coefficient has not moved a great deal with the bigger model, while

longitude has a small positive coefficient. Hence it gets a little warmer as we go west. To give an impression of the coefficients, Boston's latitude and longitude pair is about 42,71 degrees, San Francisco's is 38,123. This converts in terms of temperature to San Francisco being expected to be about 6.5 degrees Fahrenheit warmer in the winter. These are very rough numbers, as the weather in both Boston and San Francisco is highly influenced by their neighboring oceans.

This example illustrates the dangers of extrapolation. The dataset only covers U.S. cities. London is a long way away from North America! Predicting London's January temperature using this prediction model is an example of extreme extrapolation. But suspend your thinking and give it a try. It gives ridiculous results. London's latitude is about 51 degrees (Boston's is about 43), with 0 longitude. The linear regression fit predicts London should be around 22 degrees colder than Boston. But Boston has very cold winters, while London's winters are very mild due to the gulf stream (this linear fit does indicate London's weather could be severely impacted by climate change if the gulf stream is damaged).

6.6.2 Logistic regression

Much of the development of logistic regression in Section 6.2.2 allowed K predictors, plus an intercept. Here we complete the setup, starting with the score

$$\begin{aligned}\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \sum_{j=1}^{n} \frac{\partial \lambda_j}{\partial \boldsymbol{\theta}} \frac{\partial \left\{ y_j \lambda_j - \log\left(1 + e^{\lambda_j}\right)\right\}}{\partial \lambda_j}, \quad \lambda_j = \lambda(\mathbf{x}_j|\boldsymbol{\theta}) = \mathbf{x}_j^{\mathrm{T}}\boldsymbol{\theta} \\ &= \sum_{j=1}^{n} \frac{\partial \lambda_j}{\partial \boldsymbol{\theta}}(y_j - \mu_j), \quad \mu_j = e^{\lambda_j}/(1 + e^{\lambda_j}) \\ &= \sum_{j=1}^{n} \mathbf{x}_j(y_j - \mu_j).\end{aligned}$$

Thus the Fisher information in the sample, conditional on the regressors, is

$$\mathcal{I}_{\mathbf{Y}|(\mathbf{X}=\mathbf{x})}(\theta) = \sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^{\mathrm{T}} \mu_j(1 - \mu_j),$$

recalling that $\mathrm{Var}(Y_j|\mathbf{X} = \mathbf{x}, \theta) = \mu_j(1 - \mu_j)$.

The Hessian for the log-likelihood is

$$\begin{aligned}H(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} &= -\sum_{j=1}^{n} \mathbf{x}_j \frac{\partial \mu(\mathbf{x}_j|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}} = -\sum_{j=1}^{n} \mathbf{x}_j \frac{\partial \lambda_j}{\partial \boldsymbol{\theta}} \frac{\partial \mu_j}{\partial \lambda_j^{\mathrm{T}}} = -\sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^{\mathrm{T}} \mu_j(1 - \mu_j) \\ &= -\mathcal{I}_{\mathbf{Y}|(\mathbf{X}=\mathbf{x})}(\theta).\end{aligned}$$

Thus the observed Fisher information in the sample equals the Fisher information in the sample.

As $H(\boldsymbol{\theta})$ is always non-positive, the log-likelihood is globally concave in $\boldsymbol{\theta}$. This means the likelihood can be maximized quickly and reliably, even if $\boldsymbol{\theta}$ is in high dimensions. This is a major reason maximum likelihood estimation of logistic regression is so commonly used in statistics and machine learning. For large n, we have

$$\hat{\boldsymbol{\theta}}|(\mathbf{X}=\mathbf{x}) \dot{\sim} \mathcal{N}(\boldsymbol{\theta}, \mathcal{I}_{\mathbf{Y}|(\mathbf{X}=\mathbf{x})}(\theta)^{-1}).$$

6.6.3 Linear projection

The maximum likelihood estimator of the Gaussian linear predictive regression is also the least squares estimator

$$\hat{\boldsymbol{\theta}} = \underset{(a,b_1,...,b_K)\in\mathbb{R}^{K+1}}{\arg} \min \sum_{j=1}^{n} (Y_j - a - b_1 x_{j1} - ... - b_K x_{jK})^2.$$

Likewise, the descriptive regression extends to K regressors, $\mathbf{X} = (X_1, ..., X_K)^\intercal$. Then the criterion is

$$S(a, b_1, ..., b_K) = \mathrm{E}[(Y - a - b_1 X_1 - ... - b_K X_K)^2],$$

which results when minimized in the $K \times 1$ vector

$$\beta_{Y\sim\mathbf{X}} = \underset{(b_1,...,b_K)\in\mathbb{R}^K}{\arg} \min \ \min_{a\in\mathbb{R}} S(a, b_1, ..., b_K) = \{\mathrm{Var}(\mathbf{X})\}^{-1}\mathrm{Cov}(\mathbf{X}, Y),$$

where $\mathrm{Var}(\mathbf{X})$ is a $K \times K$ variance/covariance matrix and $\mathrm{Cov}(\mathbf{X}, Y)$ is a $K \times 1$ vector of covariances.

The corresponding linear projection of Y on $\mathbf{X}$ is

$$\mu_L(x) = \mathrm{E}[Y] + \beta_{Y\sim\mathbf{X}}^\intercal (x - \mathrm{E}[X]).$$

6.7 Additional regressions*

6.7.1 Regularization: ridge and Lasso

There are some advantages in shrinking regression coefficients towards zero. In Chapter 9 this is carried out using a prior and Bayes' theorem to form Bayes estimators. In statistics shrinking also appears through "penalizing" criteria like least squares. In the predictive regression context, this becomes

$$S_2(\theta) = \sum_{j=1}^{n} (Y_j - \theta x_j)^2 + \lambda\theta^2$$

where $\lambda \geq 0$ is regarded as a "penalty" parameter, discouraging large values of θ^2. Then

$$\hat{\theta}_{\text{ridge}}[\lambda] = \underset{\theta\in\mathbb{R}}{\arg\min} \quad S_2(\theta)$$

(the bracket is used to reinforce that the estimator depends upon the value of λ) which is found by looking at

$$\frac{\partial S_2(\theta)}{\partial \theta} = -2\sum_{j=1}^{n} x_j(Y_j - \theta x_j) + 2\lambda\theta$$

so

$$\hat{\theta}_{\text{ridge}}[\lambda] = \frac{\sum_{j=1}^n x_j Y_j}{\sum_{j=1}^n x_j^2 + \lambda} = w\hat{\theta}, \quad \text{where} \quad \hat{\theta} = \frac{\sum_{j=1}^n x_j Y_j}{\sum_{j=1}^n x_j^2}, \quad w = \frac{\sum_{j=1}^n x_j^2}{\sum_{j=1}^n x_j^2 + \lambda} \in [0,1].$$

Here, if $\sum_{j=1}^n x_j^2$ is very large, the data should be very informative about θ (recall the Fisher information formalizes this) and so w is close to one, so $\hat{\theta}_{\text{ridge}}[\lambda] \approx \hat{\theta}$. If $\sum_{j=1}^n x_j^2$ is small, w can be close to zero and so $\hat{\theta}_{\text{ridge}}[\lambda] \approx 0$.

In statistics this is called *ridge regression*. It nudges least squares towards 0, yielding a

$$\text{bias}(\hat{\theta}_{\text{ridge}}[\lambda]|(\mathbf{X} = \mathbf{x})) = (w-1)\theta$$

but

$$\text{Var}(\hat{\theta}_{\text{ridge}}[\lambda]|(\mathbf{X} = \mathbf{x})) = w^2\sigma_{\hat{\theta}}^2 \leq \sigma_{\hat{\theta}}^2,$$

where $\sigma_{\hat{\theta}}^2 = \text{Var}(\hat{\theta}|(\mathbf{X} = \mathbf{x}))$. Thus the ridge shrinks the variance and induces a bias. The trade-off

$$\text{MSE}(\hat{\theta}_{\text{ridge}}[\lambda]|(\mathbf{X} = \mathbf{x})) = (w-1)^2\theta^2 + w^2\sigma_{\hat{\theta}}^2,$$

is minimized by having $w = \theta^2/(\theta^2 + \sigma_{\hat{\theta}}^2)$, which implies taking $\lambda = \sigma_{\hat{\theta}}^2/\theta^2$. If θ is small, then there can be meaningful mean square reductions from using ridge regression. In practice, the gain from ridge regression become potentially important in cases where there are many predictors, where it is difficult to estimate each individual parameter.

A seemingly small twist on S_2 is to alter the penalty to

$$S_1(\theta) = \sum_{j=1}^{n}(Y_j - \theta x_j)^2 + 2\lambda|\theta|,$$

which penalizes small $|\theta|$ more and large $|\theta|$ less than the corresponding quadratic case. Then

$$\hat{\theta}_{\text{Lasso}}[\lambda] = \underset{\theta \in \mathbb{R}}{\arg\min}\ \ S_1(\theta)$$

which is found by looking at

$$\frac{\partial S_1(\theta)}{\partial \theta} = \begin{cases} -2\sum_{j=1}^n x_j(Y_j - \theta x_j) + 2\lambda & \theta > 0 \\ -2\sum_{j=1}^n x_j(Y_j - \theta x_j) - 2\lambda & \theta < 0 \end{cases}$$

so

$$\hat{\theta}_{\text{Lasso}}[\lambda] = \begin{cases} \hat{\theta} - \frac{\lambda}{\sum_{j=1}^n x_j^2} & \hat{\theta} > \frac{\lambda}{\sum_{j=1}^n x_j^2} \\ 0 & \text{elsewhere} \\ \hat{\theta} + \frac{\lambda}{\sum_{j=1}^n x_j^2} & \hat{\theta} < -\frac{\lambda}{\sum_{j=1}^n x_j^2} \end{cases}$$

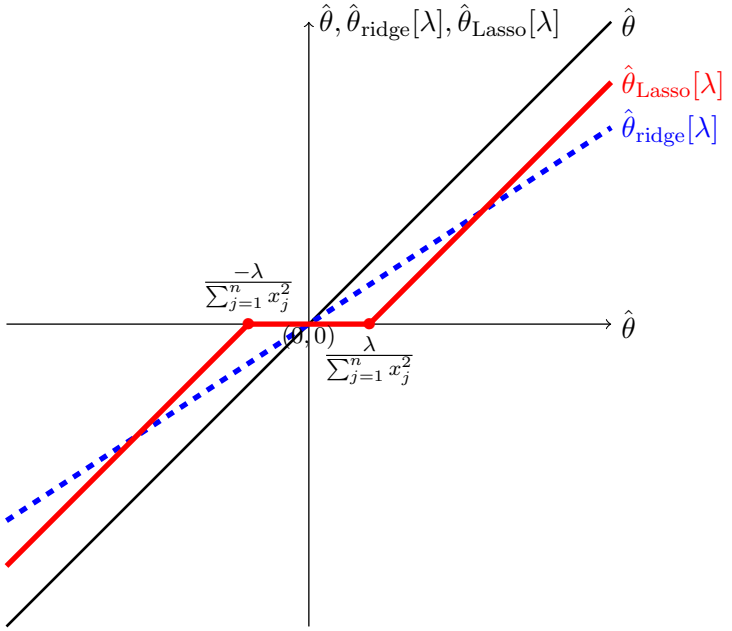


Figure 6.9: Plot of $\hat{\theta}$ (black line), $\hat{\theta}_{\text{ridge}}[\lambda]$ (blue line) and $\hat{\theta}_{\text{Lasso}}[\lambda]$ (red line) against $\hat{\theta}$, showing shrinkage of $\hat{\theta}_{\text{ridge}}, \hat{\theta}_{\text{Lasso}}$ towards 0. $\hat{\theta}_{\text{ridge}}[\lambda]$ reduces the slope. $\hat{\theta}_{\text{Lasso}}[\lambda]$ uses a threshold. Notice, outside the range of the threshold, the $\hat{\theta}$ and $\hat{\theta}_{\text{Lasso}}[\lambda]$ move in parallel — they have the same slope.

The estimator $\hat{\theta}_{\text{Lasso}}[\lambda]$ is called the Lasso (least absolute shrinkage and selection operator) regression estimator. It was introduced by Robert Tibshirani in 1996. It is like ridge regression, but instead shrinks $\hat{\theta}$ to exactly zero if it is small, otherwise shifts it towards zero by a constant.

Figure 6.9 plots $\hat{\theta}$, $\hat{\theta}_{\text{ridge}}[\lambda]$, and $\hat{\theta}_{\text{Lasso}}[\lambda]$ all against $\hat{\theta}$. For small $|\hat{\theta}|$ Lasso shrinks a great deal but as $|\hat{\theta}|$ increases the shrinkage of ridge gets larger. The properties of Lasso are beyond the scope of this book, but there will be more discussion of this in Chapter 9.

How can the penalty λ be selected? An immediate thought is to minimizing S_2 or S_1 over both θ and λ. But this always selects λ to be 0 and θ to be the least squares estimator, so that idea is useless. What to do?

A common strategy is to use cross-validation, which will be discussed in some depth in the starred Section 10.6. This selects λ by minimizing the sum of squared "out of sample" prediction errors

$$\hat{\lambda} = \underset{\lambda \in \mathbb{R}_{\geq 0}}{\arg\min} \sum_{j=1}^{n} (Y_j - x_j \hat{\theta}[\lambda]_{\setminus j})^2,$$

where $\hat{\theta}[\lambda]_{\setminus j}$ is the ridge (or Lasso) regression estimate made using all the data except the pair (x_j, Y_j) employing λ as the penalty value ($\hat{\theta}[\lambda]_{\setminus j}$ is called the j-th jackknife replicate). The resulting regression estimate is then

$$\hat{\theta}[\hat{\lambda}].$$

6.7.2 Nonparametric regression

A forward neutral network is a way of highly parameterizing a potentially non-linear predictive regression, with massive numbers of parameters. An alternative is to use non-parametric methods: an example of this is kernel regression. This is very influential for problems with low-dimensional predictors, e.g., dimensions 1 or 2.

We will focus on a single predictor and use the notation

$$\mu(x) = \mathrm{E}[Y|(X = x)], \quad \text{and} \quad \sigma^2(x) = \mathrm{Var}[Y|(X = x)],$$

while the data is $(x_1, y_1), ..., (x_n, y_n)$. For a bandwidth $h > 0$, define

$$\hat{\mu}(x) = \frac{\sum_{|x_j - x| < h/2} Y_j}{n(x, h)}, \quad n(x, h) = \sum_{|x_j - x| < h/2} 1$$

a nonparametric estimator of $\mu(x)$ — averaging the outcomes corresponding to predictors close to x. This is a special case of a "Nadaraya-Watson kernel regression" and extends the kernel density estimator we studied in Chapters 1 and 3. We will study its properties by conditioning on the predictors.

Notice when $n(x, h)$, the number of predictors close to x, is 0, then $\hat{\mu}(x)$ is not defined. Here we will look at the conditional bias and variance of $\hat{\mu}(x)$ when $n(x, h) > 0$.

Theorem 6.7.1. *Let the pairs $(X_1, Y_1), ..., (X_n, Y_n)$ be i.i.d., assume $\mu(x)$ is twice continuously differentiable, $n(x, h) > 0$, $h \approx 0$, conditioning on the observed predictors, then*

$$\mathrm{bias}(\hat{\mu}(x)) \approx \mu'(x)\bar{x}(x, h) + \frac{1}{2}\mu''(x)s^2(x, h), \quad \mathrm{Var}(\hat{\mu}(x)) \approx \frac{\sigma^2(x)}{n(x, h)},$$

where

$$\bar{x}(x, h) = \frac{\sum_{|x_j - x| < h/2}(x_j - x)}{n(x, h)}, \quad s^2(x, h) = \frac{\sum_{|x_j - x| < h/2}(x_j - x)^2}{n(x, h)}.$$

Proof. Then the conditional expectation and variance of $\hat{\mu}(x)$ are

$$\mu^*(x) = \mathrm{E}[\hat{\mu}(x)|X_1 = x_1, ..., X_n = x_n] = \frac{\sum_{|x_j - x| < h/2} \mathrm{E}[Y_j|X_1 = x_1, ..., X_n = x_n]}{n(x, h)}$$
$$= \frac{\sum_{|x_j - x| < h/2} \mu(x_j)}{n(x, h)},$$

and

$$\sigma^{2*}(x) = \mathrm{Var}[\hat{\mu}(x)|X_1 = x_1, ..., X_n = x_n] = \frac{\sum_{|x_j - x| < h/2} Var[Y_j|X_1 = x_1, ..., X_n = x_n]}{n(x, h)^2}$$
$$= \frac{1}{n(x, h)} \frac{\sum_{|x_j - x| < h/2} \sigma^2(x_j)}{n(x, h)}.$$

Use the Taylor approximations at x_j about x in the region $|x_j - x| < h/2$: $\mu(x_j) \approx \mu(x) + (x_j - x)\mu'(x) + \frac{1}{2}(x_j - x)^2\mu''(x)$ implying

$$\mu^*(x) \approx \mu(x) + \mu'(x)\bar{x}(x,h) + \frac{1}{2}\mu''(x)s^2(x,h),$$

where

$$\bar{x}(x,h) = \frac{\sum_{|x_j-x|<h/2}(x_j - x)}{\sum_{|x_j-x|<h/2} 1}, \quad s^2(x,h) = \frac{\sum_{|x_j-x|<h/2}(x_j - x)^2}{\sum_{|x_j-x|<h/2} 1}.$$

By construction, $|\bar{x}(x,h)| < h$ and $s^2(x,h) < h^2$. Likewise, $\sigma^2(x_j) \approx \sigma^2(x)$, so

$$\sigma^{2*}(x) = \frac{1}{n(x,h)} \frac{\sum_{|x_j-x|<h/2} \sigma^2(x_j)}{n(x,h)} \approx \frac{\sigma^2(x)}{n(x,h)}.$$

∎

Thus, with an initial estimate of $\mu(x)$ and $\sigma(x)$, we can approximate the conditional mean square error by

$$\text{mse}[\hat{\mu}(x)|X_1 = x_1, ..., X_n = x_n] \approx \{\mu'(x)\bar{x}(x,h) + \frac{1}{2}\mu''(x)s^2(x,h)\}^2 + \frac{\sigma^2(x)}{n(x,h)},$$

which can be use to select a reasonable bandwidth by minimizing it with respect to h, noting that the $\bar{x}(x,h)$, the $s^2(x,h)$ and the $n(x,h)$ are known constants.

Example 6.7.2. Take $n = 250$, $Y_j|(X_j = x) \overset{\text{ind.}}{\sim} \mathcal{N}(\sin(x), 0.2^2)$, while $X_j \overset{\text{ind.}}{\sim} \mathcal{N}(0, 2^2)$ to simulate pairs $(X_1, Y_1), ..., (X_n, Y_n)$. The $\hat{\mu}(x)$ is plotted as a thin black line in Figure 6.10 using $h = 1$, where $y_1, ..., y_n$ are shown using blue crosses, and the $sin(x)$ is plotted using a solid red line. The code for this example is given in Section 6.9.

6.7.3 Forward neural network

In binary predictive regression, $\mu(\mathbf{x}) = \text{E}[Y|\mathbf{X} = \mathbf{x}] = P(Y = 1|\mathbf{X} = \mathbf{x})$ is all that matters. The logistic regression parameterization of this has

$$\mu(\mathbf{x}|\boldsymbol{\theta}) = \frac{e^{\mathbf{x}^\text{T}\boldsymbol{\theta}}}{1 + e^{\mathbf{x}^\text{T}\boldsymbol{\theta}}}$$

yielding intepretable parameters which are relatively simple to effectively estimate and compute using maximum likelihood or, indeed, Bayesian methods. What about more flexible parameterizations? This might be helpful in problems with massively large n, which gives us the scope to employ a large number of parameters in order to reduce potential bias.

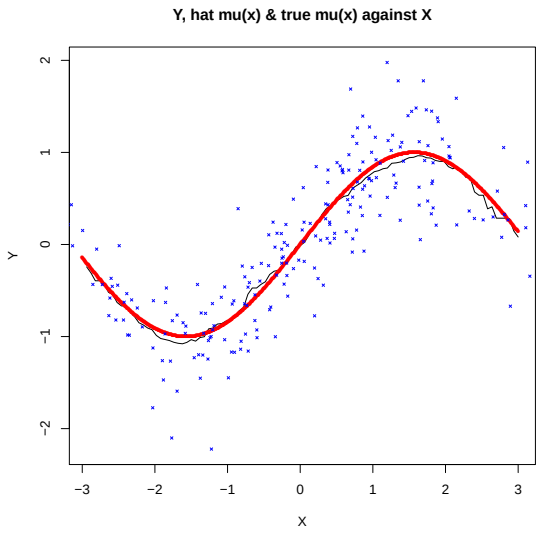


Figure 6.10: Plot of simulated $(x_1, y_1), ..., (x_n, y_n)$ (blue crosses) together with true $\mu(x)$ (red line) and $\hat{\mu}(x)$ (black line), using a kernel regression.

Write generically the logistic function as

$$g(c) = \frac{e^c}{1 + e^c}, \quad c \in R.$$

Then we repeatedly use this logistic function to parameterize particular cases of an "artificial neural network".

Definition 6.7.3 (Forward neural network)**.** Let $\mathbf{a} = (a_1, ..., a_J)^\mathsf{T}$ and $\boldsymbol{\omega} = (\omega_1, ..., \omega_J)^\mathsf{T}$. Define the parameterization, for a $K + 1$ dimensional vector of predictors (plus intercept) $\mathbf{x}$,

$$\mu(\mathbf{x}|\boldsymbol{\theta}) = g(\mathbf{a}^\mathsf{T}\boldsymbol{\omega})$$

with a single "hidden layer", based on J "hidden nodes"

$$a_j = g(\mathbf{x}^\mathsf{T}\boldsymbol{\psi}_j), \quad j = 1, ..., J,$$

index by the $(K + 1)(J + 1)$ vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\psi}_1^\mathsf{T}, ..., \boldsymbol{\psi}_J^\mathsf{T}, \boldsymbol{\omega}^\mathsf{T})^\mathsf{T}$. Then $\mu(\mathbf{x}|\boldsymbol{\theta})$ is a single layer "forward neural network", setup with a logistic "activation function". Figure 6.11 shows this structure using a graph.

This neural network generalizes to allow multiple layers, e.g., in a two layer forward neural network $\mu(\mathbf{x}|\boldsymbol{\theta}) = g(\mathbf{b}^\mathsf{T}\boldsymbol{\omega})$, where $b_j = g(\mathbf{a}^\mathsf{T}\boldsymbol{\phi}_j)$ and $a_j = g(\mathbf{x}^\mathsf{T}\boldsymbol{\psi}_j)$. There are many other types of artifical neural networks, but we will not discuss that here.

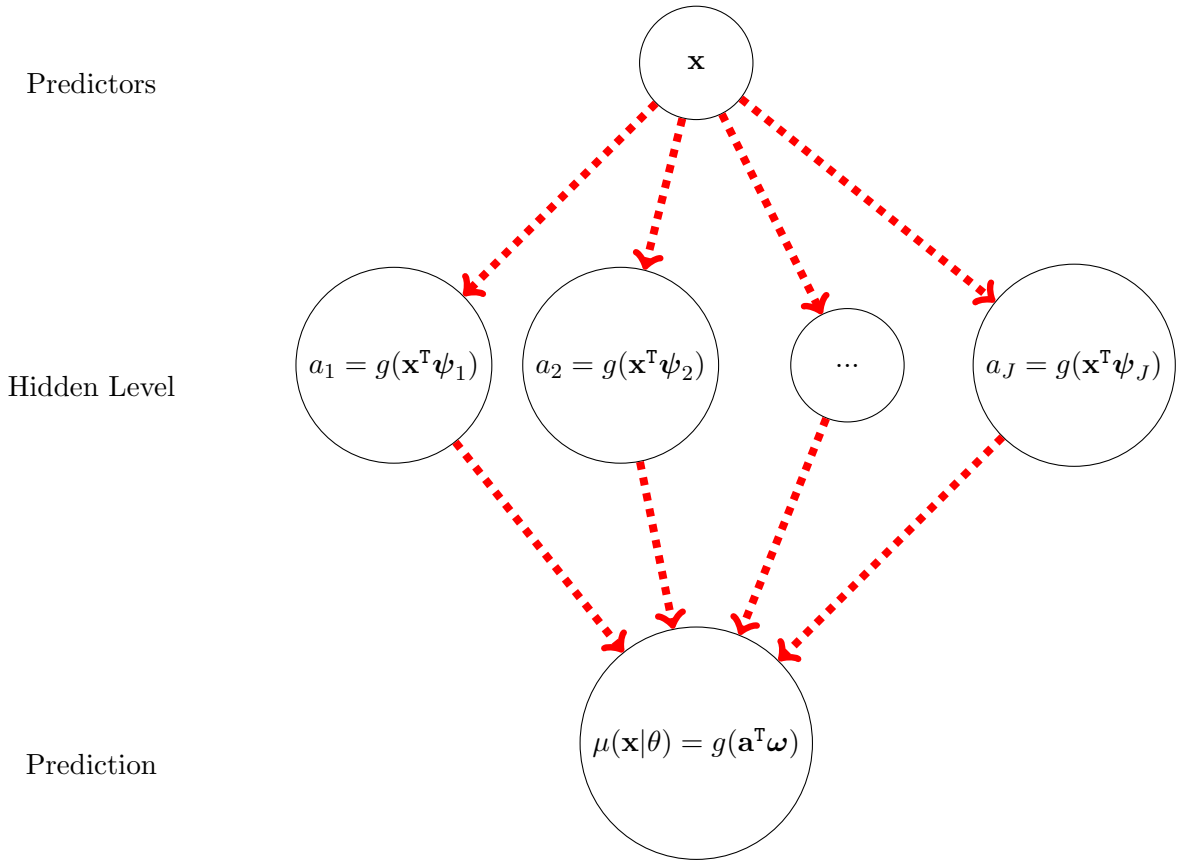


Figure 6.11: Graphical representation of $\mu(\mathbf{x}|\theta)$ built as a forward neural network with one hidden level. Dotted lines denote deterministic functions. There is nothing stochastic here!

From a statistical viewpoint, we can compute the score and information, while the likelihood could be regularized using a prior or a penalty function. Much of the challenge with this kind of model is that $\boldsymbol{\theta}$ is high dimensional and so it is can be tricky to carry out the numerical optimization of the likelihood (or regularized version) or performing numerical integration for Bayesian computations.

6.7.4 Quantile regression

In modern statistics focus is not just on $\mathrm{E}[Y|(\mathbf{X} = \mathbf{x})]$, predictive regression, but on other aspects of the predictive distribution $Y|(\mathbf{X} = \mathbf{x})$. An example of this is to model the p-quantile of the distribution of Y conditional on $\mathbf{X} = \mathbf{x}$, that is

$$Q_{Y|(\mathbf{X}=\mathbf{x})}(p), \quad p \in (0, 1).$$

To see how this works, we first need to understand a feature of quantiles, described by Theorem 6.7.5, whose proof is a small extension of the Stat 110 book proof for the median case.

Definition 6.7.4. Define the *check function* as

$$\rho_p(u) = |u|\{(1-p)I(u \leq 0) + pI(u > 0)\}.$$

The name of this loss function comes from its graph which looks like a check mark (at least for some values of p). It is drawn in Figure 6.12 for $p = 1/4$. For $p = 1/2$, the check function reduces to $|u|/2$ and then the minimization problem below is equivalent to minimizing the mean absolute error. For $p \neq 1/2$, the check function is *asymmetric*, penalizing underestimates more than overestimates (for $p < 1/2$) or vice versa (for $p > 1/2$).

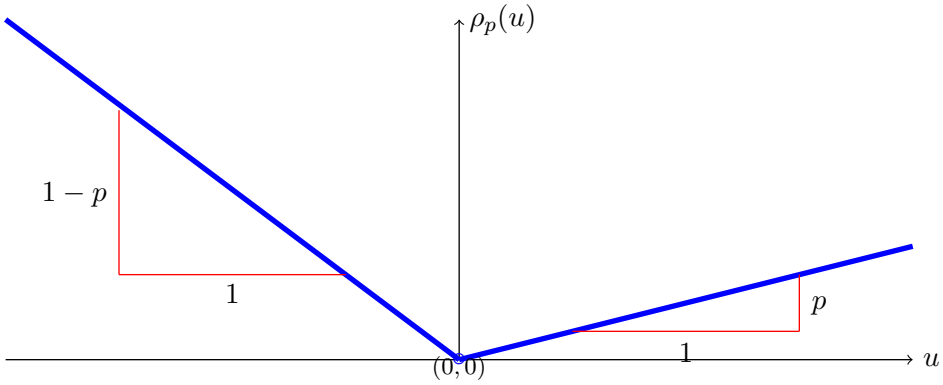


Figure 6.12: Check function $\rho_p(u)$ for p-quantile regression with $p = 1/4$.

Theorem 6.7.5. *For $p \in (0, 1)$, then*

$$Q_Y(p) = \underset{a}{\operatorname{argmin}}\ \mathrm{E}[\rho_p(Y - a)].$$

Proof. Write $q = Q_Y(p)$, and assume that $q \leq a$ (the case $q > a$ follows the same line of argument). If $Y \leq q$, then $Y \leq a$, so

$$c_a(Y) = |Y - a|I(Y < a) - |Y - q|I(Y < q) = (a - Y) - (q - Y) = a - q.$$

If $Y > q$ then (draw a picture!)

$$d_a(Y) = |Y - a|I(Y > a) - |Y - q|I(Y > q) \geq q - a.$$

Then

$$\begin{aligned}
\mathrm{E}[\rho_p(Y-a)-\rho_p(Y-q)] &= \mathrm{E}[(1-p)c_a(Y)I(Y\le q)+pd_a(Y)I(Y>q)] \\
&\ge \mathrm{E}[(1-p)(a-q)I(Y\le q)+p(q-a)I(Y>q)] \\
&= (1-p)(a-q)P(Y\le q)+p(q-a)(1-P(Y\le q)) \\
&= (a-q)\{(1-p)P(Y\le q)-p(1-P(Y\le q))\} \\
&= (a-q)\{P(Y\le q)-p\} \\
&\ge 0, \quad \text{as the definition of quantile implies } P(Y\le q)\ge p.
\end{aligned}$$

For $a=q$, $\mathrm{E}[\rho_p(Y-a)-\rho_p(Y-q)]=0$, so setting $a=q$ yields the minimum. ∎

Following a similar logic to logistic regression, *linear p-quantile regression* is defined as the parameterized model. Here we will work with a single regressor and an intercept:

$$Q_{Y|(X=x)}(p)=\theta_0+\theta_1 x,$$

where $\boldsymbol{\theta}=(\theta_0,\theta_1)'$ is the estimand. In 1978 Roger Koenker and Gilbert Bassett advocated the estimator

$$\hat{\boldsymbol{\theta}}_{KB}=\underset{(\theta_0,\theta_1)\in\mathbb{R}^2}{\arg\min}\sum_{j=1}^{n}\rho_p(Y_j-\theta_0-\theta_1 x_j),$$

motivated by Theorem 6.7.5. In practice $\hat{\boldsymbol{\theta}}_{KB}$ has to be found by numerical optimization, while its properties are studied using asymptotic methods.

The optimization can be carried out relatively reliably as minus the check function is a concave function. To see this, return to the starred Section 4.5 on the maximization of concave functions. Here we use a method which builds concave functions out of minimums.

Theorem 6.7.6. *Assume $g(x)=\min(g_1(x),...,g_n(x))$ (the pointwise minimuum), $x\in I$, and $g_1,...,g_n$ are each concave, then g is concave.*

Proof.

$$\begin{aligned}
g(px_1+(1-p)x_0) &\ge g_j(px_1+(1-p)x_0), \quad \text{for some } j\in\{1,2,...,n\} \\
&\ge pg_j(x_1)+(1-p)g_j(x_0) \\
&\ge p\min(g_1(x_1),...,g_n(x_1))+(1-p)\min(g_1(x_0),...,g_n(x_0)) \\
&\ge p\min(g_1(x_1),...,g_n(x_1))+(1-p)\min(g_1(x_1),...,g_n(x_1)) \quad = pg(x_1)+(1-p)g(x_0).
\end{aligned}$$

∎

Then the result we want is contained in the following example.

Example 6.7.7. Let $g(x) = -|x|I(x \leq 0) = \min(-x, 0)$ (which is the negative of the *hinge loss*, a loss function that often appears in classification problems in machine learning), which is the minimum of two concave functions and so q is concave by Theorem 6.7.6. For $q \in [0, 1]$, extending to minus the check function

$$g(x) = -q|x|I(x \leq 0) - (1 - q)|x|I(x > 0),$$

then g is concave by summing two concave function by Theorem 4.5.2. Applying Theorem 4.5.2 again implies

$$-\sum_{j=1}^{n} \rho_p(Y_j - \theta_0 - \theta_1 x_j)$$

is concave in θ_0, θ_1.

The details of how we carry out the optimization is beyond the scope of this book, as are the details of the asymptotics. Koenker (2005) is the classic text on quantile regression.

6.8 Recap

Regression is not one idea and we have not taught it as a single idea. Recognizing that regression covers many approaches, and placing a structure on the different methods, helps you learn about how applied scientists use the same label for many different lines of thinking.

There are predictive regressions, estimating

$$\mu(x) = \mathrm{E}[Y|(X = x)]$$

using, for example, linear, logistic and nonparametric models. There are also descriptive regressions, approximating $\mu(x)$ by the linear projection

$$\mu_L(x) = \mathrm{E}[Y] + \beta_{Y \sim X}(x - \mathrm{E}[X]). \tag{6.5}$$

Each could be extended to quantile versions and neural networks. All of these are "regressions". Some are estimated using maximum likelihood, others by least squares, or least absolute loss and others by the methods of moments. But overall, it is one of the most influential techniques in statistics, used in a myriad of application areas.

Table 6.1 provides a summary of the main ideas and notation developed in this chapter, splitting up regression into its core ideas of prediction where the predictors are conditioned on, and descriptive regression where the pair of predictors and outcomes are both regarded as random. Much of the focus

Formula or Idea	Description or name
Y	outcome
X	predictor or feature
$Y\|(X=x)$	prediction
$\mu(x)=\mathrm{E}[Y\|(X=x)]$	main summary of prediction
$\sigma^2(x)=\mathrm{Var}[Y\|(X=x)]$	heteroskedasticity
$\sigma^2=\mathrm{Var}[Y\|(X=x)]$	homoskedasticity
$\mu(x)=\theta x$	linear (in parameters) model
$\mu(x)=\frac{e^{\theta x}}{1+e^{\theta x}}$	logistic model
$\hat{\theta}=\frac{\sum_{j=1}^n x_jY_j}{\sum_{j=1}^n x_j^2}$	MLE, MoM, Least squares for linear model
$\hat{\theta}\|(\mathbf{X}=\mathbf{x})$	inference over random $Y\|X=x$
(X,Y)	Description
$\theta=\underset{a\in\mathbb{R}}{\arg\min}\,\mathrm{E}[(Y-aX)^2]=\frac{\mathrm{E}[XY]}{\mathrm{E}[X^2]}$	Estimand
$\hat{\theta}=\frac{\sum_{j=1}^n X_jY_j}{\sum_{j=1}^n X_j^2}$	MoM
$\hat{\theta}$	inference over random (X,Y)
$\sqrt{n}(\hat{\theta}-\theta)\xrightarrow{d}\mathcal{N}\left(0,\frac{\mathrm{Var}(XU)}{\mathrm{E}[X^2]^2}\right)$	asymptotic dist of $\hat{\theta}$

Table 6.1: Main ideas and notation in Chapter 6.

of the chapter has been on linear regression, specializing to a single X with no intercept.

In the linear case, a Gaussian conditional distribution $Y_j|X_j=x_j\overset{\text{ind.}}{\sim}\mathcal{N}(\theta x_j,\sigma^2)$ can allow the derivation of a simple maximum likelihood estimator

$$\hat{\theta}_{\text{MLE}}=\frac{\sum_{j=1}^n x_jY_j}{\sum_{j=1}^n x_j^2}$$

which broadly has good conditional properties like conditional unbiasedness and achieving a conditional version of the CRLB.

Instead of using MLE, the method of moments principle can be used for predictions. Assume $\mathrm{E}[Y_j|X_j=x_j]=\theta x_j$, which implies, e.g., $\mathrm{E}[x_jY_j|X_j=x_j]=\theta x_j^2$ in which case the MoM delivers

$$\hat{\theta}_{\text{MoM}}=\frac{\sum_{j=1}^n X_jY_j}{\sum_{j=1}^n X_j^2}.$$

Again, by conditioning on the predictors the properties of $\hat{\theta}_{\text{MoM}}$ are simple to find, even under heteroskedasticity.

Related to these approaches is least squares, which yields the same estimator as the MLE: this selects $\hat{\theta}_{\text{LS}}$ by minimizing $\sum_{j=1}^n(Y_j-\theta x_j)^2$ with respect to θ, then

$$\hat{\theta}_{\text{LS}}=\frac{\sum_{j=1}^n x_jY_j}{\sum_{j=1}^n x_j^2}.$$

This has the same conditional properties as the $\hat{\theta}_{\text{MoM}}$.

Note that the three estimating principles (maximum likelihood, method of moments, and least squares) deliver the same estimator

$$\hat{\theta}_{\text{MLE}} = \hat{\theta}_{\text{MoM}} = \hat{\theta}_{\text{LS}}.$$

Why? The first and third are the same as the Gaussian log-likelihood is quadratic and least squares is about minimizing a quadratic. The method of moments result is also interesting. As with all method of moments problems, we could have selected many different method of moments estimators (they just keep on coming!), but we decided to focus on the one which delivered the Gaussian maximum likelihood for this problem.

Much data is binary and in that case predictive regression is easy to carry out using a logistic regression. The MLE for this problem was discussed, as well as the relevant properties. We saw the likelihood was concave and relatively simple to maximize.

An entirely different approach to regression can be based on descriptive measures of dependence between regressors X and outcomes Y. This has $\theta = \frac{\text{E}[YX]}{\text{E}[X^2]}$, then the method of moments immediately delivers

$$\hat{\theta}_{\text{MoM}} = \frac{\sum_{j=1}^{n} X_j Y_j}{\sum_{j=1}^{n} X_j^2}.$$

The properties of $\hat{\theta}_{\text{MoM}}$ are then studied unconditionally (regarding both the regressors and the outcomes are random) using asymptotic methods. Although the descriptive approach does not require linearity of the conditional expectation, it makes strong assumptions about how the regressors behave across j, which are not needed in the predictive case.

Descriptive regression coefficient $\beta_{Y \sim X}$ drives the linear projection (6.5). This linear projection is the best affine transform of X to approximate Y in the sense that

$$\text{E}[(Y - \mu_L(X))^2] \leq \text{E}[(Y - a - bX)^2], \quad \text{for all} \quad a, b.$$

6.9 R, regression, logistic regression, and least squares

R was written by statisticians for statistics and it has, as you might expect, a wide variety of routines and packages for regression-type procedures. This section splits into three subsections. The first provides an introduction to `lm` a function for linear regression models. The second extends this to logistic regression using the function `glm`. These two subsections use simulated data to illustrate the methods. The third collects the code which produced the output used in various examples in this chapter.

6.9.1 lm: linear predictive regression

In R most regression-type procedures follow the form of lm (the name "lm" stands for "linear model"), the function for the analysis of linear in parameters predictive regression models. Hence learning about lm teaches you not only about how to run linear regressions, but sets you up nicely for some other procedures.

lm outputs a list, and there are R functions which extract information from the list, e.g., parameter estimates and standard errors, as well as regression residuals and predictions.

Here we will introduce some of this using a simulated example:

Simulate from Gaussian predictive regression

```
set.seed(111)  # simulate some data
n=400; x = rt(n,5.6); z =rnorm(n,0,1.3) # predictors
y= rnorm(n,0.2+0.4*x+0.5*z,0.8) # outputs

plot(x,y); plot(z,y) # plot some data
```

First look at the case of a single predictor x, with an intercept. R has this pretty notation for a predictive regression y ~ x, which has no reference to the intercept. If you don't want an intercept then the notation is y ~ 0 + x. The logic of the + notation will become clear in a moment.

Linear regression fit of y on x

```
> lm.Out = lm(y ~ x) # fit a linear model, store results in lm.Out
> lm.Out # produces minimal output

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
    0.20411      0.34732

> summary(lm.Out) # provides OLS estimates and SE

Call:
lm(formula = y ~ x)

Residuals:
      Min        1Q    Median        3Q       Max
-2.928147 -0.796731  0.007252  0.788227  2.775150

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 0.204114   0.054096  3.7732 0.0001857
```

```
x            0.347318   0.037807  9.1866 < 2.2e-16

Residual standard error: 1.0814 on 398 degrees of freedom
Multiple R-squared:  0.17495,   Adjusted R-squared:  0.17287
F-statistic: 84.393 on 1 and 398 DF,  p-value: < 2.22e-16

> # NOTE: SE computed under homoskedasticity
> plot(x,fitted(lm.Out),col="red") # plot predictors and fitted
> points(x,y,pch=16,cex=0.4) # plot x and y
>
```

We removed the output of "stars", which purport to discuss if results are significant at different levels.

Crucially, in `lm` the standard errors for the coefficients are computed assuming homoskedasticity. It is unfortunate that there is no switch in `lm` to compute "robust standard errors". Instead various packages can be used to carry out the relevant computation.

How about going beyond the original data, to make predictions. `lm` needs a data.frame where the new data has the same name, `x`, as the original data. Then the predictions are carried out using the `predict` function.

Prediction using new data using fitted model

```
> new.Predictors <- data.frame( # 7 predictors
+     x = rt(7,5.6)
+ )
> predict(lm.Out,new.Predictors,level = 0.9, interval = "prediction")
          fit        lwr       upr
1 -0.25834474 -2.0452537 1.5285642
2  0.50437782 -1.2816277 2.2903833
3  0.25783797 -1.5273119 2.0429879
4  0.42258938 -1.3630108 2.2081896
5  0.67933284 -1.1079441 2.4666098
6  0.77986101 -1.0083956 2.5681176
7 -0.16474690 -1.9509805 1.6214867
```

The confidence intervals for `predict` assume Gaussian prediction distribution $Y|(X = x)$, which may well not happen in practice.

One of the most useful and simple ways of using the `lm` function is a simple graphical device, this time using the `z` predictor

Plot y against x with linear regression fit imposed in plot

```
plot(z,y)
abline(lm(y ~ z),col="red")
```

which plots `y` against `z` and puts a line through the data. This is displayed in Figure 6.13.

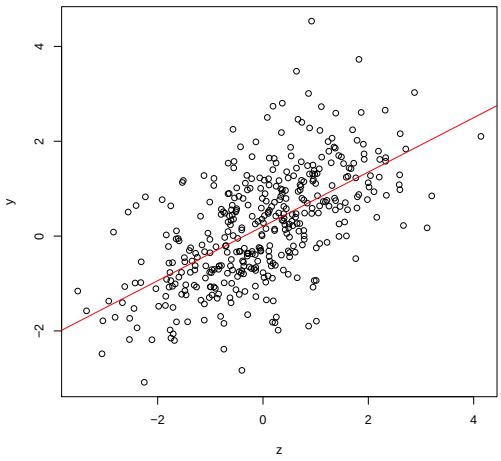


Figure 6.13: Plot of `y` data against `z` data, together with a linear fitted line

An attractive feature of `lm` is that everything generalizes to more than one variables, with no extra work, using an appealing notation.

Fitting linear regression with two predictors

```
> lm.Out = lm(y ~ x+z)
> summary(lm.Out)

Call:
lm(formula = y ~ x + z)

Residuals:
      Min         1Q     Median         3Q        Max
-2.443548 -0.541706  0.025911  0.527010  2.336119

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 0.220859   0.039995  5.5221 6.064e-08
x           0.358031   0.027951 12.8091 < 2.2e-16
z           0.581649   0.031947 18.2067 < 2.2e-16

Residual standard error: 0.79931 on 397 degrees of freedom
Multiple R-squared:  0.55037,   Adjusted R-squared:  0.54811
F-statistic: 242.98 on 2 and 397 DF,  p-value: < 2.22e-16
```

The main functions of the `lm` function are compactly shown in Table 6.2.

6.9.2 `glm`: logistic regression et al

R seamlessly extends `lm`, for linear in parameters predictive regressions, to `glm` for what are called *gen-*

Command	What it does
`lm.Out = lm(y~x+z)`	Linear model for y on x and z
`lm.Out = lm(Stock~Index,data=my.data)`	Linear model using names in my.data
`summary(lm.Out)`	Summary stats for lm.Out fit
`predict(lm.Out)`	Predictions from lm.Out for x, z
`residuals(lm.Out)`	Residuals from lm.Out fit

Table 6.2: Calling linear in parameters predictive regression model in R.

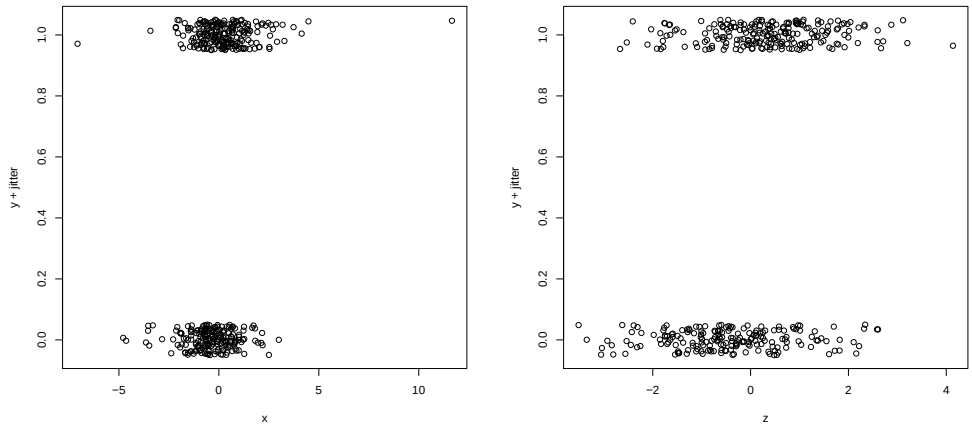


Figure 6.14: Plots of jittered y against x and y against z.

eralized linear models — basically, regression for exponential family models (e.g., Poisson, Binomial). The case we have focused on in this chapter is logistic regression and we will continue with that focus here.

Again to make everything as clear as possible we will work with a simulated dataset.

Simulate and plot logistic regression model

```
set.seed(111)  # simulate some data
n=400; x = rt(n,5.6); z = rnorm(n,0,1.3)  # predictors
lambda = 0.2+0.4*x+0.5*z
m = exp(lambda)/(1.0+exp(lambda)) # P(Y=y|X=x,Z=z)
y= rbinom(n,1,m); # outputs
jitter = -0.05 + 0.1*runif(n)

pdf("GLMFitx.pdf"); plot(x,y+jitter); dev.off() # plot some data
pdf("GLMFitz.pdf"); plot(z,y+jitter); dev.off()
```

Figure 6.14 shows the jitted y variable plotted against x and z.

The code to run logistic regression of y on x is given below, together with the output from the

code. The format of the output and the way the output is extracted has the same structure as for `lm`.

Fitting logistic regression y on x

```
> glm.Out = glm(y ~ x,family = "binomial")  # logistic regression
> glm.Out

Call:  glm(formula = y ~ x, family = "binomial")

Coefficients:
(Intercept)            x
    0.13888      0.39539

Degrees of Freedom: 399 Total (i.e. Null);  398 Residual
Null Deviance:      553.31
Residual Deviance: 529.33   AIC: 533.33
> summary(glm.Out);

Call:
glm(formula = y ~ x, family = "binomial")

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.76632  -1.15345   0.74319   1.10645   2.33413

Coefficients:
            Estimate Std. Error z value  Pr(>|z|)
(Intercept) 0.138880   0.103445  1.3425    0.1794
x           0.395394   0.088191  4.4834 7.348e-06

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 553.307  on 399  degrees of freedom
Residual deviance: 529.332  on 398  degrees of freedom
AIC: 533.332

Number of Fisher Scoring iterations: 3
```

Here numerical optimization is used to find the maximum likelihood estimates of the parameters. It converges very quickly.

The most interesting form of predictions is the fitted $P(Y = 1|X)$, and this is produced using an option within the `predict` function.

Report predictions (probabilities of a 1) for new data

```
> new.Predictors <- data.frame(x = rt(7,5.6))
> predict(glm.Out,new.Predictors,type="response")
        1         2         3         4         5         6         7
0.40429696 0.61791478 0.54984407 0.59570413 0.66371177 0.68875876 0.43019910
```

Again `glm()` generalizes to more than one variables, with no extra work.

Fitting logistic regression with two predictors

```
> glm.Out = glm(y ~ x+z, family = "binomial"); summary(glm.Out)

Call:
glm(formula = y ~ x + z, family = "binomial")

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.36286  -1.03888   0.52507   1.00765   2.05867

Coefficients:
            Estimate Std. Error z value  Pr(>|z|)
(Intercept) 0.175120   0.110295  1.5877    0.1123
x           0.464987   0.094464  4.9224 8.550e-07
z           0.636524   0.099642  6.3881 1.679e-10

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 553.307  on 399  degrees of freedom
Residual deviance: 480.316  on 397  degrees of freedom
AIC: 486.316

Number of Fisher Scoring iterations: 4
```

6.9.3 Code for various examples in this Chapter

Example 6.3.8 looks at how temperature varies by latitude. The code for this example is given below.

Code for Example 6.3.8

```
set.seed(111);
X = read.table("larry1.txt",sep=",",header=TRUE) # read in data
n = dim(X)[1] # number of cities
head(X) # check I read data
iSelect = sample(n,15, replace = FALSE, prob = NULL)
         # helps make plot pretty later
         # names some of the cities (if all, too much clutter)

x = X$Lat-mean(X$Lat); y = X$JanTemp-mean(X$JanTemp) # center

#pdf("JanTempLat.pdf")
  plot(x,y,main="Jan temp varies with latitude",
     ylab="Centered Jan temperature",
     xlab="Centered Latitude",pch=16)
```

```
  text(x[iSelect],y[iSelect],labels=X$Town[iSelect],
     cex=0.6)
  abline(a=0,b=sum(x*y)/sum(x*x),col="red"); # draw fitted reg line
#dev.off()
b= sum(x*y)/sum(x*x); u = y - b*x  # slope and residuals

#pdf("JanRes.pdf")
  plot(x,u,main="Residuals with latitude",
    ylab="Residuals",xlab="Centered Latitude",pch=16)
  text(x[iSelect],u[iSelect],labels=X$Town[iSelect],
       cex=0.6)
#dev.off()

# est, SE and robust SE
print(c(b,sqrt(var(u)/sum(x*x)),sqrt(sum(x*x*u*u)/(sum(x*x)^2))))
```

Example 6.2.15 simulates some regressors and then the corresponding binary outcomes, using a logistic regression structure. The code is below.

Code and output for Example 6.2.15

```
set.seed(111)
n=100
w = 1.0*(runif(n)>0.5)
x = w*rnorm(n,3,3) + (1.0-w)*rnorm(n,-3,2)

pi1 = exp(0.4*x)/(1.0+exp(0.4*x))
y = 1.0*(runif(n)<pi1)

#pdf("logistSim1.pdf")
  plot(x,pi1,ylim=c(0,1),xlab="x",ylab="y",
  main="Logistic regression",col="red",pch=16)
  points(x,y+(-0.02 + 0.04*runif(n)),pch=16)
#dev.off()

#pdf("logistSim.pdf")
  plot(x,pi1,ylim=c(0,1),xlab="x",ylab="y",
  main="Logistic regression",col="red",pch=16);
  points(x,y,pch=16)
#dev.off()
```

Example 6.3.9 studies how added medicare insurance varied with income in 2002.

Code and output for Example 6.3.9

```
# example from Cameron and Trivedi, Ch 14.
library(foreign)  # need to read data from Stata file
mydata <- read.dta("mus14data.dta") # read in Stata file!
```

```
n=dim(mydata)[1] # sample size
mRes = glm(mydata$private ~ mydata$hhincome,
                family = "binomial")  # fit logistic regression

summary(mRes);
#pdf("cameron.pdf")
  plot(mydata$hhincome,(mydata$private)-0.04+(runif(n,0,0.08)),pch=16,cex=0.7,
     xlab="Income per week (2002)",ylab="Have added private insurance",
     main="Medicare beneficiaries")
  points(mydata$hhincome,fitted(mRes),type="l",col="red",lwd=3)
#dev.off()
```

Example 6.5.3 uses financial data to illustrate descriptive regression. The data were downloaded from Yahoo Finance using the `R` package `Quantmod`. The data starts on 1st January 2015 and finishes 13th February 2020. The code for this example is given below. You can get more recent data by changing 2020-02-13 to a later date.

Code for Example 6.5.3

```
install.packages('quantmod') # downloads data, e.g., from Yahoo Finance
library(quantmod);

getSymbols("AAPL",from='2015-01-01',to='2020-02-13'); head(AAPL)  # Apple
getSymbols("SPY",from='2015-01-01',to='2020-02-13'); head(SPY)  # S&P500 ETF

# convert to returns and remove data frame
Y = data.matrix(100*diff(log(AAPL$AAPL.Adjusted)))[-1] # [-1] del 1st element
X = data.matrix(100*diff(log(SPY$SPY.Adjusted)))[-1]

#pdf("RegFinance.pdf");
  plot(X,Y,pch=16,main="Apple versus S&P500 daily returns",
      xlab="S&P500",ylab="Apple")
  abline(a=0,b=mean(X*Y)/mean(X*X),col="red");
#dev.off()

U = Y - (mean(X*Y)/mean(X*X))*X

# compute OLS, se under homoskedasticity and robust se
print(c(mean(X*Y)/mean(X*X),sqrt(var(U)/sum(X*X)),
      sqrt(sum(X*X*U*U)/(sum(X*X)^2))))
```

The final example is for a kernel regression, given in Example 6.7.2, which uses simulated data.

Code for Example 6.7.2

```
set.seed(111); # sets the seed of the random number generator
n = 250 # sample size
X = rnorm(n,0.0,2)
```

```
Y = rnorm(n,sin(X),0.4)

h = 1.0 # bandwidth
hhalf = h/2
muStore = rep(0,100) # store kernel regression

for (i in (1:100)){
    x = -3 + 6*i/100 # estimate at X=x
    iSel = (x-hhalf < X) & (X< x+hhalf) # X data within bandwidth
    muhat = mean(Y[iSel]) # kernel regression at x
    muStore[i] = muhat
}

plot(seq(-2.94,3,6/100),muStore,type="l",ylim=range(Y), # estimated mu(x)
     xlab="X",ylab="Y",
     main="Y, hat mu(x) & true mu(x) against X",cex=0.4)
points(seq(-3,3,0.01),sin(seq(-3,3,0.01)),col="red",pch=16,cex=0.6)
points(X,Y,col="blue",pch=4,cex=0.4) # data
```

Chapter 7

Exponential Families and Sufficiency

7.1 Natural Exponential Families

The Normal, Poisson, Gamma, Binomial, and Negative Binomial distributions are of central importance in parametric statistical modeling, and used in countless applications. These five models – and many other, less famous models – can be unified into one framework: all are examples of *natural exponential families* (NEFs).

As we will see, NEFs have many elegant, useful properties. For various results, unifying the aforementioned models allows us to provide *one* proof rather than needing to give a proof for the Normal, a separate proof for the Poisson, etc.

Perhaps more importantly from an applied statistical perspective, the common structure of these distributions allows the porting of model-building ideas and computational implementations from one application area to another. For example, linear regression, often associated with the Normal distribution, will be portable to *logistic regression* (based on the Bernoulli distribution) and to *count regression* (based on the Poisson distribution). All of these *generalized linear models* (GLMs) are based on exponential families. They play a huge role in modern statistics and machine learning.

Definition 7.1.1 (Natural exponential families)**.** A density $f(y;\theta)$ follows a *natural exponential family* (NEF) if we can write

$$f(y;\theta) = e^{\theta y - \psi(\theta)} h(y),$$

where the nonnegative function h does not depend on θ. The parameter θ is called the *natural parameter* and may be a reparameterization of how the model was originally specified.

Another way to think of a density of this form is that it factors as a function of y (not involving θ) times a function of θ (not involving y) times a function of both y and θ, where the function of both y and θ takes the simple form $e^{\theta y}$.

Example 7.1.2 (Normal as an NEF)**.** Suppose that $Y \sim \mathcal{N}(\mu, 1)$, with μ unknown. Let us show that this model is an NEF, and identify $\theta, \psi(\theta)$, and $h(y)$. Writing the PDF in NEF form,

$$f(y) = \frac{1}{\sqrt{2\pi}} e^{-(y-\mu)^2/2} = e^{\mu y - \mu^2/2} \times \frac{e^{-y^2/2}}{\sqrt{2\pi}}.$$

So we have an NEF with natural parameter $\theta = \mu$, where

$$\psi(\theta) = \frac{\theta^2}{2}, \quad h(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}.$$

The functions ψ and h are *not* uniquely determined, since we could rescale h and then adjust ψ accordingly. For example, we could have instead absorbed the Normal normalizing constant in the exponent and taken

$$\psi(\theta) = \frac{\theta^2}{2} + \frac{1}{2}\log(2\pi), \quad h(y) = e^{-y^2/2}.$$

The previous scaling has the nice feature that h is the $\mathcal{N}(0,1)$ PDF, which lets us imagine creating this NEF by starting with the standard Normal as a baseline distribution and then using it to generate an entire *family* of distributions, the $\mathcal{N}(\mu, 1)$ model. This operation, of starting from one distribution and then multiplying by an exponential function of the form $e^{\theta y - \psi(\theta)}$ to create a family of distributions, is called *exponential tilting*. With h the $\mathcal{N}(0,1)$ PDF, note that ψ is the log of the $\mathcal{N}(0,1)$ moment generating function (MGF); this is not a coincidence.

For $Y \sim \mathcal{N}(\mu, \sigma^2)$ with σ^2 known we again get an NEF, by the same calculation but with slightly more clutter due to the σ^2; then the natural parameter will be μ/σ^2 instead of just μ. If both μ and σ^2 are unknown, we still can apply exponential family techniques but need to consider the case that the parameter is multi-dimensional; in this book we focus on the case where θ is a scalar.

Example 7.1.3 (Binomial as an NEF)**.** Suppose that $Y \sim \text{Bin}(n, p)$, with n known. Let us show that this model is an NEF, and identify $\theta, \psi(\theta)$, and $h(y)$. Writing the PMF in NEF form,

$$\begin{aligned} P(Y = y; p) &= \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0, 1, 2, ..., n, \\ &= \binom{n}{y} e^{y \log p + (n-y)\log(1-p)} \\ &= \binom{n}{y} e^{y \,\text{logit}\, p + n \log(1-p)}, \end{aligned}$$

where *logit* is the following function:

$$\text{logit}\,(p) = \log\left(\frac{p}{1-p}\right).$$

The *inverse logit* (or *sigmoid*) function is

$$\text{logit}^{-1}(x) = \frac{e^x}{1+e^x}.$$

The logit and inverse logit functions are widely used in statistics and machine learning. We saw them in Section 6.2.2 on logistic regression, where the logit of $\text{E}[Y|X = x]$ was modeled as linear in parameters.

Matching the PMF up with the definition of NEF, we have that the Binomial model is an NEF with natural parameter $\theta = \text{logit}\, p$. In terms of θ, the PMF is

$$P(Y = y; \theta) = \binom{n}{y} e^{\theta y - n\log(1+e^\theta)},$$

so we can take

$$\psi(\theta) = n\log(1+e^\theta), \quad h(y) = \binom{n}{y} I(y \in \{0, 1, \ldots, n\}).$$

With a little more algebra, we could instead take h to be the $\text{Bin}(n, 1/2)$ PMF, and then ψ will turn out to be the log of the $\text{Bin}(n, 1/2)$ MGF. So the NEF construction shows how to start with a $\text{Bin}(n, 1/2)$ as a baseline distribution and then generate the entire $\text{Bin}(n, p)$ family of distributions.

Example 7.1.4 (Some famous NEFs)**.** The $\text{Pois}(\lambda)$, $\text{Gamma}(a, \lambda)$ (with a known), and $\text{NBin}(r, p)$ (with r known) models are NEFs. It follows that the Exponential model is an NEF (as it is a special case of Gamma), as is the Geometric model (as it is a special case of Negative Binomial, with the convolution parameter equal to 1).

There is a simple relationship between the mean and variance of Y from an NEF and derivatives of the function ψ.

Theorem 7.1.5 (Mean and variance in an NEF)**.** *Let Y follow the NEF $f(y; \theta) = e^{\theta y - \psi(\theta)} h(y)$. Then*

$$\text{E}[Y] = \psi'(\theta), \quad \text{Var}(Y) = \psi''(\theta).$$

Proof. We will write the proof for the case where Y is continuous; the proof for the discrete case is analogous. Densities must integrate to 1, so

$$\int_{-\infty}^{\infty} e^{\theta y} h(y) dy = e^{\psi(\theta)}.$$

By DUThIS, differentiating both sides with θ, we have

$$\int_{-\infty}^{\infty} y e^{\theta y} h(y) dy = \psi'(\theta) e^{\psi(\theta)}.$$

Therefore,

$$\psi'(\theta) = \int_{-\infty}^{\infty} y e^{\theta y - \psi(\theta)} h(y) dy = \mathrm{E}[Y],$$

by the definition of expectation. For the variance, we can DUThIS again:

$$\psi''(\theta) = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta}(y e^{\theta y - \psi(\theta)} h(y)) dy = \int_{-\infty}^{\infty} y(y - \psi'(\theta)) e^{\theta y - \psi(\theta)} h(y) dy.$$

Let $\mu = \mathrm{E}[Y] = \psi'(\theta)$. By LOTUS, we then have

$$\psi''(\theta) = \mathrm{E}[Y(Y-\mu)] = \mathrm{E}[Y^2] - \mu \mathrm{E}[Y] = \mathrm{E}[Y^2] - \mu^2 = \mathrm{Var}(Y),$$

as desired. ∎

As mentioned earlier, the choice of ψ and h is not unique, in the sense that we could rescale h and shift ψ accordingly. But adding a constant to ψ has no effect on its derivatives, so we would not get contradictory results from computing the mean and variance of Y under two different choices for ψ in a particular model.

7.1.1 NEF as a statistical model

So far we have been looking at the probabilistic properties of a single observation from an NEF. Now focus on a statistical model where the $Y_1, \ldots, Y_n$ are i.i.d. observations from an NEF with

$$f(y; \theta) = e^{\theta y - \psi(\theta)} h(y).$$

Then

$$\begin{aligned} f(\mathbf{y}; \theta) &= \prod_{j=1}^{n} e^{\theta y_j - \psi(\theta)} h(y_j), \quad \text{by independence} \\ &= \exp\left[\theta \sum_{j=1}^{n} y_j - n\psi(\theta)\right] h(y_1) \ldots h(y_n) \\ &= e^{n\{\theta \bar{y} - \psi(\theta)\}} h_{\mathbf{Y}}(\mathbf{y}), \end{aligned}$$

where $h_{\mathbf{Y}}(\mathbf{y}) = h(y_1) \ldots h(y_n)$. Note that this is very similar in form to the density of an individual observation, except with a factor of n in the exponent to account for the sample size and a sample mean $\bar{y}$ in place of the individual observation.

Maximum likelihood estimation is very convenient and well-behaved in a NEF.

Theorem 7.1.6 (MLE of an NEF)**.** *Suppose that our data are the realizations of i.i.d. $Y_1, \ldots, Y_n$ from the NEF $f(y; \theta) = e^{\theta y - \psi(\theta)} h(y)$. Let $\mu = \mathrm{E}[Y_1] = \psi'(\theta)$ be the* mean parameter *(it is a reparameterization of θ). Then the MLE of μ is its MoM estimator:*

$$\hat{\mu} = \bar{Y},$$

and so the MLE of θ (a reparameterization of μ) is uniquely given by

$$\hat{\theta} = (\psi')^{-1}(\bar{Y}).$$

The Fisher information per observation for θ is

$$\mathcal{I}_{Y_1}(\theta) = \psi''(\theta) = \mathrm{Var}(Y_1),$$

and the asymptotic distribution of the MLE of θ is

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \psi''(\theta)^{-1}).$$

The corresponding results for μ are

$$\mathcal{I}_{Y_1}(\mu) = \left(\frac{\partial\theta}{\partial\mu}\right)^2 \mathcal{I}_{Y_1}(\theta) = \frac{1}{\psi''(\theta)} = \frac{1}{\mathrm{Var}(Y_1)},$$

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, \psi''(\theta)).$$

Furthermore, the MLE of μ achieves the CRLB (with equality, not just asymptotically).

Proof. The log-likelihood function is

$$l(\theta; \mathbf{y}) = n\{\theta\bar{y} - \psi(\theta)\},$$

and the score function is

$$s(\theta; \mathbf{y}) = n\{\bar{y} - \psi'(\theta)\}.$$

Setting the score equal to 0, we have that the MLE of θ is as claimed. To check that we have found the maximum, use the second derivative test:

$$\frac{\partial}{\partial\theta} s(\theta; y) = -n\psi''(\theta) < 0,$$

since $\psi''(\theta) = \mathrm{Var}(Y) > 0$. So the log-likelihood function is concave and the Fisher information is as stated in the theorem by the information equality. The MLE is unique as the function ψ' has an inverse since it is continuous (because it is differentiable) and strictly increasing (because $\psi''(\theta) = \mathrm{Var}(Y) > 0$).

By invariance, the MLE of $\mu = \psi'(\theta)$ is

$$\hat{\mu} = \psi'(\hat{\theta}) = \psi'((\psi')^{-1}(\bar{Y})) = \bar{Y}.$$

The asymptotic properties are implied by the standard MLE properties using the Fisher information. To show that $\hat{\mu} = \bar{Y}$ achieves the CRLB, note that $\hat{\mu}$ is unbiased, with

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{n}\mathrm{Var}(Y_1) = \frac{1}{n\mathcal{I}_{Y_1}(\mu)}.$$

∎

NEFs also serve as building blocks (via transformation) for various other important models, such as the Log-Normal (obtained by exponentiating a Normal) and Weibull (obtained by raising an Exponential to a known power). An *exponential family* (EF) is obtained by transforming the variable in a natural exponential family. Every NEF is an EF, but not conversely.

Definition 7.1.7. A density $f(y;\theta)$ follows an *exponential family* (EF) if we can write

$$f(y;\theta) = e^{\theta T(y) - \psi(\theta)} g(y),$$

where g does not depend on θ.

The difference between an NEF and an EF is that the observation appears as itself in the exponent in an NEF, whereas it appears in some transformed form in the exponent in an EF. If $T(y) = y$ then we recover the definition of an NEF.

EFs, and their extension to where the natural parameter is a vector, are general enough to include many of the most widely used models in practice, but specific enough to still have nice structure and exclude strange models like the Cauchy location model

$$f(y;\theta) = \frac{1}{\pi\{1 + (y-\theta)^2\}}$$

and the Unif$(0,\theta)$ model.

7.2 Sufficient statistics

7.2.1 Data compression

Much of statistics deals in one way or another with the idea of *compressing* data, extracting the essential information from a possibly complicated dataset. Data compression of massive datasets plays a crucial role in modern technology.

Some of data compression methods are *lossless compression*, which compressing the data (so it can be transferred quickly) but then the compressed file can be restored without any loss. A ZIP file is an example of this.

In comparison, *lossy compression* degrades the information in the file, but not materially for the intended use. Examples of this are JPEG files and MP3 files, which contain image and audio data, respectively.

7.2.2 Principles

One way of quantifying what compressing a dataset means from a statistical perspective is through the notion of a *sufficient statistic*. Reducing the dataset down to a sufficient statistic is lossless.

Informally, finding a sufficient statistic asks the question: to learn about some estimand θ for a parametric statistical model, which aspects or summaries of the data are needed, and which are irrelevant?

Definition 7.2.1 (Sufficient statistic)**.** For $Y_1, \ldots, Y_n$ from the parametric statistical model $F_{\mathbf{Y};\boldsymbol{\theta}}$, a statistic $T(\mathbf{Y})$ is *sufficient* for θ if the conditional distribution of $(Y_1, \ldots, Y_n)|T$ does not depend on θ.

Here the conditional distribution of $\mathbf{Y}$ given T does not involve θ, so once we know T there is no further statistical information to be gained about θ from looking at the entire vector $(Y_1, \ldots, Y_n)$. A sufficient statistic always exists, since the full dataset is always sufficient (though useless in terms of compression).

- Sufficient statistics are not unique. In particular, it T is a one-dimensional sufficient statistic for θ then so is $g(T)$ for any one-to-one function g, e.g., if $\sum_{j=1}^n Y_j$ is sufficient then so is $n^{-1}\sum_{j=1}^n Y_j$.
- Usually we strive to find a sufficient statistic that is as low-dimensional as possible to maximize compression. The k-dimensional sufficient statistic allows us to compress our long list of n data points $Y_1, \ldots, Y_n$ into just k pieces of essential information. If the sufficient statistic is of the smallest dimension possible then it is called a *minimal sufficient statistic*.

In the rest of this subsection we will look at three special cases where we can show a particular statistic T is sufficient. In the next subsection we will give a powerful and simple recipe for finding sufficient statistics in practice.

Example 7.2.2 (Sufficiency with Bernoulli trials)**.** Let $Y_1, \ldots, Y_n$ be i.i.d. Bern(p). Then

$$T = \sum_{j=1}^n Y_j$$

is a sufficient statistic. To prove this, we can use Bayes' rule or the definition of conditional probability: for $y_1 + \cdots + y_n = t$,

$$\begin{aligned} P(\mathbf{Y} = \mathbf{y}|T = t) &= \frac{P(T = t|\mathbf{Y} = \mathbf{y})P(\mathbf{Y} = \mathbf{y})}{P(T = t)} \\ &= \frac{1 \cdot p^t(1-p)^{n-t}}{\binom{n}{t}p^t(1-p)^{n-t}} = \frac{1}{\binom{n}{t}}, \end{aligned}$$

which does not depend on p. What matters is the total *number* of success in the n trials, not exactly which particular trials were successes and which were failures. This does *not* mean that we should throw out the dataset and retain only T. For example, the full dataset is also important for

reproducibility, so that others (possibly even yourself in the future) can check through your work. But sufficiency *does* say that, assuming the model is correct, we are justified in basing our inferences on T rather than having to use all of $\mathbf{Y}$.

Example 7.2.3 (Sufficiency with independent Binomials)**.** Let $X \sim \text{Bin}(n, p)$, $Y \sim \text{Bin}(m, p)$ be independent observations and $T = X + Y$. By the Fisher exact test (from Chapter 3 of the Stat 110 book) or by direct calculation using Bayes' rule,

$$X|(T = t) \sim \text{HGeom}(n, m, t).$$

That is, the conditional distribution given T is a Hypergeometric distribution, whose parameters do not involve p. So T is sufficient.

Example 7.2.4 (Sufficiency with independent Poissons)**.** Let X, Y be i.i.d. Pois(λ). We will again show that $T = X + Y$ is a sufficient statistic. By the chicken-egg story or by direct calculation using Bayes' rule,

$$X|(T = n) \sim \text{Bin}(n, 1/2).$$

This conditional distribution does not depend on λ, so T is sufficient. On the other hand, if we generalize the model to the case where X and Y are independent but $X \sim \text{Pois}(\lambda_1)$ and $Y \sim \text{Pois}(\lambda_2)$, then T is no longer sufficient if $\lambda_1 \neq \lambda_2$. The conditional distribution given T is then

$$X|(T = n) \sim \text{Bin}\left(n, \frac{\lambda_1}{\lambda_1 + \lambda_2}\right),$$

which does depend on (λ_1, λ_2).

7.2.3 Finding sufficient statistics in practice

In the above examples we directly computed the conditional distribution given T to check whether T was sufficient, but often such calculations are difficult. Furthermore, it may be far from obvious what T to try in the first place. Fortunately, there is a simple criterion for determining whether a statistic T is sufficient. Typically this criterion is far easier to use than the definition, both in deciding what T to consider and in determining whether a particular T is sufficient.

Theorem 7.2.5 (Factorization criterion)**.** *For θ from the parametric statistical model $F_{\mathbf{Y};\boldsymbol{\theta}}$ the statistic T is sufficient if and only if we can factor*

$$f(\mathbf{y}; \theta) = g_\theta(t)h(\mathbf{y}),$$

where t is the observed value of T and the function h does not depend on θ.

Proof. We will prove the factorization criterion in the discrete case. The continuous case is analogous but more technical to prove. Suppose that T is sufficient. Then

$$f(\mathbf{y};\theta) = P(\mathbf{Y} = \mathbf{y};\theta) = P(\mathbf{Y} = \mathbf{y}, T = t;\theta),$$

since T is a deterministic function of $\mathbf{Y}$. So

$$f(\mathbf{y};\theta) = P(T = t;\theta)P(\mathbf{Y} = \mathbf{y}|T = t) = g_\theta(t)h(\mathbf{y}),$$

where $g_\theta(t) = P(T = t;\theta)$ and $h(\mathbf{y}) = P(\mathbf{Y} = \mathbf{y}|T = t)$. This is a valid choice for h since it does not depend on θ (since the conditional distribution of $\mathbf{Y}$ given T does not involve θ) and since t is a deterministic function of $\mathbf{y}$.

Conversely, suppose that

$$f(\mathbf{y};\theta) = g_\theta(t)h(\mathbf{y}).$$

Let $T = s(\mathbf{Y})$. The conditional PMF of $\mathbf{Y}|T$ is

$$P(\mathbf{Y} = \mathbf{y}|T = t;\theta) = \frac{P(\mathbf{Y} = \mathbf{y}, T = t;\theta)}{P(T = t;\theta)} = \frac{P(\mathbf{Y} = \mathbf{y};\theta)}{P(T = t;\theta)}$$

for $t = s(\mathbf{y})$, and 0 otherwise. We need to show that this conditional distribution does not depend on θ. To do so, we can expand the denominator based on all possible values of $\mathbf{Y}$ that are compatible with the observed t:

$$\frac{P(\mathbf{Y} = \mathbf{y};\theta)}{P(T = t;\theta)} = \frac{P(\mathbf{Y} = \mathbf{y};\theta)}{\sum\limits_{\tilde{\mathbf{y}}:s(\tilde{\mathbf{y}})=t} P(\mathbf{Y} = \mathbf{y};\theta)} = \frac{g_\theta(t)h(\mathbf{y})}{\sum\limits_{\tilde{\mathbf{y}}:s(\tilde{\mathbf{y}})=t} g_\theta(t)h(\tilde{\mathbf{y}})} = \frac{h(\mathbf{y})}{\sum\limits_{\tilde{\mathbf{y}}:s(\tilde{\mathbf{y}})=t} h(\tilde{\mathbf{y}})},$$

which does not depend on θ. Thus, T is a sufficient statistic for θ. ∎

Example 7.2.6 (Factorization criterion in an NEF)**.** Let the observations $Y_1, \ldots, Y_n$ be i.i.d. random variables from the NEF

$$f(y;\theta) = e^{\theta y - \psi(\theta)}h(y).$$

The joint density of $\mathbf{Y} = (Y_1, \ldots, Y_n)$ is

$$f(\mathbf{y};\theta) = e^{n(\theta\bar{y} - \psi(\theta))}h_n(\mathbf{y}).$$

This is exactly in the form needed for the factorization criterion: the $h_n(\mathbf{y})$ factor does not depend on θ, and the exponential factor depends on $\mathbf{y}$ only through the sample mean $\bar{y}$. Hence the sample mean $\bar{Y}$ is a sufficient statistic.

Example 7.2.7 (Continuing Example 4.4.3)**.** For a two-state Markov chain, using the distribution of

$$\begin{aligned}&\log P(Y_1 = y_1, ..., Y_n = y_n; \theta)\\ &= y_1 \log \pi + (1 - y_1) \log(1 - \pi) + n_{11} \log p_{11} + n_{10} \log(1 - p_{11}) + n_{01} \log p_{01} + n_{00} \log(1 - p_{01}),\end{aligned}$$

the sufficient statistics are y_1, n_{00}, n_{01}, n_{10} and n_{11}, recalling $n_{lk} = \sum_{j=2}^{n} I(y_{j-1} = l)I(y_j = k)$. Alternatively, if inference is based on the conditional distribution $P(Y_2 = y_2, ..., Y_n = y_n | Y_1 = y_1; \theta)$, not the unconditional distribution, then the sufficient statistics would be n_{00}, n_{01}, n_{10} and n_{11}.

7.2.4 Sufficient statistics and likelihoods

Sufficiency is an incredibly powerful concept for likelihood-based methods as *the data only enter the likelihood through the sufficient statistic.* As the focus is on the likelihood, which varies with θ an the data fixed, we write $g(\theta; t) = g_\theta(t)$.

Theorem 7.2.8 (Likelihood function based on a sufficient statistic)**.** *Let t be a sufficient statistic for the model $f(\mathbf{y}; \theta)$. Then the likelihood function can be expressed as a function of t (up to a multiplicative constant that does not depend on θ). In particular, knowing the sufficient statistic suffices for knowing the likelihood function, and we can take the $g(\theta; t)$ appearing in the factorization criterion as our likelihood function.*

Proof. Note that if t is sufficient then, with notation as in the factorization criterion,

$$\begin{aligned}L(\theta; \mathbf{y}) &= f(\mathbf{y}; \theta)\\ &= g(\theta; t)h(\mathbf{y}),\end{aligned}$$

where $h(\mathbf{y})$ does not depend on θ. Since for likelihood purposes we can drop multiplicative constants (including functions of the data), we can take $g(\theta; t)$ as our likelihood function. In this function, the data only appears through t. ∎

Corollary 1. *Suppose that T is a sufficient statistic for θ and that we have factored the likelihood function as*

$$L(\theta; \mathbf{y}) = cg(\theta; t).$$

Then the MLE depends upon the data only through T, and the Fisher information in the sample is

$$\mathcal{I}_{\mathbf{Y}}(\theta) = \mathrm{Var}\left[\frac{\partial \log g(\theta; T)}{\partial \theta}\right].$$

Corollary 2. *Suppose that T is a sufficient statistic for θ. Then the posterior for θ, $\pi(\theta|\mathbf{y})$, depends upon the data only through T:*

$$\pi(\theta|\mathbf{y}) = \pi(\theta|t),$$

which depends on the data only through t.

Proof. Let $\pi(\theta)$ be the prior density. Then the posterior density is

$$\pi(\theta|\mathbf{y}) \propto g(\theta; t)\pi(\theta).$$

The data enter the right-hand side only through t, and normalizing $g(\theta; t)\pi(\theta)$ so that it integrates to 1 (with respect to θ) does not introduce any additional involvement of the data. ∎

7.2.5 Sufficient statistics and Rao-Blackwellization

Another application of sufficient statistics is through the Rao-Blackwell theorem. This theorem is a general recipe for improving an estimator that is not a function of a sufficient statistic.

Theorem 7.2.9 (Rao-Blackwell)**.** *Let $\hat{\theta}$ be an estimator for θ and T be a sufficient statistic for θ for the parametric model $F_{\mathbf{Y};\theta}$. Then the* Rao-Blackwellized *estimator*

$$\hat{\theta}_{\text{RB}} = \text{E}[\hat{\theta}|T]$$

is better than or equal to $\hat{\theta}$ in MSE. The estimator $\hat{\theta}_{\text{RB}}$ is strictly better than $\hat{\theta}$ in MSE unless $\hat{\theta}$ is already a deterministic function of T.

Proof. The result follows from Adam's law and Eve's law. By Adam's law,

$$\text{E}[\hat{\theta}_{\text{RB}}] = \text{E}[\text{E}[\hat{\theta}|T]] = \text{E}[\hat{\theta}],$$

so the bias of $\hat{\theta}_{\text{RB}}$ is the same as that of $\hat{\theta}$. Therefore, if $\hat{\theta}_{\text{RB}}$ has lower variance than $\hat{\theta}$, then it also has lower MSE. To compare the variances, we can use Eve's law:

$$\text{Var}(\hat{\theta}) = \text{E}[\text{Var}(\hat{\theta}|T)] + \text{Var}(\text{E}[\hat{\theta}|T]) = \text{E}[\text{Var}(\hat{\theta}|T)] + \text{Var}(\hat{\theta}_{\text{RB}}) \geq \text{Var}(\hat{\theta}_{\text{RB}}),$$

with strict inequality unless $\text{Var}(\hat{\theta}|T) = 0$ with probability 1. If $\text{Var}(\hat{\theta}|T) = 0$ then, given T, the estimator $\hat{\theta}$ is *constant*, which means that $\hat{\theta}$ is a deterministic function of T. ∎

Example 7.2.10. Each page of a certain book has a Pois(λ) number of typos, independently. A sample of $n \geq 2$ pages is proofread extremely carefully, resulting in i.i.d. $Y_1, \ldots, Y_n \sim \text{Pois}(\lambda)$, where

Y_j is the number of typos on the jth page that was proofread. The estimand is the probability of a page having no typos, which is

$$\theta = P(Y_j = 0) = e^{-\lambda}.$$

A silly estimator for θ is obtained by only looking at one page of the book:

$$\hat{\theta} = I(Y_1 = 0).$$

This estimator is unbiased, but stupidly throws out all but one data point. Rao-Blackwell provides an automated way to improve the estimator. Since the model is an NEF, we know that $T = \sum_{j=1}^{n} Y_j$ is a sufficient statistic (equivalently, $\bar{Y}$ is a sufficient statistic). By the chicken-egg story or directly calculating the conditional distribution from Bayes' rule,

$$Y_1|(T = t) \sim \text{Bin}\left(t, \frac{1}{n}\right).$$

Hence, the Rao-Blackwellized version of $\hat{\theta}$ is

$$\hat{\theta}_{\text{RB}} = \text{E}[\hat{\theta}|T] = P(Y_1 = 0|T) = \left(1 - \frac{1}{n}\right)^{T} = \left(1 - \frac{1}{n}\right)^{n\bar{Y}}.$$

This is a much more sensible estimator than $\hat{\theta}$. For comparison's sake, note that the MLE of λ is $\bar{Y}$ (by NEF properties or direct calculation), so by invariance the MLE of θ is $\hat{\theta}_{\text{MLE}} = e^{-\bar{Y}}$. The estimators $\hat{\theta}_{\text{RB}}$ and $\hat{\theta}_{\text{MLE}}$ are more similar than they may appear to be at first glance, since for fixed x and large n we have $\left(1 + \frac{x}{n}\right)^n \approx e^x$, which, letting $x = -1$, implies that for large n we have

$$\left(1 - \frac{1}{n}\right)^{n\bar{Y}} \approx e^{-\bar{Y}}.$$

An alternative to $\hat{\theta}$ is

$$\tilde{\theta} = \frac{1}{n}\sum_{j=1}^{n} I(Y_j = 0),$$

a MoM estimator of $P(Y_j = 0)$. The $\tilde{\theta}$ uses the whole of the dataset but does not exploit the Poissonness of the data. Then the Rao-Blackwellized version

$$\tilde{\theta}_{\text{RB}} = \text{E}[\tilde{\theta}|T] = \frac{1}{n}\sum_{j=1}^{n} \text{E}[I(Y_j = 0)|T] = \frac{1}{n}\sum_{j=1}^{n}\left(1 - \frac{1}{n}\right)^{T} = \left(1 - \frac{1}{n}\right)^{n\bar{Y}},$$

again.

7.3 Recap

The natural exponential family includes many of the most common parametric models in statistics. It has a common structure which clarifies a great deal of statistical thinking and prompts new kinds of models, such as *generalized linear models* (GLMs).

The main points of the chapter are summarized in Table 7.1.

Formula or idea	Description or name
$f(y;\theta) = e^{\theta y - \psi(\theta)} h(y)$	natural exponential family (NEF)
$\mathrm{E}[Y] = \psi'(\theta), \mathrm{Var}(Y) = \psi''(\theta)$	mean and variance of NEF
MLE: $\hat{\theta} = (\psi')^{-1}(\bar{Y})$	MLE of θ in NEF
distribution of $\mathbf{Y}\|T(\mathbf{Y})$ is parameter free	then $T(\mathbf{Y})$ is a sufficient statistic
$f(\mathbf{y};\theta) = g(t;\theta)h(\mathbf{y})$	then $t = t(\mathbf{y})$ sufficient statistic: factorization theorem
$L(\theta;\mathbf{y}) \propto g(t;\theta)$	likelihood dependent on data through sufficient statistic
minimal sufficient statistic	is smallest-dimensional sufficient statistic
$\hat{\theta}_{RB} = \mathrm{E}[\hat{\theta}\|T]$	Rao-Blackwellized estimator improves MSE

Table 7.1: Main ideas and notation in Chapter 6.

Sufficiency is one way of formalizing how much data compression is possible when the goal is to carry out statistical inference on a parametric model. The factorization theorem allows the easy discovery of sufficient statistics. Finally, Rao-Blackwell theory shows how conditioning can improve the statistical properties of an initial inefficient estimator.

7.4 R

We have already introduced techniques for working with various exponential families and various sufficient statistics in R. So we take this opportunity to discuss a structure which is used extensively in R: a list.

7.4.1 Lists

A *list* is an ordered collection of objects. We will see that when models are fitted in R it typically outputs the model fit as a list: parameter estimates are one object, predictions are another, something called residuals are a separate object, etc. This will come up with models used in Chapter 7.

A list is a flexible storage container. If we have a bunch of numbers or a bunch of strings, it may well be simpler to store them as a vector. If we have a bunch of vectors *of the same length*, it

may well be simpler to store them as a matrix or data frame (with the latter offering the additional flexibility that some columns can be numerical and other columns can be strings). But with a list we can combine objects of many different types into one collection. For example, we can have numbers, strings, vectors of various lengths, functions, and even other lists together in one list.

How are lists set up? Here is an example:

Example of a list

```
a = c(31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31)
b = c("Stat 110","Stat 111")
d = 2*(1:5) - 1
f = list(a,b,d)
```

So the list `f` contains 3 objects. We used `d` rather than `c` for the third object to avoid clashing with the concatenation function `c`, but `R` *does* allow using `c` as a variable name (without breaking the concatenation function). There are only a few reserved words, such as `if` and `TRUE`, that can't be used as variable names; type `?Reserved` to get the reserved words.

To see what is in the `f` list, or lists made up by objects within `f`, we use the commands

Example of accessing objects within a list f

```
> f
[[1]]
[1] 31 28 31 30 31 30 31 31 30 31 30 31

[[2]]
[1] "Stat 110" "Stat 111"

[[3]]
[1]  1 3 5 7 9

> f[2]; # the list made up of object 2
[[1]]
[1] "Stat 110" "Stat 111"

> f[3]; # the list made up of object 3
[[1]]
[1]  1 3 5 7 9

> f[2:3]; # the list made up of objects 2 and 3
[[1]]
[1] "Stat 110" "Stat 111"

[[2]]
[1]  1 3 5 7 9
```

Crucially, `f[2]` is not `b`, the second *object* in `f`. It is a *list* with only one object, `b`. This can be confusing at first. If you type f[2][1] in the vain attempt to get the first element of the second object, you will produce

Failing at getting elements of a list

```
> f[2][1] # incorrect attempt to get 1st element of object 2
[[1]]
[1] "Stat 110" "Stat 111"
> f[2][2] # incorrect attempt to get the 2nd element of object 2
[[1]]
NULL
```

To get hold of the object within a list, *double* brackets are needed:

Accessing objects within a list

```
> f[[2]] # get object 2
[1] "Stat 110" "Stat 111"
> f[[2]][1] # 1st element of object 2
[1] "Stat 110"
```

Once you have used the double brackets to extract an object, all the usual `R` functions can be applied to the object.

Finally, list members can be named, which can help make code more readable (so whoever is reading the code does not have to look up or remember, e.g., what the interpretation of the third object in a list is). Then we can refer to a member of the list by name rather than number, as illustrated below.

Naming objects within lists

```
> a = c(31, 28, 31, 30, 31, 30, 31, 31, 30, 31, 30, 31)
> b = c("Stat 110","Stat 111")
> d = 2*(1:5) - 1
> f = list(days_in_months = a, courses = b, odd_digits = d)
> f["odd_digits"] # the list made up of object "odd_digits"
$odd_digits
[1] 1 3 5 7 9

> f[["odd_digits"]] # object "odd_digits"
[1] 1 3 5 7 9

> f$odd_digits # simpler way to get object "odd_digits"
[1] 1 3 5 7 9
```

Chapter 8

Hypothesis Testing

8.1 Introduction

Often in research, technology, and public policy we need to make a decision based on data.

Example 8.1.1 (Testing a new treatment)**.** In pharmaceutical drug trials, the treatment is usually a new drug regime and the control is the current best standard of care. The outcome is usually a particular aspects of the patient's health, e.g., HIV RNA viral load, blood pressure, or A1C. How should the researchers at the drug company report their results to the FDA (U.S. Food and Drug Administration), which regulates drugs for sale in the US, so they can objectively judge if the new treatment outperforms the standard of care?

If we used decision theory, we would specify the possible *actions* available to us, then calculate the expected *utility* (or the negative of this, the expected *loss*) of the action, averaging out the randomness and take the decision based on the action which delivers the highest expected utility. Many decisions are carried out this way, particularly when they are based on the decision makers' personal utility (e.g., deciding how to build an investment portfolio, or building a spam filter).

☣ **8.1.2.** Many decisions have major consequences for human health, funding decisions, industrial success, etc. Approving a drug or not can impact the health of many thousands of people and be a multimillion dollar decision for a drug company (e.g., in March 2019 when Biogen announced the initial failure of the potential Alzheimer's drug "aducanumab" in a phase III clinical trial, its company valuation immediately fell by over $10 billion. Having an effective treatment for Alzheimer's is a pressing social need since this devastating disease impacts millions of people.)

Researchers often passionately believe in the usefulness of their projects to society or business, even if the evidence is (initially) not very strongly on their side. Assuming all researchers will always behave

ethically is naive. Having rigorous standardized procedures can aid us in avoiding being manipulated as well as ease communication.

An alternative approach, which loses the flexibility of using a utility function but avoids the burden of having to specify a utility function, is *frequentist hypothesis testing*. This approach is very widely used throughout the natural, medical, and social sciences. It is also known as *null hypothesis significance testing* (NHST).

In frequentist hypothesis testing, we specify in advance (before collecting data) a *null hypothesis* and an *alternative hypothesis*. After getting the data, we take one of two actions: we either *reject* the null hypothesis or *retain* the null hypothesis. For brevity, we sometimes say "reject the null" or even just "reject" when we mean "reject the null hypothesis".

Typically, the null hypothesis corresponds to the status quo and the alternative hypothesis is something that would be more interesting or surprising. In the context of a new treatment that is being studied, the null hypothesis is typically that the treatment does not work (compared to a placebo or the current standard treatment) and the alternative hypothesis is that the treatment does work.

8.2 Hypotheses, tests, critical values, and power

First let us formalize the hypothesis testing framework.

Definition 8.2.1 (Statistical hypothesis)**.** Partition the parameter space Θ into two disjoint pieces: $\Theta = \Theta_0 \cup \Theta_1$. Then test

$$H_0 : \theta \in \Theta_0 \text{ versus } H_1 : \theta \in \Theta_1.$$

$H_0 : \theta \in \Theta_0$ is called the *null hypothesis* and $H_1 : \theta \in \Theta_1$ is called the *alternative hypothesis*. The null hypothesis is *simple* if it consists of a single point, say $\Theta_0 = \{\theta_0\}$; we then write the null hypothesis as $H_0 : \theta = \theta_0$. The null hypothesis is *composite* if it is not simple. Similarly, the alternative hypothesis can be simple or composite.

Definition 8.2.2 (Statistical hypothesis test)**.** A *hypothesis testing procedure*, or *test* for short, specifies which values of $\mathbf{y}$ lead to H_0 being rejected and which lead to H_0 being retained.

The outcome of a statistical hypothesis test is binary. Rejection or retention are the only possibilities. "Not sure what to do" is not allowed! We say "retain" rather than "accept" since if we retain the null hypothesis we are *not* definitively concluding that it is true; rather, we are concluding that there is not sufficient evidence to reject it. The lack of evidence could be because the null hypothesis

is true, but it could also be the case that the null hypothesis is false and our sample size was too small or we were not using the optimal test or we were just unlucky.

Hypothesis testing is carried out by specifying the region of possible values of the data where we would retain the null, and the region of possible values where we would reject the null. Once we have the data, we check which of these two regions it falls into, and then decide accordingly whether to retain or reject the null.

Definition 8.2.3 (Critical region)**.** The *retention region* A is the set of possible values of the data $\mathbf{y}$ such that we retain the null hypothesis: $\mathbf{y} \in A$. The *rejection region* or *critical region* of a hypothesis test is the complement of the set A, written A^C. So the hypothesis is rejected if $\mathbf{y} \notin A$.

Instead of working with the entire dataset $\mathbf{y}$, statisticians often base their test on a statistic $T(\mathbf{y})$, known as a *test statistic*. The rejection region is then given by $A^C = \{\mathbf{y} : T(\mathbf{y}) \in D\}$. In practice, $T(\mathbf{y})$ is often univariate and we often consider tests with a rejection region of the form $A^C = \{\mathbf{y} : T(\mathbf{y}) > c\}$ or $A^C = \{\mathbf{y} : T(\mathbf{y}) < c_L \text{ or } T(\mathbf{y}) > c_U\}$. In words, we say that such a test rejects the null for large or small values of the test statistic.

Definition 8.2.4 (Critical values)**.** Let $T(\mathbf{y})$ be a test statistic. If the rejection region A^C is of the form $\{\mathbf{y} : T(\mathbf{y}) > c\}$ or of the form $\{\mathbf{y} : T(\mathbf{y}) < c_L \text{ or } T(\mathbf{y}) > c_U\}$ for fixed numbers c, c_L, c_U, then the numbers, c, c_L, c_U are called *critical values*. The critical values provide the thresholds such that crossing the thresholds changes the outcome of the test.

One of the most important quantities to look at for a test is its *power*, which is the probability of rejecting the null hypothesis. This probability depends on θ (some values of θ are easier to reject than others), which gives us the notion of the *power function* for a test.

Definition 8.2.5 (Power function)**.** Assume the data are generated by the model $F_{\mathbf{Y};\theta}$. Suppose we have formulated our null and alternative hypotheses and selected a retention region A. The *power function* of our test is

$$\beta(\theta) = P_{\mathbf{Y};\theta}(\mathbf{Y} \notin A) = \int_{A^C} f_{\mathbf{Y};\theta}(\mathbf{y}) d\mathbf{y},$$

for $\theta \in \Theta$, recalling A^c, the rejection region, is the complement of A.

Example 8.2.6 (Power in a Normal example)**.** Assume that

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{\sigma} \sim \mathcal{N}(0, 1),$$

a distribution which is parameter-free and known — so we have a pivot! For simplicity, let σ^2 be known and focus on

$$H_0 : \theta = \theta_0 = 0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0 = 0.$$

Build a test statistic

$$T(\mathbf{Y}) = \frac{\sqrt{n}\hat{\theta}}{\sigma},$$

A physicist, Gabriella, knows that the physics community only takes empirical results seriously with extremely strong evidence. She suggests a testing procedure of rejecting H_0 if $T(\mathbf{y}) < c_L$ or $T(\mathbf{y}) > c_U$, where the critical values are $c_L = -4$ and $c_U = 4$. Then Gabriella's test has power function

$$\begin{aligned} P_{\mathbf{Y};\theta}(\mathbf{Y} \in A^C) &= P_{\mathbf{Y};\theta}(T(\mathbf{Y}) > c_U) + P_{\mathbf{Y};\theta}(T(\mathbf{Y}) < c_L) \\ &= 1 - F_{T(\mathbf{Y});\theta}(c_U) + F_{T(\mathbf{Y});\theta}(c_L). \end{aligned}$$

To calculate the power, first note that $\sqrt{n}\hat{\theta} = \sqrt{n}(\hat{\theta} - \theta) + \sqrt{n}\theta$ so that

$$T(\mathbf{Y}) \sim \mathcal{N}\left(\frac{\sqrt{n}\theta}{\sigma}, 1\right).$$

This means that,

$$\begin{aligned} F_{T(\mathbf{Y});\theta}(c_L) &= P(T(\mathbf{Y}) \leq c_L) = P(T(\mathbf{Y}) - \sqrt{n}\theta/\sigma \leq c_L - \sqrt{n}\theta/\sigma) \\ &= F_{\mathcal{N}(0,1)}\left(c_L - \frac{\sqrt{n}\theta}{\sigma}\right), \end{aligned}$$

which implies

$$P_{\mathbf{Y};\theta}(\mathbf{Y} \notin A) = 1 - F_{\mathcal{N}(0,1)}\left(c_U - \frac{\sqrt{n}\theta}{\sigma}\right) + F_{\mathcal{N}(0,1)}\left(c_L - \frac{\sqrt{n}\theta}{\sigma}\right).$$

Gabriella's experiment turns out to have $\sigma^2 = 1$. Figure 8.1 sketches the power function of her test against θ for $n = 10$, $n = 100$, and $n = 1,000$. As the sample size increases the power function rises everywhere except at $\theta = \theta_0 = 0$, the null value. This means that as the sample size increases Gabriella has more of a reasonable chance to reject finer and finer departures from her null. However, the power function rises only slowly with n as it enters the function through $\sqrt{n}$. Hence to get large power, large n and large departures from the null are useful!

Figure 8.1 looks like the power is 0 when $\theta = \theta_0 = 0$, but that is not correct. When $\theta = 0$, the power is

$$P_{\mathbf{Y};\theta}(\mathbf{Y} \notin A) = 1 - F_{\mathcal{N}(0,1)}(c_U) + F_{\mathcal{N}(0,1)}(c_L),$$

which is roughly 0.000063, which holds for all n. This is the probability of rejecting the null if the null is true. Gabriella's test is highly unlikely to make that kind of mistake, but the chance is still there — it never goes away however large the sample is. This point will become important in a moment.

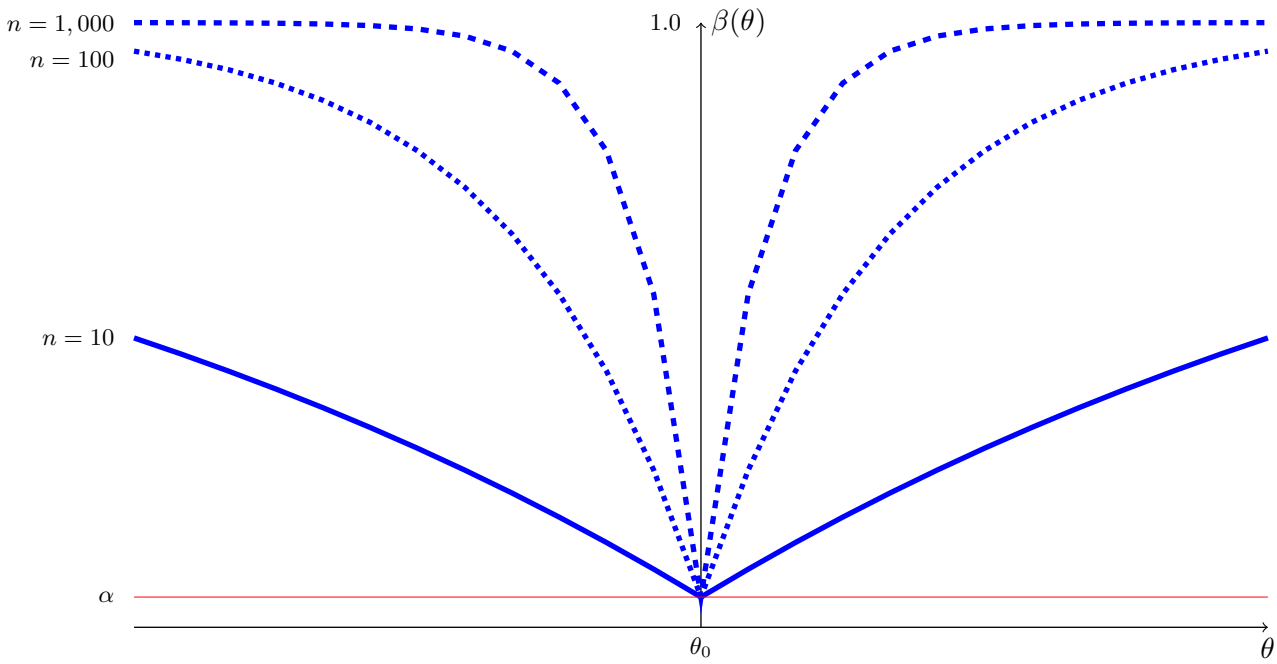


Figure 8.1: Power function $\beta(\theta)$ for a two-sided test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ as sample size n increases.

☣ **8.2.7.** Gabriella reflects the culture of physics, being very careful in their empirical claims. It is not always a good idea to be conservative in testing; thought is always needed. A technology company may regard the null and alternative basically symmetrically, making small changes to a website based on only weak evidence of improvement may make sense. Perhaps less obviously, it might make sense to approve drugs even if they are only equal to the current standard of care, for although approving medically equivalent drugs does not improve medical treatment, the availability of more treatment options may drive down the price of the drugs, in turn improving patient or societal welfare (by potentially reducing the cost of health insurance and the cost to the tax payer of Medicare and Medicaid).

We now introduce some common naming conventions in statistical hypothesis testing.

Definition 8.2.8 (One-sided tests and two-sided tests)**.** A test of the hypotheses

$$H_0 : \theta = \theta_0 \text{ versus } H_1 : \theta \neq \theta_0$$

is called a *two-sided* test. Tests of the hypotheses

$$H_0 : \theta \leq \theta_0 \text{ versus } H_1 : \theta > \theta_0$$

or

$$H_0 : \theta \geq \theta_0 \text{ versus } H_1 : \theta < \theta_0$$

are called *one-sided* tests.

Note the two-sided tests are a simple null hyopthesis against a composite alternative hypothesis, while the one-sided test is a composite against a composite.

☣ **8.2.9.** It is common to write a one-sided test as $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$ rather than $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, but this is illogical if $\Theta = \mathbb{R}$ since then the parameter space allows $\theta < \theta_0$ but neither the null hypothesis nor the alternative hypothesis allows $\theta < \theta_0$. There are a lot of confusing explanations in the literature about how to choose whether to do a one-sided test or a two-sided test, but using the null $H_0 : \theta \leq \theta_0$ rather than $H_0 : \theta = \theta_0$ for a one-sided test clarifies that for a one-sided test the goal is to test *directionality*, whereas the two-sided test is testing whether θ *equals* a pre-specified value.

Example 8.2.10. In finance, the monthly returns on individual stocks are often regressed on the monthly returns large market index, such as the S&P500. The resulting (descriptive regression) slope is called a "beta" in finance — it is our θ here. High beta stocks magnify moves in the index, so have $\theta > 1$. Some investors like holding such stocks. Suppose we wish to discover high beta stocks. Then one way of doing this is to setup the null as having a beta of 1 or less and then the alternative is bigger than one:

$$H_0 : \theta \leq 1 \text{ versus } H_1 : \theta > 1.$$

Thus here the default going into the hypothesis test is that the stock is not a high beta stock: we need empirical evidence to reject that null to say we have "discovered" a high beta stock.

8.3 Hypothesis testing errors and size

The outcome of a statistical hypothesis test is binary: retain the null or reject it. This binary structure implies, in turn, there are two types of mistakes we can make when we conduct a statistical hypothesis test. We can reject H_0 when H_0 is true (this is called a Type I error) and we can fail to reject H_0 when H_0 is false (this is called a Type II error). Writing this in terms of the critical region, we have:

Definition 8.3.1 (False positive, false negative)**.** For statistical hypothesis testing there are two types of errors: Type I error: $\theta \in \Theta_0$ but $\mathbf{y} \notin A$; Type II error: $\theta \in \Theta_1$ but $\mathbf{y} \in A$. A Type I error is also known as a *false positive* or *false discovery*, while a Type II error is also known as a *false negative*.

The probabilities of these errors, for each element of θ in Θ_0 and Θ_1, can be read off from the power function:

$$P_{\mathbf{Y};\theta}(\text{Type I error}) = P_{\mathbf{Y};\theta}(\mathbf{Y} \in A^C), \quad \text{for} \quad \theta \in \Theta_0,$$

and

$$P_{\mathbf{Y};\theta}(\text{Type II error}) = P_{\mathbf{Y};\theta}(\mathbf{Y} \in A), \quad \text{for} \quad \theta \in \Theta_1.$$

Example 8.3.2 (Continuing Example 8.2.6)**.** The null is simple $\theta = 0$, so for Gabriella's two-sided test $P_{\mathbf{Y};\theta}(\text{Type I error}) \approx 0.000063$, while

$$P_{\mathbf{Y};\theta}(\text{Type II error}) = F_{\mathcal{N}(0,1)}\left(c_U - \frac{\sqrt{n}\theta}{\sigma}\right) - F_{\mathcal{N}(0,1)}\left(c_L - \frac{\sqrt{n}\theta}{\sigma}\right),$$

which varies over θ.

For simple null hypotheses $H_0 : \theta = \theta_0$ and a particular testing procedure based on $\mathbf{Y}$, the $P_{\mathbf{Y};\theta}(\text{Type I error})$ is called the *size* or *level* of the test. The more general version of this is more complicated, but captures the same spirit.

Definition 8.3.3 (Size)**.** The *size* or *level* of the test is the maximum possible Type I error probability:

$$\alpha = \max_{\theta \in \Theta_0} \beta(\theta).$$

An α-sized test is said to be *valid* if the size is indeed α and *invalid* if not.

This general case is needed for typical one-sided tests where the null is, for example, $H_0 : \theta \leq \theta_0$ against the alternative $H_1 : \theta > \theta_1$. Figure 8.2 sketches a typical power function in this case. Here the power function monotonically increases in θ. Hence

$$\max_{\theta \in \Theta_0} \beta(\theta) = \beta(\theta_0),$$

so we can ensure such a one-sided test has α size by just checking that $\alpha = \beta(\theta_0)$.

In this chapter we will mostly be concerned with simple null hypotheses, in which case the size is just the probability of a Type I error, given the null hypothesis. The size of a hypothesis test is an important concept. Understanding it and its role allows us to avoid making the following lazy mistake.

Example 8.3.4 (Continuing Example 8.2.6)**.** Flash boy Billy has recently started work in physics. He knows a more aggressive test using $c_L = -1.96$ and $c_U = 1.96$ will drive up his statistical power, since there is a wide region on which Billy will reject the null but Gabriella will retain the null (and there are no points where the reverse is true). Billy then boasts that he is a better tester than Gabriella, who is celebrated in the physics community for her careful work. Billy's claim is absurd. Billy has a test with size approximately 0.05, while Gabriella's test has size approximately 0.000063. Indeed, it is easy to come up with a test whose power is equal to 1 for all values of θ: always reject the null hypothesis, regardless of the data. But such a test is useless. Instead, we would like to develop tests that have high power *and* have low Type I error probabilities.

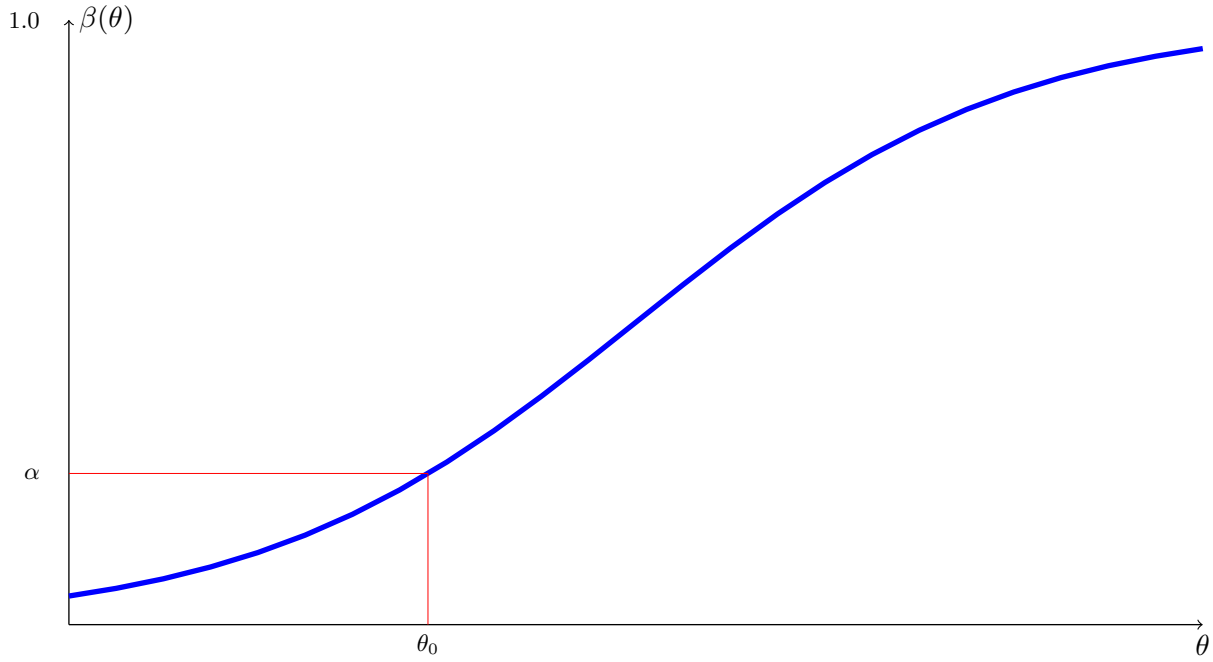


Figure 8.2: Draws $\beta(\theta)$, the power function, as θ varies for one-sided test $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$. α is the size of the test.

The tradition in statistics is to compare testing procedures with equal size, to make for a fair comparison. Choosing the size of the testing procedure depends upon the scientific, technological, or policy question at hand. The selection of the size is more of a policy decision than a statistical decision. A default setting that became common in applied statistics is to aim for a size of 5%, leading to an old joke about statisticians:

> A statistician is someone who wants to be wrong 5% of the time.

The frequentist hypothesis testing version of this joke is:

> A statistician is someone who wants to commit Type I errors 5% of the time in situations where the null hypothesis is true.

☣ **8.3.5.** Using 0.05 as the desired size has become conventional in many fields. However, it is an arbitrary convention, not a "magic number". Many data analyses in the literature put far too much emphasis on whether a statistic known as the *p-value* is less than 0.05, even though there is no principle that supports using 0.05 as a threshold and it is often not necessary to choose a threshold at all.

8.4 Calibrating the size of testing procedures

Hypothesis testing procedures can be compared across commonly sized procedures, favoring those which are most powerful under the alternative. But how can the tests be calibrated to hit a pre-specified size α?

Let us focus on a simple null hypothesis: $H_0 : \theta = \theta_0$. Then the size is

$$\beta(\theta_0) = P_{\mathbf{Y};\theta_0}(\text{Type I error}) = P_{\mathbf{Y};\theta_0}(\mathbf{Y} \in A^C).$$

A correctly sized test finds a rejection region A^C such that

$$\beta(\theta_0) = \alpha.$$

Typically, many such regions exist; selecting a good one will be context-specific.

It is often convenient to use a rejection region of the form $T(\mathbf{Y}) > c$, where $T(\mathbf{Y})$ is a continuous test statistic. In words, we say that "the test rejects for large values of $T(\mathbf{Y})$". This sounds very vague at first (how large is large?), but in fact it is a precise statement as long as α is fixed. The critical value c is whatever it needs to be in order to make $\beta(\theta_0) = \alpha$.

Furthermore, if we have a test that rejects for $T(\mathbf{Y}) > c$ but $T(\mathbf{Y})$ looks complicated, we can often find a much simpler "nugget" in the expression for $T(\mathbf{Y})$, where the nugget is a simpler statistic such that $T(\mathbf{Y})$ is a continuous, strictly increasing function of the nugget. Then the test that rejects when $T(\mathbf{Y})$ is large is the same as the test that rejects when the nugget is large, so we can just use the nugget as our test statistic. It may be much easier finding the critical value for the nugget than the critical value for $T(\mathbf{Y})$. For example, suppose that

$$T(\mathbf{Y}) = \{1 + 2\exp(\bar{Y})\}^3.$$

Then $T(\mathbf{Y})$ is a continuous, strictly increasing function of $\bar{Y}$, so using the nugget $\bar{Y}$ as our test statistic is equivalent to the original test.

Example 8.4.1 (Continuing Example 8.2.6)**.** Gabriella's two-sided test fixed c_L and c_U, which determined the size of her test. She could instead have fixed the size and let that choice determine c_L and c_U. That way of going, from size to critical region, is more easily communicated to other scientists and also allows different tests to be more easily compared. So how does it work? Recall that

$$\beta(\theta_0) = 1 - F_{\mathcal{N}(0,1)}(c_U) + F_{\mathcal{N}(0,1)}(c_L).$$

Setting this equal to α yields an infinite number of solutions for c_L, c_U. In the symmetric case where $c_U = -c_L$, there is a unique solution:

$$\alpha = 1 - F_{\mathcal{N}(0,1)}(c_U) + F_{\mathcal{N}(0,1)}(-c_U) = 2 - 2F_{\mathcal{N}(0,1)}(c_U)$$

by symmetry of the Normal, so

$$F_{\mathcal{N}(0,1)}(c_U) = 1 - \frac{\alpha}{2},$$

which yields

$$c_U = Q_{\mathcal{N}(0,1)}(F_{\mathcal{N}(0,1)}(c_U)) = Q_{\mathcal{N}(0,1)}(1 - \alpha/2).$$

Then the test rejects H_0 if and only if

$$T(\mathbf{y}) < -c_U \text{ or } T(\mathbf{y}) > c_U.$$

If $\alpha = 0.05$, 0.01, 0.001, 0.0001, 0.00001, and 0.000001 then, respectively, the critical values are $Q_{\mathcal{N}(0,1)}(1 - \alpha/2) = 1.96$, 2.58, 3.29, 3.89, 4.42, and 4.89. This delivers a correctly sized test.

☣ **8.4.2.** Although it is mathematically correct to say that a $\alpha = 0.000001$ sized test of this type needs a critical value which is 4.89, it is nerve-racking and potentially humiliating to carry out a statistical hypothesis test at that size. The size relies on the pivotal test statistic, which here is assumed to be Gaussian. Any small deviations away from Gaussianity (caused by small modeling errors or some mis-recording of the data) are most likely to be impactful in the extreme quantiles of the distribution of the pivotal test statistics.

Being wrong on the size of your test by, for example, 0.01 may not matter so much if your desired size is 0.05, but 0.01 totally dominates 0.000001, potentially opening you up to overly precise (e.g., some DNA based testing is carried out at very small α, but sometimes errors in the data collection, storage, or statistical modeling can lead to errors not encoded in the pivotal statistical test). It is better practice simply not to use extremely small sizes unless absolutely necessary. If it is necessary, then very careful statistical modeling is mandatory, as is the issuance of warnings about hard to quantify problems with the test!

8.4.1 Approximate size control

To exactly control the size of a test we need to know the distribution of the test statistic under the null. For many statistics this is too difficult, e.g. we know the asymptotic distribution of the p-quantile and MLE, but typically not their exact distribution. Hence, in many applications statisticians aim to

achieve correct size, asymptotically. That is

$$\lim_{n\to\infty} P_{\mathbf{Y};\theta_0}(\mathbf{Y} \in A^C) = \alpha,$$

as the sample size $n \to \infty$. In this situation statisticians will say the test has actual size $P_{\mathbf{Y};\theta_0}\mathbf{Y}(\mathbf{Y} \in A^C)$ and "nominally size α".

If you develop a new test with nominal size α, it would be expected that the actual size of new tests will be calculated using simulation for some realistic cases. Sometimes the asymptotics provide a great approximation, but other times it will take a very large n for the nominal size to be close to the actual size.

☣ **8.4.3.** Asymptotic approximations to the distribution of test statistics are often poor in the tails and so nominal α tests with very small α are quite likely to deviate substantially from actually having size α. Why worry about poor performance in the tails? Most asymptotics in statistics are based on Taylor approximations (including the delta method) and/or the CLT. A Taylor approximation is good near the point you are expanding around and may be terrible further away, e.g., $e^x \approx 1 + x$ is good near $x = 0$ but terrible further away (for $x < -1$ the sign is not even correct and for $x > 1$ we are comparing exponential growth to linear growth). In statistics we often do a Taylor expansion at the mean of some distribution, so should expect the approximation to be better near the mean and worse in the tails of that distribution.

An alternative to the use of asymptotic arguments is a technique known as *resampling*. We will discuss resampling in Chapter 10.

Example 8.4.4 (t-test statistic)**.** Suppose that

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{\hat{\sigma}} \xrightarrow{d} \mathcal{N}(0, 1).$$

That is, the left-hand side is an asymptotic pivot. For testing the simple null $H_0 : \theta = \theta_0$ against the composite alternative $H_1 : \theta \neq \theta_0$, define the *t-test* statistic

$$T(\mathbf{Y}) = \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\hat{\sigma}}$$

and reject the null if

$$T(\mathbf{Y}) < Q_{\mathcal{N}(0,1)}(\alpha/2) \text{ or } T > Q_{\mathcal{N}(0,1)}(1 - \alpha/2),$$

where the Gaussian quantiles are calculated under the null using the *asymptotic distribution* (giving up on finding the actual distribution of $T(\mathbf{Y})$). Then the test has nominal size of α. Most applied statistical testing problems have this form.

8.4.2 One-sided tests

Recall that a one-sided test looks at a hypothesis of the form $H_0 : \theta \leq \theta_0$ and $H_1 : \theta > \theta_0$. Suppose we wish to use a statistic $T(\mathbf{y})$ which rejects the null if $T(\mathbf{y}) > c$ where c is "large", how do we select c so the size of such a test is correct? Recall $\alpha = \max_{\theta \in \Theta_0} \beta(\theta)$, which looks complicated.

Example 8.4.5 (Controlling the size, with a Normal pivot)**.** Assume the pivot

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{\sigma} \sim \mathcal{N}(0, 1),$$

when σ is known, and define the test statistic

$$T(\mathbf{Y}) = \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sigma} \sim \mathcal{N}\left(\frac{\sqrt{n}(\theta - \theta_0)}{\sigma}, 1\right).$$

Then the power function is

$$\beta(\theta) = P_{\mathbf{Y};\theta}(T(\mathbf{Y}) > c) = 1 - P_{\mathbf{Y};\theta}(T(\mathbf{Y}) \leq c) = 1 - F_{\mathcal{N}(0,1)}\left(c - \frac{\sqrt{n}(\theta - \theta_0)}{\sigma}\right).$$

This function is strictly increasing in θ since the $\mathcal{N}(0, 1)$ CDF is a strictly increasing function. Hence controlling the probability of Type I error for $\theta = \theta_0$ delivers control of the size over the entire range of the null hypothesis space.

Thus, reject the null if

$$T(\mathbf{Y}) = \frac{\sqrt{n}(\hat{\theta} - \theta_0)}{\sigma} > Q_{\mathcal{N}(0,1)}(1 - \alpha),$$

placing all the probability α in the right-hand tail of the distribution. If $\alpha = 0.05$, 0.01, 0.001, 0.0001, 0.000005 and 0.0000005 then, respectively, $Q_{\mathcal{N}(0,1)}(1-\alpha) = 1.64$, 2.32, 3.09, 3.71, 4.26 and 4.75. This delivers a correctly sized one-sided test.

Typically we do not have such control and need to use either asymptotic arguments or resampling. In the asymptotic case we have exactly the same setup as before, controlling nominal size, rather than size.

8.4.3 t-tests

So far we have built tests using a pivot or an asymptotic pivot

$$\frac{\sqrt{n}(\hat{\theta} - \theta)}{\sigma} \sim \mathcal{N}(0, 1), \quad \text{and} \quad \frac{\sqrt{n}(\hat{\theta} - \theta)}{\hat{\sigma}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Both use the Normal distribution to manipulate the pivot to control size or nominal size.

A rather different example, with deep roots in statistical history, is similar to the case above but avoids asymptotics and only studies the sample mean. First restate a result we used in Chapter 5 on confidence intervals.

Theorem 8.4.6 (t-statistic has a t-distribution)**.** *Let* $\mathbf{Y} = (Y_1, \ldots, Y_n)$ *be i.i.d.* $\mathcal{N}(\mu, \sigma^2)$*. Then*

$$T(\mathbf{Y}) = \frac{\sqrt{n}(\bar{Y} - \mu)}{\hat{\sigma}} \sim t_{n-1},$$

where t_{n-1} *is the Student-t distribution with* $n-1$ *degrees of freedom (see Definition 10.4.4 in the Stat 110 book) and* $\hat{\sigma}^2 = (n-1)^{-1} \sum_{j=1}^{n} (Y_j - \bar{Y})^2$*.*

The key to the usefulness of this theorem is that $T(\mathbf{Y})$ is a pivotal quantity. Hence we can carry out one-sided testing of $H_0 : \mu \leq \mu_0$ versus $H_1 : \mu > \mu_0$ with a size of α, by rejecting the null for large t-statistics:

$$T(\mathbf{Y}) = \frac{\sqrt{n}(\bar{Y} - \mu_0)}{\hat{\sigma}} > Q_{t_{n-1}}(1 - \alpha).$$

That this achieves the stated size is simply a repeat of the arguments we made before in the Gaussian case. No new ideas arose once we saw the pivot. Likewise, for the two-sided test: $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ with a size of α, by rejecting the null for large absolute t-statistics:

$$T(\mathbf{Y}) < Q_{t_{n-1}}(\alpha/2) \text{ or } T > Q_{t_{n-1}}(1 - \alpha/2).$$

The challenge with using this exact result in applied statistics is that it relies on the data being exactly Gaussian. However, if n is moderately large the t-distribution converts to the Gaussian, so it offers two chances of being right. One if the data is Gaussian, another using the CLT. On the downside it would usually be unwise in applied work to entirely rely on the precise t-distribution result, again it is better to think of it as an elegant approximation whose nominal size is likely to be a bit closer to the correct size for smallish n than the Gaussian approximation. All of these points become much more pressing if the size α of the test is small, for the tails of the t are much fatter (unless n is above, say, 30) than those of the Gaussian.

8.5 Duality between hypothesis tests and confidence intervals

Hypothesis tests and confidence intervals are closely connected. Recall a $1 - \alpha$ confidence interval $C(\mathbf{y})$ has the property that for all θ,

$$P_{\mathbf{Y};\theta}(\theta \in C(\mathbf{Y})) = 1 - \alpha,$$

while a test retains the null $H_0 : \theta = \theta_0$ if $\mathbf{y} \in A(\theta_0)$ and has size α if

$$P_{\mathbf{Y};\theta_0}(\mathbf{Y} \in A(\theta_0)) = 1 - \alpha.$$

Crucially, the retention region $A(\theta_0)$ is fixed and the confidence interval $C(\mathbf{Y})$ is random.

If you state a $1-\alpha$ confidence interval, can I use it to form a size α test? If I setup a size α test, can you use it to form a $1-\alpha$ confidence interval? Here we will answer those questions in the case of a simple null hypothesis (the result holds for composite nulls, but the proof is beyond this book). First, we give a simple example to review confidence intervals.

Example 8.5.1. Think of a simple pivot

$$\frac{\hat{\theta}-\theta}{\sigma} \sim \mathcal{N}(0,1)$$

where σ is known. Recall that we retain the null $H_0 : \theta = \theta_0$ if $\mathbf{y} \in A(\theta_0)$, which can be implemented by retaining the null if

$$\hat{\theta} \in \left[\theta_0 + Q_{\mathcal{N}(0,1)}(\alpha/2)\sigma, \theta_0 + Q_{\mathcal{N}(0,1)}(1-\alpha/2)\sigma\right].$$

This is an α sized test. The corresponding $1-\alpha$ confidence interval is

$$C(\mathbf{Y}) = \left[\hat{\theta} + Q_{\mathcal{N}(0,1)}(\alpha/2)\sigma, \hat{\theta} + Q_{\mathcal{N}(0,1)}(1-\alpha/2)\sigma\right].$$

The first theorem will show that we can use a confidence interval to test.

Theorem 8.5.2 (Inverting a confidence interval)**.** *Suppose $C(\mathbf{Y})$ is a $1-\alpha$ confidence interval for θ. Retaining the null if $\theta_0 \in C(\mathbf{Y})$ is a α sized test of the null $H_0 : \theta = \theta_0$.*

Proof. The definition of a $1-\alpha$ CI is that $P_{\mathbf{Y};\theta}(\theta \in C(\mathbf{Y})) = 1-\alpha$. Specializing this result to $\theta = \theta_0$ gives the desired size for the hypothesis test. ∎

The other way around is slightly more complicated. It says find a $1-\alpha$ confidence interval by finding all the null hypothesis values of θ which are not rejected by a size α test of θ.

Theorem 8.5.3 (Inverting a test)**.** *For any θ_0, define the set $A(\theta_0)$ as the retention region (i.e., H_0 is retained if $\mathbf{y} \in A(\theta_0)$) for a α sized test of the null $H_0 : \theta = \theta_0$. Then for a fixed $\mathbf{y}$, let*

$$C(\mathbf{y}) = \{\theta : \mathbf{y} \in A(\theta)\},$$

the set of all parameter values θ which this data $\mathbf{y}$ would lead to the null being retained. Then $C(\mathbf{Y})$ is a $1-\alpha$ confidence interval for θ.

Proof. By construction $\mathbf{y} \in A(\theta)$ if and only if $\theta \in C(\mathbf{y})$. This implies

$$P_{\mathbf{Y};\theta}(\theta \in C(\mathbf{Y})) = P_{\mathbf{Y};\theta}(\mathbf{Y} \in A(\theta)).$$

But the test has size α, so this establishes the theorem. ∎

Example 8.5.4 (Continuing Example 8.5.1)**.** Suppose that $\mathbf{y} \in A(\theta)$ happens if and only if

$$\hat{\theta} \in \Big[\theta + Q_{\mathcal{N}(0,1)}(\alpha/2)\sigma, \theta + Q_{\mathcal{N}(0,1)}(1-\alpha/2)\sigma\Big].$$

Inverting this test fixes $\hat{\theta}$ and finds all the θ where this interval contains $\hat{\theta}$. Writing this out in longhand,

$$\theta + Q_{\mathcal{N}(0,1)}(\alpha/2)\sigma < \hat{\theta} < \theta + Q_{\mathcal{N}(0,1)}(1-\alpha/2)\sigma,$$

and solving for θ yields

$$\hat{\theta} - Q_{\mathcal{N}(0,1)}(1-\alpha/2)\sigma < \theta < \hat{\theta} - Q_{\mathcal{N}(0,1)}(\alpha/2)\sigma,$$

delivering a $1-\alpha$ sized $C(\mathbf{Y})$ as previously stated in Example 8.5.1 using the symmetry of the Normal distribution.

Both of us are rather skeptical of the routine use of hypothesis testing in statistics, but are fans of confidence intervals. Given they are equivalent, how can that make sense? Hypothesis testing results in a binary outcome: rejection or retention. Confidence intervals give so much more, providing the information to carry out an infinite number of hypothesis tests — telling you a whole range of θ where the null would be retained. To us confidence intervals are a more potent communication device.

8.6 Testing using likelihood-based quantities

So far we have taken the test statistics for granted, worked out how to calibrate it so the size or nominal size is correct, and calculate the power of the procedure. Here we discuss likelihood based procedures which generate new types of tests based on specifying a parametric statistical model.

As usual, we will start by writing the data as $\mathbf{y}$ and assuming a statistical model $F_{\mathbf{Y};\theta}$. To focus on statistical ideas rather than clutter we will assume θ is scalar. In this Section we return to writing $\hat{\theta}$ as the MLE. Recall from Chapter 4 that under some regularity assumptions

$$\hat{\theta} \dot{\sim} \mathcal{N}(\theta, \mathcal{I}_{\mathbf{Y}}^{-1}(\theta)),$$

where $\mathcal{I}_{\mathbf{Y}}(\theta)$ is the Fisher information in the sample.

Throughout this section, our focus will be on a two-sided test with a simple null hypothesis:

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

8.6.1 Wald test

A direct way to turn the MLE $\hat{\theta}$ into a test statistic is to use the asymptotic pivot under the null distribution

$$T(\mathbf{Y}) = \sqrt{\mathcal{I}_{\mathbf{Y}}(\theta_0)}(\hat{\theta} - \theta_0) \dot{\sim} \mathcal{N}(0,1),$$

noting that under the null the true value of θ is $\theta^* = \theta_0$ and so the information is evaluated at θ_0. This is called the *Wald test* in statistics. The approximate pivot property means that the test that rejects if

$$T(\mathbf{Y}) < Q_{\mathcal{N}(0,1)}(\alpha/2) \text{ or } T(\mathbf{Y}) > Q_{\mathcal{N}(0,1)}(1-\alpha/2)$$

has a nominal α size. Wald tests are often reported using the square of $T(\mathbf{Y})$, with rejection if

$$W(\mathbf{Y}) = \mathcal{I}_{\mathbf{Y}}(\theta_0)(\hat{\theta} - \theta_0)^2 > Q_{\chi^2_1}(1-\alpha),$$

using the fact that the square of a $\mathcal{N}(0,1)$ random variable is χ^2_1. The result is a test which has nominal size of α.

8.6.2 Score test

The score test is based on the score function evaluated at the null value θ_0:

$$s(\theta_0; \mathbf{y}) = \frac{\partial \log L(\theta_0; \mathbf{y})}{\partial \theta}.$$

The expected value of the score is 0 at the true parameter value, which is $\theta = \theta_0$ under the null. So intuitively, a very large value of $|s(\theta_0, \mathbf{y})|$ makes it seem implausible that $\theta = \theta_0$.

To calibrate a test based on the score, recall that, under the null hypothesis and some regularity conditions, $s(\theta_0; \mathbf{Y}) \dot{\sim} \mathcal{N}(0, \mathcal{I}_{\mathbf{Y}}(\theta_0))$. The *score test* uses the asymptotic pivot under the null hypothesis

$$T(\mathbf{Y}) = \frac{s(\theta_0; \mathbf{Y})}{\sqrt{\mathcal{I}_{\mathbf{Y}}(\theta_0)}} \dot{\sim} \mathcal{N}(0,1)$$

as the test statistic. The null is rejected if

$$T(\mathbf{Y}) < Q_{\mathcal{N}(0,1)}(\alpha/2) \text{ or } T(\mathbf{Y}) > Q_{\mathcal{N}(0,1)}(1-\alpha/2),$$

which gives a test with nominal α size. Score tests are often stated as the square of $T(\mathbf{Y})$, rejecting the null if

$$\frac{s(\theta_0; \mathbf{Y})^2}{\mathcal{I}_{\mathbf{Y}}(\theta_0)} > Q_{\chi^2_1}(1-\alpha).$$

Again this has nominal size of α.

8.6.3 Likelihood ratio (LR) test

To start thinking about testing directly using the values of likelihoods, assume the parameters space $\Theta = \{\theta_0, \theta_1\}$, that is it contains only two points. This abstraction looks initially bizarrely specialized, but good general ways of thinking come as a result of this initial setup.

For a hypothesis test of the simple null $H_0 : \theta = \theta_0$ against the simple alternative $H_1 : \theta = \theta_1$, the likelihood ratio

$$LR(\mathbf{y}) = \frac{L(\theta_1; \mathbf{y})}{L(\theta_0; \mathbf{y})},$$

looks interesting, rejecting the null if $LR(\mathbf{y}) > c$, where c is determined so that $P_{\mathbf{Y};\theta_0}(LR(\mathbf{Y}) > c) = \alpha$. The ratio of likelihoods is familiar, e.g. it appeared at the core of the Kullback-Leibler divergence in Chapter 4, providing a justification that the MLE is a consistent estimator.

Further, it is not difficult to implement, for under the null we know $F_{\mathbf{Y};\theta_0}$. This means that even if we do not know analytically the distribution of $LR(\mathbf{Y})$ we can use the known density of $\mathbf{Y}$ to simulate under the null R distinct i.i.d. copies $\mathbf{Y}^{(1)}, \ldots, \mathbf{Y}^{(R)}$ and so R distinct i.i.d. copies $LR(\mathbf{Y}^{(1)}), \ldots, LR(\mathbf{Y}^{(R)})$, where R is very large (e.g., $100,000$). Then we could select the smallest c so that

$$R^{-1} \sum_{j=1}^{R} I(LR(\mathbf{Y}^{(j)}) > c) = \alpha.$$

Can we justify focusing on the likelihood ratio test construction? Yes!

Theorem 8.6.1 (Neyman-Pearson lemma)**.** *For a simple null $H_0 : \theta = \theta_0$ against a simple alternative $H_1 : \theta = \theta_1$ the most powerful test, at size α, is the likelihood ratio $LR(\mathbf{y})$.*

This fundamental theorem, the proof of which is outside the scope of this book, says that in parametric testing with a simple null and simple alternative, the likelihood ratio is the best test possible. The line of argument typically extends to some one-sided tests (delivering "uniformly most powerful tests"), $H_0 : \theta = \theta_0$ against the simple alternative $H_1 : \theta > \theta_0$, but not typically to two-sided tests.

Here we explore a different kind of extension. Go back to the two-sided test problem: $H_0 : \theta = \theta_0$, versus $H_1 : \theta \neq \theta_0$. Then the *likelihood ratio* test statistic is

$$\mathrm{LR}(\mathbf{y}) = \frac{L(\hat{\theta}; \mathbf{y})}{L(\theta_0; \mathbf{y})}.$$

Note that $\mathrm{LR}(\mathbf{y})$ is always at least 1. The likelihood ratio test rejects H_0 if $\mathrm{LR}(\mathbf{y})$ is "large", which makes sense intuitively since a large value of $\mathrm{LR}(\mathbf{y})$ means that some parameter value other than θ_0 has much higher likelihood than θ_0. Like the Wald test, the LR test is based on the MLE $\hat{\theta}$.

As usual it is better to work with log-likelihoods, which we access by working with the nonnegative quantity

$$\log\{\mathrm{LR}(\mathbf{y})\} = \log L(\hat{\theta}; \mathbf{y}) - \log L(\theta_0; \mathbf{y}).$$

The asymptotic distribution of the log of the likelihood ratio is reported in this elegant theorem.

Theorem 8.6.2 (Asymptotic distribution of the log likelihood ratio)**.** *Let*

$$\Lambda(\mathbf{y}) = 2\Big[\log L(\hat{\theta}; \mathbf{y}) - \log L(\theta_0; \mathbf{y})\Big]$$

be twice the log of the likelihood ratio. Under the null $H_0 : \theta = \theta_0$ and some mild regularity conditions, rejecting the null if

$$\Lambda(\mathbf{y}) > Q_{\chi^2_1}(1 - \alpha)$$

has a nominal α size.

Proof. Under the null, $\hat{\theta} \xrightarrow{p} \theta_0$, so using a Taylor expansion of θ_0 about $\hat{\theta}$ (we do it this way to knock out a term in the expansion)

$$\log L(\theta_0; \mathbf{y}) \approx \log L(\hat{\theta}; \mathbf{y}) + (\theta_0 - \hat{\theta}) s(\hat{\theta}; \mathbf{y}) + \frac{1}{2}(\hat{\theta} - \theta_0)^2 s'(\hat{\theta}; \mathbf{y}).$$

Now for regular problems, by the definition of the MLE, $s(\hat{\theta}; \mathbf{y}) = 0$, so rearranging produces

$$\Lambda(\mathbf{y}) \approx (\hat{\theta} - \theta_0)^2 [-s'(\hat{\theta}; \mathbf{y})] = \{\sqrt{n}(\hat{\theta} - \theta_0)\}^2 \{-n^{-1} s'(\hat{\theta}; \mathbf{y})\}.$$

To be clear about our logic, think about the i.i.d. case. Then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, 1/\mathcal{I}_{Y_1}(\theta_0))$$

and, by the strong law of large numbers,

$$-n^{-1} s'(\theta_0; \mathbf{Y}) = -n^{-1} \sum_{j=1}^{n} s'(\theta_0; Y_j) \xrightarrow{p} -\mathrm{E}[s'(\theta_0; Y_1)] = \mathcal{I}_{Y_1}(\theta_0),$$

where the last equality holds by the information equality. So by the continuous mapping theorem, under the null,

$$-n^{-1} s'(\hat{\theta}; \mathbf{Y}) \xrightarrow{p} \mathcal{I}_{Y_1}(\theta_0).$$

Hence the result holds by Slutsky's Theorem. ∎

☣ **8.6.3.** The asymptotic result stated above for the log of the likelihood ratio breaks down if θ_0 is on the boundary of the parameter space. For example, suppose the parameter space is $[0, \infty)$ and we are testing $H_0 : \theta_0 = 0$ vs. $H_1 : \theta > 0$, with i.i.d. $Y_1, \ldots, Y_n \sim \mathcal{N}(\theta, 1)$. Then the MLE of θ is $\hat{\theta} = \max(\bar{Y}, 0)$. But this implies that under the null hypothesis, with probability $1/2$, $\hat{\theta} = 0$. If this happens, the likelihood ratio is 1, and the log of the likelihood ratio is 0. Thus,

$$P_{\mathbf{Y};\theta_0}(\Lambda(\mathbf{Y}) = 0) = \frac{1}{2}$$

no matter how large n is, which implies that the distribution of $\Lambda(\mathbf{Y})$ cannot converge to a χ^2_1 distribution (or any continuous distribution).

8.6.4 Relationship between the three tests

The three tests that we just introduced are linked through a single picture, Figure 8.3.

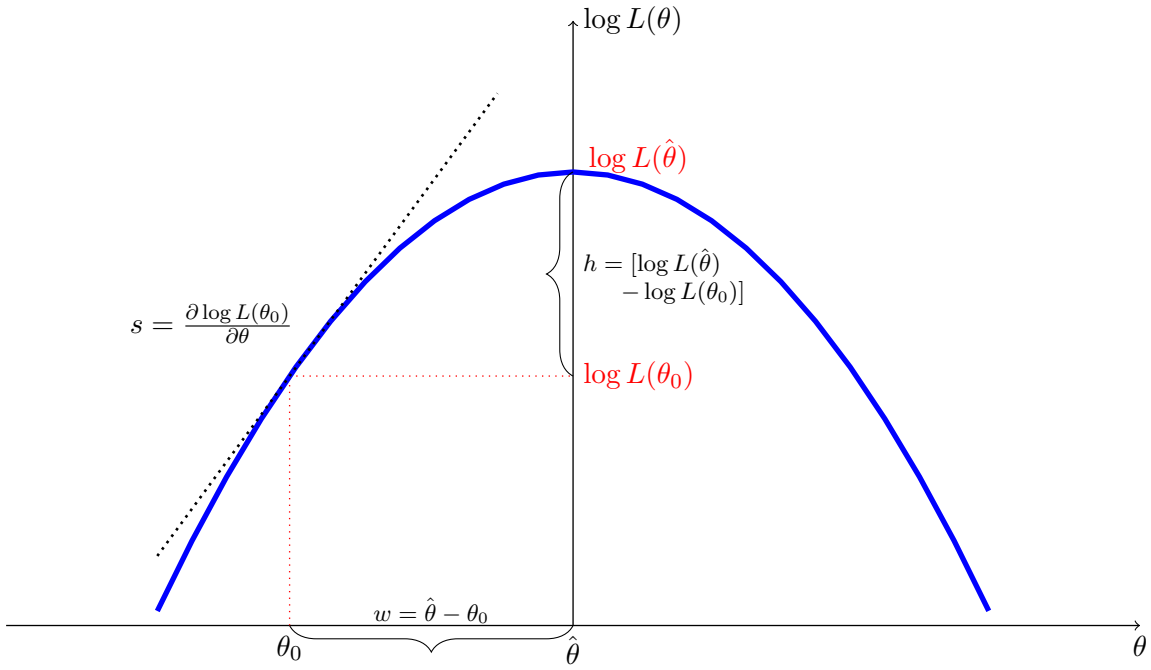


Figure 8.3: Comparison of the likelihood ratio test (based on height $h = \log L(\hat{\theta}) - \log L(\theta_0)$), score test (based on slope $s = s(\theta_0) = \partial \log L(\theta_0)/\partial\theta$) and Wald test (based on distance $w = \hat{\theta} - \theta_0$).

The log-likelihood function is depicted as a parabola for concreteness, since that is the ideal case. As labeled in the diagram, we use the:

- width $w = \hat{\theta} - \theta_0$ for the Wald test;
- height $h = \log L(\hat{\theta}; \mathbf{y}) - \log L(\theta_0; \mathbf{y}) = \log \frac{L(\hat{\theta};\mathbf{y})}{L(\theta_0;\mathbf{y})}$ for the likelihood ratio test;

- slope $s = s(\theta_0; \mathbf{y})$ for the score test.

Each of these three tests has nominal α size. The score is a relatively simple statistic, just averaging the individual data points' scores. Hence you might expect it to have its actual size to be relatively close to the nominal value. However, in practice the Fisher information is often not available in closed form, and having to estimate it might cause some additional distortion. The Wald test is usually viewed as the worst behaving of the three tests, as the MLE is not the sum of items so the asymptotic distribution can take quite a while to kick in. Additionally, the Fisher information may again have to be estimated. Fascinatingly, the likelihood ratio test is usually expected to have the size closest to the nominal value. It looks quite complicated but it does not require finding or estimating the Fisher information, and it enjoys a nice invariance property due to the invariance property of likelihood functions.

Example 8.6.4 (Wald test, score test, likelihood ratio test for Exponential data)**.** Assume a statistical model where $Y_1, \ldots, Y_n$ are i.i.d. Expo(θ), so $f_{Y_1}(y) = \theta \exp(-y\theta)$. Then

$$\log L(\theta) = n \log(\theta) - \theta n \bar{y}, \quad s(\theta) = \frac{n}{\theta} - n\bar{y}, \quad \mathcal{I}_{\mathbf{Y}}(\theta) = \frac{n}{\theta^2}, \quad \hat{\theta} = \frac{1}{\bar{y}},$$

and $\log L(\hat{\theta}) = -n\{\log(\bar{y}) + 1\}$. Then for a test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$,

$$W = n\left(\frac{1}{\theta_0 \bar{y}} - 1\right)^2, \quad S = n(1 - \theta_0 \bar{y})^2, \quad \Lambda(\mathbf{y}) = 2n\{-\log(\theta_0 \bar{y}) + (\theta_0 \bar{y} - 1)\},$$

which are all asymptotically χ^2_1 under the null. For the simple null $H_0 : \theta = \theta_0$ against the simple alternative $H_1 : \theta = \theta_1$, where $\theta_1 > \theta_0$, the optimal Neyman-Pearson type test is based on

$$\log L(\theta_1) - \log L(\theta_0) = -(\theta_1 - \theta_0) n \bar{y} + n\{\log(\theta_1) - \log(\theta_0)\},$$

which is large if $\bar{y}$ is *small*. The critical value can be determined by noting that

$$n\bar{Y} \sim \text{Gamma}(n, \theta_0)$$

(by Section 8.2 of the Stat 110 book), so the optimal level α test against $H_1 : \theta = \theta_1$ rejects if $n\bar{y} < Q_{\text{Gamma}(n,\theta_0)}(\alpha)$.

Example 8.6.5 (Wald test, score test, likelihood ratio test for Gaussian linear model)**.** Assume a statistical model where the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and model $Y_1 | X_1 = x, \theta \sim \mathcal{N}(\theta x, \sigma^2)$. Then, assuming σ^2 is known,

$$\log L(\theta) = -\frac{1}{2\sigma^2} \sum_{j=1}^{n} (y_j - x_j \theta)^2, \quad s(\theta) = \frac{1}{\sigma^2} \sum_{j=1}^{n} x_j (y_j - x_j \theta), \quad \mathcal{I}_{\mathbf{Y}}(\theta) = \frac{1}{\sigma^2} \sum_{j=1}^{n} x_j^2,$$

and

$$\hat{\theta} = \frac{\sum_{j=1}^n y_j x_j}{\sum_{j=1}^n x_j^2}.$$

For a test of $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$, we have

$$W = \frac{\hat{\theta}^2}{\sigma^2} \sum_{j=1}^n x_j^2, \quad S = W, \quad \Lambda(\mathbf{y}) = \frac{1}{\sigma^2}\left[\sum_{j=1}^n y_j^2 - \sum_{j=1}^n (y_j - x_j\hat{\theta})^2\right] = W.$$

8.7 p-values

One of the most widely used and controversial communication devices in statistics is the *p-value*.

Definition 8.7.1 (p-value for simple null)**.** For a simple null hypothesis H_0, suppose that there is a test statistic $T(\mathbf{Y})$, such that the test rejects H_0 when $T(\mathbf{Y})$ is large, i.e., there is a critical value c such that the rejection region is $\{\mathbf{y} : T(\mathbf{y}) > c\}$. Let $\mathbf{y}$ be the observed data and t be the observed value of T. Then the observed *p-value* is

$$p(\mathbf{y}) = P(T \geq t | H_0).$$

Similarly, if the test rejects H_0 when $T(\mathbf{Y})$ is small, then the observed *p-value* is

$$p(\mathbf{y}) = P(T \leq t | H_0).$$

That is, the p-value is the probability of a result *at least as extreme* as what was actually observed, assuming the null hypothesis.

In the literature, $p(\mathbf{y})$ is typically written more simply as p. But when studying p-values as statistical objects it is clearer to carry around the $\mathbf{y}$ in the notation. In applied data analysis we fix $\mathbf{y}$ to be the observed data, but in understanding the statistical properties of p-values, it is important think about the random variable, $p(\mathbf{Y})$, which is random if the data are regarded as random.

In the applied literature a very small p-value is often interpreted as very strong evidence against a null, as a formal test would need a tiny α before the null is retained. This approach has the virtue and downside of avoiding carefully selecting an α before the experiment is conducted.

Note that the p-value is a *statistic*: it is something we compute based on the data that we observe. Before we have data, the p-value is a random variable. It is then natural to wonder what the *distribution* of the p-value is, under the null. There is a simple answer to this question, at least for a continuous test statistic.

Theorem 8.7.2 (p-values are Uniform under the null)**.** *Let $T(\mathbf{Y})$ be a continuous test statistic, and suppose we are using a hypothesis test that rejects $H_0 : \theta = \theta_0$ when $T(\mathbf{y})$ is large. Then the p-value is Uniform under the null:*

$$p(\mathbf{Y}) \sim \text{Unif}(0, 1),$$

under H_0.

Proof. Let t_0 be the observed value of T. The p-value that gets computed from the data is then $P(T \geq t_0) = 1 - F_T(t_0)$, where F_T is the CDF of T under H_0. So as a random variable, the p-value is $1 - F_T(T)$. By universality of the Uniform,

$$F_T(T) \sim \text{Unif}(0, 1)$$

if H_0 is true. Since $1 - U \sim \text{Unif}(0, 1)$ for $U \sim \text{Unif}(0, 1)$, it follows that the p-value is $\text{Unif}(0, 1)$ under the null hypothesis. ∎

☣ **8.7.3.** The p-value is often confused with the probability that the null hypothesis is true or incorrectly described in ways such as "the probability that the results of the study were just due to chance". Also, results with p-values less than 0.05 are often called "statistically significant", but it is essential to keep in mind that 0.05 is an arbitrary threshold and that statistical significance does not imply scientific or practical significance. See `https://www.amstat.org/asa/files/pdfs/p-valuestatement.pdf` for a statement from the American Statistical Association about the proper use and interpretation of p-values.

☣ **8.7.4.** Companies and researchers running a hypothesis test sometimes have an incentive to obtain "significant" results, and sometimes attempt to achieve this by engaging in "p-hacking" (or "data dredging"). A simple version of this would be for a researcher to run many experiments where each time the null is true, but only report the results of ones with low p-values, pretending to offer important discoveries. This is dangerous nonsense. See `https://fivethirtyeight.com/features/science-isnt-broken/` for some interesting discussion and an interactive visualization.

In practice, most p-values are computed assuming a simple null. However, arguably simple nulls are overused for this very reason: even if a composite null would have made more scientific sense, it is easier to define and calculate p-values under a simple null. If the null is composite, then it's unclear what $P(T \geq t|H_0)$ means, since then conditioning on H_0 typically does not determine the distribution of T. To come up with a single number $P(T \geq t|H_0)$, we would need to know which specific distribution to use for T.

Instead, we will give a much more general definition of p-value, that allows the null to be composite and that does not require having a one-dimensional test statistic. A further advantage of the general definition is that it relates the notion of a p-value to the notion of an α-level.

Definition 8.7.5 (p-value for general null)**.** For a simple null hypothesis H_0, let A_α be a retention region for each α, such that the test has size α: $P_{\mathbf{Y};\theta_0}(\mathbf{Y} \in A_\alpha^C) = \alpha$. For data $\mathbf{y}$, the *p-value* is the smallest α (Type I error rate) at which we could have rejected H_0. That is,

$$p(\mathbf{y}) = \min\{\alpha : \mathbf{y} \in A_\alpha^C\}.$$

Let's check that, if the null hypothesis is simple and the test statistic T is a continuous r.v., then the two definitions of p-value are equivalent. Let F_0 be the CDF of T under the null (note that this would not even be well-defined, if the null had been composite). The critical value $c(\alpha)$ is obtained by setting

$$1 - F_0(c(\alpha)) = \alpha,$$

to make the Type I error rate equal to α. Then

$$\min(\alpha : T(\mathbf{y}) \geq c(\alpha)) = \min(\alpha : F_0(T(\mathbf{y})) \geq F_0(c(\boldsymbol{\alpha}))).$$

So the p-value with the general definition is

$$\min(\alpha : F_0(T(\mathbf{y})) \geq 1 - \alpha) = \min(\alpha : \alpha \geq 1 - F_0(T(\mathbf{y}))) = 1 - F_0(T(\mathbf{y})),$$

which agrees with the definition of the p-value for a simple null.

Example 8.7.6. Suppose the null H_0 is simple, that we have a test statistic $T(\mathbf{Y})$ that is $\mathcal{N}(0,1)$ under the null, and that the test rejects for T large. Let $t = T(\mathbf{y})$. Then the p-value is

$$p(\mathbf{y}) = P(T(\mathbf{Y}) \geq t | H_0) = 1 - F_{\mathcal{N}(0,1)}(T(\mathbf{y})).$$

If we want the test to have size α, then the test rejects when $T(\mathbf{y}) > Q_{\mathcal{N}(0,1)}(1-\alpha)$. Then rejection is the same event as $\alpha > 1 - F_{\mathcal{N}(0,1)}(T(\mathbf{y}))$, by applying the $\mathcal{N}(0,1)$ CDF to both sides. So the p-value of this test is the same under both definitions. For example, if $T(\mathbf{y}) = 0.375$, then $p \approx 0.36$. If $T(\mathbf{y}) = 1.64$, then $p \approx 0.05$. The same numerical p-value would be obtained regardless of which of the two definitions was used.

8.8 Multiparameter testing*

Suppose $\boldsymbol{\theta}$ is a $K \times 1$ vector of estimands. Assume $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ and, under the null,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, \Sigma).$$

Here we will assume that Σ is a positive definite matrix, so it can be inverted. Then if the invertible estimator $\hat{\Sigma} \xrightarrow{p} \Sigma$, then by Slutsky's Theorem,

$$\mathbf{v} = \hat{\Sigma}^{-1/2}\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(0, I_K),$$

which means that the statistic

$$T(\mathbf{Y}) = \mathbf{v}^\intercal\mathbf{v} = n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\intercal\hat{\Sigma}^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \chi^2_K,$$

so it is an asymptotic pivot and can be used calibrated to have α nominal size, rejecting if the observed $T(\mathbf{y})$ is bigger than $Q_{\chi^2_K}(1-\alpha)$.

In the context of testing parametric statistical models $f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta})$, then this is typically carried out using the asymptotic distribution of the MLE, score, and likelihood ratio. In particular, the multiparameter Wald test is

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\intercal\mathcal{I}_{\mathbf{Y}}(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \chi^2_K$$

while the score test is

$$s(\boldsymbol{\theta}_0; \mathbf{Y})^\intercal\mathcal{I}_{\mathbf{Y}}(\boldsymbol{\theta}_0)^{-1}s(\boldsymbol{\theta}_0; \mathbf{Y}) \xrightarrow{d} \chi^2_K$$

and the likelihood ratio test is

$$2\Big[\log L(\hat{\boldsymbol{\theta}}; \mathbf{Y}) - \log L(\boldsymbol{\theta}_0; \mathbf{Y})\Big] \xrightarrow{d} \chi^2_K.$$

Each distribution given here is calculated under the null hypothesis.

It turns out that in most applied testing, the testing is carried out element-by-element, $H_0 : \theta_j = \theta_{j0}$ versus, for example, $H_1 : \theta_j \neq \theta_{j0}$. In that context, for the test of the jth estimand, θ_j is often called the "parameter of interest" and $\theta_1, ..., \theta_{j-1}, \theta_{j+1}, \ldots, \theta_K$ are called "nuisance parameters". Then for the jth parameter we can use the asymptotic pivot under the null

$$T(\mathbf{Y}) = \frac{\sqrt{n}(\hat{\theta}_j - \theta_{j0})}{\hat{\Sigma}_{jj}^{1/2}} \xrightarrow{d} \mathcal{N}(0,1)$$

to calibrate the test, where we assume $\hat{\Sigma}_{jj} \xrightarrow{p} \Sigma_{jj}$, the (j,j) entry of the $K \times K$ matrix Σ.

A more general version of the element-by-element testing is to test that if $\boldsymbol{\theta}$ obeys p constraints under the null, $H_0 : R(\boldsymbol{\theta}) = 0$ against the unconstrained alternative $H_1 : R(\boldsymbol{\theta}) \neq 0$. Define the constrained ML estimator, $\tilde{\boldsymbol{\theta}}$, which computes the value of $\boldsymbol{\theta}$ which has the highest likelihood of all the values of $\boldsymbol{\theta} \in \Theta$ which obeys $R(\boldsymbol{\theta}) = 0$ (the null). Mathematically we write this as

$$\tilde{\boldsymbol{\theta}}_0 = \underset{\theta:R(\boldsymbol{\theta})=0}{\operatorname{argmax}} \log L(\boldsymbol{\theta}; y).$$

Then the likelihood ratio test of this kind of constraint has the property that, under the null,

$$2\Big[\log L(\hat{\boldsymbol{\theta}}; \mathbf{Y}) - \log L(\tilde{\boldsymbol{\theta}}_0; \mathbf{Y})\Big] \xrightarrow{d} \chi^2_p,$$

suggesting rejecting if

$$2\Big[\log L(\hat{\boldsymbol{\theta}}; \mathbf{y}) - \log L(\tilde{\boldsymbol{\theta}}_0; \mathbf{y})\Big] > Q_{\chi^2_p}(1-\alpha),$$

at nominal α size. The proof that this test has a χ^2_p asymptotic distribution under the null is largely linear algebra rather than statistical, and so we will not discuss it here. Wald and score tests can also be constructed for this type of constrained null, although they are less elegant than the LR test.

8.9 Testing when model approximates the truth*

Returning to Section 4.7, suppose once again that a researcher builds a parameterized CDF of

$$\mathbf{Y} = (Y_1, \ldots, Y_n),$$

which she denotes $G_{\mathbf{Y};\theta}$, where $\theta \in \Theta$. However, the actual data has the CDF $F_{\mathbf{Y}}$. This combination means the MLE estimates θ^*, the pseudo-true value.

How do the Wald, score, and likelihood ratio tests change if we use them to test

$$H_0 : \theta^* = \theta_0 \quad \text{versus} \quad H_1 : \theta^* \neq \theta_0?$$

First recall that for $Y_1, .., Y_n$ i.i.d. from F_{Y_1}, under some regularity conditions,

$$\sqrt{n}\left(\widehat{\theta} - \theta^*\right) \xrightarrow{d} N(0, \mathcal{J}_{Y_1}^{-2}\mathcal{I}_{Y_1}), \quad \text{where} \quad \mathcal{I}_{Y_1}(\theta^*) = \operatorname{Var}(s(\theta^*; Y_1)), \quad \mathcal{J}_{Y_1}(\theta^*) = -\mathrm{E}[s'(\theta^*; Y_1)].$$

Typically $\mathcal{I}_{Y_1}(\theta^*)$ is estimated by $\widehat{\mathcal{I}}_{Y_1}$, the sample variance of $s(\widehat{\theta}; Y_1), ..., s(\widehat{\theta}; Y_n)$, while $\mathcal{J}_{Y_1}(\theta^*)$ is estimated by $\widehat{\mathcal{J}}_{Y_1}$, the sample average of $-s'(\widehat{\theta}; Y_1), ..., -s'(\widehat{\theta}; Y_n)$.

The Wald test requires a more complicated normalization, becoming based on

$$\sqrt{n\widehat{\mathcal{J}}_{Y_1}\widehat{\mathcal{I}}_{Y_1}^{-1/2}}\left(\widehat{\theta} - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, 1),$$

under the null hypothesis. The score test becomes based on

$$\frac{1}{\sqrt{n\widehat{\mathcal{I}}_{Y_1}}} s(\theta_0; \mathbf{Y}) \xrightarrow{d} \mathcal{N}(0,1),$$

under the null, so is unaffected by the data being potentially outside the parametric model.

The likelihood ratio test becomes based on

$$2\frac{\widehat{\mathcal{J}}_{Y_1}}{\widehat{\mathcal{I}}_{Y_1}} \left\{ \log L(\widehat{\theta}) - \log L(\theta_0) \right\} \xrightarrow{d} \chi^2_1,$$

under the null hypothesis. This has an adjustment, the ratio of $\widehat{\mathcal{J}}_{Y_1}/\widehat{\mathcal{I}}_{Y_1}$.

Using a Taylor expansion, under the null where $\theta^* = \theta_0$, and again assuming the data is i.i.d. from F_{Y_1}, then

$$\begin{aligned}
\log L(\theta^*) &\simeq \log L(\widehat{\theta}) + \left(\theta^* - \widehat{\theta}\right) \frac{\partial \log L(\widehat{\theta})}{\partial \theta} + \frac{1}{2}\left(\theta^* - \widehat{\theta}\right)^2 \frac{\partial^2 \log L(\widehat{\theta})}{\partial \theta^2} \\
&= \log L(\widehat{\theta}) + \frac{1}{2}\left(\theta^* - \widehat{\theta}\right)^2 \frac{\partial^2 \log L(\widehat{\theta})}{\partial \theta^2} \\
&= \log L(\widehat{\theta}) - \frac{1}{2}\left\{\sqrt{n}\left(\theta^* - \widehat{\theta}\right)\right\}^2 \left\{ -\frac{1}{n}\frac{\partial^2 \log L(\widehat{\theta})}{\partial \theta^2} \right\} \\
&\simeq \log L(\widehat{\theta}) - \frac{1}{2}\left\{\sqrt{n}\left(\theta^* - \widehat{\theta}\right)\right\}^2 \left\{ -\frac{1}{n}\frac{\partial^2 \log L(\theta^*)}{\partial \theta^2} \right\}.
\end{aligned}$$

Now $\sqrt{n}\left(\widehat{\theta} - \theta^*\right) \xrightarrow{d} N(0, \mathcal{J}_{Y_1}^{-1}\mathcal{I}_{Y_1}\mathcal{J}_{Y_1}^{-1})$, while

$$-\frac{1}{n}\frac{\partial^2 \log L(\theta^*)}{\partial \theta^2} \xrightarrow{p} \mathcal{J}_{Y_1},$$

which yields the result by Slutsky's theorem.

8.10 Recap

Hypothesis testing forces the researcher to be clear about the null and alternative hypothesis. At the end of the test only one will be standing: the null is retained or rejected. This binary conclusion allows statisticians to compute the size and power of different testing procedures, calibrating potential tests to have given size, which should be selected according to the scientific question at hand.

The main ideas and notation for the chapter are listed in Table 8.1.

The key steps of generating hypothesis testing are: set up the null and alternative hypothesis, deciding on a test statistic, determine what size α is appropriate for the scientific problem you are working on, calilbrating the test statistic by selecting critical values so it has size control or asymptotic

size control. Compute the test statistics and report the resulting decision: retain or reject the null hypothesis.

The distribution of the test statistic under the null is often setup to be pivotal under the null or asymptotically pivotal under the null. This pivotal properties allows us to deliver nominal size control or asymptotic control.

As well as basic tests, based on using a natural estimator (or a standardized version of it) as a test statistic such as a sample mean or sample quantile, in parametric testing the three tests Wald, score, and likelihood ratio are popular.

Instead of setting α from the outset, some researchers avoid α choice by reporting not a rejection or retain decision by a p-value. This is the smallest value of α which would have lead to the rejection of the null. Very small p-values, e.g., 0.001, are often thought to represent some form of evidence against the null — noting the p-value under the null is distributed as a standard Uniform.

Unfortunately, p-values are often misinterpreted, misused through p-hacking, or used in a mechanical, unthoughtful way. They can be useful as diagnostics, as a measure of how surprised we should be to see the results we obtained if the null hypothesis is true, but we discourage making p-values the primary objects on which inferences or decisions are made.

Formula or idea	Description or name
$\Theta_0 \cup \Theta_1 = \Theta, \quad \Theta_0 \cap \Theta_1 = \emptyset$	partition of parameter space
$H_0 : \theta \in \Theta_0, \quad H_1 : \theta \in \Theta_1$	null and alternative hypotheses
$\Theta_0 = \{\theta_0\}$, i.e. single value	simple null hypothesis
composite hypothesis	not a simple hypothesis
e.g. $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$	simple null, composite alternative
e.g. $H_0 : \theta \leq \theta_0$, $H_1 : \theta > \theta_0$	one-sided hypothesese
statistical hypothesis test	reject or retain null (binary decision)
$\mathbf{y} \in A$	retention region
$\mathbf{y} \notin A^C$	rejection region
$T(\mathbf{y})$	test statistic
reject if $T(\mathbf{y}) > c$	one sided critical value, c
reject if $c_L > T(\mathbf{y}) > c_U$	two sided critical values, c_L, c_U
$\beta(\theta) = P_{\mathbf{Y};\theta}(\mathbf{Y} \notin A)$	power
type I error or false discovery or false positive	reject null, null true
type II error or false negative	retain null, null false
$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta)$	set size of test, find A or c, c_L, c_U to make equality
$\lim_{n\to\infty} \beta(\theta_0) = \alpha$	asymptotic size control
$\beta(\theta_0)$	nominal size of asymptotically controlled test
if $\theta_0 \in C(\mathbf{Y})$ retain null	duality between testing and $1-\alpha$ CI $C(\mathbf{Y})$
if $c(\mathbf{y}) = \{\theta : \mathbf{y} \in A\}$. Then $C(\mathbf{Y})$ is CI	duality between testing and $1-\alpha$ CI C($\mathbf{Y}$)
$p(\mathbf{y})$	p-value: given data, smallest α lead to rejection
$p(\mathbf{Y}) \sim U(0,1)$	p-value is uniform under null
p-hacking	fraud by reporting only favorite results from multiple tests,
$(\bar{Y} - \theta_0)/\hat{SE}(\bar{Y}) \xrightarrow{d} \mathcal{N}(0,1)$	t-statistic
$\sqrt{\mathcal{I}_{\mathbf{Y}}(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0,1)$	Wald test
$s(\theta_0; \mathbf{y})/\sqrt{\mathcal{I}_{\mathbf{Y}}(\theta_0)} \xrightarrow{d} \mathcal{N}(0,1)$	score test
$2\{\log L(\hat{\theta}) - \log L(\theta))\} \xrightarrow{d} \chi^2_1$	LR test

Table 8.1: Main ideas and notation in Chapter 8.

Chapter 9

Bayesian Inference

9.1 Introduction

In the *Bayesian* approach to statistics, for the problem we are studying we set up a full probability model for both the observable and the unobservable quantities of interest. That is, we specify a *joint* distribution for the data and the parameters. This is in contrast with the frequentist approach, for which the parameters are fixed but unknown constants, not having a distribution. We then condition on the observed data, to obtain the *posterior distribution*, which is the distribution of the unobserved quantities, given what we did observe. The name of this branch of statistics comes from the fact that Bayes' rule is normally the method used to compute the posterior distribution, but a better name might have been *full probability modeling* since the key distinction between Bayesian and non-Bayesian statistics is that in Bayesian statistics, the parameter θ of the model is given a distribution, to reflect our uncertainty about it.

The famous Bayesian statistician Dennis V. Lindley, as recounted in Christensen, Johnson, Branscum, and Hanson (2011), explained that there are only two rules for Bayesian inference:

1. *Always obey the laws of probability.*

2. *All uncertainty is to be modeled using probability.*

It is rather uncontroversial that probability is the language of uncertainty, and Bayes' rule is an uncontroversial theorem in probability, so why is there any controversy at all about the Bayesian approach? Some people object to giving θ a distribution at all, saying that it just an unknown constant and should not be modeled as a random variable. However, it would be extremely limiting to only use probability only when there is physical randomness. Most people are comfortable with statements like "There is a 50% chance of the coin landing Heads" or "There is a 30% chance that it will rain tomorrow", rather than saying the probability is 0 or 1 (we just do not know which).

If we knew enough about the coin and how it was tossed, and had vast knowledge of physics and vast computational power, we would know deterministically which way it would land, but we do not (and even a small amount of uncertainty about the initial conditions with which the coin was tossed propagates into great uncertainty about how it will land). If we knew enough about the weather conditions, and had vast knowledge of meteorology and vast computational power, we would know deterministically whether it would rain tomorrow, but we do not. Instead, we use probability to quantity our uncertainty.

A more serious objection to the Bayesian approach is that a *prior* distribution on θ must be specified. Different people may have different priors, and the choice of prior often seems very subjective. However, the choice of the *likelihood* is also often subjective and, if the prior is not too extreme and the sample size is not too small, the likelihood will usually be much more influential than the prior in determining our inferences. Still, the choice of prior is a very important issue, which deserves careful thought when applying Bayesian methods.

9.2 Prior to posterior

Definition 9.2.1 (Prior and posterior; marginal likelihood)**.** Consider a parametric model $F_{Y|\theta}$ for data $\mathbf{y}$ with parameter θ. In taking a Bayesian approach, we posit a joint distribution for the pair

$$\mathbf{Y}, \theta.$$

Then $f(\mathbf{y}; \theta) = f(\mathbf{y}|\theta)$ is the conditional distribution of $\mathbf{y}$ given θ. The *prior* for θ is the marginal distribution of θ. The *posterior* for θ is the conditional distribution of θ given $\mathbf{y}$. The prior density for θ is often denoted by $\pi(\theta)$, in which case the posterior density is $\pi(\theta|\mathbf{y})$. The *marginal likelihood* $f(\mathbf{y})$ is the marginal distribution of the data.

In general if random variables X and Z have a joint distribution, it is often useful to look at their marginal distributions, the conditional distribution of $Z|X$, and the conditional distribution of $X|Z$. Here we are working with a joint distribution for $\mathbf{Y}$ and θ, and all of the above distributions turn out to be important. Recapping the above definition, the marginal distribution of $\mathbf{Y}$ is the *marginal likelihood*, the marginal distribution of θ is the *prior*, and the conditional distribution of $\theta|\mathbf{Y}$ is the *posterior*. As for the conditional distribution of $\mathbf{Y}|\theta$, that is the data-generating distribution we've been thinking about throughout the earlier chapters. And the density of $\mathbf{Y}|\theta$, regarded as a function of θ, is our old friend the likelihood function.

We will now restate Bayes' rule as a fundamental relationship between likelihood, prior, and posterior.

Theorem 9.2.2 (Bayes' rule)**.** *Consider a parametric model $F_{Y|\theta}$ for data $\mathbf{y}$, and let $\pi(\theta)$ be the prior density on the parameter θ. Let $L(\theta;\mathbf{y}) = f(\mathbf{y}|\theta)$ be the likelihood function. Then the posterior density for θ is proportional to the likelihood times the prior:*

$$\pi(\theta|\mathbf{y}) \propto L(\theta;\mathbf{y})\pi(\theta).$$

Including the constant of proportionality, we can write this as

$$\pi(\theta|\mathbf{y}) = \frac{L(\theta;\mathbf{y})\pi(\theta)}{f(\mathbf{y})},$$

where the normalizing constant is the marginal likelihood $f(\mathbf{y})$, which is regarded as a constant here since we are considering the distribution of θ given the $\mathbf{y}$, so we are fixing $\mathbf{Y}$. The marginal likelihood can be found by the law of total probability:

$$f(\mathbf{y}) = \int_{-\infty}^{\infty} L(\tilde{\theta};\mathbf{y})\pi(\tilde{\theta})d\tilde{\theta}.$$

Example 9.2.3 (Bernoulli with Beta prior)**.** Let Y be binary, and model Y as $P(Y = 1|\theta) = \theta$. Suppose that we use the prior $\theta \sim \text{Beta}(\alpha, \beta)$ (which is by far the most common choice of prior in this setting). Then the posterior density is

$$\pi(\theta|y) \propto L(\theta;y)\pi(\theta) \propto \theta^y(1-\theta)^{1-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{\alpha+y-1}(1-\theta)^{\beta+1-y-1},$$

so

$$\theta|y \sim \text{Beta}(\alpha + y, \beta + 1 - y).$$

This is an example of *Beta-Binomial conjugacy*, which is also discussed in Story 8.3.3 of the Stat 110 book and later in this chapter.

The marginal likelihood is

$$f(y) = \text{Be}(\alpha + y, \beta + 1 - y),$$

where Be is the *beta function*, defined by

$$\text{Be}(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

for $a > 0$ and $b > 0$. But even if we did not know the normalizing constant of the Beta distribution, we could have worked up to proportionality and recognized that the posterior distribution was $\text{Beta}(\alpha + y, \beta + 1 - y)$. The left-hand side of Figure 9.1 plots this in the case where $\alpha = \beta = 2$, so the prior is symmetric about $1/2$. The posterior shifts towards 1 if $y = 1$, becoming asymmetric. The mirror

of this happens if $y = 0$. In this tiny data case, the prior plays a large role in the posterior. If the posterior is based on n independent Bernoulli trials $P(Y_j = 1|\theta) = \theta$, then the posterior density is

$$\pi(\theta|\mathbf{y}) \propto \theta^{n\bar{y}}(1-\theta)^{n(1-\bar{y})}\theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{\alpha+n\bar{y}-1}(1-\theta)^{\beta+n(1-\bar{y})-1},$$

which implies

$$\theta|\mathbf{y} \sim \text{Beta}(\alpha + n\bar{y}, \beta + n(1-\bar{y})).$$

The right hand side of Figure 9.1 shows the result when $n = 37$ and $n\bar{y} = 29$, again having $\alpha = 2$ and $\beta = 2$. The likelihood peaks, of course, at the MLE, which is $\bar{y} = 29/37$ and has much less spread. The posterior, in this richer data case, is larger determined by the shape of the likelihood, not the prior. The prior nudges the mode of the posterior slightly towards 0.5, but it is dominated by the likelihood.

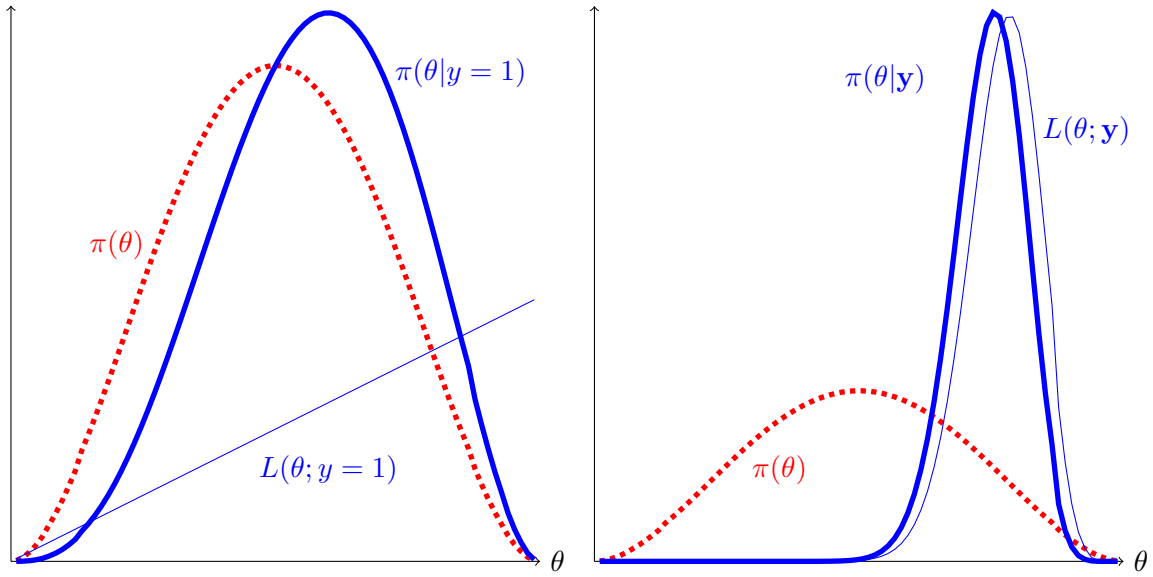


Figure 9.1: Left-hand side: Bayesian updating with a single data point y. Thin red line is the prior density. The likelihood function (thin line) and posterior density (thick line) is given in blue when $y = 1$. Right-hand side: Bayesian updating with $\mathbf{y}$, based on $n = 37$ and $\bar{y} = 29/37$. Thin red line is the prior density. The likelihood function (thin line) and posterior density (thick line).

At first glance, frequentist inference and Bayesian inference may seem diametrically opposed: the former focuses on the data-generating model $f(\mathbf{y}|\theta)$, whereas the latter focuses on the posterior distribution $\pi(\theta|\mathbf{y})$. However, the divide between Bayesian and frequentist approaches is smaller than that between model-based and design-based approaches in sampling and causal inference — which will be the focus of our last two chapters. Both Bayesians and frequentists agree that we should posit

a model $f(\mathbf{y}|\theta)$ for the data; the Bayesian goes one step further and also posits a model $\pi(\theta)$ for the parameter. This does *not* mean the Bayesian believes θ to have been randomly generated; θ is still an unknown constant, it's just that we are using probability to quantify our uncertainty about it. At first sight, Bayesian methods look fundamentally different than non-parametric methods (e.g., kernel density estimators), but this is not the case — there are Bayesian approaches to estimating infinite dimensional parameters, such as density functions. Bayesian nonparametrics are, however, beyond the scope of this book.

Brad Efron (2013) described a "250-year argument" between Bayesians and frequentists, and indeed there have been many bitter debates between these camps, but increasingly many statisticians are pragmatic about the debate, using whichever approach seems more well-suited to a particular problem. Here are two bridges toward a reconciliation between Bayesian and frequentist thinking:

- In a sense, there is no inherent conflict between Bayesian and frequentist thinking, since Bayesian thinking is an approach to *constructing* a procedure (by building a full probability model and finding the posterior distribution) and frequentist thinking is an approach to *evaluating* a procedure (by assessing the long-run performance of a procedure, under repeated sampling).

  For example, we can think like a Bayesian to come up with a point estimator and then like a frequentist to evaluate its MSE, or we can think like a Bayesian to come up with an interval estimator and then like a frequentist to evaluate its coverage probability. Often (but not always), if the prior is reasonable then Bayesian procedures will have good frequentist properties.

- Both Bayesians and frequentists recognize the central role of *likelihood*. In our study of frequentist approaches, we have encountered likelihood extensively when studying maximum likelihood estimation and likelihood ratio tests. If the parameter space is a bounded interval (a, b) and we use the *flat prior* $\theta \sim \text{Unif}(a, b)$, then the posterior distribution *is* the likelihood function, normalized to integrate to 1. If the parameter space is the whole real line then it is not possible to have a Uniform distribution across the entire parameter space, but some statisticians are still willing to use the *improper prior* $\pi(\theta) \propto 1$ anyway, if formally plugging this in to Bayes' theorem yields a proper (i.e., normalizable) posterior distribution.

☣ **9.2.4.** If $\pi(\theta) = 0$ then Bayes' theorem implies $\pi(\theta|\mathbf{y}) = 0$, no matter what the data $\mathbf{y}$ is. Hence if you wrongly a priori rule something out you can never recover from it whatever the evidence in the data. Lindley said it was good practice to avoid putting a prior probability of 0 or 1. He named this *Cromwell's rule*, inspired by Oliver Cromwell's letter to the Church of Scotland in 1650, where he

wrote:

> I beseech you, in the bowels of Christ, think it possible that you may be mistaken.

9.3 Point estimation

From a Bayesian point of view, the essential object of interest is the posterior distribution $\pi(\theta|\mathbf{y})$. One of the strengths of Bayesian inference is that we obtain an entire *distribution* that quantifies our uncertainty about θ, rather than being limited to, say, a point estimate and an estimate of the standard error of the point estimate. It is often valuable to visualize the posterior distribution (by plotting the posterior density or posterior CDF), and then we can glean various information such as the probability that $\theta > 3$ (given the data).

Often we do still want simpler summaries too, such as a point estimate $\hat{\theta}$, and we can extract these from the posterior distribution. Three especially natural point estimates to consider are the *posterior mean*, *posterior median*, and *posterior mode*.

Definition 9.3.1 (Prior mean, median, and mode; posterior mean, median, and mode)**.** Let the estimand θ have a continuous prior density $\pi(\theta)$. Then define

$$\begin{aligned}
&\text{prior mean} = \mathrm{E}[\theta] = \int_{-\infty}^{\infty} \theta\, \pi(\theta) d\theta, \quad \text{if this integral exists,} \\
&\text{prior median} = Q_\theta(0.5), \quad \text{which always exists,} \\
&\text{prior mode} = \operatorname*{argmax}_\theta \pi(\theta), \quad \text{if this value exists and is unique.}
\end{aligned}$$

Likewise, let θ have a continuous posterior density $\pi(\theta|\mathbf{y})$. Then define

$$\begin{aligned}
&\text{posterior mean} = \mathrm{E}[\theta|\mathbf{y}] = \int_{-\infty}^{\infty} \theta\, \pi(\theta|\mathbf{y}) d\theta, \quad \text{if this integral exists,} \\
&\text{posterior median} = Q_{\theta|\mathbf{y}}(0.5), \quad \text{which always exists,} \\
&\text{posterior mode} = \operatorname*{argmax}_\theta \pi(\theta|\mathbf{y}), \quad \text{if this value exists and is unique.}
\end{aligned}$$

The posterior mode is also known as the *maximum a posteriori* (MAP) estimator. There are also, of course, corresponding *estimators* (also called the posterior mean, posterior median, and posterior mode), defined with $\mathbf{Y}$ in place of $\mathbf{y}$; from the context it should be clear whether we are taking a pre-data or post-data vantage point.

Example 9.3.2 (Continued from Example 9.2.3)**.** Recall that we had the prior $\theta \sim \text{Beta}(\alpha, \beta)$ and the posterior $\theta|\mathbf{y} \sim \text{Beta}(\alpha + n\bar{y}, \beta + n(1-\bar{y}))$. Assume that $\alpha > 1$ and $\beta > 1$. Table 9.1 gives the prior and posterior mean, median, and mode for θ. Since the posterior distribution is in the same

Prior estimates	Posterior estimates
Prior mean $= \frac{\alpha}{\alpha+\beta}$	Posterior mean $= \frac{\alpha+n\bar{y}}{\alpha+\beta+n}$
Prior median $\approx \frac{\alpha-(1/3)}{\alpha+\beta-(2/3)}$	Posterior median $\approx \frac{\alpha+n\bar{y}-(1/3)}{\alpha+\beta+n-(2/3)}$
Prior mode $= \frac{\alpha-1}{\alpha+\beta-2}$	Posterior mode $= \frac{\alpha+n\bar{y}-1}{\alpha+\beta+n-2}$

Table 9.1: Prior and posterior estimates for a Beta parameter with Bernoulli data as in Example 9.2.3.

family as the prior distribution, once we know the results for the prior mean, median, and mode we can immediately write down the corresponding results for the posterior mean, median, and mode, just by updating the parameters.

In the absence of a loss function, it is unclear what it means to say that one estimator is better than another. Once we specify a loss function $\text{Loss}(\theta, \hat{\theta})$, which represents the loss incurred if we estimate θ as $\hat{\theta}$, we can try to find the $\hat{\theta}$ that minimizes the *posterior expected loss*, $\text{E}[\text{Loss}(\theta, \hat{\theta})|\mathbf{y}]$. In this expectation, $\hat{\theta}$ is fixed (since it must be a function of the data $\mathbf{y}$, which we are conditioning on), whereas θ is distributed according to its posterior distribution.

Theorem 9.3.3 (Posterior mean minimizes squared error loss; posterior median minimizes absolute error loss)**.** *For squared error loss* $\text{Loss}(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$*, the posterior mean is*

$$\text{E}[\theta|\mathbf{y}] = \operatorname*{argmin}_c \text{E}[(\theta - c)^2|\mathbf{y}].$$

For absolute error loss $\text{Loss}(\theta, \hat{\theta}) = |\theta - \hat{\theta}|$*, the posterior median is*

$$Q_{\theta|\mathbf{y}}(0.5) = \operatorname*{argmin}_c \text{E}[|\theta - c||\mathbf{y}].$$

Proof. The result follows from Theorem 6.1.4 in the Stat 110 book, which says that for any r.v. X,

$$\text{E}[X] = \operatorname*{argmin}_c \text{E}[(X - c)^2],$$

and

$$Q_X(0.5) = \operatorname*{argmin}_c \text{E}[|X - c|].$$

These results and a more general result for $Q_X(p)$ are also proven in Chapter 6. ∎

The most celebrated of these estimators is the posterior mean. Example 9.3.4 gives the posterior mean when the likelihood is Gaussian but the prior is arbitrary.

Example 9.3.4. (Gaussian data, arbitrary prior). Assume that

$$f(y|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\theta)^2\right\}, \quad \text{so note} \quad \frac{\partial f(y|\theta)}{\partial y} = \frac{\partial f(y|\theta)}{\partial (y-\theta)} = \frac{(\theta - y)}{\sigma^2} f(y|\theta).$$

Then, of course, $f(y) = \int f(y|\theta)\pi(\theta)\mathrm{d}\theta$ and $\pi(\theta|y) = f(y|\theta)\pi(\theta)/f(y)$, so

$$\begin{aligned} f(y)\left\{\mathrm{E}[\theta|y] - y\right\} &= \int (\theta - y) f(y|\theta)\pi(\theta)\mathrm{d}\theta \\ &= \sigma^2 \int \frac{\partial f(y|\theta)}{\partial y}\pi(\theta)\mathrm{d}\theta \\ &= \sigma^2 \frac{\partial}{\partial y}\int f(y|\theta)\pi(\theta)\mathrm{d}\theta, \quad \text{using DUThIS} \\ &= \sigma^2 \frac{\partial f(y)}{\partial y}. \end{aligned}$$

So, rearranging

$$\mathrm{E}[\theta|y] = y + \sigma^2 \frac{\partial \log f(y)}{\partial y},$$

which is called Tweedie's formula in statistics. Our derivation follows the argument of Peracchi and Smith (1992). The corresponding

$$\mathrm{Var}(\theta|y) = \sigma^2 + \sigma^4 \frac{\partial^2 \log f(y)}{\partial y^2}.$$

Tweedie's formula is central to an active area of Bayesian research called *empirical Bayes* (see Efron and Morris (1977) and Efron (2012)). As an example of this formula, if $\theta \sim N(\mu_0, \sigma_0^2)$, then $Y \sim N(\mu_0, \sigma^2 + \sigma_0^2)$ and so

$$\frac{\partial \log f(y)}{\partial y} = \frac{\partial}{\partial y}\left\{-\frac{(y-\mu_0)^2}{2\left(\sigma^2+\sigma_0^2\right)}\right\} = -\frac{y-\mu_0}{\sigma^2+\sigma_0^2}, \quad \text{and} \quad \frac{\partial^2 \log f(y)}{\partial y^2} = -\frac{1}{\sigma^2+\sigma_0^2}$$

implying

$$\mathrm{E}[\theta|y] = y - \frac{\sigma^2}{\sigma^2+\sigma_0^2}(y-\mu_0) = \frac{\sigma_0^2}{\sigma^2+\sigma_0^2} y + \frac{\sigma^2}{\sigma^2+\sigma_0^2}\mu_0,$$

and

$$\mathrm{Var}(\theta|y) = \sigma^2 - \frac{\sigma^4}{\sigma^2+\sigma_0^2} = \frac{\sigma_0^2}{\sigma^2+\sigma_0^2}.$$

We will see these expressions again in Section 9.6.1.

9.3.1 The MAP

The MAP (the posterior mode) has the very convenient property that it can be computed without knowing the normalizing constant in Bayes theorem, since multiplying a function by a positive constant has no effect on where the function is maximized. Of course, *convenient* does not necessarily imply *good*; it is possible that the additional effort needed to compute the posterior mean or posterior median would be worthwhile (depending on the problem and the loss function). We have

$$\begin{aligned}\hat{\theta}_{\text{MAP}} &= \operatorname*{argmax}_{\theta} \pi(\theta|\mathbf{y}) \\ &= \operatorname*{argmax}_{\theta} L(\theta;\mathbf{y})\pi(\theta) \\ &= \operatorname*{argmax}_{\theta} \{\log L(\theta;\mathbf{y}) + \log \pi(\theta)\}\end{aligned}$$

The MAP is closely connected to the MLE $\operatorname*{argmax}_{\theta} \log L(\theta;\mathbf{y})$: it just has the additional $\log \pi(\theta)$ term in the optimization, to take the prior into account. Note that if the parameter space is a finite interval $[a, b]$ and a flat (Uniform) prior is used, then the MAP reduces to the MLE.

Example 9.3.5 (Laplace prior and LASSO)**.** Suppose that $Y_1, \ldots, Y_n$ are i.i.d. $\mathcal{N}(\theta, \sigma^2)$ and the prior for θ is

$$\pi(\theta) = \frac{d}{2}\exp(-d|\theta|), \quad d = \frac{\sqrt{2}}{\tau_0},$$

a *Laplace distribution* centered at 0 with variance of τ_0^2 (the Laplace is a symmetrized Exponential). Then

$$\begin{aligned}\log \pi(\theta|\mathbf{y}) &= -\frac{1}{2\sigma^2}\sum_{j=1}^{n}(y_j-\theta)^2 - d|\theta| = c_1 - \frac{1}{2\sigma^2}n(\bar{y}-\theta)^2 - d|\theta| \\ &= c_1 - \frac{n}{\sigma^2}\left[\frac{1}{2}(\bar{y}-\theta)^2 - c|\theta|\right], \quad c = d\sigma^2/n.\end{aligned}$$

The normalizing constant, the posterior mean and posterior median have relatively simple forms, e.g. Peracchi and Smith (1992), Mitchell (1994) and De Luca, Magnus, and Peracchi (2021), but the posterior mode is even easier to find. Note that for $\theta \neq 0$,

$$\frac{\partial \log \pi(\theta|\mathbf{y})}{\partial \theta} = \begin{cases} \frac{n}{\sigma^2}\left[(\bar{y}-\theta)-c\right], & \theta > 0, \\ \frac{n}{\sigma^2}\left[(\bar{y}-\theta)+c\right], & \theta < 0, \end{cases}$$

which is drawn on the left hand side of Figure 9.2. So the MAP is

$$\hat{\theta}_{\text{MAP}} = \begin{cases} \bar{y}-c, & \bar{y} > c, \\ 0, & \bar{y} \in [-c, c] \\ \bar{y}+c, & \bar{y} < -c, \end{cases}$$

which has the same form as the LASSO estimator which appeared in a starred section of Chapter 6. Note that the MAP $\hat{\theta}_{\text{MAP}}$ sets $\bar{y}$ exactly to 0 if $|\bar{y}| < c$. The $c = d\sigma^2/n = \sqrt{2}\sigma^2/(n\tau_0)$ is called a *threshold* in statistics. This is shown in the right hand side of Figure 9.2.

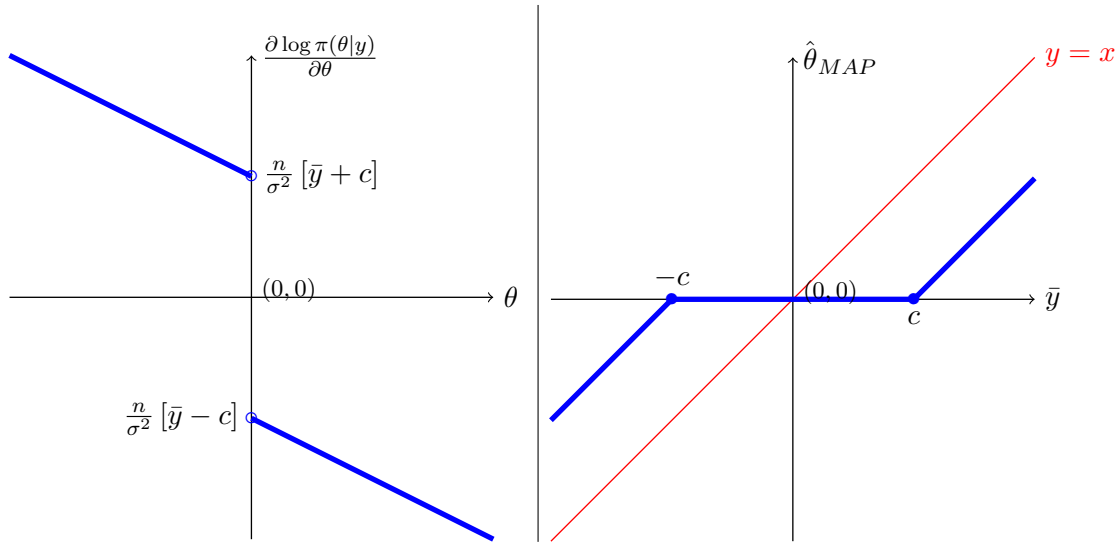


Figure 9.2: Finding MAP, for Gaussian likelihood with Laplace prior centered on 0. LHS: derivative of log-posterior. RHS: $\hat{\theta}_{\text{MAP}}$ plotted against $\bar{y}$.

9.4 Computing Bayesian estimators

9.4.1 The power of simulation

Sometimes we can compute quantities like the posterior mean and posterior median analytically, while other times it is much easier to use simulation.

Suppose that we want to use the posterior mean $\mathrm{E}[\theta|\mathbf{y}]$. We can write

$$\mathrm{E}[\theta|\mathbf{y}] = \int_{-\infty}^{\infty} \theta\, \pi(\theta|\mathbf{y}) d\theta,$$

and in some problems we can do this integral, but in many other problems this integral may be hard or infeasible to do analytically. There are principally two reasons for this.

First, we might know everything about $\pi(\theta|\mathbf{y})$ but to do the integral to compute $\mathrm{E}[\theta|\mathbf{y}]$ may be too difficult.

Second, we may not know the normalizing constant in $\pi(\theta|\mathbf{y})$. Bayes' theorem says that

$$\pi(\theta|\mathbf{y}) \propto L(\theta; \mathbf{y})\pi(\theta)$$

but in many models of interest in modern statistics and machine learning, the normalizing constant is difficult to compute. The posterior expectation

$$\mathrm{E}[\theta|\mathbf{y}] = \frac{\int_{-\infty}^{\infty} \theta\, L(\theta;\mathbf{y})\pi(\theta)d\theta}{\int_{-\infty}^{\infty} L(\theta;\mathbf{y})\pi(\theta)d\theta}$$

may then leave us with *two* difficult integrals to solve, not just one.

A powerful approach to use when such integrals are difficult is *Markov chain Monte Carlo* (MCMC), such as the Metropolis-Hastings algorithm (see Chapter 12 of the Stat 110 book for an overview of MCMC). The input into MCMC is the *unnormalized* posterior

$$L(\theta;\mathbf{y})\pi(\theta).$$

Then MCMC creates a simulated path from a Markov chain whose stationary distribution is $\pi(\theta|\mathbf{y})$. If this chain is run for a long time, then we can use the simulated θ values

$$\theta^{[1]}, \ldots, \theta^{[B]}$$

from the chain to approximate $\mathrm{E}[\theta|\mathbf{y}]$ or $Q_{\theta|\mathbf{y}}(0.5)$ (or other summaries of the posterior distribution) using sample versions, e.g., use the sample mean of $\theta^{[1]}, \ldots, \theta^{[B]}$ to approximate the posterior mean, and use the sample median of $\theta^{[1]}, \ldots, \theta^{[B]}$ to approximate the posterior median.

This simulation strategy is much more powerful than it appears at first sight. Suppose that $\boldsymbol{\theta}$ is multivariate, as it often will be in applications, but that one component ψ of $\boldsymbol{\theta}$ is of primary interest to us. Split $\boldsymbol{\theta}$ into two parts: $\boldsymbol{\theta} = (\psi, \boldsymbol{\lambda})$, where without loss of generality we list ψ first in the vector. Simulate draws $\boldsymbol{\theta}^{[1]}, \ldots, \boldsymbol{\theta}^{[B]}$ from $\boldsymbol{\theta}|\mathbf{y}$, and let

$$\boldsymbol{\theta}^{[b]} = (\psi^{[b]}, \boldsymbol{\lambda}^{[b]}), \quad b = 1, \ldots, B.$$

To find the posterior distribution of ψ mathematically, we marginalize out the components $\boldsymbol{\lambda}$:

$$\pi(\psi|\mathbf{y}) = \int \pi(\psi, \boldsymbol{\lambda}|\mathbf{y})d\boldsymbol{\lambda},$$

where the integral is over all possible values of $\boldsymbol{\lambda}$. This may be a hard integral even if λ is a scalar, and if, say, $\boldsymbol{\lambda}$ is a 4-dimensional vector then it is a quadruple integral (very likely to be tedious or intractable).

Via simulation though, instead of integration we can marginalize out $\boldsymbol{\lambda}$ in an incredibly simple way: just throw out the $\boldsymbol{\lambda}^{[b]}$ draws! This gives us simulated draws

$$\psi^{[1]}, \ldots, \psi^{[B]}$$

from $\psi|\mathbf{y}$, and then, for example, the sample mean and sample median of $\psi^{[1]}, \ldots, \psi^{[B]}$ to approximate the marginal distribution quantities $\mathrm{E}[\psi|\mathbf{y}]$ and $Q_{\psi|\mathbf{y}}(0.5)$, respectively.

9.5 Credible Intervals

Confidence intervals are interpreted in terms of *repeated sampling*. If we get some data and compute, say, $[3, 7]$ as our 95% confidence interval for an estimand θ, we can not say that there is a 95% chance that θ is in $[3, 7]$; the randomness comes only from the *interval*, not from the parameter. There is a 95% chance the random interval covers θ.

With a Bayesian approach, we can instead construct a *credible interval*, which has a direct probability interpretation. For example, if our 95% credible interval is $[3, 7]$, then we *can* say that $P(3 \leq \theta \leq 7|\mathbf{y}) = 0.95$.

Definition 9.5.1 (Credible interval)**.** Let $0 < \alpha < 1$. A $1-\alpha$ *credible interval* or *posterior probability interval* for an estimand θ is an interval estimate $[a(\mathbf{y}), b(\mathbf{y})]$ such that

$$P(a(\mathbf{y}) \leq \theta \leq b(\mathbf{y})|\mathbf{y}) = 1 - \alpha.$$

Note that, in the above definition, once we condition on $\mathbf{y}$ the endpoints $a(\mathbf{y})$ and $b(\mathbf{y})$ are fixed; the only randomness comes from our uncertainty about θ.

In constructing confidence intervals, we often needed clever tricks such as pivotal quantities. For credible intervals, we can just look at the posterior distribution and choose endpoints so that encompassing the desired amount of area. For example, for a 90% credible interval we can just choose endpoints a and b such that the area under the posterior PDF curve from a to b is 0.9. Specifically, we can cut off 5% of the area from each tail, and our credible interval then runs from the 0.05 quantile to the 0.95 quantile of the posterior distribution. Figure 9.3 illustrates this when $\theta|\mathbf{y} \sim \text{Beta}(4, 2)$.

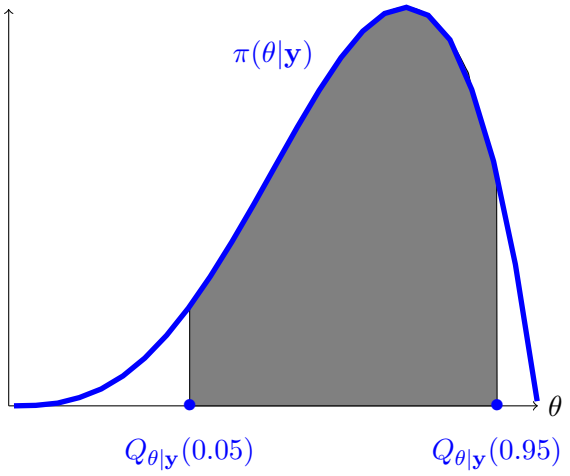


Figure 9.3: 90% credible interval for $\theta|\mathbf{y} \sim \text{Beta}(4, 2)$, placing 5% in each tail so shaded area contains 90% of the posterior.

In general, for a $1-\alpha$ credible interval we can choose the interval

$$[Q_{\theta|\mathbf{y}}(\alpha/2), Q_{\theta|\mathbf{y}}(1-\alpha/2)],$$

the $\alpha/2$ and $1-\alpha/2$ quantiles of the posterior. This is not a unique choice for the credible interval, and is not necessarily the shortest credible interval, but we will usually use this construction since it is easy to interpret and compute. If the computation is carried out using by sampling B times from the posterior, then this type of credible interval can be implemented by reporting the interval based on two sample quantiles of the draws

$$[\theta_{(\lceil 0.025B \rceil)}, \theta_{(\lceil 0.975B \rceil)}].$$

There is no guarantee that a 95% credible interval will be a 95% confidence interval, nor is there any guarantee that a 95% confidence interval will be a 95% credible interval. Once we have developed a credible interval estimator though, we can assess its coverage probability (often via simulation).

On *average* though, the coverage probability of a 95% credible interval will be 95%, where the averaging is over both θ and $\mathbf{Y}$. To see this, we can just apply Adam's law: letting I be the indicator of the credible interval covering θ,

$$P(I=1) = \mathrm{E}[I] = \mathrm{E}[\mathrm{E}[I|\mathbf{Y}]] = \mathrm{E}[P(I=1|\mathbf{Y})] = \mathrm{E}[0.95] = 0.95.$$

9.6 Conjugate Priors

The notion of a conjugate prior provides a very widely used, convenient way to select a prior.

Definition 9.6.1 (Conjugate prior)**.** A family of priors is *conjugate* for a particular statistical model if choosing a prior in the family always results in a posterior that is in the same family.

Example 9.6.2 (Beta-Binomial conjugacy)**.** It follows from Example 9.2.3 that the conjugate prior for the Binomial statistical model is the Beta distribution (see also Story 8.3.3 in the Stat 110 book). In that example we just had one Bernoulli as data, but if $Y|p \sim \mathrm{Bin}(n,p)$ and $p \sim \mathrm{Beta}(a,b)$, then we can apply the result of Example 9.2.3 repeatedly (thinking of the Binomial as being the sufficient statistic for n i.i.d. $\mathrm{Bern}(p)$ r.v.s) to obtain

$$p|(Y=y) \sim \mathrm{Beta}(a+y, b+n-y).$$

Intuitively, if a and b are positive integers and we interpret $a-1$ as the number of prior successes and $b-1$ as the number of prior failures (in previous experiments, which could be real or hypothetical),

then after we observe $Y = y$, to update the parameters of the Beta we just add on the additional successes to the first parameter and the additional failures to the second parameter. For further intuition as to why Beta is conjugate to Binomial, note that the Beta density mimics a Binomial likelihood: both look like p to a power times $1-p$ to a power, and multiplying two expressions of this form is still of this form.

Example 9.6.3 (Gamma-Poisson conjugacy)**.** Another example of conjugacy is that Gamma is the conjugate prior for the Poisson (see also Story 8.4.5 in the Stat 110 book). Specifically, if

$$Y|\lambda \sim \text{Pois}(\lambda), \quad \lambda \sim \text{Gamma}(r_0, b_0),$$

then

$$\lambda|(Y = y) \sim \text{Gamma}(r_0 + y, b_0 + 1).$$

This result makes sense intuitively since a Gamma density for λ and a Poisson likelihood for λ both look like an exponential function of λ times λ to a power, so multiplying likelihood times prior density will still be of this form. Writing this out, the posterior density is

$$\pi(\lambda|y) \propto L(y;\lambda)\pi(\lambda) \propto e^{-\lambda}\lambda^y \lambda^{r_0-1} e^{-b_0\lambda} = e^{-(b_0+1)\lambda}\lambda^{r_0+y-1}.$$

Likewise, if $Y_1, \ldots, Y_n$ are i.i.d. with $Y_j|\lambda \sim \text{Pois}(\lambda)$, then

$$\lambda|(\mathbf{Y} = \mathbf{y}) \sim \text{Gamma}(r_0 + n\bar{y}, b_0 + n).$$

9.6.1 Five cases of Normal-Normal conjugacy

Our third example of conjugacy is Normal-Normal. We will present this as five different cases as the Normal-Normal conjugacy is so helpful in thinking about many statistical problems.

Example 9.6.4 (Normal)**.** The Normal distribution is the conjugate prior for the Normal likelihood. We will start with the case of a single Normal observation, and then generalize the result to the case of n Normal observations. If

$$Y|\mu \sim \mathcal{N}(\mu, \sigma^2), \quad \mu \sim \mathcal{N}(\mu_0, \tau_0^2),$$

with $\sigma^2, \mu_0, \tau_0^2$ known, then

$$\mu|(Y = y) \sim \mathcal{N}(\mu_1, \tau_1^2),$$

where

$$\tau_1^{-2} = \sigma^{-2} + \tau_0^{-2}, \quad \mu_1 = \tau_1^2\left(\frac{y}{\sigma^2} + \frac{\mu_0}{\tau_0^2}\right).$$

To verify this result, let us work with the log-density (to reduce clutter):

$$\begin{aligned}\log \pi(\mu|y) &= c + \log L(\mu; y) + \log \pi(\mu) \\ &= c_1 - \frac{1}{2\sigma^2}(y-\mu)^2 - \frac{1}{2\tau_0^2}(\mu-\mu_0)^2 \\ &= c_2 - \frac{1}{2\sigma^2}\mu^2 + \frac{1}{\sigma^2}y\mu - \frac{1}{2\tau_0^2}\mu^2 + \frac{1}{\tau_0^2}\mu\mu_0 \\ &= c_2 - \frac{1}{2}\mu^2\left(\frac{1}{\sigma^2}+\frac{1}{\tau_0^2}\right) + \mu\left(\frac{y}{\sigma^2}+\frac{\mu_0}{\tau_0^2}\right),\end{aligned}$$

where c, c_1, c_2 do not depend on μ. This is a quadratic function in μ so the posterior distribution is Gaussian. Writing the mean and variance of that Gaussian distribution as μ_1 and τ_1^2, we have

$$\log \pi(\mu|y) = d - \frac{1}{2\tau_1^2}(\mu-\mu_1)^2 = d_1 - \frac{1}{2}\mu^2\tau_1^{-2} + \mu\tau_1^{-2}\mu_1,$$

where d and d_1 do not depend on μ. Matching this expression up with the previous expression for the log density delivers the stated result.

Next, consider the more general situation where the sample size is n.

Theorem 9.6.5 (Normal-Normal conjugacy with general sample size)**.** *Let $Y_1, \ldots, Y_n$ be random variables with*

$$Y_j|\mu \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2).$$

Let the prior be $\mu \sim \mathcal{N}(\mu_0, \tau_0^2)$, with $\sigma^2, \mu_0, \tau_0^2$ known. Then the posterior distribution for μ is

$$\mu|(\mathbf{Y}=\mathbf{y}) \sim \mathcal{N}(\mu_n, \tau_n^2), \quad \tau_n^{-2} = n\sigma^{-2} + \tau_0^{-2}, \quad \mu_n = \tau_n^2\left(\frac{n\bar{y}}{\sigma^2} + \frac{\mu_0}{\tau_0^2}\right).$$

Write the posterior variance divided by the prior variance as

$$b_n = \frac{\tau_n^2}{\tau_0^2} = \frac{\tau_0^{-2}}{n\sigma^{-2}+\tau_0^{-2}} = \frac{\sigma^2}{\sigma^2 + n\tau_0^2}.$$

Then we can also write the posterior distribution in the following nice form:

$$\mu|(\mathbf{Y}=\mathbf{y}) \sim \mathcal{N}((1-b_n)\bar{y} + b_n\mu_0, b_n\tau_0^2).$$

Proof. To show this, we can calculate the posterior directly as we did in the $n = 1$ case, but a more elegant way is to note that $\bar{Y}$ is a sufficient statistic and

$$\bar{Y}|\mu \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

so the conditional distribution of $\mu|(\mathbf{Y}=\mathbf{y})$ is the same as the conditional distribution of $\mu|(\bar{Y}=\bar{y})$, which we already know from the $n = 1$ case. ∎

Due to the symmetry of the Normal distribution, the usual Bayesian estimators for μ in the model from the above theorem (the posterior mean, posterior median, and posterior mode) are all

$$\hat{\mu} = (1 - b_n)\bar{y} + b_n\mu_0 = \bar{y} - b_n(\bar{y} - \mu_0).$$

This says that the posterior mean is a compromise between the sample mean and the prior mean. The quantity b_n, which is called the *shrinkage factor*, says how much weight to put on $\bar{y}$ versus μ_0. The more data we have, the smaller b_n will be and so the more weight we will put on $\bar{y}$ relative to μ_0. If τ_0^2 is large then we only have weak prior information about μ, and so we will put more weight on $\bar{y}$ than we would have with stronger prior information. If σ^2 is large (which corresponds to the data being noisy), we will put more weight on μ_0 than we would have if the data had been less noisy.

The special case where $\mu_0 = 0$ is important. Then

$$\hat{\mu} = (1 - b_n)\bar{y},$$

which *shrinks* $\bar{y}$ to zero, but $\hat{\mu}$ is very unlikely to be exactly zero — so this relates to ridge regression from Chapter 6. This case contrasts with the MAP we obtained from the Laplace prior case in Example 9.3.5 (which was setup to have the same prior mean and variance as the Gaussian case), where the MAP thresholded $\bar{y}$ all the way to 0, whereas here the shrinkage just scales $\bar{y}$ towards zero.

The extension to allow for heteroskedasticity, $Y_j|\mu \sim \mathcal{N}(\mu, \sigma_j^2)$, is also important in many problems. The proof follows exactly the same lines as before so is not written out here.

Theorem 9.6.6 (Normal model with heteroskedasticity)**.** *Suppose* $Y_j|\mu \overset{\text{ind.}}{\sim} \mathcal{N}(\mu, \sigma_j^2)$ *and* $\mu \sim \mathcal{N}(m_0, \tau_0^2)$. *Then* $\mu|y_1, \ldots, y_n \sim \mathcal{N}(m_n, \tau_n^2)$ *,where*

$$\tau_n^{-2} = \tau_0^{-2} + \sum_{j=1}^{n} \sigma_j^{-2}, \quad m_n = \tau_n^2 \left(\tau_0^{-2} m_0 + \sum_{j=1}^{n} \sigma_j^{-2} y_j \right),$$

where we assume that $\sigma_1^2, \ldots, \sigma_n^2, m_0, \tau_0$ *are known.*

Hence data with lower variances are more highly weighted in the posterior mean. Note that in the posterior mean $y_1, \ldots, y_n$ all have nonnegative weights, which sum to be at most 1. Imagining that μ_0 is a prior observation with its own weight, the weights sum to 1.

Example 9.6.7 (Repeated measurements)**.** It may look contrived to model $Y_j|\mu \overset{\text{ind.}}{\sim} \mathcal{N}(\mu, \sigma_j^2)$, where the σ_j^2 are *known* but differ. But think of the case where there are repeated measurements on the jth individual. The simplest version of this is where we model the ith measurement on the jth individual as

$$Y_{i,j}|\mu \overset{\text{ind.}}{\sim} \mathcal{N}(\mu, \sigma^2), \quad i = 1, ..., n_j, \quad j = 1, ..., n,$$

which means we have n_j observations on the jth individual. Then write

$$Y_j = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{i,j},$$

which implies $Y_j|\mu \sim \mathcal{N}(\mu, \sigma_j^2)$, where $\sigma_j^2 = \sigma^2/n_j$. The n_j are known, so only one parameter σ^2 is needed to determine the variances. Many applied problems are of this type.

Repeated measurements generalize to regressions. Chapter 6 focused on predictive and descriptive regression. Here we will look at Bayesian inference for the Gaussian linear predictive regression with a single predictor. Again the proof has the same structure so we do not write it out here.

Theorem 9.6.8 (Bayesian Gaussian linear regression)**.** *Assume that*

$$Y_j|(\mathbf{X} = \mathbf{x}, \theta) \overset{\text{ind.}}{\sim} \mathcal{N}(\theta x_j, \sigma^2), \quad j = 1, \ldots, n,$$

and $\theta|(\mathbf{X} = \mathbf{x}) \sim \mathcal{N}(m_0, \tau_0^2)$*. Then* $\theta|(\mathbf{Y} = \mathbf{y}, \mathbf{X} = \mathbf{x}) \sim \mathcal{N}(m_n, \tau_n^2)$*, where*

$$\tau_n^{-2} = \tau_0^{-2} + \sigma^{-2}\sum_{j=1}^{n} x_j^2, \quad m_n = \tau_n^2\left(\tau_0^{-2}m_0 + \sigma^{-2}\sum_{j=1}^{n} x_j y_j\right).$$

where we assume that m_0, τ_0 *are known.*

☠ **9.6.9.** Note that the assumption $\theta|(\mathbf{X} = \mathbf{x}) \sim \mathcal{N}(m_0, \tau_0^2)$ implies that θ is independent of $X_1, \ldots, X_n$ (if they are not independent then the X_j would appear in the conditional distribution). In this context, many people write the prior in the form $\pi(\theta)$ instead of $\pi(\theta|\mathbf{x})$, but for this to be correct requires the major assumption that θ and $X_1, \ldots, X_n$ are independent.

9.7 Bayesian model choice

Suppose that Bill advocates for a statistical model and Jose prefers a very different model. How can we learn from the data which researcher (if either) to believe? Is Bill a better statistical modeler, or is Jose? In statistics, this type of problem is called *model choice*.

In Bayesian model choice, we follow Lindley's advice and quantify the uncertainty about model choice using probability. To formalize this, write Bill's likelihood and prior as $f(\mathbf{y}|\theta, \text{Bill})$ and $f(\theta|\text{Bill})$, and Jose's likelihood and prior as $g(\mathbf{y}|\lambda, \text{Jose})$ and $g(\lambda|\text{Jose})$. There is no need for there to be any connection between Bill's model and Jose's model, which we have emphasized by using different letters and parameters for their models. Importantly, there is no need for the dimensions of θ and λ to be the same.

We can then compare the relative support the data gives to the two models, with the help of the following ratio.

Definition 9.7.1 (Bayes factor)**.** The *Bayes factor* is defined as the ratio of marginal likelihoods for the two models:

$$\frac{f(\mathbf{y}|\text{Bill})}{g(\mathbf{y}|\text{Jose})}.$$

The Bayes factor is the factor that converts the ratio of prior probabilities of the two models to the ratio of posterior probabilities:

$$\frac{P(\text{Bill}|\mathbf{y})}{P(\text{Jose}|\mathbf{y})} = \frac{P(\text{Bill})}{P(\text{Jose})}\frac{f(\mathbf{y}|\text{Bill})}{g(\mathbf{y}|\text{Jose})},$$

The above equation follows from applying Bayes' rule separately to the numerator and denominator; note that the $f(\mathbf{y})$ cancels.

In the case where exactly one of the two models is true, the above result says

Posterior odds of Bill's model = Prior odds of Bill's model × Bayes factor.

To find the marginal likelihood of Bill's model, we can use the law of total probability (and likewise for Jose's model):

$$f(\mathbf{y}|\text{Bill}) = \int f(\mathbf{y}|\theta, \text{Bill}) f(\theta|\text{Bill}) d\theta.$$

9.7.1 Marginal likelihood

A challenge in using Bayes factors is calculating the marginal likelihood, which typically involves integration. Sometimes this can be sidestepped by using the so-called "candidate's formula" (which takes Bayes' rule $P(B|A) = P(A|B)P(B)/P(A)$ and swaps $P(B|A)$ and $P(A)$),

$$P(A) = \frac{P(A|B)P(B)}{P(B|A)}$$

which holds for any random variable B. You get to choose B to make your life easiest! Why is this helpful here? It means, for any value of t, the marginal likelihood

$$f(\mathbf{y}) = \frac{f(\mathbf{y}|\theta = t)f(\theta = t)}{f(\theta = t|\mathbf{y})}.$$

Example 9.7.2. Model $\mathbf{Y}$ i.i.d. with $Y_j|\theta \sim \mathcal{N}(\theta, \sigma^2)$, σ^2 known, and the prior $\theta \sim \mathcal{N}(m_0, \tau_0^2)$. Then

$$\theta|(\mathbf{Y} = \mathbf{y}) \sim \mathcal{N}(m_n, \tau_n^2).$$

By the candidate's formula at $\theta = 0$,

$$f(\mathbf{y}) = \frac{f(\mathbf{y}|\theta = 0)f(\theta = 0)}{f(\theta = 0|\mathbf{y})} = \left(\frac{1}{2\pi\sigma^2}\right)^{-n/2} \tau_0^{-1}\tau_n \ \exp\left(-\frac{1}{2\sigma^2}\sum_{j=1}^{n} y_j^2 - \frac{m_0^2}{2\tau_0^2} + \frac{m_n^2}{2\tau_n^2}\right).$$

Other expressions can be obtained by setting $\theta = m_0$ or at $\theta = m_n$ or indeed at $\theta = \bar{y}$.

Example 9.7.3. Jose used the model in Example 9.7.2, but takes $m_0 = 0$ and writes $\tau_0^2 = \sigma^2/c_0$, where $c_0 > 0$. This implies $\tau_n^2 = \sigma^2/(n + c_0)$ and $m_n = n\bar{y}/(n + c_0)$. Bill is rather confident so $Y_j|\text{Bill} \sim \mathcal{N}(0, \sigma^2)$, knowing σ^2 (this is comparable to a hypothesis test, when Bill has a simple null that the mean is zero, while Jose has a composite alternative). Then

$$\frac{f(y_1, \ldots, y_n|\text{Bill})}{g(y_1, ..., y_n|\text{Jose})} = \tau_0\tau_n^{-1} \exp\left(- \frac{m_n^2}{2\tau_n^2}\right) = c_0^{-1/2}(n + c_0)^{1/2} \exp\left(- \frac{nt_n^2}{2(n + c_0)}\right),$$

where

$$t_n = n^{1/2}\bar{y}/\sigma$$

is the usual t-statistic. Hence the Bayes factor will be small, and so favor Jose's model, if $|t_n|$ is large. But the same level of $|t_n|$ will impact the Bayes factor more when n is small than when n is large. Thus Bayesians are more demanding to move away from Bill's simpler model if n is large — they need to see a larger $|t_n|$ than frequentists who do not penalize the t-statistic for n.

Example 9.7.3 shows that Bayesian model choice is different than some aspects of the frequentist conception of hypothesis testing. This clash is called *Lindley's paradox*, in honor of a 1957 paper by Dennis Lindley on this topic, or the *Jeffreys-Lindley paradox*, since it was discussed earlier by Sir Harold Jeffreys in his 1939 book on probability.

9.7.2 Bayesian hypothesis testing

Bayesians have the entire $\pi(\theta|\mathbf{y})$, everything there is to know about θ given the data. But sometimes it is helpful to provide a summary, like the Bayes estimator $\text{E}[\theta|\mathbf{y}]$ or the credible interval $[Q_{\theta|\mathbf{y}}(\alpha/2), Q_{\theta|\mathbf{y}}(1 - \alpha/2)]$. Another type of summary is the posterior probability of a null hypothesis.

To set this up, return to Chapter 8 with a null $H_0 : \theta \in \Theta_0$ and alternative $H_1 : \theta \in \Theta_1$, then the posterior of the null is

$$P(\theta \in \Theta_0|\mathbf{y}).$$

By Cromwell's rule this setup only makes sense to use if $1 > P(\theta \in \Theta_0) > 0$.

☣ **9.7.4.** For a simple null $H_0 : \theta = \theta_0$ then if the prior is, e.g. $\theta \sim \mathcal{N}(m_0, \tau_0^2)$, then $P(\theta \in \Theta_0) = 0$. This is not allowed by Cromwell's rule. For a Bayesian, a simple null can only possibly make any sense if the prior has strictly positive probability exactly at the null value. Composite nulls are much safer to work with from a Bayesian perspective.

Example 9.7.5. Suppose $\theta \sim \mathcal{N}(0, \sigma^2/c_0)$ and $Y_j|\theta \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\theta, \sigma^2)$, so $\theta|\mathbf{y} \sim \mathcal{N}(m_n, \tau_n^2)$, where

$$\frac{m_n}{\tau_n} = \tau_n \frac{n\bar{y}}{\sigma^2} = \sqrt{\frac{n}{n + c_0}} t_n,$$

with, again, $t_n = n^{1/2}\bar{y}/\sigma$. Suppose $H_0 : \theta \leq 0$ and $H_1 : \theta > 0$. Then the posterior of the null is

$$P(\theta \leq 0|\mathbf{y}) = P\left(\frac{\theta - m_n}{\tau_n} \leq \frac{-m_n}{\tau_n}|\mathbf{y}\right) = \Phi\left(-\sqrt{\frac{n}{n+c_0}}t_n\right),$$

where Φ is the $\mathcal{N}(0,1)$ CDF. The posterior probability is close to the p-value for a one sided t-test of this problem, which is $\Phi(-t_n)$, if c_0 is small.

Here the classical and Bayesian solutions are quite similar, but they are rather different from the Bayesian model choice Example 9.7.3, which compares a model $\mathcal{N}(\theta, \sigma^2)$ with the special case $\mathcal{N}(0, \sigma^2)$. Why? The following example bridges the gap.

Example 9.7.6. (continuing Example 9.7.5) Now set $H_0 : |\theta| \leq d$ and $H_1 : |\theta| > d$, where $d > 0$ and is small. Then the posterior of the null is

$$P(|\theta| \leq d|\mathbf{y}) = P\left(\frac{-d - m_n}{\tau_n} \leq \frac{\theta - m_n}{\tau_n} \leq \frac{d - m_n}{\tau_n}|\mathbf{y}\right) \tag{9.1}$$

$$= \Phi\left(\frac{d}{\tau_n} - \sqrt{\frac{n}{n+c_0}}t_n\right) - \Phi\left(\frac{-d}{\tau_n} - \sqrt{\frac{n}{n+c_0}}t_n\right), \tag{9.2}$$

Using a Taylor expansion for small d, then

$$P(|\theta| \leq d|\mathbf{y}) \approx 2\frac{d}{\tau_n}\Phi'\left(-\sqrt{\frac{n}{n+c_0}}t_n\right) = \frac{2}{\sqrt{2\pi}}\frac{d}{\tau_n}\exp\left(-\frac{nt_n^2}{2(n+c_0)}\right)$$

$$\approx \frac{\tau_0}{\tau_n}P(|\theta| \leq d)\exp\left(-\frac{nt_n^2}{2(n+c_0)}\right),$$

which is closely related to the result in Example 9.7.3 and so is different in spirit from the p-value associated with the classical point null hypothesis.

9.8 Bayesian prediction

Bayesian thinking can also be applied to prediction. Having a joint probability model for observed data, unobserved data, and parameters empowers us to consider whichever conditional distribution we are interested in, e.g., the conditional distribution of a future observation, given all the data we have so far.

Definition 9.8.1 (Posterior predictive distribution)**.** Let Y be a variable that we wish to predict, given some observed data. In the Bayesian framework, the *posterior predictive distribution* of Y is the conditional distribution of Y, given the observed data.

In Chapter 6 we introduced predictive regression models, based on modeling the conditional distribution of an outcome variable Y given a predictor variable X. In that chapter we took a frequentist approach.

The predictive models we have looked at already focus on the conditional distribution of $Y|X,\theta$ (where the conditioning on θ is often left implicit). But from a Bayesian point of view, it is more relevant to find the conditional distribution of a not-yet-observed Y, given all the data that have been observed, leaving θ as random (to account for our uncertainty) rather than conditioned on. It then makes sense to introduce a prior $\pi(\theta|\mathbf{X})$, as discussed in Example 9.6.8.

Suppose that we have observed $(X_1,Y_1),\ldots,(X_n,Y_n)$, and would like to predict Y_{n+1} for a new individual whose predictor variable is known to be x_{n+1}. Let $\mathbf{x}=(x_1,\ldots,x_{n+1})$ and $\mathbf{y}=(y_1,\ldots,y_n)$. By the law of total probability, with everything conditioned on $\mathbf{x},\mathbf{y}$, the posterior predictive distribution is

$$f(y_{n+1}|\mathbf{x},\mathbf{y})=\int f(y_{n+1}|\mathbf{x},\mathbf{y},\theta)\pi(\theta|\mathbf{x},\mathbf{y})d\theta.$$

Note that, whether or not θ itself is of scientific interest, the posterior $\pi(\theta|\mathbf{x},\mathbf{y})$ is a key ingredient for finding the posterior predictive distribution.

The $f(y_{n+1}|\mathbf{x},\mathbf{y},\theta)$ term can be found through an assumed model (e.g., a Gaussian regression model, with θ treated as fixed), while $\pi(\theta|\mathbf{x},\mathbf{y})$ can be found using techniques from earlier in this chapter for finding posterior distributions. The resulting integral may or may not be tractable to do mathematically. Simulation-based methods would be an attractive alternative: we can simulate draws from $f(y_{n+1}|\mathbf{x},\mathbf{y})$ by drawing $\theta^{[1]},\ldots,\theta^{[B]}$ from $\pi(\theta|\mathbf{x},\mathbf{y})$, and then drawing $y_{n+1}^{[j]}$ from $f(y_{n+1}|\mathbf{x},\mathbf{y},\theta^{[j]})$ for $j=1,\ldots,B$.

Here are a couple useful examples of posterior predictive distributions that are available in closed form.

Example 9.8.2 (Posterior predictive in Normal-Normal model)**.** Consider again the Normal-Normal model from Section 9.6.1: let $Y_1,Y_2,\ldots$ be random variables with

$$Y_j|\mu \overset{i.i.d.}{\sim} \mathcal{N}(\mu,\sigma^2),\quad \mu\sim\mathcal{N}(\mu_0,\tau_0^2),$$

with σ^2,μ_0,τ_0^2 known. Suppose that we have observed $Y_1,\ldots,Y_n$, and would like to predict Y_{n+1}. We don't just want a single estimated value for Y_{n+1}; we want the *distribution* of Y_{n+1} given the data. For example, we can then obtain say what the probability that $Y_{n+1}>3$ is, given the data, or we can give an interval that has a 95% chance of containing Y_{n+1}, given the data.

In a purely frequentist perspective, we can't even write down something like the posterior predictive distribution. In frequentist notation, $Y_{n+1}|Y_1,\ldots,Y_n$ presumably means $Y_{n+1}|Y_1,\ldots,Y_n,\mu$ implicitly, which by the i.i.d. assumption just reduces to $Y_{n+1}|\mu$. But from a Bayesian perspective, $Y_{n+1}|Y_1,\ldots,Y_n$ makes sense, and propagates our uncertainty about μ correctly. Note that $Y_1,Y_2,\ldots$ are conditionally

independent given μ, but not independent. Observing $Y_1, \ldots, Y_n$ gives information about μ, which in turn gives information about Y_{n+1}.

It can be shown that $(Y_1, Y_2, \ldots, Y_{n+1}, \mu)$ is Multivariate Normal. A nice property of the Multivariate Normal is that within a Multivariate Normal, conditional distributions are also Multivariate Normal. So the posterior predictive distribution is Normal. Then we just need to find its parameters! By Adam's law,

$$\mathrm{E}[Y_{n+1}|Y_1, \ldots, Y_n] = \mathrm{E}[\mathrm{E}[Y_{n+1}|Y_1, \ldots, Y_n, \mu]|Y_1, \ldots, Y_n].$$

By conditional independence, this reduces to

$$\mathrm{E}[Y_{n+1}|Y_1, \ldots, Y_n] = \mathrm{E}[\mu|Y_1, \ldots, Y_n] = (1 - b_n)\bar{Y}_n + b_n\mu_0,$$

with

$$b_n = \frac{\sigma^2}{\sigma^2 + n\tau_0^2},$$

by the posterior mean result from Theorem 9.6.5.

By Eve's law and the posterior variance result from Theorem 9.6.5.

$$\begin{aligned}
\mathrm{Var}(Y_{n+1}|Y_1, \ldots, Y_n) &= \mathrm{E}[\mathrm{Var}(Y_{n+1}|Y_1, \ldots, Y_n, \mu)|Y_1, \ldots, Y_n] + \mathrm{Var}(\mathrm{E}[Y_{n+1}|Y_1, \ldots, Y_n, \mu]|Y_1, \ldots, Y_n) \\
&= \mathrm{E}[\mathrm{Var}(Y_{n+1}|\mu)|Y_1, \ldots, Y_n] + \mathrm{Var}(\mu|Y_1, \ldots, Y_n) \\
&= \mathrm{E}[\sigma^2|Y_1, \ldots, Y_n] + \tau_n^2 \\
&= \sigma^2 + \tau_n^2,
\end{aligned}$$

where

$$\tau_n^{-2} = n\sigma^{-2} + \tau_0^{-2}.$$

The result for the posterior predictive variance of Y_{n+1} makes sense intuitively, since there are two sources of uncertainty for Y_{n+1}: uncertainty about μ, and the randomness of each observation, even conditioned on μ. Interestingly, these two variances (the posterior variance of μ, and the variance for each new observation, given μ) simply add to give the overall posterior predictive variance.

Thus, the posterior predictive distribution is

$$Y_{n+1}|Y_1, \ldots Y_n \sim \mathcal{N}((1 - b_n)\bar{Y}_n + b_n\mu_0, \sigma^2 + \tau_n^2).$$

Example 9.8.3 (Posterior predictive for Poisson data)**.** Suppose that we have observed Poisson data $Y_1, \ldots, Y_n$, and want to predict a future data point Y_{n+1}. Taking a Bayesian perspective, model

$$Y_j|\theta \overset{\text{ind.}}{\sim} \mathrm{Pois}(c_j\theta),$$

with conjugate prior

$$\theta \sim \text{Gamma}(r_0, b_0),$$

where $r_0, b_0, c_1, \ldots c_{n+1}$ are known, $r_0 > 0, b_0 > 0$, and r_0 is an integer. By Gamma-Poisson conjugacy (or directly applying Bayes' rule), the posterior for θ is

$$\theta|Y_1, \ldots, Y_n \sim \text{Gamma}\left(r_0 + \sum_{j=1}^{n} Y_j, b_0 + \sum_{j=1}^{n} c_j\right).$$

But now our goal is to find the posterior predictive distribution of $Y_{n+1}|Y_1, \ldots, Y_n$.

An elegant approach is to first find the *marginal* (unconditional) distribution of Y_{n+1}. By Story 8.4.5 of the Stat 110 book, or using the candidate's formula along with the results for Gamma-Poisson conjugacy from earlier in this chapter, the marginal is Negative Binomial:

$$Y_{n+1} \sim \text{NBin}\left(r_0, \frac{b_0}{b_0 + c_{n+1}}\right).$$

We can then obtain the posterior predictive distribution without needing any additional calculations! This is because the usefulness of conditioning on $Y_1, \ldots, Y_n$, for purposes of helping us predict Y_{n+1}, is purely to help us learn about θ. We use $Y_1, \ldots, Y_n$ to update our prior for θ to a posterior for θ. Then our posterior for θ becomes our new prior, and we have no further use for $Y_1, \ldots, Y_n$ for predicting Y_{n+1} (due to the conditional independence structure). Hence, the posterior predictive distribution is

$$Y_{n+1}|Y_1, \ldots Y_n \sim \text{NBin}\left(r_0 + \sum_{j=1}^{n} Y_j, \frac{b_0 + \sum_{j=1}^{n} c_j}{b_0 + \sum_{j=1}^{n+1} c_j}\right).$$

9.9 Hierarchical models

One of the most important distinctly Bayesian approaches to model building is *hierarchical modeling*. This approach has been extremely influential in statistics over the last 50 years. Together with the computational revolution which places simulation at the heart of Bayesian statistics, it is the main driver of the modern resurgence of Bayesian methods in statistics. Here we will focus on a couple of Gaussian versions.

Definition 9.9.1 (Two-level Gaussian hierarchical model)**.** Assume that

$$Y_j|\mu_1, \ldots, \mu_K \overset{\text{ind.}}{\sim} \mathcal{N}(\mu_j, \sigma^2), \quad j = 1, \ldots, K,$$

where each conditional mean is

$$\mu_j \overset{\text{ind.}}{\sim} \mathcal{N}(\gamma, \lambda_0^2).$$

Then the pair $(\mathbf{Y}, \boldsymbol{\mu})$ follow a *two-level Gaussian hierarchical model*, also known as a *two-level multilevel model*. The model is indexed by $\psi = (\sigma, \gamma, \lambda_0)$, which are called *hyperparameters*. Here we will think of them as known.

In this hierarchical model the Y_j are conditionally independent given $\boldsymbol{\mu} = (\mu_1, ..., \mu_K)$, and the μ_j are i.i.d. By the result from Theorem 9.6.5,

$$\mu_j|(\mathbf{Y} = \mathbf{y}) \sim \mathcal{N}(m_j, \lambda_K^2), \quad m_j = \lambda_K^2(\lambda_0^{-2}\gamma + \sigma^{-2}y_j), \quad \lambda_K^{-2} = \lambda_0^{-2} + \sigma^{-2}. \tag{9.3}$$

Marginally,

$$Y_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\gamma, \sigma^2 + \lambda_0^2).$$

We will use this two-level model in the next section when we discuss Stein's paradox. First, we go one level deeper into Bayesian hierarchical models.

Definition 9.9.2 (Three-level Gaussian hierarchical model)**.** Assume that

$$Y_j|\mu_1, \ldots, \mu_K, \gamma \overset{\text{ind.}}{\sim} \mathcal{N}(\mu_j, \sigma^2), \quad j = 1, \ldots, K,$$

where each conditional mean is

$$\mu_j|\gamma \overset{\text{ind.}}{\sim} \mathcal{N}(\gamma, \lambda_0^2),$$

and

$$\gamma \sim \mathcal{N}(g_0, \tau_0^2).$$

Then the triple $(\mathbf{Y}, \boldsymbol{\mu}, \gamma)$ follow a *three-level Gaussian hierarchical model*. It is indexed by hyperparameters $\psi = (\sigma, \lambda_0, g_0, \tau_0)$.

This model links the data through having γ in common, as the parent of all of $\mu_1, \ldots, \mu_K$. Each μ_j delivers a conditionally independent Y_j. Figure 9.4 is a graphical representation of this, where the directed arrows, called *directed edges*, show the tree-like structure of the model, connecting the random variables, which are expressed on the graph as *nodes*.

Theorem 9.9.3. *Assume a three-level Gaussian hierarchical model. Then*

$$\gamma|(\mathbf{Y} = \mathbf{y}) \sim \mathcal{N}(g_K, \tau_K^2), \quad g_K = \frac{\tau_0^{-2}g_0 + (\sigma^2 + \lambda_0^2)^{-2}K\bar{y}}{\tau_0^{-2} + K(\sigma^2 + \lambda_0^2)^{-2}}, \tag{9.4}$$

and

$$\mu_j|(\mathbf{Y} = \mathbf{y}) \sim \mathcal{N}\left(\frac{\lambda_0^{-2}g_K + \sigma^{-2}y_j}{\lambda_0^{-2} + \sigma^{-2}}, \frac{\lambda_0^{-4}\tau_K^2}{(\lambda_0^{-2} + \sigma^{-2})^2} + \frac{1}{\lambda_0^{-2} + \sigma^{-2}}\right),$$

where $\tau_K^{-2} = \tau_0^{-2} + K(\sigma^2 + \lambda_0^2)^{-2}$.

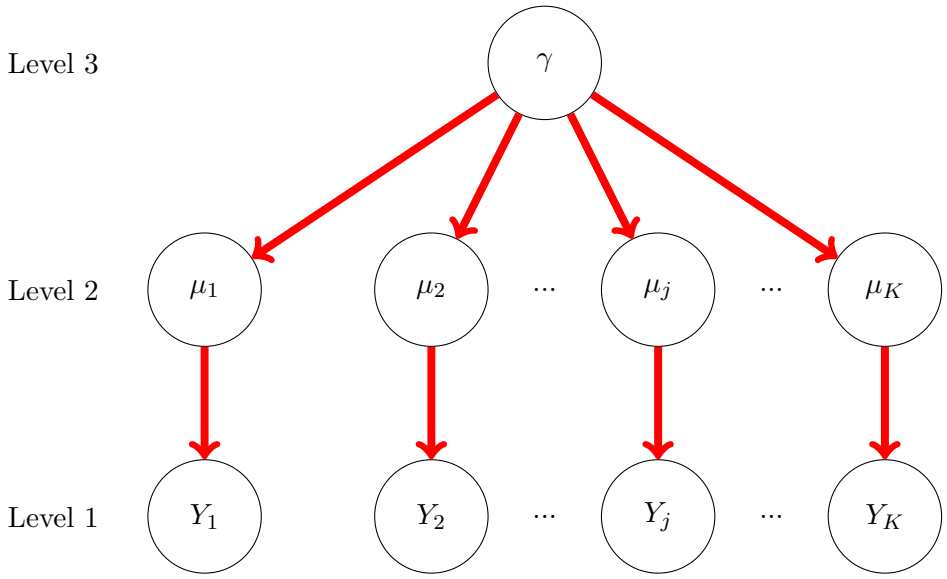


Figure 9.4: Graphical representation of a level 3 hierarchical Bayesian model.

Proof. Conditional on γ, $(\mathbf{Y}, \boldsymbol{\mu})|\gamma$ is a two-level Gaussian hierarchical model, so

$$Y_j|\gamma \overset{\text{ind.}}{\sim} \mathcal{N}(\gamma, \sigma^2 + \lambda_0^2).$$

Then (9.4) holds by the result for the Normal-Normal posterior. Now recall from the result for the two-level model that

$$\mu_j|(\mathbf{Y} = \mathbf{y}, \gamma) \sim \mathcal{N}\left(\frac{\lambda_0^{-2}\gamma + \sigma^{-2}y_j}{\lambda_0^{-2} + \sigma^{-2}}, \frac{1}{\lambda_0^{-2} + \sigma^{-2}}\right),$$

The result for the parameters of $\mu_j|(\mathbf{Y} = \mathbf{y})$ then follow from Adam's law and Eve's law:

$$\mathrm{E}[\mu_j|(\mathbf{Y} = \mathbf{y})] = \mathrm{E}[\mathrm{E}[\mu_j|(\mathbf{Y} = \mathbf{y}, \gamma)],$$

$$\mathrm{Var}(\mu_j|(\mathbf{Y} = \mathbf{y})) = \mathrm{E}[\mathrm{Var}(\mu_j|(\mathbf{Y} = \mathbf{y}, \gamma))] + \mathrm{Var}\left(\mathrm{E}[\mu_j|(\mathbf{Y} = \mathbf{y}, \gamma)]\right).$$

∎

Here the posterior mean of the individual mean μ_j is

$$\frac{\lambda_0^{-2}g_K + \sigma^{-2}y_j}{\lambda_0^{-2} + \sigma^{-2}},$$

a weighted average of y_j and g_K, the posterior mean of $\gamma|y_1, \ldots, y_K$, which in turn is a weighted average of the prior mean g_0 and sample average $\bar{y}$.

9.10 Stein's Paradox

9.10.1 Risk function and inadmissibility

To finish this chapter, let us return to some frequentist concepts for evaluating estimators developed in Chapter 3. Throughout this section, suppose that we are working with respect to a loss function $L(\theta, \hat{\theta})$ and recall the *risk function* of the estimator $\hat{\theta}$ is its expected loss,

$$R(\theta) = \mathrm{E}_{\mathbf{Y};\theta}[\mathrm{Loss}(\theta, \hat{\theta})].$$

This is the average loss incurred by using $\hat{\theta} = T(\mathbf{Y})$, averaged over the random $\mathbf{Y}; \theta$. This was first introduced in Definition 3.3.1, which lead up to our discussion of mean square error.

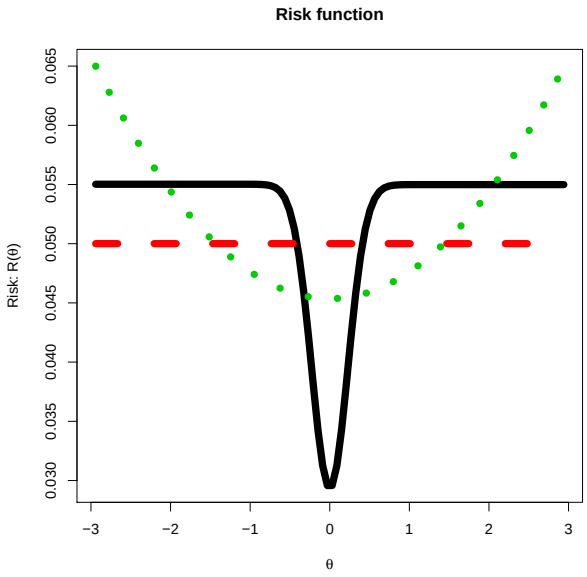


Figure 9.5: Risk functions $R(\theta)$ for $\bar{\theta}$ (black line), $\tilde{\theta}$ (green dotted line) and $\bar{Y}$ (red dotted line), respectively. $\bar{\theta}$ is MAP estimator from the Normal-Laplace model, the LASSO estimator. $\bar{\theta}$ is posterior mean estimator from Normal-Normal, the ridge estimator.

In the Bayesian approach we can look at the expected loss *given the data*,

$$\mathrm{E}_{\theta|\mathbf{Y}=\mathbf{y}}[\mathrm{Loss}(\theta, \hat{\theta})] = \mathrm{E}[\mathrm{Loss}(\theta, \hat{\theta})|y],$$

averaging over the random θ given $\mathbf{Y} = \mathbf{y}$; if we look at this quantity for several different estimators, we have a natural way to decide which is better. In the frequentist approach, we face the challenge that the risk function for an estimator depends on θ, which is unknown. So instead of computing $R(\theta)$ at the true value of θ, we need to consider the entire *function* $R(\theta)$.

Example 9.10.1 (Three risk functions)**.** Suppose that $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$, then Figure 9.5 shows the risk function $R(\theta)$ for three different estimators of θ for $\theta \in [-3, 3]$ when $n = 20$. The first is $\bar{Y}$, the MLE. The second is the Bayesian estimator $\tilde{\theta}$, the posterior mean from the Normal-Normal model, from Example 9.6.4. The third, is the posterior mode (MAP) estimator $\bar{\theta}$ from the Normal-Laplace model given in Example 9.3.5. For both Bayesian estimators the prior mean and variance are set at 0 and τ_0^2, respectively, so $\tilde{\theta}$ corresponds to a ridge estimator and $\bar{\theta}$ the LASSO estimator. Recall, both estimators shrink $\bar{Y}$ towards 0. In this example we take $n = 20$ and $\tau_0^2 = 1$. The code which produced Figure 9.5 is given in Section 9.12.

In comparing two different estimators, it will often be the case that one has lower risk for some values of θ, while the other has lower risk for other values of θ. It may then be far from obvious which of the two estimators (if either) to use. Example 9.10.1 illustrates this point with three estimators, two of which have been shrunk towards zero. In some situations, however, one estimator dominates another *across the board*; this leads to the notion of *admissibility*.

Definition 9.10.2 (Admissibility)**.** An estimator $\hat{\theta}$ is *inadmissible* if there exists another estimator whose risk function is less than or equal to that of $\hat{\theta}$ for all possible θ, with strict inequality for at least one possible value of θ. An estimator is *admissible* if it is not inadmissible.

If we are confident about our choice of risk function, admissibility seems like a rather minimal feature we would like our estimator to have. If we have an estimator $\hat{\theta}$ that turns out to be inadmissible, then there is some other estimator $\tilde{\theta}$ that is at least as good (in terms of risk) for all θ, and sometimes strictly better, so why not use $\tilde{\theta}$ instead? One reason could be that we may not know how to compute $\tilde{\theta}$ efficiently, and another reason could be that we may be concerned about model misspecification. Let us assume though that our model is correct, and that we can compute $\tilde{\theta}$ efficiently. Then it does not seem to make much sense to use an inadmissible estimator rather than an estimator we know to be better.

9.10.2 James-Stein estimator

Charles Stein shocked the world (or at least the world of statisticians) in 1956 by proving the following result, in a seemingly innocuous setting where we have a bunch of independent Normal random variables, each with its own mean, and we want to estimate the means.

Theorem 9.10.3 (Stein's theorem)**.** *Let*

$$Y_j \sim \mathcal{N}(\mu_j, \sigma^2)$$

for $j = 1, \ldots, K$ be independent, where $K \geq 3$, the μ_j are unknown, and σ^2 is known. Let the estimand be the vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$ and the loss function be the total squared error loss,

$$\text{Loss}(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \sum_{j=1}^{K} (\mu_j - \hat{\mu}_j)^2.$$

Then the MLE, which is $\mathbf{Y} = (Y_1, \ldots, Y_K)$ itself, is inadmissible.

A few years later, Stein and Willard James gave a specific formula for an estimator, the *James-Stein estimator*, that has strictly lower risk than the MLE $\mathbf{Y}$ for all $\boldsymbol{\mu}$. The estimator may seem bizarre at first, considering that the r.v.s $Y_1, \ldots, Y_K$ are *independent*, yet the James-Stein estimator shrinks $\mathbf{Y}$ toward 0, using a quantity that depends on *all* of $Y_1, \ldots, Y_K$ to decide how far toward 0 to go.

Theorem 9.10.4 (James-Stein estimator)**.** *In the setup of the previous theorem, let*

$$S = \sum_{j=1}^{K} Y_j^2.$$

Let $\hat{\boldsymbol{\mu}}_{\text{JS}} = (\hat{\mu}_{\text{JS},1}, \ldots, \hat{\mu}_{\text{JS},K})$ be the James-Stein estimator*, defined by*

$$\hat{\mu}_{\text{JS},\, j} = \left[1 - \frac{(K-2)\sigma^2}{S}\right] Y_j.$$

Then $\hat{\boldsymbol{\mu}}_{\text{JS}}$ has strictly lower risk than $\mathbf{Y}$ for all $\boldsymbol{\mu} \in \mathbb{R}^K$. Specifically, the risk function of $\mathbf{Y}$ is the constant $K\sigma^2$, whereas the risk function of $\hat{\boldsymbol{\mu}}_{\text{JS}}$ is

$$\left[K - (K-2)^2 \sigma^2 \text{E}\left(\frac{1}{S}\right)\right] \sigma^2.$$

We will not prove the above theorem, but we will show how a Bayesian perspective makes estimators such as the James-Stein estimator more natural.

For a two-level hierarchical model write

$$Y_j | \mu_1, \ldots, \mu_K \overset{\text{ind.}}{\sim} \mathcal{N}(\mu_j, \sigma^2), \quad \mu_j \overset{\text{ind.}}{\sim} \mathcal{N}(\mu_0, \sigma_\mu^2), \quad j = 1, \ldots, K,$$

so the Y_j are conditionally independent given $\boldsymbol{\mu} = (\mu_1, ..., \mu_K)$, the μ_j are i.i.d., and $\sigma^2, \mu_0, \sigma_\mu^2$ are known. Throughout, suppose that $\mu_0 = 0$. Then the posterior mean of μ_j is

$$\text{E}[\mu_j | \mathbf{y}] = (1 - b) y_j, \quad \text{where} \quad b = \frac{\sigma^2}{\sigma^2 + \sigma_\mu^2}.$$

But what if we do not know σ_μ^2, and thus do not know b?

A natural approach is then to try to *estimate* b. Marginally, we have

$$Y_j \sim \mathcal{N}(0, \sigma^2 + \sigma_\mu^2),$$

so

$$\frac{S}{\sigma^2 + \sigma_\mu^2} = \sum_{i=1}^{K} \left(\frac{Y_j}{\sqrt{\sigma^2 + \sigma_\mu^2}} \right)^2 \sim \chi_K^2, \quad \text{and then} \quad \mathrm{E}[S] = (\sigma^2 + \sigma_\mu^2)K.$$

Therefore,

$$\hat{\sigma}_\mu^2 = \frac{S}{K} - \sigma^2$$

is an unbiased estimator for σ_μ^2.

Plugging $\hat{\sigma}_\mu^2$ in for σ_μ^2 in the definition of b yields the estimator

$$\hat{b} = \frac{\sigma^2}{\sigma^2 + \hat{\sigma}_\mu^2} = \frac{K\sigma^2}{S} \sim \frac{K\sigma^2}{\sigma^2 + \sigma_\mu^2} \chi_K^{-2}.$$

However, $\mathrm{E}[\chi_K^{-2}] = 1/(K-2)$ (as shown using LOTUS in the footnote),[1] so $\hat{b}$ is a biased estimator of b. Instead, the James-Stein estimator uses the unbiased estimator

$$\hat{b}_{\mathrm{JS}} = \frac{(K-2)\sigma^2}{S}.$$

Replacing b by the unbiased estimator $\hat{b}_{\mathrm{JS}}$ in $\mathrm{E}[\mu_j|\mathbf{y}] = (1-b)y_j$ produces the James-Stein estimate for μ_j,

$$(1 - \hat{b}_{\mathrm{JS}})y_j = \left(1 - \frac{(K-2)\sigma^2}{S}\right) y_j.$$

The James-Stein estimator thus has a natural Bayesian interpretation in terms of the two-level model, in which the μ_j are i.i.d. with the same mean μ_0. Amazingly, Stein's theorem holds even without assuming the two-level model: it assumes nothing whatsoever about the μ_j, other than that they are real numbers!

9.11 Recap

The Bayesian approach to statistics is conceptually simple: apply the laws of probability and model all uncertainty using probability. The central object of interest is the posterior distribution of the

[1] If $X \sim \chi_K^2$, then

$$\mathrm{E}[X^{-1}] = \frac{1}{2^{K/2}\Gamma(K/2)} \int_0^\infty x^{-1} x^{K/2-1} e^{-x/2} dx = \frac{1}{2^{K/2}\Gamma(K/2)} \int_0^\infty x^{(K-2)/2-1} e^{-x/2} dx$$
$$= \frac{2^{(K-2)/2}\Gamma((K-2)/2)}{2^{K/2}\Gamma(K/2)} = \frac{\Gamma(K/2-1)}{2\Gamma(K/2)},$$

using the fact that $\Gamma(a+1) = a\Gamma(a)$ for any $a > 0$, so $\Gamma(K/2)/\Gamma(K/2-1) = (K/2-1) = (K-2)/2$. Thus $\mathrm{E}[X^{-1}] = \frac{1}{K-2}$, as stated.

estimand θ, given the data. That is, we can focus on probabilities for unknown quantities given known quantities.

The main ideas and notation for the chapter are listed in Table 9.2.

Formula or idea	Description or name
obey probability laws & model uncertainty using prob.	key Bayesian thinking
$\pi(\theta)\|\mathbf{y} \propto L(\theta;\mathbf{y})\pi(\theta)$	Bayes' theorem
$\pi(\theta)\|\mathbf{y}$	posterior
$\pi(\theta)$	prior
Cromwell's rule	don't put prior of 0 unless entirely certain
E$[\theta\|\mathbf{y}]$, $Q_{\theta\|\mathbf{y}}(0.5)$, MAP (posterior mode)	popular Bayesian estimates
MCMC: simulate $\theta\|\mathbf{y}$. Summarize simulations	computing Bayesian quantities
$P(\theta \in C(\mathbf{y})\|Y = \mathbf{y})$, e.g. $[Q_{\theta\|\mathbf{y}}(0.025), Q_{\theta\|\mathbf{y}}(0.975)]$	credible interval
Normal/Normal, Beta/Binomial, etc.	conjugate priors
$P(\text{Model}\|\mathbf{y})$	Bayesian model choice
$P(\theta \in \Theta_0\|\mathbf{y})$	posterior of null
Lindley's paradox	simple null: p-value not close to posterior of null
$Y_j\|\mu_j$, while $\mu_j\|\theta$ and θ	hierarchical models
improving over sample mean by shrinkage	Stein's paradox, inadmissibility, risk

Table 9.2: Main ideas and notation in Chapter 9.

Bayes' rule tells us how to update our prior for θ to the posterior given the data. Bayes' rule itself is uncontroversial, but many controversies have arisen over the choice of the prior. A conjugate prior is a very convenient choice in many situations, but convenient does not necessarily mean good, and even when using a conjugate prior the parameters of the prior distribution, which are known as hyperparameters, must be chosen.

The most widely used point estimators in Bayesian inference are the posterior mean, posterior median, and posterior mode. For interval estimation, we obtain credible intervals directly from the quantiles of the posterior distribution. A 95% credible interval may not be a 95% confidence interval, nor conversely; the interpretations of the two kinds of intervals are very different.

Bayesian computation made Bayesian inference intractable for centuries, but recent advances in Monte Carlo computation plus the availability of fast computers in the last few decades have made Bayesian inference a powerful, practical tool for a wide variety of applications. Markov chain Monte Carlo (MCMC) has had an especially large impact on the feasibility of carrying out Bayesian computations. The strategy is to simulate a large number of draws that are approximately from the posterior distribution and then, for example, approximate the posterior mean with the sample mean of the draws rather than doing difficult integrals.

9.12 R and Bayesian statistical inference

This chapter has looked at the broad principles of Bayesian inference. When discussing examples, we focused on prior distributions which are conjugate and so the computations were carried out analytically.

Section 9.4 outlined simulation tools which can be used to go beyond the conjugate case. These are crucial ideas for Bayesian work in practice, allowing the methods to be used in fantastically large variety of applications. These methods are the Bayesian equivalent to the routine use of numerical optimization which extends the applicability of maximum likelihood estimation beyond relatively simple models (e.g., moving from regression to logistic regression).

If MCMC simulation is needed to simulate from the posterior distribution, then it is often very helpful to use generic packages which embed professional knowledge in designing good MCMC methods for your problem in hand (just like in numerical optimization, we typically use a function to do the optimization for us, rather than coding up your own algorithm). An excellent platform for MCMC is `Stan`, which can be called from `R`, `Python`, or `Julia`. It is free to use. The page `https://mc-stan.org/` provides introductory material for `Stan`.

This section is completed with code from one example in the chapter.

Code and output for Example 9.10.1

```
set.seed(111);
n = 20; sigma=1; tau=1
d = sqrt(2)*sigma*sigma/tau
B = 1000000; mAns  = matrix(0,100,5)

Ubar = rnorm(B,0.0,1.0/sqrt(n))

for (i in (1:100)){
  mu = -3.0 + 6.0*i/101

  muhat = mu + Ubar
  mubar = (muhat-(d/n))*(muhat>=(d/n)) + (muhat+(d/n))*(muhat<=(-d/n))
  mutilde = muhat*(n*tau*tau)/(n*tau*tau + sigma*sigma)

  mAns[i,] = c(mu,mean(mubar),mean((mubar-mu)^2),
               mean((muhat-mu)^2),mean((mutilde-mu)^2))
}

#pdf("lasso.pdf")
  matplot(mAns[,1],mAns[,3:5],type="l",main="Risk function",
          xlab="mu",ylab="Risk",lwd=3)
#dev.off()
```

Chapter 10

Sampling and Resampling

10.1 Introduction

Statisticians describe, predict, and make causal statements using data that they view as random. Often, randomness is viewed as annoying noise that interferes with the precision of our inferences. However, there are also many problems where intentionally *creating* randomness is a powerful tool. Two such examples that we have seen so far are jittering of data to improve a visualization (see Section 6.2.2) and sampling from a posterior distribution to avoid difficult integrals (see Section 9.4).

The above-mentioned examples of deliberately introducing randomness are for *after* the data have been collected. In this chapter and the next, we introduce some important concepts for using randomness at the *design stage*, to help us gather better data for the problem at hand. In this chapter, we introduce tools for drawing a *random sample* from a finite population of interest, where the randomness helps make the sample resemble the population. In the next chapter, we consider *randomized experiments*, where the randomness helps us to draw valid causal conclusions.

In addition to sampling, this chapter introduces a technique called *resampling*, which (as the name suggests) involves generating new samples from an existing dataset. The most notable example of resampling is the *bootstrap*, which is a computationally intensive method for approximating the distribution of estimators and test statistics by resampling, rather than having to resort to asymptotic approximations. The bootstrap is conceptually similar to sampling from a posterior distribution, providing an attractive unity between Bayesian and frequentist inference methods.

Together, these methods provide fundamentally new ways of approaching statistical challenges.

10.2 Design-based inference

Throughout statistics and its applications, we encounter problems where a random *sample* is drawn from a *population* of interest, and we wish to use the sample to learn about an aspect of the population.

Sometimes the population is a hypothetical, infinite population, such as when we consider i.i.d. draws from a parametric statistical model $F_{Y;\theta}$. So far in this book we have focused on *model-based inference*, where we introduce a model for the data $Y_1, Y_2, \ldots, Y_n$, with unknown CDF $F_{Y;\theta}$.

In contrast, *design-based inference* involves sampling from a specific, finite population. The values in the population of the variable of interest are fixed numbers $y_1, y_2, \ldots, y_N$, where N is the population size. The randomness comes entirely from drawing a random sample from the population, not from, say, modeling the data as following a Normal distribution. In this section, we will introduce the design-based framework in detail. For some motivating examples where sampling from a finite population comes up, consider the following:

- A public opinion survey is being conducted in some city, state, or country, to assess views on some public policy issue or political candidate.
- In studying how common a certain disease or medical procedure is, medical records from a sample of hospitals may be inspected.
- When a company is performing quality control for a product, it may be infeasible to thoroughly test each and every unit they produce, so they may take a random sample of units in a batch, test the sample in depth, and then draw inferences about the entire product.
- To learn about the abundance of some kind of animal in a region (e.g., elk in a forest), an ecologist may count how many of the animals there are in some randomly chosen sites, and then attempt to infer how many there are in the entire region of interest.
- In auditing the financial records of a large company, a random sample of records may be studied very carefully to check for errors or fraud. Even if would have been feasible to study all the records, there are tradeoffs to consider: investigating a sample with great care may yield better results than a cursory inspection of a voluminous pile of records.

Throughout this section and the next, consider the following setup. There is a fixed, finite population of interest, consisting of N individuals. The individuals have been given ID numbers $1, 2, \ldots, N$, so that each individual is uniquely labeled. Let y_i be the quantity of interest for individual i (e.g., the ith person's age or income), so the entire finite population is

$$y_1, \ldots, y_N.$$

This is just a list of N numbers: the y_i are *not* regarded as coming from a statistical model. A model-based approach would have endowed the y_i's with a model; a design-based approach declines to do

so. Both approaches have strengths and weaknesses. A model-based approach with an approximately correct model may use the data more efficiently, while also relying less on assumptions about how the sampling was done. But sometimes we do not think we have enough information to come up with a reasonable model for the y_i, or we have a very tentative model, whose adequacy is highly debatable. A design-based approach can reduce concerns about whether we have a decent model and about subjectivity in the choice of the model.

Definition 10.2.1 (Finite sample estimands)**.** A *finite sample estimand* is an estimand that is defined as a function of $y_1, \ldots, y_N$. Some notable examples of finite sample estimands are

$$\mu = \frac{1}{N}\sum_{j=1}^{N} y_j, \quad \sigma^2 = \frac{1}{N}\sum_{j=1}^{N} (y_j - \mu)^2, \quad F(y) = \frac{1}{N}\sum_{j=1}^{N} I(y_j \leq y),$$

which are called the *population mean*, *population variance*, and *population CDF* of the y variable, respectively.

☣ **10.2.2.** Note that F is reminiscent of the empirical CDF (ECDF), which we introduced in Chapter 1. But there is a crucial conceptual difference. For the ECDF, we were *estimating* the true distribution of some random variable based on a sample. Here, F is the *true distribution* of the entire finite population. The same point holds for μ and σ^2, which at first glance look like estimates.

If we are able to observe *all* of the y_i's then we can compute μ, σ, and F deterministically. This situation is called a *census*.

Definition 10.2.3 (Census)**.** In a *census*, we get to observe all of $y_1, y_2, \ldots, y_N$.

☣ **10.2.4.** Some people claim that if we are able to do a census, there is no more need for statistics. However, although quantities like μ, σ, and F would then be known, there are still various relevant statistical tasks. For example, we may need statistical thinking to help *describe* trends and relationships in the population accurately, to *predict* future values or values for other populations, and to draw *causal* inferences.

Very often it is infeasible to conduct a census (e.g., too expensive or too logistically difficult). So instead we select a *random sample* of the y_i's.

Example 10.2.5. Let y_i be binary, taking the value 1 if the ith person in a city would test positive for antibodies to a particular virus if tested on March 2, 2020, and 0 otherwise. Knowing the infection rate,

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i,$$

is important for many reasons, e.g., since it impacts the hospitalization rate and fatality rate for this virus. It may be infeasible to test everyone in the population, so a sample may be needed. There are many complications in how to obtain a sample, e.g., how to select people to be tested, how to find those people, and what to do if those people do not consent to be tested.

In the next section, we will delve further into the design-based approach to sampling. We will consider various possible sampling schemes and, having specified a sampling scheme, we can try to come up with a good estimator and evaluate its properties.

10.3 Sampling design

In design-based inference, to compute F, μ, or σ exactly, we would need access to all of $y_1, \ldots, y_N$. This is often difficult or expensive to achieve. Instead, a powerful, widely-used technique is to draw a *random sample* and compute an estimate. Since we are taking a design-based perspective, the randomness comes only from the *sampling design*, the joint distribution of the IDs that get sampled.

☣ **10.3.1.** Why do we want a *random* sample? A *convenience sample* would, as the name suggests, involve just observing some set of y_i's that are convenient to access. A class of convenience samples which are influential in the social and medical sciences are *administrative data*, which are vast datasets collected for non-research purposes. Influential examples are income tax records and electronic medical records.

Convenience samples will be, typically, unacceptable for design-based inference. Suppose that $N = 100$ and, out of convenience, we choose to observe $y_1, y_2, \ldots, y_{10}$. There is no known link between this sample and the rest of the population, so we have no information about $y_{11}, y_{12}, \ldots, y_{100}$. Moreover, we are unable to make any statement along the lines of "the estimator is close to the estimand with high probability", since we have neither randomness from the data nor controlled randomness from the sample.

Random sampling is illustrated in Figure 10.1. This picture is inspired by Figure 1.11 of Diez, Cetinkaya-Rundel, and Barr (2020).

To specify and hopefully control the randomness of a random sample, we need the notion of a *sampling design*.

Definition 10.3.2 (Sampling design)**.** Suppose that we will collect a random sample of size n. Let I_j be the ID number of the jth individual selected. The *sampling design* is the joint probability mass

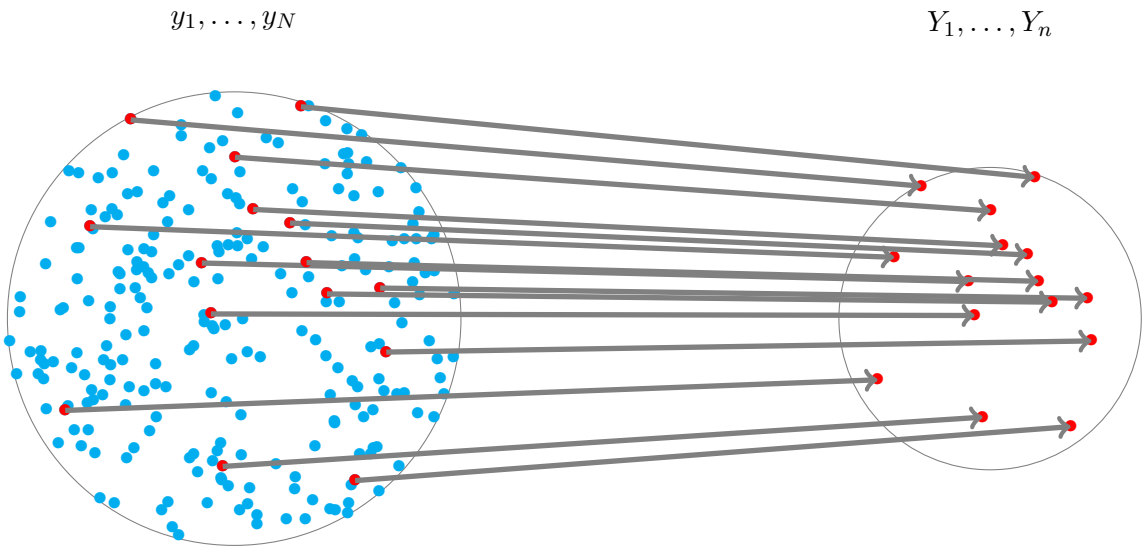


Figure 10.1: Random sample $Y_1, \ldots, Y_n$ of size $n = 15$ from population $y_1, \ldots, y_N$.

function of $I_1, \ldots, I_n$. That is, writing $\mathbf{i}_{1:n} = (i_1, \ldots, i_n)$, the sampling design specifies

$$P(I_1 = i_1, \ldots, I_n = i_n), \quad \text{for all} \quad i_{1:n} \in \{1, \ldots, N\}^n .$$

This joint probability can be written more compactly as $P(\mathbf{I}_{1:n} = \mathbf{i}_{1:n})$.

The sampling design determines the statistical properties of the outcomes

$$\mathbf{Y}_{1:n} = (Y_1, \ldots, Y_n) = (y_{I_1}, \ldots, y_{I_n}),$$

as we regard $y_1, \ldots, y_N$ as fixed, but the sampled IDs $\mathbf{I}_{1:n}$ as random. If we choose a sampling design and then collect data in accordance with the sampling design, then we will be able to make probabilistic statements about various estimators we can construct. A few natural estimators to consider are the descriptive estimators

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j, \quad S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2, \quad \hat{F}(y) = \frac{1}{n}\sum_{j=1}^{n} I(Y_j \leq y),$$

for μ, σ^2, and F, respectively.

Definition 10.3.3 (Equal probability sample)**.** A sampling design such that the marginal PMFs satisfy

$$P(I_j = k) = 1/N, \quad \text{for all } j = 1, \ldots, n, \quad k = 1, \ldots, N$$

is called an *equal probability sample.*

Let's first look at what happens when we have an equal probability sample of size 1, also known as a *uniformly random draw*.

Definition 10.3.4 (Uniformly random draw)**.** Let I be a random variable whose possible values are $1, 2, \ldots, N$, with equal probabilities. That is, I is Discrete Uniform on $\{1, 2, \ldots, N\}$. Let

$$Y = y_I.$$

Then Y is the y value for an individual chosen at random from the population. We call I a *uniformly random draw* from the population of ID numbers and Y a uniformly random draw from the population of y values. The values $y_1, \ldots, y_N$ are regarded as fixed, so all of the randomness in Y comes from the randomness in I. This is the essential perspective of design-based inference.

In the design-based approach, a useful interpretation of μ, σ^2 and F is that they are the mean, variance, and CDF of a uniformly random draw from the population. That is, letting Y and I be as in the previous definition, we have

$$\mathrm{E}[Y] = \sum_{j=1}^{N} P(I = j) y_j = \mu,$$

$$\mathrm{Var}(Y) = \mathrm{E}\left[(Y - \mu)^2\right] = \sum_{j=1}^{N} P(I = j)(y_j - \mu)^2 = \sigma^2,$$

$$\mathrm{E}[I(Y \leq y)] = \sum_{j=1}^{N} P(I = j) I(y_j \leq y) = F(y).$$

These results follow immediately from LOTUS, viewing Y as a function of I.

In a moment we will give the two most famous sampling designs: simple random sampling *with replacement* and simple random sampling *without replacement*. Both yield equal probability samples. But for now, the equal probability sample assumption is enough to generate some fundamental results.

Theorem 10.3.5. *All equal probability samples have the property that, for $j = 1, \ldots, n$, the*

$$\mathrm{E}_I[Y_j] = \mu, \quad \mathrm{Var}_I(Y_j) = \sigma^2, \quad \mathrm{E}_I[\bar{Y}] = \mu, \quad \mathrm{E}_I[\hat{F}(y)] = F(y),$$

where the expectation is over the sampling design, regarding $y_1, \ldots, y_N$ as fixed.

Proof. By linearity,

$$\mathrm{E}_I[\bar{Y}] = \frac{1}{n} \sum_{j=1}^{n} \mathrm{E}_I[y_{I_j}].$$

But

$$\mathrm{E}_I[Y_j] = \mathrm{E}_I[y_{I_j}] = \sum_{k=1}^{N} P(I_j = k) y_k = \frac{1}{N} \sum_{k=1}^{N} y_k = \mu,$$

by the equal probability sample assumption. The same argument yields $\mathrm{E}_I[\hat{F}(y)] = F(y)$. ∎

However, the equal probability sample assumption is not enough to tell us what the variance of $\bar{Y}$ or $\hat{F}(y)$ are, nor the expectation of S^2.

10.3.1 Sampling with replacement

Among the most widely used sampling designs in statistics is simple random sampling (SRS) with replacement. It appears in survey sampling, bootstrapping, and many other areas of application.

Definition 10.3.6 (SRS with replacement)**.** *Simple random sampling (SRS) with replacement* draws i.i.d. $I_1, \ldots, I_n$ from a Discrete Uniform distribution $\{1, 2, \ldots, N\}$, and sets

$$Y_j = y_{I_j}, \quad j = 1, \ldots, n.$$

The sampling is with replacement since the same ID number can be chosen multiple times. Conveniently, $Y_1, \ldots, Y_n$ are i.i.d., with each being a uniformly random draw from the population. The sampling design is

$$P(I_1 = i_1, \ldots, I_n = i_n) = \prod_{j=1}^{n} P(I_j = i_j) = 1/N^n, \quad i_{1:n} \in \{1, \ldots, N\}^n.$$

When calculating an expectation under SRS with replacement, we sometimes write E_{with} where the "with" serves as a reminder that the sampling is with replacement.

So SRS with replacement is an equal probability sample.

Example 10.3.7. Let $N = 7$. We sampled with replacement $n = 5$ times from $(y_1, \ldots, y_N) = (37, 6, 36, 42, 28, 31, 10)$. The result was $(31, 36, 42, 36, 10)$, which corresponds to the sampled IDs being $(6, 3, 4, 3, 7)$. Note that individual 3 got sampled twice, resulting in the value 36 appearing twice among the sampled y values. The `R` code for this example is given in Section 10.9.

Example 10.3.8 (Implementing SRS with replacement)**.** A simple way to implement SRS with replacement is to draw i.i.d. $U_1, \ldots, U_n \sim \text{Unif}(0, 1)$ and then let the jth sampled ID number be $\lceil NU_j \rceil$. This works because $NU_j \sim \text{Unif}(0, N)$ and the ceiling function rounds everything in $(0, 1]$ to 1, everything in $(1, 2]$ to 2, etc., and for a Uniform r.v., probability is proportional to length. It is even easier

when working in a language with a direct sampling with replacement command, e.g., in `R` we can use s`sample(y,n,replace=TRUE)`.

Many properties of SRS with replacement follow immediately (as $Y_1, \ldots, Y_n$ are i.i.d. from F) from the results from Chapter 3 on the bias, variance, and asymptotic distribution of descriptive statistics. For example, the estimators $\bar{Y}$, S^2 and $\hat{F}(y)$ are all unbiased for their respective estimands, and

$$\mathrm{Var}_{\mathrm{with}}(\bar{Y}) = \sigma^2/n, \quad \mathrm{Var}_{\mathrm{with}}(\hat{F}(y)) = F(y)(1-F(y))/n.$$

Furthermore, the CLT is always applicable to $\bar{Y}$ (since $\mathrm{Var}(Y_j) = \sigma^2$ is finite), and we have the following asymptotic results:

$$\sqrt{n}(\bar{Y}-\mu) \xrightarrow{d} \mathcal{N}(0,\sigma^2), \quad \sqrt{n}(S^2-\sigma^2) \xrightarrow{d} \mathcal{N}(0,\mu_4-\mu_2^2),$$
$$\sqrt{n}\{\hat{F}(y)-F(y)\} \xrightarrow{d} \mathcal{N}(0,F(y)\{1-F(y)\}).$$

Example 10.3.9 (Continuing Example 10.2.5)**.** Let μ, the proportion of the population who have antibodies to the virus, be the estimand. Suppose that a SRS with replacement is obtained, with sample size n. Then

$$\mathrm{E}_{\mathrm{with}}[\bar{Y}] = \mu, \quad \mathrm{Var}_{\mathrm{with}}(\bar{Y}) = \frac{\mu(1-\mu)}{n}, \quad \frac{\sqrt{n}(\bar{Y}-\mu)}{\sqrt{\mu(1-\mu)}} \xrightarrow{d} \mathcal{N}(0,1),$$

where the subscript is a reminder that the sampling was with replacement. By the law of large numbers, continuous mapping theorem, and Slutsky's theorem, we can also replace the denominator by an estimated version of it, obtaining

$$\frac{\sqrt{n}(\bar{Y}-\mu)}{\sqrt{\bar{Y}(1-\bar{Y})}} \xrightarrow{d} \mathcal{N}(0,1).$$

10.3.2 Sampling without replacement

Sampling with replacement allows the same individual to be sampled multiple times. The consequences of this depend upon the scientific context and the sizes of n and N.

Remark 6. The chance that at least one individual is sampled more than once, in a SRS with replacement of size n from a population of size N, is isomorphic to the birthday problem (see sections 1.4 and 4.7 of the Stat 110 book). Imagine that there are N days in a year (on some planet) and n people whose birthdays are being compared. By Poisson approximation, the probability of at least one repetition is approximately

$$1 - \exp\left\{-\frac{n(n-1)}{2N}\right\}.$$

Applying this approximation to the sampling with replacement problem, this probability of a repetition is remarkably high in many cases, e.g., for $n = 2,000$ and $N = 10^6$ it is about 95%, even though $2,000$ is tiny compared to a million.

In many applications, it makes sense to avoid sampling the same individual repeatedly (e.g., it makes no sense to ask the same person multiple times whether they are unemployed, unless we are studying changes over time). In sampling *without* replacement, we do not allow choosing the same individual more than once. Often, sampling with replacement is more *mathematically* convenient to analyze than sampling without replacement, but worse in terms of *statistical* efficiency (and less sensible in practice).

By the multiplication rule, if we choose a sample of size n without replacement, there are

$$N(N-1)\dots(N-n+1) = \frac{N!}{(N-n)!}$$

possibilities. For example, if $N = 3$ and $n = 2$, then there are 6 permutations:

$$(1,2),(1,3),(2,1),(2,3),(3,1),(3,2).$$

If we do not care about the ordering of the IDs, then there are

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

possible combinations. For example, if $N = 3$ and $n = 2$, then there are 3 combinations of IDs:

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

Simple random sampling with replacement samples from the $N!/(N-n)!$ permutations with equal probability on each permutation. It follows by symmetry that the $\binom{N}{n}$ combinations (where we ignore the order in which the sample was collected) are equally likely.

Definition 10.3.10 (SRS without replacement)**.** A *simple random sample without replacement* of size $n \leq N$ from the population is a sample of size n, chosen without replacement, such that all $N!/(N-n)!$ possible permutations are equally likely. Then set

$$Y_j = y_{I_j}, \quad j = 1,\dots,n.$$

When calculating an expectation under SRS without replacement, we sometimes write $\mathrm{E}_{\mathrm{w/o}}$ where the "w/o" is short for "without" and serves as a reminder that the sampling is without replacement.

In terms of the sampling design, the SRS without replacement scheme has

$$P(I_1 = i_1, \ldots, I_n = i_n) = \frac{1}{N!/(N-n)!}$$

for all $i_{1:n}$ without repetitions, since all $i_{1:n}$ without repetitions are equally likely. A natural way to implement SRS without replacement is to choose the first individual uniformly at random, the second individual uniformly at random from everyone other than the first sampled individual, etc. This is a valid way to obtain a SRS without replacement since then, by the prediction decomposition,

$$\begin{aligned} P(I_1 = i_1, \ldots, I_n = i_n) &= P(I_1 = i_1) \prod_{j=2}^{n} P(I_j = i_j | I_{1:j-1} = i_{1:j-1}) \\ &= \frac{1}{N} \prod_{j=2}^{n} \frac{1}{N-j+1} = \frac{1}{N(N-1)\ldots(N-n+1)} = \frac{(N-n)!}{N!} \\ &= \frac{1}{N!/(N-n)!}, \end{aligned}$$

for all possible $i_{1:n}$.

By symmetry, SRS without replacement is an equal probability sampling scheme. So Theorem 10.3.5 again applies, giving

$$\mathrm{E}_{\mathrm{w/o}}[Y_j] = \mu, \, \mathrm{Var}_{\mathrm{w/o}}(Y_j) = \sigma^2.$$

However, we now have

$$\mathrm{Cov}_{\mathrm{w/o}}(Y_j, Y_k) \neq 0, \quad j, k \in \{1, 2, \ldots, N\},$$

since the sampled IDs are *dependent*. We will return to this crucial covariance soon.

Remark 7. The distinction between with replacement and without replacement parallels the distinction between the Binomial and Hypergeometric distributions. Indeed, if each member of the population is labeled as either *tagged* or *untagged*, and there are t tagged individuals and $N-t$ untagged individuals, then the number of tagged individuals in a simple random sample without replacement of size n is $\mathrm{HGeom}(t, N-t, n)$, while the number of tagged individuals in a simple random sample with replacement of size n is $\mathrm{Bin}(n, \frac{t}{N})$. See Section 3.9 of the Stat 110 book for discussion of the relationship between the Binomial and the Hypergeometric.

Example 10.3.11. In Example 10.3.11, we sampled with replacement 12 times from $(y_1, \ldots, y_N) = (37, 6, 36, 42, 28, 31, 10)$. Now let's sample *without* replacement. Now we need $n \leq N$ (else we will run out of individuals to sample!). We sampled without replacement $n = 5$ times from $(y_1, \ldots, y_N)$. The result was $(31, 36, 42, 10, 28)$, corresponding to the sampled IDs being $(6, 3, 4, 7, 5)$. There are no repeated IDs here, by design. The `R` code for this example is given in Section 10.9.

Let $Y_1, \ldots, Y_n$ be a SRS without replacement. Intuitively, the Y_j should be negatively correlated since, for example, if we observe a value of Y_1 that is greater than μ then we should predict that Y_2 will be less than μ, since the average of y_i's with the first sampled individual removed is less than the average with the first sampled individual included. By linearity, we still have $\mathrm{E}_{\mathrm{w/o}}[\bar{Y}] = \mu$, but to find the variance of $\bar{Y}$ requires more work since the Y_j are negatively correlated.

Theorem 10.3.12. *For SRS without replacement, for $j, k \in \{1, \ldots, n\}$ and $j \neq k$,*

$$\mathrm{E}_{\mathrm{w/o}}[Y_j] = \mu, \quad \mathrm{Var}_{\mathrm{w/o}}(Y_j) = \sigma^2, \quad \mathrm{Cov}_{\mathrm{w/o}}(Y_j, Y_k) = -\frac{\sigma^2}{N-1}.$$

Consequently,

$$\mathrm{E}_{\mathrm{w/o}}[\bar{Y}] = \mu, \quad \mathrm{Var}_{\mathrm{w/o}}(\bar{Y}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}.$$

Proof. The results for $\mathrm{E}_{\mathrm{w/o}}[Y_j]$ and $\mathrm{Var}_{\mathrm{w/o}}(Y_j)$ were shown above. For the covariance, we first compute the expected value of $Y_1 Y_2$. Note that

$$\mathrm{E}_{\mathrm{w/o}}[Y_2|Y_1] = \frac{1}{N-1}(N\mu - Y_1),$$

since given Y_1, when considering Y_2 we can imagine that the individual corresponding to Y_1 has been removed from the population and then a uniformly random individual is chosen from the $N-1$ remaining individuals. Then by Adam's law,

$$\begin{aligned}
\mathrm{E}_{\mathrm{w/o}}[Y_1 Y_2] &= \mathrm{E}[\mathrm{E}[Y_1 Y_2|Y_1]] \\
&= \mathrm{E}[\mathrm{Y_1}\mathrm{E}[Y_2|Y_1]] \\
&= \frac{1}{N-1}(N\mu^2 - \mathrm{E}_{\mathrm{w/o}}[Y_1^2]) \\
&= \frac{1}{N-1}\{(N-1)\mu^2 - \sigma^2\} \\
&= \mu^2 - \frac{\sigma^2}{N-1}.
\end{aligned}$$

Thus,

$$\mathrm{Cov}_{\mathrm{w/o}}(Y_1, Y_2) = \mathrm{E}_{\mathrm{w/o}}[Y_1 Y_2] - \mathrm{E}_{\mathrm{w/o}}[Y_2]\mathrm{E}_{\mathrm{w/o}}[Y_1] = -\sigma^2/(N-1).$$

By symmetry, the covariance is the same for any Y_j and Y_k with $j \neq k$.

Alternatively, we can derive the covariance by leveraging the fact that in the design-based framework, $y_1, \ldots, y_N$ are fixed. Imagine that the sampling continues until all N individuals get sampled (it is valid to consider this since $Y_1, Y_2, \ldots, Y_n$ do not depend on whether or not the sampling continues after sampling n individuals). Note that

$$Y_1 + Y_2 + \cdots + Y_N = y_1 + y_2 + \cdots + y_N,$$

since the sampling is without replacement. But the right-hand side is a constant, so its variance is zero! Taking the variance of both sides, we have

$$\begin{aligned} 0 &= \mathrm{Var}_{\mathrm{w/o}}(Y_1 + \cdots + Y_N) = N\mathrm{Var}_{\mathrm{w/o}}(Y_1) + 2\binom{N}{2}\mathrm{Cov}_{\mathrm{w/o}}(Y_1, Y_2) \\ &= N\sigma^2 + N(N-1)\mathrm{Cov}_{\mathrm{w/o}}(Y_1, Y_2), \end{aligned}$$

which again yields

$$\mathrm{Cov}_{\mathrm{w/o}}(Y_1, Y_2) = -\sigma^2/(N-1).$$

Lastly, to obtain $\mathrm{Var}_{\mathrm{w/o}}(\bar{Y})$, we expand the variance in terms of variances and covariances:

$$\mathrm{Var}_{\mathrm{w/o}}(\bar{Y}) = n^{-2}\left[n\sigma^2 + n(n-1)\mathrm{Cov}_{\mathrm{w/o}}(Y_1, Y_2)\right] = \frac{\sigma^2}{n}\frac{N-n}{N-1}.$$

∎

Definition 10.3.13 (Finite population correction)**.** The factor $(N-n)/(N-1)$ that appears in the variance for sampling without replacement is called the *finite population correction*.

This factor is less than 1 for $n > 1$, so the sample average has lower variance for sampling without replacement than for sampling with replacement. This fact makes sense intuitively due to the negative correlation between the Y_j, and since it seems redundant to sample the same individual more than once. Of course, if N is massive compared to n, then the finite population correction is minor since it is then approximately equal to 1. Also note that in the extreme case $n = N$, the finite population correction is 0, which also makes sense since then we have a census, and $\bar{Y}$ will always equal μ.

10.3.3 Stratified sampling

In sampling, it is often desirable to partition the population $y_1, \ldots, y_N$ into subpopulations. This is called *stratification*, and each subpopulation is called a *stratum*. In *stratified sampling*, a sample is drawn from each stratum, with these samples independent across strata.

Typically, the strata are subgroups of interest to the investigator. In terms of statistical efficiency, it is desirable that individuals *within* each stratum be more similar to each other than individuals *between* different strata. In addition to helping us learn about subpopulations of interest, it is often possible to use stratified sampling to lower the overall variance of our estimators for the entire population.

Example 10.3.14. Geographic areas such as cities or counties are often used as strata, as are age, gender, ethnicity, and industry. For example, suppose that we are conducting a public opinion poll in the state of Massachusetts. We may then want to stratify by city (so each stratum is the people

who live in a particular city within Massachusetts). Suppose that in addition to being interested in the results for Massachusetts, we are also interested in results for Cambridge, the city within greater Boston which houses Harvard and MIT. With a SRS, by chance we may end up with no one or almost no one in Cambridge being in the sample. With stratified sampling, we can guarantee a prescribed sample size for Cambridge. This may be useful both for learning about Cambridge (and other specific cities) and for improving our estimators overall.

Definition 10.3.15 (Strata)**.** Partition the IDs into L subsets, called *strata*. Index the strata by ℓ, and refer to the ID within the ℓth stratum by the pair

$$(j,l)\,, \quad j=1,\ldots,N_l, \quad \ell=1,\ldots,L,$$

where N_l is the size of the ℓth stratum. Assume that each $N_\ell \geq 1$ and $\sum_{\ell=1}^{L} N_\ell = N$. Within the ℓth stratum, the fixed population is

$$y_{1,\ell},\ldots,y_{N_\ell,\ell}.$$

Define the population quantities within the ℓth stratum as:

$$\mu_\ell = N_\ell^{-1}\sum_{j=1}^{N_\ell} y_{j,\ell}, \quad \sigma_\ell^2 = N_\ell^{-1}\sum_{j=1}^{N_\ell}(y_{j,\ell}-\mu_\ell)^2, \quad F_\ell(y) = N_\ell^{-1}\sum_{j=1}^{N_\ell} I(y_{j,\ell}\leq y).$$

The population-specific quantities can be recovered from these stratum-specific quantities.

Theorem 10.3.16. *The stratum-specific quantities can be used to produce the corresponding population-specific quantities via*

$$\mu = \sum_{\ell=1}^{L}\frac{N_\ell}{N}\mu_\ell, \quad \sigma^2 = \sum_{\ell=1}^{L}\frac{N_\ell}{N}\sigma_\ell^2 + \sum_{\ell=1}^{L}\frac{N_\ell}{N}(\mu_\ell-\mu)^2, \quad F(y) = \sum_{\ell=1}^{L}\frac{N_\ell}{N}F_\ell(y).$$

Proof. The estimand μ involves the sum of y values across the entire population. Since we have divided the population into strata, it is then natural to re-group this sum so that we add up the y values within each stratum, and then sum up these stratum totals. Writing this out, we have

$$\mu = \frac{1}{N}\sum_{\ell=1}^{L}\sum_{j=1}^{N_\ell} y_{j,\ell} = \sum_{\ell=1}^{L}\frac{N_\ell}{N}\left(\frac{1}{N_\ell}\sum_{j=1}^{N_\ell} y_{j,\ell}\right) = \sum_{\ell=1}^{L}\frac{N_\ell}{N}\mu_\ell,$$

as desired. The result for $F(y)$ follows from the same argument. The result for σ^2 takes some more

work to show. The algebraic strategy is to add 0, in the form $\mu_\ell - \mu_\ell$.

$$\begin{aligned}
\sum_{j=1}^{N}(y_{j,\ell}-\mu)^2 &= \sum_{\ell=1}^{L}\sum_{j=1}^{N_\ell}\{y_{j,\ell}-\mu_\ell+(\mu_\ell-\mu)\}^2 \\
&= \sum_{\ell=1}^{L}\sum_{j=1}^{N_\ell}(y_{j,\ell}-\mu_\ell)^2 + 2\sum_{\ell=1}^{L}\sum_{j=1}^{N_\ell}(y_{j,\ell}-\mu_\ell)(\mu_\ell-\mu) + \sum_{\ell=1}^{L}\sum_{j=1}^{N_\ell}(\mu_\ell-\mu)^2 \\
&= \sum_{\ell=1}^{L}N_l\sigma_\ell^2 + 2\sum_{\ell=1}^{L}(\mu_\ell-\mu)\sum_{j=1}^{N_\ell}(y_{j,\ell}-\mu_\ell) + \sum_{\ell=1}^{L}N_\ell(\mu_\ell-\mu)^2 \\
&= \sum_{\ell=1}^{L}N_\ell\sigma_\ell^2 + \sum_{\ell=1}^{L}N_\ell(\mu_\ell-\mu)^2.
\end{aligned}$$

Dividing by N then delivers the stated result. ∎

A useful way of thinking of these results is to think of $N_1/N, \ldots, N_L/N$ as positive weights which sum to one. Hence μ is a weighted average of the $\mu_1, \ldots, \mu_L$, giving larger weights to larger strata.

Next, let's consider a design for *sampling* from a stratified finite population. First split the sample size n, into strata sample sizes

$$n_1, \ldots, n_L, \quad n_\ell \geq 1, \quad n = \sum_{\ell=1}^{L} n_l,$$

determined by the researcher. The sampled jth ID in the ℓth stratum is written as

$$I_{j,\ell} \in \{1, \ldots, N_\ell\}, \quad j = 1, \ldots, n_\ell, \quad \ell = 1, \ldots, L.$$

Then a stratified sampling design carries out sampling independently across strata.

Definition 10.3.17 (Stratified sampling design)**.** Collect all the sampled IDs in the ℓth stratum as

$$\mathbf{I}_{1:n_\ell,\ell} = (I_{1,\ell}, \ldots, I_{n_\ell,\ell}), \quad \ell = 1, \ldots, L.$$

A sampling design is a *stratified sampling design* if the sampling is done independently across strata. If $P(I_{j,\ell} = k) = 1/N_\ell$ for all $k = 1, \ldots, N_\ell$, $j = 1, \ldots, n_\ell$ and $\ell = 1, \ldots, L$, then the design is an *equal probability stratified sample*.

The outcomes can be used as the input into stratum-specific estimators (of μ_l and $F_l(y)$), such as

$$\overline{Y}_\ell = n_\ell^{-1}\sum_{j=1}^{n_\ell} Y_{j,\ell}, \quad \widehat{F}_\ell(y) = n_\ell^{-1}\sum_{j=1}^{n_\ell} I(Y_{j,\ell} \leq y), \quad \ell = 1, \ldots, L,$$

while they can be pooled, using the weights from Theorem 10.3.16, as the following population estimators of μ and $F(y)$:

$$\widehat{\mu}_{\text{strat}} = \sum_{\ell=1}^{L}\frac{N_\ell}{N}\overline{Y}_\ell, \quad \widehat{F}_{\text{strat}}(y) = \sum_{\ell=1}^{L}\frac{N_\ell}{N}\widehat{F}_\ell(y).$$

☣ **10.3.18.** Imagine that researcher A conducts stratified random sampling (with a SRS without replacement for each stratum), and hands over the data to researcher B without bothering to mention how the data were sampled. Then researcher B carelessly assumes that the data were an SRS without replacement, and uses the naive sample mean $\bar{Y}$ to estimate the population mean μ. Let $n = n_1 + \cdots + n_L$ be the total sample size, and $Y_{j,\ell}$ be the jth observation in the SRS for stratum ℓ, for $j = 1, 2, \ldots, n_\ell$. Then the naive sample mean is

$$\bar{Y} = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{j=1}^{n_\ell} Y_{j,\ell},$$

which has expected value

$$\mathrm{E}[\bar{Y}] = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{j=1}^{n_\ell} \mathrm{E}[Y_{j,\ell}] = \frac{1}{n} \sum_{\ell=1}^{L} \sum_{j=1}^{n_\ell} \mu_\ell = \frac{1}{n} \sum_{\ell=1}^{L} n_\ell \mu_\ell.$$

This shows that $\bar{Y}$ could be terribly biased, even if n is large. The researcher needs to use a weighted estimator instead, such as $\widehat{\mu}_{\text{strat}}$.

For any stratified sampling design,

$$\mathrm{E}(\widehat{\mu}_{\text{strat}}) = \sum_{\ell=1}^{L} \frac{N_\ell}{N} \mathrm{E}[\overline{Y}_\ell], \quad \mathrm{Var}(\widehat{\mu}_{\text{strat}}) = \sum_{\ell=1}^{L} \frac{N_\ell^2}{N^2} \mathrm{Var}(\overline{Y}_\ell), \tag{10.1}$$

where the latter result used the independence of the sampling between strata, which implies the $\overline{Y}_\ell$ are independent over l. Assuming equal probability stratified sampling, $\mathrm{E}[\overline{Y}_\ell] = \mu_\ell$, so $\mathrm{E}[\widehat{\mu}_{\text{strat}}] = \mu$. The variances vary depending on the ℓth stratum sampling design, with

$$\begin{aligned} \mathrm{Var}(\overline{Y}_\ell) &= \frac{1}{n_\ell} \sigma_\ell^2, \quad \text{under SRS with replacement,} \\ \mathrm{Var}(\overline{Y}_\ell) &= \frac{1}{n_\ell} \frac{N_\ell - n_\ell}{N_\ell - 1} \sigma_\ell^2, \quad \text{under SRS without replacement,} \end{aligned}$$

which determine the corresponding $\mathrm{Var}(\widehat{\mu}_{\text{strat}})$, through the weighted sum in (10.1). Notice that $N_\ell \geq 2$ for all ℓ is needed for the $\mathrm{Var}(\overline{Y}_\ell)$ to exist under SRS without replacement.

An important consideration in designing a stratified sampling scheme is the choice of the n_ℓ. A common choice in practice is *proportional allocation*, which sets n_ℓ to be proportional to N_ℓ. In principle, it would be better (in terms of minimizing the variance of μ_{strat} to use *Neyman allocation*, which sets n_ℓ to be proportional to $N_\ell \sigma_\ell$, but this may not be feasible since σ_ℓ is typically unknown (and when deciding the stratum sample sizes we haven't yet collected data, so don't have data with which to estimate σ_ℓ).

10.4 Horvitz–Thompson estimator

The Horvitz–Thompson estimator is a very general way to construct an unbiased estimator of μ, for *any* sampling design such that for each individual there is a known, positive probability that the individual will be included in the sample.

Definition 10.4.1 (Horvitz–Thompson estimator)**.** Let the sampling design be

$$P(I_1 = i_1, \ldots, I_n = i_n).$$

Let

$$C_j = I(I_1 = j) + \cdots + I(I_n = j), \quad j = 1, ..., N,$$

be the number of times that ID j is selected for the sample. (So $C_j \leq 1$ if the sampling is without replacement.) Let the *inclusion probability* of ID j be

$$\pi_j = P(C_j \geq 1) = 1 - P(C_j = 0).$$

Assume that N and π_j are known, with $\pi_j > 0$ for all j. The *Horvitz–Thompson estimator* of μ is

$$\hat{\mu}_{\text{HT}} = \frac{1}{N} \sum_{j=1}^{N} \frac{I(C_j \geq 1)}{\pi_j} y_j,$$

At first sight this looks infeasible to compute, as we don't know all $y_1, \ldots, y_N$ (if we did, we could calculate μ and would not need an estimator). However, in this sum $I(C_j \geq 1) = 0$ if y_j is not in the sample, and then that term in the sum is zeroed out! Hence this estimator is simple to compute: it says to sum up y_j/π_j over all individuals j that appear at least once in the sample, and then divide by N.

The Horvitz–Thompson is an unbiased estimator of the population mean.

Theorem 10.4.2 (Unbiasedness of Horvitz–Thompson estimator)**.** *With notation as above,*

$$\mathrm{E}[\hat{\mu}_{HT}] = \mu.$$

Proof. By linearity,

$$\mathrm{E}[\hat{\mu}_{\text{HT}}] = \frac{1}{N} \sum_{j=1}^{N} \frac{y_j \mathrm{E}[I(C_j \geq 1)]}{\pi_j} = \frac{1}{N} \sum_{j=1}^{N} \frac{y_j \pi_j}{\pi_j} = \frac{1}{N} \sum_{j=1}^{N} y_j = \mu.$$

∎

Example 10.4.3. For SRS without replacement,

$$\pi_j = \frac{n}{N},$$

since the events $I_1 = j, I_2 = j, \ldots, I_n = j$ are disjoint, each with probability $1/N$, so the probability of their union is n/N. The Horvitz–Thompson estimator is then

$$\hat{\mu}_{\text{HT}} = \frac{1}{N}\sum_{j=1}^{n}\frac{Y_j}{\pi_{I_j}} = \frac{1}{N}\sum_{j=1}^{n}\frac{N}{n}Y_j = \frac{1}{n}\sum_{j=1}^{n}Y_j = \bar{Y},$$

the ordinary sample mean.

For SRS with replacement, the inclusion probability changes to

$$\pi_j = 1 - \{1 - P(I_1 = j)\}^n = 1 - \left(\frac{N-1}{N}\right)^n.$$

Then

$$\hat{\mu}_{\text{HT}} = \frac{T}{N\left(1 - \left(\frac{N-1}{N}\right)^N\right)},$$

where T is the total of the y values of the *distinct* IDs in the sample, e.g., even if ID 3 gets sampled 4 times, only one y_3 term appears in T. For example, consider the sample from Example 10.3.7, where we sampled with replacement $n = 5$ times from $(y_1, \ldots, y_N) = (37, 6, 36, 42, 28, 31, 10)$. The result was $(31, 36, 42, 36, 10)$. The population mean is

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i = 27.14286.$$

The observed value of the sample mean is

$$\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j = 31.$$

The value 36 happened to appear twice, so the sample mean (unwisely) puts twice as much weight on the 36 as on a value that appeared only once. For the Horvitz–Thompson estimator, we instead calculate

$$\hat{\mu}_{\text{HT}} = \frac{31 + 36 + 42 + 10}{N\left(1 - \left(\frac{N-1}{N}\right)^N\right)} = 25.75432,$$

which turns out to be much closer than $\bar{Y}$ to μ. Note that in the numerator of $\hat{\mu}_{\text{HT}}$, there is only one term corresponding to each ID that got sampled, even if that ID got sampled more than once.

The Horvitz–Thompson estimator is quite general, but it does rely on the major assumption that the inclusion probabilities $\pi_1, \ldots, \pi_N$ are known. In practice the π_j are often unknown, or known

only approximately, and replacing the π_j by estimators typically introduces bias. For known π_j, the Horvitz–Thompson estimator is always unbiased, but in some situations it has a very large variance (and hence very large MSE). Careful thought is required in designing how to collect the sample as well as in deciding how to analyze the data.

10.5 The bootstrap

The bootstrap is a powerful framework introduced by Bradley Efron in 1979, in which computer intensive work rather than mathematical intensive work is used for tasks such as studying the standard error of an estimator or constructing approximate confidence intervals for a parameter. Using the bootstrap is often much easier than a more mathematical approach, since it entails writing a couple lines of code and letting the computer do the calculations, rather than doing difficult or tedious mathematical derivations of asymptotic distributions by hand. In addition to saving time and effort, the bootstrap often yields *better* answers than classical approaches that may involve strong parametric assumptions or questionable asymptotic approximations.

The bootstrap is based on *resampling*: we have a sample from a population but then treat the sample itself as a new population, from which we draw new samples via simulation. This idea is both simple and subtle: it is easy to describe and implement the procedure, but far from obvious that it is a useful procedure. In fact, the bootstrap turns out to be amazingly useful for a wide variety of problems.

Definition 10.5.1 (Bootstrap)**.** Let $\mathbf{y} = (y_1, \ldots, y_n)$ be the observed dataset, assumed to be i.i.d. from an unknown CDF F. We make no parametric assumptions about F (parametric versions of the bootstrap also exist and can be useful, but here we will focus on the nonparametric bootstrap). Create a synthetic dataset

$$\mathbf{Y}^* = (Y_1^*, \ldots, Y_n^*),$$

by performing a simple random sample with replacement from $(y_1, \ldots, y_n)$. Equivalently, let

$$Y_j^* \overset{\text{i.i.d.}}{\sim} \hat{F}, \quad j = 1, \ldots, n$$

where $\hat{F}$ is the ECDF of $\mathbf{y}$:

$$\hat{F}(y) = \frac{1}{n} \sum_{j=1}^{n} I(y_j \leq y).$$

Then $(Y_1^*, \ldots, Y_n^*)$ is called a *bootstrap sample*. The *bootstrap* entails generating some large number B of independent bootstrap samples (each of size n), and then using the bootstrap samples for inferential tasks such as approximating the standard error of an estimator.

In particular, suppose that we have an estimand θ and an estimator $\hat{\theta}$, and we want to approximate the standard error of $\hat{\theta}$. This is often difficult to do mathematically (once we go beyond especially tractable estimators like the sample mean). The bootstrap provides a simulation-based approach.

Generate B bootstrap samples with B large, and for each bootstrap sample compute a replicate $\hat{\theta}^*$ based on that sample, using the same procedure as was used to compute $\hat{\theta}$ based on the real data. That is, if $\hat{\theta}$ is the statistic $T(\mathbf{y})$, then for a bootstrap sample $\mathbf{Y}^*$ we calculate $\hat{\theta}^* = T(\mathbf{Y}^*)$. In this way, repeating B independent times, we obtain B *bootstrap replications*

$$\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B.$$

The bootstrap approximation of the standard error of $\hat{\theta}$ is then

$$\mathrm{SE}_{\mathrm{boot}}(\hat{\theta}^*) = \sqrt{\mathrm{E}_{\mathrm{boot}}[(\hat{\theta}^* - \mathrm{E}_{\mathrm{boot}}[\hat{\theta}^*])^2]}, \tag{10.2}$$

which can be estimated by the *bootstrap standard error*

$$\hat{\mathrm{SE}}_{\mathrm{boot}}(\hat{\theta}^*) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^*_b - \bar{\theta^*}\right)^2}, \quad \bar{\theta^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_b.$$

The bootstrap expectation, $\mathrm{E}_{\mathrm{boot}}$, in (10.2), is calculated as sampling with replacement from $\hat{F}$, conditional on the underlying data $(y_1, \ldots, y_n)$.

Example 10.5.2. Return to the duration of birth example from Chapter 1. For ease of printing on the page in this example, we will think of the subset of the data

```
19.00  9.50  3.40  7.30 16.00  8.50  9.75 10.00  2.10  4.25
```

will be the whole dataset and denote it here as $\mathbf{y}$. We will run the bootstrap with $B = 20$ bootstrap samples (in practice, we would take B to be much larger, such as $B = 10^4$). The bootstrap samples are:

```
 7.30  3.40  2.10 16.00  3.40 10.00  4.25 19.00  4.25  7.30
10.00  4.25  2.10 10.00 19.00  9.75 16.00 19.00  2.10 10.00
 2.10  3.40  9.75  7.30  7.30  8.50  2.10  7.30  7.30  8.50
16.00 16.00 16.00  8.50  7.30  2.10  4.25  9.50  4.25 10.00
 9.75 16.00 19.00  4.25  4.25  9.75  7.30  7.30 19.00  2.10
 8.50 10.00  3.40  4.25 10.00  4.25 19.00  2.10  4.25  4.25
 9.50 16.00  3.40 19.00  2.10  8.50  9.75  8.50  9.50  2.10
 9.50  8.50  7.30  7.30 16.00  8.50 10.00  8.50  8.50  9.75
19.00  4.25  7.30 10.00  9.75  3.40  4.25  9.75 16.00  9.50
 3.40  2.10  8.50 10.00  7.30  7.30  4.25  8.50  9.50  3.40
 9.50  3.40  8.50 16.00  3.40  2.10  4.25  7.30  3.40  4.25
19.00  2.10  4.25  9.75  3.40 10.00  3.40  4.25  4.25  4.25
```

```
 3.40 10.00 19.00 16.00  8.50  3.40  9.50  2.10  9.75 10.00
 3.40  7.30  8.50  4.25  7.30 10.00 10.00  8.50  8.50  9.75
19.00  2.10  3.40  9.75  7.30 10.00  3.40  4.25  4.25 16.00
10.00  9.75  2.10 19.00  7.30  2.10  8.50  8.50  4.25  8.50
 3.40  2.10 16.00 10.00 19.00  9.50 19.00 19.00 10.00  9.50
 9.75  2.10 16.00  3.40 19.00  3.40  3.40  4.25  3.40 10.00
 9.50  8.50 16.00  2.10 10.00  8.50  9.50  9.75 19.00  9.75
 9.50  9.50  9.50 10.00 10.00  8.50  7.30 19.00  9.75 19.00
```

Note that it is common to see repeated values within a bootstrap sample, and that every value obtained above was one of the original 10 data points. It was easy to create these bootstrap samples, and would have been just about as easy to create thousands of them, but it is not obvious why it is *useful* to do so. Finally, the R code for this Example is in Section 10.9.

☣ **10.5.3.** In Example 10.5.2 repeated values appear in all the bootstrap samples and that every value obtained above was one of the original 10 data points. The flip side of this is that within each bootstrap sample $Y_1^*, \ldots, Y_n^*$ (with incredibly high probability) some elements of original sample will not appear. We can quantify the latter, noting the chance the first real data point y_1 is excluded from the first bootstrap sample is

$$P(Y_j^* \neq y_1, j = 1, \ldots, n) = \left(\frac{n-1}{n}\right)^n = \exp\{n \log(1 - n^{-1})\} \to e^{-1} \approx 1/3,$$

as n gets large.

Having generated a lot of bootstrap samples, what can we *do* with them? The bootstrap got its name from the expression "pulling oneself up by one's bootstraps", and indeed the bootstrap seems at first too good to be true: data is *expensive* to collect, whereas here we get a seemingly endless stream from a fountain of synthetic datasets. We only have one *real* dataset though, and the synthetic datasets do not contain any numbers not in the original, so how can this be useful? We will first do a quick example of how the bootstrap can be used to approximate standard errors, and then we will delve more into the math and intuition behind the bootstrap.

Example 10.5.4. Returning to the duration of birth example from Example 10.5.2 (but using all $n = 95$ data points from now on), suppose that our estimand θ is the median duration. The sample median is 7.5. But how can we assess the accuracy of this estimate (without knowing θ^*)? The asymptotic distribution of the sample median was derived in Chapter 4, but there are two major issues with applying this result:

- The asymptotic distribution of the sample median requires knowing $f(\theta^*)$, where f is the PDF of one of the data points. We know neither f nor θ^*, nor do we have a parametric model for

f. We could estimate θ^* by $\hat{\theta}$ and f through kernel density estimation, but it is unclear how accurate these approximations would be (especially the latter).

- This is an asymptotic result, based on both the CLT and the delta method, and it is unclear how large the sample size needs to be for the asymptotics to yield good approximations.

The bootstrap method for approximating the standard error of $\hat{\theta}$ is simple: generate B bootstrap samples for some large B, say $B = 10^4$. Let $\hat{\theta}_j^*$ be the sample median of the jth bootstrap samples. The first ten of these bootstrap samples of the median for the birth data are

```
7.30 7.30 7.80 7.50 7.20 7.30 7.50 8.20 7.25 6.80
```

Overall we have B simulated sample medians. The bootstrap approximation to the standard error of $\hat{\theta}$ is then the standard deviation of the B simulated sample medians. Running this gave 0.406 as the approximate standard error of $\hat{\theta}$. `R` code for this bootstrap procedure is given in Section 10.9.

Moreover, if the estimator had been something more complicated than the sample median, it may have been very difficult mathematically to obtain its asymptotic distribution, whereas for the bootstrap we are approximating the standard error by simulation rather than using math.

10.5.1 Simulating from a known distribution

To start gaining insight into why the bootstrap works (and when it can fail), let us consider the case where we *know* the true distribution F_Y and can sample from it. This is rare in applications but often a useful idea, such as when running a simulation to study the performance of a method.

Let $T(\mathbf{Y})$ be a statistic we are interested in. Sample n times from F_Y, and collect these draws as $\mathbf{Y}^{(1)} = (Y_1^{(1)}, \ldots, Y_n^{(1)})$. This allows us to compute

$$T^{(1)} = T(\mathbf{Y}^{(1)}).$$

Doing all of this B times, independently, we obtain B copies or *replications*

$$T^{(1)}, \ldots, T^{(B)}$$

from the distribution of $T(\mathbf{Y})$. We can use these replications to approximate the distribution of $T(\mathbf{Y})$, or summary measures of the distribution such as the expectation, standard deviation, or quantiles. Recall that the same idea was used in Chapter 9 to study posterior distributions via simulation rather than via integrals. With enough computational power, we can ramp up B to be very large, so these estimates should be very close to the population quantities.

Example 10.5.5. Suppose we observe i.i.d. $Y_1, \ldots, Y_n$ with

$$Y_j \sim \theta + 0.5t_3,$$

where t_3 denotes the Student-t distribution with 3 degrees of freedom. Suppose that we wish to test the null hypothesis that $H_0 : \theta = 3$ against the alternative $H_1 : \theta \neq 3$. Fix α, the desired size of the test. Consider the test statistic

$$T(\mathbf{Y}) = \frac{\sqrt{n}(\bar{Y} - 3)}{S},$$

where $S^2 = (n-1)^{-1} \sum_{j=1}^{n} (Y_j - \bar{Y})^2$ is the unbiased sample variance. Under H_0,

$$T(\mathbf{Y}) \xrightarrow{d} \mathcal{N}(0, 1),$$

so this asymptotic Gaussian pivot suggests a nominal level α test that rejects the null if

$$T(\mathbf{y}) \notin [Q_{\mathcal{N}(0,1)}(\alpha/2), Q_{\mathcal{N}(0,1)}(1 - \alpha/2)].$$

However, this test is based on asymptotics and the sample size is small. It would be better to work with the *actual* distribution of $T(\mathbf{Y})$ under the null, not an asymptotic approximation. Then our test would reject the null if

$$T(\mathbf{y}) \notin [Q_{T(\mathbf{Y})}(\alpha/2), Q_{T(\mathbf{Y})}(1 - \alpha/2)].$$

But we do not know these quantiles; the t_3 distribution is complicated, and the distribution of the sum of t_3 random variables is even more complicated.

Instead of using the asymptotic Normal quantiles, we can use a *simulation-based* approximation based on B replications of $T(\mathbf{Y})$ under the null, computing the $\alpha/2$ and $1 - \alpha/2$ empirical quantiles of the B simulations as $\hat{Q}_{T(\mathbf{Y})}(\alpha/2)$ and $\hat{Q}_{T(\mathbf{Y})}(1 - \alpha/2)$. This leads to the null being rejected if

$$T(\mathbf{y}) \notin [\hat{Q}_{T(\mathbf{Y})}(\alpha/2), \hat{Q}_{T(\mathbf{Y})}(1 - \alpha/2)].$$

Typically, B might be selected as $10^4, 10^5$, or 10^6, depending on the computational cost of the simulations and the importance of the scientific problem. When performing a simulation, it is a good idea to run the entire simulation a few times to get a sense of the stability of the results.

As B gets large, $\hat{Q}_{T(\mathbf{Y})}(\alpha/2) \xrightarrow{p} Q_{T(\mathbf{Y})}(\alpha/2)$, so the nominal size of this test becomes α, as B increases. This looks like another asymptotic approximation, but now the asymptotics is in terms of computational effort B rather than sample size. It is usually *much* easier to increase B (by generating more replications on the computer) than to increase the sample size (by going out and collecting more data).

10.5.2 Real world vs. bootstrap world

Suppose we have data $y_1, \ldots, y_n$ and that these were i.i.d. draws from an unknown CDF F. We are taking a *nonparametric* approach here rather than assuming that F is in some parametric family of distributions such as Normal or Gamma.

Let θ be the estimand (e.g., the median) and $\hat{\theta}$ be an estimator (e.g., the sample median) for θ. We would like to learn about the properties of $\hat{\theta}$, e.g., find its standard error. This is often difficult to do mathematically, so it is natural to try a simulation-based approach. A fundamental challenge is that we only have *one* dataset and thus only *one* observed value of $\hat{\theta}$.

Ideally, we could generate simulated data

$$\{Y_1^{(1)}, \ldots, Y_n^{(1)}\}, \ldots, \{Y_1^{(B)}, \ldots, Y_n^{(B)}\},$$

each $Y_j^{(b)} \overset{\text{i.i.d.}}{\sim} F$, for $b = 1, \ldots, B$, as in the previous subsection, and compute replications

$$\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(B)}.$$

For B large, the standard deviation of these replications will be close to the standard error of $\hat{\theta}$. However, F is unknown! So we do not know how to generate such replications.

The essential idea in the bootstrap is to use the empirical CDF $\hat{F}$ as an approximation to the CDF F. The empirical CDF, which is given by

$$\hat{F}(y) = \frac{1}{n} \sum_{j=1}^{n} I(y_j \leq y).$$

is known once we have the data in hand. And as discussed in Chapter 1, by the strong law of large numbers, for each $y \in \mathbb{R}$ we have

$$\hat{F}(y) \to F(y)$$

with probability 1. So if n is large, $\hat{F}$ is likely to approximate F well.

Note that generating n i.i.d. draws from the empirical CDF is equivalent to generating a simple random sample of size n with replacement from $y_1, \ldots, y_n$, since the empirical distribution puts mass $1/n$ on each of the n observations.

Figure 10.2 illustrates how the "bootstrap world" parallels the "real world". The core point is:

- In the real world, we only have *one* dataset, that was generated from the *unknown* CDF F.
- In the bootstrap world, we generate *many* simulated datasets from the *known* CDF $\hat{F}$.

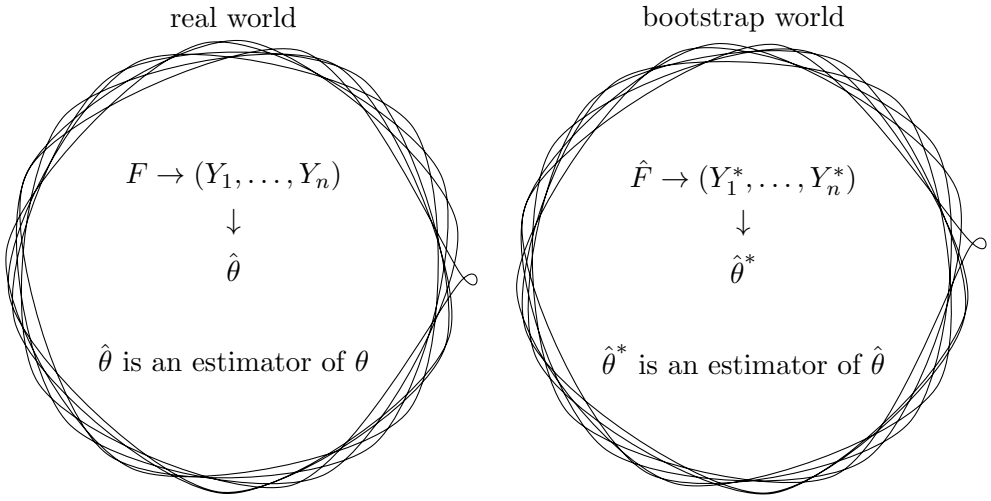


Figure 10.2: The bootstrap world parallels the real world. In the real world, data comes from CDF F and used to compute an estimator $\hat{\theta}$ for the estimand θ. In the bootstrap world, bootstrap samples are generated from the known CDF $\hat{F}$ and used to compute an estimator $\hat{\theta}^*$ for $\hat{\theta}$.

We write one of these simulated datasets generically as

$$(Y_1^*, \ldots, Y_n^*).$$

In the real world we only have one estimate $\hat{\theta}$ for θ. In the bootstrap world we have many estimates

$$\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*,$$

which we regard as estimates of $\hat{\theta}$ rather than of θ. Of course, we know $\hat{\theta}$ so we do not need to estimate it; the goal is to use the bootstrap samples to learn about the properties of the estimator $\hat{\theta}$, e.g., if we do not know $SE(\hat{\theta})$, or have a good $\widehat{\text{SE}}$ based directly on the data.

In the bootstrap world, the bias is

$$\text{Bias}_{\text{boot}}(\hat{\theta}^*) = \text{E}_{\text{boot}}[\hat{\theta}^*] - \hat{\theta}. \tag{10.3}$$

Again, the bootstrap expectation, E_{boot}, is calculated by sampling with replacement from $\hat{F}$, conditional on the underlying data $(y_1, \ldots, y_n)$. Likewise

$$\text{Var}_{\text{boot}}(\hat{\theta}^*) = \text{E}_{\text{boot}}[(\hat{\theta}^* - \text{E}_{\text{boot}}[\hat{\theta}^*])^2] = \text{E}_{\text{boot}}[\hat{\theta}^{*2}] - \{\text{E}_{\text{boot}}[\hat{\theta}^*]\}^2, \tag{10.4}$$

and the corresponding mean square error $\text{MSE}_{\text{boot}}(\hat{\theta}^*) = \text{E}_{\text{boot}}[(\hat{\theta}^* - \hat{\theta})^2]$. Further, the bootstrap world cumulative distribution function is

$$F_{\hat{\theta}^*}(\theta) = P_{\text{boot}}(\hat{\theta}^* \leq \theta) = \text{E}_{\text{boot}}[I(\hat{\theta}^* \leq \theta)],$$

while the associated bootstrap world p-quantile is written as

$$Q_{\hat{\theta}^*}(p), \quad p \in (0,1).$$

Example 10.5.6. For a single element of a bootstrap sample and arbitrary function g,

$$\mathrm{E}_{\mathrm{boot}}[g(Y_1^*)] = \frac{1}{n}\sum_{j=1}^{n} g(y_j).$$

Hence $\mathrm{E}_{\mathrm{boot}}$ is exactly the same expectation we saw in Section 10.2 when sampling with replacement from a finite population, which here is $(y_1, \ldots, y_n)$. This implies, in particular, that

$$\mathrm{E}_{\mathrm{boot}}[Y_1^*] = \frac{1}{n}\sum_{j=1}^{n} y_j = \bar{y}, \quad \mathrm{Var}_{\mathrm{boot}}(Y_1^*) = \frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2 = s^2.$$

This means that, in the example where $\bar{Y^*} = \frac{1}{n}\sum_{j=1}^{n} Y_j^*$,

$$\mathrm{E}_{\mathrm{boot}}[\bar{Y^*}] = \mathrm{E}_{\mathrm{boot}}[Y_1^*] = \bar{y}, \quad \mathrm{Var}_{\mathrm{boot}}(\bar{Y^*}) = n^{-1}\mathrm{Var}_{\mathrm{boot}}(Y_1^*) = \frac{1}{n}s^2.$$

Furthermore,

$$T^* = \frac{\bar{Y^*} - \bar{y}}{\sqrt{\mathrm{Var}_{\mathrm{boot}}(\bar{Y^*})}},$$

has mean 0 and standard deviation 1 in the bootstrap world, so it behaves roughly like

$$T = \frac{\bar{Y} - \mathrm{E}[Y_1]}{\sqrt{\mathrm{Var}(\bar{Y})}},$$

which has mean 0 and standard deviation 1 in the real world.

☣ **10.5.7.** For more involved statistics, analytic calculations with the bootstrap can become tedious or infeasible. For example, to calculate the bias of $\hat{\theta}^* = T(\mathbf{Y}^*)$ in the bootstrap world, we need

$$\mathrm{E}_{\mathrm{boot}}[T(\mathbf{Y}^*)] = \frac{1}{n^n}\sum_{j_1=1}^{n}\cdots\sum_{j_n=1}^{n} T(y_{j_1}, \ldots, y_{j_n}).$$

which becomes incredibly expensive to calculate if n is even moderately large. We will turn to simulation to overcome this problem, very much paralleling the use of simulation in Bayesian inference in Chapter 9.

In the statistics literature the bootstrap world bias, variance, MSE, CDF, and p-quantile are nearly always estimated using simulation: using the bootstrap replications. The corresponding estimated "bootstrap bias", is

$$\widehat{\mathrm{Bias}}_{\mathrm{boot}}(\hat{\theta}^*) = \bar{\theta^*} - \hat{\theta}, \quad \text{where} \quad \hat{\mathrm{E}}_{\mathrm{boot}}[\hat{\theta}^*] = \bar{\theta^*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_j^*,$$

while the *bootstrap standard error*, which is used a great deal in modern statistics, is

$$\hat{\mathrm{SE}}_{\mathrm{boot}}(\hat{\theta}^*) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_j^* - \bar{\theta^*}\right)^2}.$$

The $\hat{F}_{\hat{\theta}^*}(\theta) = \frac{1}{B}\sum_{j=1}^{B} I(\hat{\theta}_j^* \leq \theta)$ is the *bootstrap empirical cumulative distribution function*, while $\hat{Q}_{\hat{\theta}^*}(p) = \hat{\theta}^*_{(\lceil Bp \rceil)}$ is the *bootstrap p-quantile*.

Crucially, there are two different sources of bootstrap error in estimating, for example, the standard error:

- The more fundamental one, is the difference between $\mathrm{SE}(\hat{\theta})$ and $\mathrm{SE}_{\mathrm{boot}}(\hat{\theta}^*)$. This is caused by $\hat{F}$ not being F, that is it is caused by the data. This error falls as n increases.
- The less troublesome one, is the difference between $\mathrm{SE}_{\mathrm{boot}}(\hat{\theta}^*)$ and $\hat{\mathrm{SE}}_{\mathrm{boot}}(\hat{\theta}^*)$ caused by using simulation. This error falls as B increases, so is under our control.

Figure 10.2 can also be summarized as an analogy:

Population is to sample as sample is to bootstrap sample (that is F is to Y as $\hat{F}$ is to Y^*)

The bootstrap is powerful but not magic. It should work well if $\hat{F}$ is a good approximation for F but there is no reason to expect it to work well if n is not large enough to make this approximation a good one. Also, we assumed that the data are i.i.d. There is a large literature that attempts to extend the bootstrap to non-i.i.d. settings such as regression models or time series models, with varying degrees of success.

10.5.3 Bootstrap confidence intervals

The bootstrap can also be used to construct approximate confidence intervals. Many such bootstrap procedures have been proposed. They vary in their complexity (both in computing them and understanding where they come from) and in their performance (in terms of having the desired coverage probability). We will introduce three versions. Two are "quick and dirty"; the third is more computationally intensive but also tends to have better performance.

Let θ be the estimand and $\hat{\theta}$ be an estimator for θ. Suppose that we want to use the bootstrap to get an approximate $(1-\alpha)$ confidence interval. To simplify notation, take $\alpha = 0.05$; general α can be handled in the same manner.

1. *Normal interval with bootstrap standard error.* Recall from Chapter 5 the confidence interval

$$\hat{\theta} \pm 1.96 \cdot \hat{\text{SE}}(\hat{\theta}),$$

where $\hat{\text{SE}}$ is an estimate for the standard error of $\hat{\theta}$. In classical methods, the standard error is estimated using mathematical calculations, often requiring asymptotic theorems such as the CLT and delta method, or the result for the asymptotic distribution of the MLE in terms of Fisher information.

The Normal interval with bootstrap standard error $\hat{\text{SE}}_{\text{boot}}(\hat{\theta})$ uses the bootstrap estimate of the standard error of $\hat{\theta}$. Then we have the approximate 95% CI

$$\hat{\theta} \pm 1.96 \cdot \hat{\text{SE}}_{\text{boot}}(\hat{\theta}^*).$$

For this to work well, we want the distribution of $\hat{\theta}$ to approximately be Normal with mean θ. The only advantage of this method over trying the classical method is that it avoids needing to do any math to approximate the standard error.

2. *Percentile method.* Another simple method is to directly use the sample quantiles of the bootstrap replications of $\hat{\theta}$. The percentile interval goes from the 0.025 sample quantile to the 0.975 sample quantile of the bootstrap replications

$$[\hat{\theta}^*_{(\lceil 0.025B \rceil)}, \hat{\theta}^*_{(\lceil 0.975B \rceil)}],$$

very much like how Bayesian credible intervals were estimated by simulation in Chapter 9.3. The percentile interval has the advantages that it is very easy to compute, easy to explain and does not require $\hat{\theta}$ to be approximately Normal. However, the percentile method can have poor performance if the distribution of $\hat{\theta}$ is skewed or if $\hat{\theta}$ is biased. For example, suppose that $\hat{\theta}$ has a severe positive bias, and tends to be substantially larger than θ. Then the bootstrap *amplifies* the bias, since a replication $\hat{\theta}^*$ will tend to be even larger than $\hat{\theta}$.

3. *Bootstrap t interval, also known as the studentized bootstrap interval.* Lastly, we introduce a method that is more complicated but tends to perform better than the two methods discussed above. Despite the name "bootstrap t interval", the interval does not use a t distribution. It has this name because it is analogous to the classical method of creating a *pivot* that has a t distribution (see Section 4 of Chapter 5). But instead of needing parametric assumptions such as Normality, an approximate pivot is constructed and the bootstrap is used to obtain its distribution.

As in Chapter 5, consider a quantity of the form

$$T = \frac{\hat{\theta} - \theta}{\hat{\text{SE}}(\hat{\theta})}.$$

Unlike in Chapter 5, we will approximate the distribution of T without making any parametric assumptions. In the bootstrap world, the quantity corresponding to T is

$$T^* = \frac{\hat{\theta}^* - \hat{\theta}}{\text{SE}_{\text{boot}}(\hat{\theta}^*)}.$$

Then the *bootstrap t interval* uses the distribution of T^* to approximate the distribution of T. Recall that for T^*, the data $y_1, \ldots, y_n$ are fixed (so $\hat{\theta}$ is also fixed); the randomness comes from the resampling, resulting in a random $\hat{\theta}^*$.

The distribution of T may be difficult or infeasible to find, but the distribution of T^* can be found using the bootstrap. Generate a large number B of bootstrap samples, and then compute the corresponding bootstrap replications

$$\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$$

of $\hat{\theta}$ and the corresponding bootstrap replications

$$T_1^*, \ldots, T_B^*$$

of T, so

$$T_j^* = \frac{\hat{\theta}_j^* - \hat{\theta}}{\text{SE}_{\text{boot}}(\hat{\theta}_j^*)}.$$

A computational challenge here is that we may not know $\text{SE}_{\text{boot}}(\hat{\theta}_j^*)$ so we could replace it by computing the bootstrap SE estimator $\hat{\text{SE}}_{\text{boot}}(\hat{\theta}_j^*)$ for each T_j^*. In principle this is easy to code and run. But it means that, like a dream within a dream, we may need to generate B bootstrap replications

$$\hat{\theta}_1^{**}, \ldots, \hat{\theta}_B^{**},$$

associated with each and every $\hat{\theta}_j^*$ (that is generating B bootstrap samples from each of the original B bootstrap samples $Y_1^*, \ldots, Y_n^*$ — not going back to the original data), yielding

$$\hat{\text{SE}}_{\text{boot}}(\hat{\theta}_j^*) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_j^{**} - \overline{\theta^{**}}\right)^2}, \quad \text{where} \quad \overline{\theta^{**}} = \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}_j^{**}.$$

This sounds difficult (it does involve B^2 estimations) but this approach is commonly used in practice.

Let Q^* be the quantile function of T^* and $\hat{Q}^*(p) = T^*_{(\lceil Bp \rceil)}$ be the p-sample quantile of the $T^*_1, \ldots, T^*_B$. By letting B be large, we can make $\hat{Q}$ very close to Q. Then

$$\begin{aligned}
0.95 &\approx P(\hat{Q}^*(0.025) \leq T^* \leq \hat{Q}^*(0.975)) \\
&\approx P(\hat{Q}^*(0.025) \leq T \leq \hat{Q}^*(0.975)) \\
&= P(\hat{Q}^*(0.025)\hat{\mathrm{SE}}_{\mathrm{boot}}(\hat{\theta}) \leq \hat{\theta} - \theta \leq \hat{Q}^*(0.975)\hat{\mathrm{SE}}_{\mathrm{boot}}(\hat{\theta})) \\
&= P(\hat{\theta} - \hat{Q}^*(0.975)\hat{\mathrm{SE}}_{\mathrm{boot}}(\hat{\theta}) \leq \theta \leq \hat{\theta} - \hat{Q}^*(0.025)\hat{\mathrm{SE}}_{\mathrm{boot}}(\hat{\theta})).
\end{aligned}$$

The bootstrap t interval is then

$$[\hat{\theta} - \hat{Q}^*(0.975)\hat{\mathrm{SE}}_{\mathrm{boot}}(\hat{\theta}), \hat{\theta} - \hat{Q}^*(0.025)\hat{\mathrm{SE}}_{\mathrm{boot}}(\hat{\theta})].$$

Example 10.5.8 (Continuing Example 10.5.4)**.** How do these bootstrap based CI methods work with the duration of birth data from Example 10.5.4? Recall the sample median was 7.5, while the asymptotic nominal 95% confidence interval was reported as [6.51,8.48] in Table 5.1. The Normal CI based on the bootstrap standard error is [6.67,8.32], while the percentile bootstrap CI is [6.90,8.30] and the bootstrap t interval is [6.55,8.25]. The differences between these intervals turned out to be fairly small.

Table 10.1 gives results for other descriptive statistics, comparing the asymptotic methods, with the normal based on the bootstrap SE and the percentile bootstrap. The results line up strongly. Again the asymptotic results are reprinted from Table 5.1 as points of comparison. There is some evidence that the asymptotic based CIs are particularly close to the bootstrap versions for $\overline{y}$, s_n, $\widehat{F}(5)$, $\widehat{F}(10)$ and $\widehat{F}(15)$ — which are all averages of one type or other. The corresponding results for the sample quantiles, which are not simple averages, are less lined up. The bootstrap t interval for the $\widehat{F}(15)$ fails. Why? Only 3 data points in the sample are larger than 15, which means there is around a $e^{-3} \approx 0.03$ chance none of the 3 are a bootstrap sample $Y^*_1, \ldots, Y^*_n$. If none of them are in the bootstrap sample, then in the second stage bootstrap there will be no $Y^{**}_1, \ldots, Y^{**}_n$ above 15 — so driving $\widehat{F}(15)^{**}$ to 0 in each replication and so the $\hat{SE}(\hat{\theta}^*_j)$ will be 0. This is enough to destroy the left hand side of the confidence interval even though $\hat{SE}(\hat{\theta})$ is a well behaved. It is the estimated quantile Q^* which is problematic.

We have highlighted this last example, for $\hat{F}(15)$, to reinforce that it is important to always think about what is happening when you apply the bootstrap. It is a remarkably simple and powerful device, but it does not always work. Here we are applying it too far in the tail of the distribution for it to be reliable.

Estimate	value	Nominal 95% CI							
		Asymptotic Based		Normal based bootstrap SE		Percentile bootstrap		Bootstrap t interval	
	$\hat{\theta}$	Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
$\overline{y}$	7.72	7.00	8.44	6.96	8.47	7.02	8.43	6.98	8.54
s_n	3.57	2.98	4.07	3.01	4.12	2.99	4.08	3.07	4.33
$\hat{Q}(0.05)$	2.57	1.69	3.44	1.87	3.26	2.07	3.40	1.66	3.19
$\hat{Q}(0.5)$	7.50	6.51	8.48	6.67	8.32	6.80	8.30	6.55	8.25
$\hat{Q}(0.95)$	14.5	12.1	16.9	12.4	16.6	11.2	16.1	12.7	17.6
$\widehat{F}(5)$	0.263	0.174	0.352	0.172	0.353	0.178	0.357	0.171	0.365
$\widehat{F}(10)$	0.768	0.684	0.853	0.682	0.853	0.683	0.852	0.668	0.847
$\widehat{F}(15)$	0.968	0.933	1.003	0.933	1.003	0.926	1.000	NA	0.995

Table 10.1: Extension of Table 5.1. Nominal 95% Confidence intervals, based on asymptotic approximations and various bootstrap methods, for summary statistics for the duration of birth example from Chapter 1. The table also contains the summary statistics for the data. The bootstrap t interval fails for the $\widehat{F}(15)$ as there are only 3 data points larger than 15. Throughout, $B = 1,000$.

An attractive feature of the percentile bootstrap is that the confidence interval for $\widehat{F}(15)$ sits between 0 and 1. This is an automatic feature of this method, reflecting the constraint imposed on the estimator itself.

10.6 Jackknife*

A second type of resampling method, the *jackknife*, was introduced by Maurice H. Quenouille in 1949 and further developed into its modern form by John Tukey in 1956. The jackknife inspired Efron to introduce the bootstrap a couple of decades later.

☣ **10.6.1.** Placing our discussion of the jackknife in this sampling-focused Chapter can be confusing as jackknifing does not involve sampling from a fixed data set $\mathbf{y} = (y_1, \ldots, y_n)$. Instead the jackknife deterministically selects from the random data (returning to the population line of thinking from Chapters 1 to 9). Even though there is no design-based sampling, in the statistics literature the jackknife is categorized as a "resampling method", as it reuses the data.

Let $T = T(\mathbf{Y})$ be a statistic based on the random $\mathbf{Y} = (Y_1, \ldots, Y_n)$.

Definition 10.6.2. The jth jackknife sample is $\mathbf{Y}_{[j]} = (Y_1, \ldots, Y_{j-1}, Y_{j+1}, \ldots., Y_n)$, which drops the jth data point Y_j from the sample. The corresponding jth *jackknife replicate* is the statistic applied

to $\mathbf{Y}_{[j]}$,

$$T_j^* = T(\mathbf{Y}_{[j]}), \quad j \in \{1, 2, \ldots, n\}.$$

The sample jackknife average is $\bar{T}^* = n^{-1} \sum_{j=1}^n T_j^*$, the average of the jackknife replicates.

The bias terms of T, T_j^* and $\bar{T}^*$ for the estimand θ are defined as

$$b(n) = \mathrm{E}[T] - \theta, \quad b_j(n) = \mathrm{E}[T_j^*] - \theta, \quad \bar{b}(n) = \mathrm{E}[\bar{T}^*] - \theta,$$

writing the notation for bias to explicitly note the role of n. For many problems the bias term for an estimator based on n data point is roughly b/n. The following theorem shows that this n^{-1} bias can be eliminated using the jackknife.

Theorem 10.6.3. *Assume $Y_1, \ldots, Y_n$ are i.i.d. from F_Y and $b(n) = \mathrm{E}[T] - \theta = b/n$ for each n. Then the estimator*

$$\hat{T} = T - (n-1)(\bar{T}^* - T)$$

is unbiased for θ.

Proof. By the i.i.d. assumption,

$$b_j(n) = \frac{b}{n-1}, \quad \bar{b}(n) = \frac{1}{n} \sum_{j=1}^n b_j(n) = \frac{b}{n-1}.$$

This means that

$$(n-1)\{\bar{b}(n) - b(n)\} = b(n-1)\left(\frac{1}{n-1} - \frac{1}{n}\right) = \frac{b}{n},$$

the bias of T. Taking the expectation of $\hat{T}$ delivers the result. ∎

In practice the bias $b(n)$ is not exactly b/n since it includes other bias terms like c/n^2, which go to zero faster. So we should think of this theorem as saying the largest part of the bias can be removed for many statistics by using the jackknife.

☣ **10.6.4.** Unbiasedness is a nice property but does not make sense as a main criterion to focus on, in view of the bias-variance tradeoff and the fact that in some problems an unbiased estimator does not exist, or exists but is absurd. So the above argument does not imply that $\hat{T}$ is a better estimator than T. For example, it may have a higher MSE than T. Still, it is helpful having the jackknife as a versatile bias reducer.

Cross-validation

Jackknifing quantifies the impact of dropping of the jth data point on a statistic. That idea can be applied to linear predictive regression (Section 7.2). The jackknifed replicate of the linear regression estimator

$$\widehat{\theta}_{\setminus j} = \frac{\sum_{i\neq j}^{n} x_i Y_i}{\sum_{i\neq j}^{n} x_i^2}, \quad j \in \{1, \ldots, n\},$$

excludes the j-th outcome from $\widehat{\theta} = \sum_{i=1}^{n} x_j Y_j / \sum_{i=1}^{n} x_j Y_j$. Using $x_j \widehat{\theta}_{\setminus j}$ to predict the j-th outcome Y_j yields the j-th "leave-one-out residual"

$$\widehat{U}_{\setminus j} = Y_j - x_j \widehat{\theta}_{\setminus j}.$$

The $\widehat{U}_{\setminus j}$ is, probabilistically, an out of sample prediction — for the predictor $x_j \widehat{\theta}_{\setminus j}$ does not involve Y_j. In particular, if the outcomes are independent given the predictors, then

$$\mathrm{E}[\widehat{U}_{\setminus j}^2 | \mathbf{X} = \mathbf{x}] = \sigma_j^2 + \left(\mu_j - x_j \theta_{\setminus j}\right)^2 + x_j^2 \frac{\sum_{i\neq j}^{n} x_i^2 \sigma_i^2}{\left(\sum_{i\neq j}^{n} x_i^2\right)^2},$$

where $\mu_j = \mathrm{E}[Y_j | \mathbf{X} = \mathbf{x}]$, $\sigma_j^2 = \mathrm{Var}(Y_j | \mathbf{X} = \mathbf{x})$ and

$$\theta_{\setminus j} = \mathrm{E}[\widehat{\theta}_{\setminus j} | \mathbf{X} = \mathbf{x}] = \frac{\sum_{i\neq j}^{n} x_i \mu_i}{\sum_{i\neq j}^{n} x_i^2}.$$

The average squared leave-one-out residuals

$$\widehat{LOOCV} = \frac{1}{n} \sum_{j=1}^{n} \widehat{U}_{\setminus j}^2$$

is called the "leave-one-out" cross-validation (LOOCV) measure. It is a conditionally unbiased estimator of the average squared conditional loss

$$LOOCV = \frac{1}{n} \sum_{j=1}^{n} \mathrm{E}[\widehat{U}_{\setminus j}^2 | \mathbf{X} = \mathbf{x}],$$

which is impacted by the conditional variance of the noise, the squared conditional bias and the influence of the conditional variance of parameter estimation. $\widehat{LOOCV}$ is a popular estimator of the likely out of sample squared error of the prediction.

Leave-one-out cross-validation extends to dropping multiple outcomes (e.g. "leave-p-out" and "k-fold" cross-validation), the use of other estimators (e.g., ridge, Lasso and neural nets), to select bandwidth and shrinkage type parameters (e.g., kernel density estimator) and other criteria than squared loss. The details of this are beyond the scope of this book.

Example 10.6.5. In ridge regression the prediction of Y_j is carried out using $cx_j\widehat{\theta}$, where $c \in [0,1]$ is a shrinkage parameter (e.g. Section 7.2.2). But what should c be? One way of selecting c is to minimize "leave-one-out" cross-validation (selecting a c to do well for out of sample prediction), which in this case is

$$\widehat{c} = \underset{c\in[0,1]}{\arg\min} \sum_{j=1}^{n} \left(Y_j - \ cx_j\widehat{\theta}_{\setminus j}\right)^2,$$

where again $\widehat{\theta}_{\setminus j}$ is the jackknifed replicate of the linear regression estimator. Now let

$$\widetilde{c} = \frac{\sum_{j=1}^{n} Y_j x_j \widehat{\theta}_{\setminus j}}{\sum_{j=1}^{n} x_j^2 \widehat{\theta}_{\setminus j}^2},$$

the linear regression estimator of Y_j on $x_j\widehat{\theta}_{\setminus j}$, so then, after some work, it turn out that

$$\widehat{c} = \begin{cases} 0 & \text{if } \widetilde{c} < 0, \\ \widetilde{c} & \text{if } \widetilde{c} \in [0,1], \\ 1 & \text{if } \widetilde{c} > 1. \end{cases}$$

10.7 Permutation tests

Resampling ideas can also be applied to hypothesis testing. Suppose that we are comparing two groups, group 0 and group 1, and wish to test whether the distribution that generated the data in group 0 is the same as the distribution that generated the data in group 1. *Permutation tests* provide an elegant approach to this problem. As with the bootstrap, computational effort replaces mathematical effort.

Let the data for group 0 be the observed values of

$$X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} F_X,$$

and the data for group 1 be the observed values of

$$Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} F_Y.$$

Also assume that $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are independent. The CDFs F_X and F_Y are unknown, and no parametric assumptions are made about them.

A permutation test for

$$H_0 : F_X = F_Y \quad \text{vs.} \quad H_1 : F_X \neq F_Y,$$

can be conducted as follows.

1. Let T be a test statistic (chosen before looking at the data). For simplicity, assume that T is nonnegative. The test statistic T should be chosen so that large values of T are evidence against

H_0. For example, a natural choice is $T = |\bar{Y} - \bar{X}|$. If we are more interested in medians than means, we could instead let T be the absolute difference between the sample medians.

2. Compute the observed value t_0 of T from the data.

3. Generate B random permutations of the data, for some large number B. Each random permutation is a completely random shuffle of $X_1, \ldots, X_m, Y_1, \ldots, Y_n$. In particular, this scrambles which data points belong to which group. The random permutations should be a simple random sample without replacement, though for computational convenience a simple random sample with replacement can be used instead.

4. For each of the B random permutations, compute the test statistic T. Call these $t_1, \ldots, t_B$.

5. The p-value is

$$P_0(T \geq t_0) \approx \frac{1}{B} \sum_{j=1}^{B} I(t_j \geq t_0),$$

where P_0 denotes probability under the *permutation distribution* of T, i.e., the distribution under random shuffles of the data rather than under repeated sampling.

In very small problems we may be able to look at *all* permutations rather than random permutations, but this quickly becomes infeasible as m and n grow since the factorial function grows so fast. For example, for $m = n = 30$, there are $60! \approx 8.3 \times 10^{81}$ ways to permute a list of 60 distinct values. In comparison, the number of atoms in the entire known universe has been estimated as 10^{80}.

Happily, we can run a permutation test by generating a lot of *random* permutations rather than going through all permutations. A typical choice is $B = 10^4$ or $B = 10^5$. The number 10^4 may be tiny compared with the total number of possible permutations but it is the absolute size that matters, not the relative size.

Intuitively, the idea behind the test is that under the null, the X_i's and Y_j's are completely interchangeable, so it would be surprising if the observed value of the test statistic were extreme compared with simulated values of the test statistic obtained by randomly shuffling the data. Note that the null hypothesis is that the *distributions* F_X and F_Y are the same, not just that the means are the same. This null hypothesis is rather strong and inflexible, whereas the choice of test statistic is very flexible.

10.8 Recap

Sampling methods regard the finite population as fixed, and randomly choosing amongst them to produce new samples to estimate the properties of the fixed population. No model of the finite population is needed. Such methods are at the core of survey sampling methods, but appear in many areas of statistics, e.g. Bayesian statistics and causal inference. There are two main versions of sampling: sampling with replacement and sampling without replacement.

The main ideas and notation for the chapter are listed in Table 10.2.

Formula or idea	Description or name
finite population or design based	fixed $y_1, \ldots, y_N$
$\mu = \bar{y}$, $\sigma^2 = N^{-1}\sum_{j=1}^{n}(y_j - \bar{y})^2$	examples of estimands
sampling with replacement	$Y_1, \ldots, Y_n$ iid from $y_1, \ldots, y_N$
sampling without replacement	sample with $n \leq N$ point equally likely out of $y_1, \ldots, y_N$
stratified sampling	split $y_1, \ldots, y_N$ into subpops (e.g., by city). Sample in subpops
resampling	split the data up, reuse in some way
bootstrap	simulation based approx, alternative to asymptotics
$\mathbf{Y}^* = (Y_1^*, \ldots, Y_n^*) \overset{iid}{\sim} \hat{F}$, ECDF	1 bootstrap sample
$\hat{\theta}^* = T(\mathbf{Y}^*)$	1 bootstrap replication
$\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$	B bootstrap replications
$\hat{SE}(\hat{\theta}^*) = \sqrt{\frac{1}{N-1}\sum_{b=1}^{B}(y_b^* - \bar{y^*})^2}$	bootstrap SE
$\hat{\theta} \pm 1.96 \times \hat{SE}(\hat{\theta}^*)$	normal CI interval using bootstrap SE
$[\hat{\theta}^*_{(\lceil 0.025B \rceil)}, \hat{\theta}^*_{(\lceil 0.025B \rceil)}]$	percentile bootstrap CI
jackknife	compute $\hat{\theta}$ on all data but y_j
cross-validation	out of sample prediction, dropping each time 1 data point
permutation test	testing nonparametrically $H_0 : F_X = F_Y$ against $H_0 : F_X \neq F_Y$

Table 10.2: Main ideas and notation in Chapter 10.

Bootstrapping conditions on the data, to produce a finite population $\hat{F}$, and produce new bootstrap replications by sampling with replacement from $\hat{F}$. Each bootstrap replication is used to produce a new estimate, the replications produce a collection of such estimates which can then approximate the distribution of the estimator underlying the estimate based on the data. The bootstrap is an example of a resampling method. Other resampling methods include the jackknife and permutation tests.

10.9 R, sampling, and the bootstrap

10.9.1 Sampling

In R, sampling with or without replacement can be implemented using the `sample` function. The core function for sampling with replacement Example 10.3.7 is `sample(y, n, replace = TRUE)`. Reproducing the results for this example, we interlace code and output, as

Code and output for Example 10.3.7

```
set.seed(111)
N <- 7; n <- 5
y <- c(37, 6, 36, 42, 28, 31, 10)
sample(y,n,replace = TRUE)
[1] 31 36 42 36 10
```

Example 10.9.1. Another example is where $N = 7$, drawing each data point from $\mathcal{N}(0, 1)$. Then these are fixed and they are sampled with replacement to produce a sample of size $n = 12$.

Code and output for Example 10.9.1

```
set.seed(111); n <- 12; y <- rnorm(7,0,1);
print(y,digits = 2);
[1] 0.24 -0.33 -0.31 -2.30 -0.17  0.14 -1.50
ySamp = sample(y,n,replace=TRUE)
print (ySamp, digits = 2)
[1] -0.17 -0.33 -2.30 -1.50 0.14 -0.33 0.24 -0.31 0.24 -0.31 -1.50 -0.17
```

The code and output for the sampling without replacement Example!10.3.11 has the same structure, but `replace` is changed to `FALSE`. Here $N = 7$ and $n = 5$.

Code and output for Example 10.3.11

```
N <- 7; n <- 5
y <- c(37, 6, 36, 42, 28, 31, 10)
sample(y,n,replace = FALSE)
[1] 28 10 6 42 36
```

10.9.2 Bootstrapping

The bootstrap Example 10.5.2 uses the duration of birth dataset, producing $B = 20$ replications. The code is here:

Code and output for Example 10.5.2

```
set.seed (111)
y <- sample(births$time, 10, replace = FALSE);  # setup the subset
print(y)
```

```
B <- 20;  # draw 20 bootstrap samples
boot_samples <- t(replicate(B, sample (y, 10, replace = TRUE )))
print(boot_samples)
```

The resulting output is recorded in Example 10.5.2.

This code uses the powerful `replicate` command — it has more flexibility than we are demonstrating here. This is a simple way of generating lots of bootstrap samples — it is told how many times to do the same operation, here B times, and then what the core command is, here sampling with replacement. `sample(y, 10, replace = TRUE)` produces a 10-dimensional vector, so `replicate(B, sample(y, 10, replace = TRUE))` outputs a $10 \times B$ matrix. Here the matrix transpose command `t` implies we output the transpose of the matrix.

Example 10.9.2. Now let us use this structure to carry out some statistical calculations by turning to the code for Example 10.5.4. There is a new operation in this code. The apply function is used. The apply function applies the quantile calculation to each column of the $n \times B$ matrix `yBoot` separately. The number "2" tells apply to work with columns, if we had used "1" it would work with rows.

Code for Example 10.5.4

```
set.seed(111)
y <- births$time; B = 10^4; n = length(y)
quantile(y,probs=0.5); # median of data

yBoot = replicate(B,sample(y,n,replace=TRUE))
quantBoot = apply(yBoot,2,quantile,probs=0.5) # median for each rep

print(quantBoot[1:10]) # print first 10 bootstrap medians

sd(quantBoot) # compute SD of medians
```

Our final bootstrapping code focuses on Example 10.5.8, which extends the code from Example 10.5.4 above. The task in the example is to generate 95% confidence intervals. This is carried out in three ways: Normal, percentile, and the double bootstrap.

Code for Example 10.5.8

```
set.seed(111)
y <- births$time; n = length(y)
Mresults = matrix(nrow=8,ncol=7) # storage for results
colnames(Mresults) = c("est","2.5% Normal","97.5% Normal","2.5% percent",
                       "97.5% percent","2.5% t","97.5% t")
rownames(Mresults) = c("mean","sd","5Q","50Q","95Q","F(5)","F(10)","F(15)")
B=1000 # bootstrap reps
```

```
yBoot = replicate(B, sample(y, n, replace = TRUE)) # n*B matrix
boot_est_SDs = numeric(B); t_reps <- numeric(B) # build storage

for (i in 1:8){

  if (i==1){estY=mean(births$time); estBoot = apply(yBoot,2,mean)}
  if (i==2){estY=sd(births$time); estBoot = apply(yBoot,2,sd)}

  if (i==3){estY=quantile(births$time,probs=0.05);
            estBoot=apply(yBoot,2,quantile,probs=0.05)}
  if (i==4){estY=quantile(births$time,probs=0.5);
            estBoot=apply(yBoot,2,quantile,probs=0.5)}
  if (i==5){estY=quantile(births$time,probs=0.95);
            estBoot=apply(yBoot,2,quantile,probs=0.95)}

  if (i==6){estY=mean(1.0*(births$time<=5));
    estBoot=apply(1.0*(yBoot<=5),2,mean)}
  if (i==7){estY=mean(1.0*(births$time<=10));
    estBoot=apply(1.0*(yBoot<=10),2,mean)}
  if (i==8){estY=mean(1.0*(births$time<=15));
    estBoot=apply(1.0*(yBoot<=15),2,mean)}

  Mresults[i,1:5] = c(estY,estY-1.96*sd(estBoot),estY+1.96*sd(estBoot),
                          quantile(estBoot,prob=c(0.025,0.975)))

  for(j in 1:B) {
    boot_in_boot_samples = replicate(B,sample(yBoot[,j],n,
            replace = TRUE))

    if (i==1){boot_in_boot_est=apply(boot_in_boot_samples,2,mean)}
    if (i==2){boot_in_boot_est=apply(boot_in_boot_samples,2,sd)}

    if (i==3){boot_in_boot_est=apply(boot_in_boot_samples,2,
        quantile,probs=0.05)}
    if (i==4){boot_in_boot_est=apply(boot_in_boot_samples,2,
        quantile,probs=0.5)}
    if (i==5){boot_in_boot_est=apply(boot_in_boot_samples,2,
        quantile,probs=0.95)}

    if (i==6){boot_in_boot_est=apply(1.0*(boot_in_boot_samples<=5),2,
        mean)}
    if (i==7){boot_in_boot_est=apply(1.0*(boot_in_boot_samples<=10),2,
        mean)}
    if (i==8){boot_in_boot_est=apply(1.0*(boot_in_boot_samples<=15),2,
        mean)}

    boot_est_SDs[j] = sd(boot_in_boot_est)
  }
```

```
  t_reps = (estBoot - estY)/boot_est_SDs
  # boot t int
  boot_t_int=estY-quantile(t_reps,probs=c(0.975,0.025))*sd(estBoot)
  Mresults[i,6:7] = boot_t_int
}

print(Mresults)
```

Chapter 11

Experiments and Causality

11.1 Causality

As we discussed in Chapter 1, the discipline of statistics centers around three main goals:

1. exploring and describing data and a phenomenon of interest;

2. predicting one variable using another variable;

3. drawing causal conclusions about the effect of changing one variable on another.

Much of Chapter 6 focused on predictive linear and logistic regression, which are two of the central approaches that statisticians use to carry out Goal 2. Building and fitting models, as well as generating summary statistics and data visualizations, are the main ways that statisticians deliver on Goal 1. In this final chapter, we focus on Goal 3.

Prediction is very important for many areas of health care (e.g., the number of ER visits at 11 pm on a Saturday night, so that a hospital can plan staffing levels well), public administration (e.g., tax revenue, so a government can plan a budget well), and business (e.g., the number of bags of basmati rice that customers will buy this week from a store, so the store manager can plan inventory levels well). But often *causality* is the core issue, e.g., what will be the *effect* of a change in how the health insurance system works, a change in the tax rates, or a new advertising campaign by a store.

It is helpful to keep a concrete example in mind, so we often focus on the case where the individuals are patients with an illness, and they can be given a new drug (the *treatment*) or the current standard drug (the *control*). Does the new drug improve the patient's health within a week or not? This is a challenging question to answer. If the patient's health improves within a week after taking the new drug, how can we know whether the improvement was *caused* by the new drug or whether the patient's health would have improved anyway?

11.2 Causal framework

11.2.1 Potential outcomes and treatment effects

To address issues like this clearly, it is important to formalize notation and terminology.

Definition 11.2.1 (Assignment)**.** Consider n patients in a study, labeled from 1 through n. Define the jth *assignment*

$$W_j \in \{0, 1\}$$

to be 1 if patient j receives the treatment (e.g., new drug) and 0 if patient j receives the control (e.g., standard of care). Collect all the assignments in the study as the *treatment assignment vector*

$$\mathbf{W} = (W_1, \ldots, W_n) \in \{0, 1\}^n.$$

For simplicity, we will assume throughout this chapter that assignments are binary (so there is one treatment group and one control group, but the concepts can be extended to situations where there are several different treatments under consideration in the same study). We are also assuming that the patient receives the treatment they were assigned to. In practice, complications sometimes arise, such as that a patient might be assigned a drug but then might forget to take it (or stop taking it due to side effects). This issue is called *non-compliance*. In some situations this is a crucial issue, but we ignore this issue here for simplicity.

In the statistical literature, less context-specific terminology is often used in setting up these types of assumptions, referring to the jth *individual* or *unit*, rather than the jth patient. The word "individual" or "unit" could refer to, e.g., a patient, person, family, animal, object, webpage, university, company, or country. In some applied contexts where the individuals are humans, the term "unit" can appear dehumanizing and is inappropriate.

Definition 11.2.2 (Potential outcome and treatment effect)**.** Let $Y_j(w_1, \ldots, w_n)$ be a *potential outcome* for patient j, defined to be the (possibly random) outcome for patient j if the assignments for all the patients were $w_1, \ldots, w_n$. There are 2^n potential outcomes for patient j, corresponding to the 2^n possible different assignments.

For example, in the context of a drug trial we often let the potential outcomes be binary, with an outcome of 1 indicating a health improvement and 0 indicating a lack of health improvement.

Jerzy Neyman introduced potential outcomes in causal studies in 1923.

11.2.2 Interference and treatment effects

It can be very unwieldy having 2^n potential outcomes for each patient, e.g., for $n = 10$ there are over 1000 potential outcomes to deal with for each patient. Fortunately, in many situations it is reasonable to assume that the outcome of a patient is unaffected by what treatments the other patients received.

Assumption 11.2.3 (Non-interference)**.** The non-interference assumption says that for each j, the treatment of others has no impact on the outcome for the jth individual:

$$Y_j(w_1, \ldots, w_{j-1}, w_j, w_{j+1}, \ldots, w_n) = Y_j(w'_1, \ldots, w'_{j-1}, w_j, w'_{j+1}, \ldots, w'_n),$$

for all $(w_1, \ldots, w_{j-1}, w_{j+1}, \ldots, w_n) \in \{0,1\}^{n-1}$ and $(w'_1, \ldots, w'_{j-1}, w'_{j+1}, \ldots, w'_n) \in \{0,1\}^{n-1}$.

Under the non-interference assumption, we will write the potential outcomes in the vastly simpler notation

$$Y_j(w_j),$$

and there are only two of them, $Y_j(0)$ and $Y_j(1)$, rather than 2^n of them. We then write the collection of potential outcomes under the control as

$$\mathbf{Y}(0) = \{Y_1(0), \ldots, Y_n(0)\},$$

and the corresponding collection under the treatment as

$$\mathbf{Y}(1) = \{Y_1(1), \ldots, Y_n(1)\}.$$

The non-interference assumption was introduced by Sir David R. Cox in 1958.

☣ **11.2.4.** We regard $\mathbf{Y}(1), \mathbf{Y}(0)$ as random variables. Later, we will sometimes carry out inference by conditioning on them — making them fixed. Some researchers always assume $\mathbf{Y}(1), \mathbf{Y}(0)$ are fixed, which simplifies some points of exposition and complicates others. As always in statistics, it is important to be clear about what is being conditioned on, and what is being viewed as fixed and what is being viewed as random.

In causal inference, we often want to estimate *treatment effects*: what is the causal impact on person j's outcome if person j is moved from control to treatment? What is the average causal impact for n individuals if all of them are moved from control to treatment?

Definition 11.2.5 (Treatment effect)**.** The *treatment effect* for the jth patient of moving from assignment $(w_1, \ldots, w_n)$ to assignment $(w'_1, \ldots, w'_n)$ is the difference

$$\tau_j = Y_j(w'_1, \ldots, w'_n) - Y_j(w_1, \ldots, w_n).$$

Under the non-interference assumption, the treatment effect for the jth person simplifies to

$$\tau_j = Y_j(1) - Y_j(0).$$

The *average treatment effect* for the n individuals is the sample mean of their individual treatment effects:

$$\bar{\tau} = n^{-1} \sum_{j=1}^{n} \tau_j.$$

In most applications, there seems little reason to expect each person's reaction to a drug to be the same. In the rather extraordinary situation where they are all the same, i.e., $\tau_j = \tau$ for all j for some constant τ, then the treatment effects are said to be *homogeneous*; otherwise, they are said to be *heterogeneous*.

The average treatment effect $\bar{\tau}$ is the average causal effect in the *sample*. This is called a *finite sample* estimand since it applies only to the specific group of patients in the study. Another causal quantity is $\mathrm{E}[\tau_1]$, where the expectation averages over some population. This population quantity will be explained more in the next section, where we define a population to average over.

☣ **11.2.6.** Non-interference is a very powerful assumption. For cancer drugs it may be reasonable to assume that treating one individual does not impact the medical outcome of other individuals. However, if the drug impacts a virus, e.g., giving the patient immunity, then this treatment may not only improve the health of that patient, but also improve the health of other patients to whom that patient may have spread the virus. In such a situation, the non-interference assumption would be violated. In applying the potential outcomes framework in practice, thought is always needs about whether non-interference is a plausible assumption.

Let Y_j be the actual outcome for person j. The following simple but fundamental identity, which is called the *switching equation*, links person j's actual outcome Y_j to their potential outcomes $(Y_j(0), Y_j(1))$.

Theorem 11.2.7 (Switching equation)**.**

$$Y_j = Y_j(W_j) = W_j Y_j(1) + (1 - W_j) Y_j(0), \quad j = 1, \ldots, n.$$

Proof. The identity just says that if $W_j = 1$ then the actual outcome is $Y_j(1)$, while if $W_j = 0$, then the actual outcome is $Y_j(0)$. ∎

☣ **11.2.8.** An additional assumption we are making is that the form of the treatment is the same across units and that the outcome does not vary with how the treatment was administered, e.g., the

same form of drug is administered each time treatment is assigned (and it has not expired!) and the packaging of the tablet makes no difference to the health outcome. For example, if the treatment is a surgery and several different surgeons perform the surgery, then we are assuming it does not matter which surgeon performs the surgery. If there are major differences in how the surgeons perform the surgery, we could consider having one treatment group for each surgeon (rather than having W_j be binary), or clarifying that the treatment protocol includes being randomly assigned to a surgeon, in which case the causal effect would be a mixture of the effects for the various surgeons. Likewise, we are assuming that there is only one form of control for all the units.

Definition 11.2.9 (SUTVA)**.** The *Stable Unit Treatment Value Assumption (SUTVA)* is shorthand for the assumptions of non-interference and, as discussed in the previous biohazard, that the form of the treatment is the same across all units (and likewise for the form of the control).

Statisticians will typically get to observe the assignment W_j and outcome Y_j, the realized values of which, consistent with our usual notation, we write using the corresponding lowercase letters: (w_j, y_j). The output tells us one of the potential outcomes, $Y_j = Y_j(W_j)$, but not the other. A major challenge in causal inference, sometimes called the *fundamental problem of causal inference*, is that we only get to observe one of the two potential outcomes for an individual. It follows that we can never directly observe an individual treatment effect $Y_j(1) - Y_j(0)$. The *missing* potential outcome, $Y_j(1 - W_j)$, is sometimes called a *counterfactual* — it *could* have happened, but did not. To connect the observed outcome to the observed assignment, we use the realized potential outcomes $y_j(0)$ and $y_j(1)$: as in the switching equation,

$$y_j = y_j(w_j) = w_j y_j(1) + (1 - w_j) y_j(0), \quad j = 1, \ldots, n.$$

It is essential to consider the probabilistic link between the potential outcomes and the assignments. For example, is each individual randomized to treatment or control by flipping a fair coin (independent of all the potential outcomes)? Or does a doctor with decades of experience examine individuals before assignment and then assign an individual to treatment only if the doctor believes the individual would not respond well to the control? This link is formalized through the notion of an *assignment mechanism*.

Definition 11.2.10 (Assignment mechanism)**.** The *assignment mechanism* is the joint probability mass function of the assignments given the potential outcomes:

$$P(\mathbf{W} = \mathbf{w} | \{\mathbf{Y}(0), \mathbf{Y}(1)\}).$$

11.3 Ethics of experimentation

Experimentation can lead to great improvements in welfare, allowing researchers to efficiently and systematically evaluate new medicines and advance scientific understanding. However, not all forms of experimentation are ethical. The American Statistical Association has a code of conduct for statisticians, which emphasizes that:

> The ethical statistician protects and respects the rights and interests of human and animal subjects at all stages of their involvement in a project. This includes respondents to the census or to surveys, those whose data are contained in administrative records, and subjects of physically or psychologically invasive research.

A researcher should not carry out a form of experimentation where they know or expect that individuals in the study will be seriously harmed by the experiment. Intention matters, being a pure utilitarian is not enough. In drug trials, the control is typically the best standard of medical care currently available. Likewise, a trial for a new treatment should only be conducted if there is a reasonable chance that the treatment will improve patient welfare.

Typically tiny groups are initially experimented on, to ensure the safety of the drug, to minimize any surprising sources of damage. Once basic safety has been established larger studies are carried out to see if the drugs have substantial treatment effects. After drugs are approved, often large observational studies (e.g., using electronic medical records) are conducted to try to identify if there are any side effects which impact small subsets of the wider population which the trials might not be large enough to accurately capture.

Statisticians are expected to ensure that each individual in the experiment has given "informed consent" over the experiment. Each individual is expected to be briefed about how the experiment will be carried out and for what reason it is being conducted, including possible hazards. Making the briefing overly complicated to extract consent is not viewed as yielding informed consent.

Some technology companies seem not to take informed consent particularly seriously, while carrying out many experiments with the functioning of their websites. Often consent is claimed by adding a catch-all experimentation clause to their long and rarely read user agreements. It is not at all clear whether such clauses represent what is usually meant by informed consent.

When humans are the subject of experimentation in academic or medical institutions, prior approval of the project generally needs to be obtained by an *Institutional Review Board* (IRB).

In contrast, many technology companies are performing experiments on human subjects without

needing IRB approval or oversight. For example, Facebook and Cornell researchers published an article in the Proceedings of the National Academy of Sciences of the United States of America (PNAS) in 2014 analyzing an experiment that Facebook conducted on manipulating emotions of users, manipulating the news feeds of about 700,000 users to show especially positive or especially negative content in order to assess the effect on the user's emotions. See

`https://journals.sagepub.com/doi/full/10.1177/1747016115599568`

for a commentary on the ethics of this Facebook experiment.

11.4 Randomized control trials

The following kind of assignment mechanism, in which the assignments are independent of the potential outcomes, is extremely important in both theory and practice.

Definition 11.4.1 (Randomization)**.** If the assignment mechanism satisfies

$$P(\mathbf{W} = \mathbf{w}|\{\mathbf{Y}(0), \mathbf{Y}(1)\}) = P(\mathbf{W} = \mathbf{w}),$$

i.e., the assignments are independent of the potential outcomes, then the assignments have been *randomized*. Using the independence symbol $\perp\!\!\!\perp$, this assumption can be written compactly as

$$\mathbf{W} \perp\!\!\!\perp \{\mathbf{Y}(0), \mathbf{Y}(1)\}.$$

How can we randomize the assignments in a study? To create the random assignments, we can generate them using a random number generator on a computer, without allowing any dependence on the individuals in the study. For example, we can (metaphorically) flip a coin for each patient to decide whether to assign them to treatment or control.

Definition 11.4.2 (RCT)**.** Experiments where the assignments are randomized are called *randomized experiments* or *randomized control trials* (RCTs).

RCTs are one of statistical science's most important contributions to improving human welfare.

Example 11.4.3 (RCT for chronic fatigue)**.** In 1991, I.M. Cox, M.J. Campbell, and D. Dowson published in *The Lancet* results from a RCT which compared intra-muscular magnesium injections to placebo injections for 32 people diagnosed with chronic fatigue syndrome. The outcomes and assignments are given in Table 11.1. There were 15 people who were given the magnesium injections, and there were 17 people who were given the placebo. In total, 15 people reported feeling better after the injection, and 17 reported not feeling better.

		Placebo	Magnesium
		$w = 0$	$w = 1$
Did not feel better	$y = 0$	14	3
Feel better	$y = 1$	3	12

Table 11.1: Treatment of chronic fatigue syndrome through intra-muscular magnesium compared to placebo.

Randomized experiments are enormously important in modern society and research. They are being applied in many areas of science, social science, medicine, technology, and public policy. For example:

- Most pharmaceutical drug approvals are based upon the results from RCTs.
- Technology companies such as Google and LinkedIn run thousands of RCTs each day, tuning their websites, to encourage, for example, click-through to advertisers' websites or community engagement. Typically, large technology companies carry out the experiments in stages, much like drug trials: first, to mitigate the risk of new product release; second, to find degradations before releasing to the wider public; third, to obtain a measure of how liked a new product is through a gradual release. In technology companies and in data science, these tests are often called *A/B tests*; in statistics terms, they are randomized experiments with binary assignments.
- Abhijit Banerjee, Esther Duflo, and Michael Kremer received the Nobel Prize in Economics in 2019 for their work on randomized experiments that studied interventions aiming to reduce global poverty. The press release for their award states, "The research conducted by this year's Laureates has considerably improved our ability to fight global poverty. In just two decades, their new experiment-based approach has transformed development economics."

11.4.1 Causal estimands: finite sample and population-based

The above setup delivers the observed data in pairs:

$$(w_1, y_1), \ldots, (w_n, y_n).$$

Recall that the observed outcomes and observed treatments are linked through $y_j = y_j(w_j)$ and, crucially, only the pairs

$$(w_1, y_1), \ldots, (w_n, y_n)$$

are observed. In particular, we only observe half of the full set of potential outcomes

$$\{y_1(0), y_1(1)\}, \ldots, \{y_n(0), y_n(1)\}.$$

So it is not obvious whether it is even possible to obtain a good estimate of the finite sample quantity $\bar{\tau}$ or the population quantity $\mathrm{E}[\tau_1]$, even if the sample size is large.

It is important to distinguish carefully between the estimands $\bar{\tau}$ and $\mathrm{E}[\tau_1]$. This distinction is similar to the distinction (discussed in Chapter 10) between the design-based perspective and the population model-based perspective in sampling.

1. The **finite sample**, or design-based, quantity $\bar{\tau}$ is specific to the units in the study, i.e., the average outcome if all the n units in the study are given the treatment minus the average outcome if all the n units in the study are given the control. The finite sample quantity says, in principle, nothing about the possible causal effect on any units which are not in the study. The analysis is entirely self-contained, making conclusions only about the n units. This style of study is limited but powerful.

   On the one hand, it can be carried out without making strong assumptions such as how the units in the study were sampled from the wider population. On the other hand, it is severely limited by the fact that it provides no general knowledge about how the treatment might work beyond these n units. If we are focusing on a finite sample estimand, we typically take an approach similar to the *design-based* perspective from Chapter 10, treating the potential outcomes as fixed and letting the randomness be only due to randomness in the assignments.

2. The **population quantity** $\mathrm{E}[\tau_1]$ is the causal quantity for all units in a wider population beyond the sample. This is extrapolative: inference will take data from the n units and extrapolate to the entire population. This kind of inference is of course highly desirable to help inform decisions about, e.g., whether a new drug should be approved and widely prescribed. But it requires stronger assumptions than does inference for the finite sample quantity. If we are focusing on a population estimand, we typically take a model-based approach, building a statistical model for $\{W_1, Y_1(0), Y_1(1)\}, \ldots, \{W_n, Y_n(0), Y_n(1)\}$, such as assuming that they are i.i.d. draws from some parametric or nonparametric model.

These two statistical strategies are also summarized in Table 11.2.

Example 11.4.4 (Pharmaceutical drug trial)**.** Most pharmaceutical drug trials are carried out using the finite sample approach, making no conclusion as to the effect of the drug on people not in the

Random: $\mathbf{W}, \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{Y}$ Data: $\mathbf{W}, \mathbf{Y}$ Randomization: $\mathbf{W} \perp\!\!\!\perp \{\mathbf{Y}(0), \mathbf{Y}(1)\}$		
Statistical strategy	Estimand	Inferential framework
Finite sample condition on $\mathbf{Y}(0) = \mathbf{y}(0), \mathbf{Y}(1) = \mathbf{y}(1)$	$\overline{\tau} = \frac{1}{n} \sum_{j=1}^{n} \{y_j(1) - y_j(0)\}$	$\mathbf{W} \mid \{\mathbf{Y}(0) = \mathbf{y}(0), \mathbf{Y}(1) = \mathbf{y}(1)\}$
Population unconditional	$\mathrm{E}[\tau_1] = \mathrm{E}[Y_1(1)] - \mathrm{E}[Y_1(0)]$	Unconditional or $\mathbf{Y} \mid \mathbf{W}$

Table 11.2: Statistical strategies, estimands and inferential frameworks for randomized control trials.

trial. In practice, studies like this are rarely carried out in isolation; typically sequences of studies are conducted in a variety of circumstances, and then the results are pooled using *meta-studies* (e.g., through hierarchical modeling).

Example 11.4.5 (experiment about technology design)**.** In technology experiments, e.g., to see if a change to a website yields more click throughs, the population quantity $\mathrm{E}(\tau_1)$ is typically what researchers are interested in — what would happen if they made the treated website the default. This may sound straightforward, but can be tricky as the population can be difficult to define (e.g., the population might adjust as the technology is adopted by new users).

11.5 A population-based statistical model for experiments

In population-based modeling, we assume a statistical model. Here the model will be for the triples

$$\{W_1, Y_1(0), Y_1(1)\}, \ldots, \{W_n, Y_n(0), Y_n(1)\},$$

and we assume that $\{W_j, Y_j(0), Y_j(1)\}$ are independent across the subscript j. This implies that the pairs

$$(W_1, Y_1), \ldots, (W_n, Y_n)$$

are independent. If this is combined with the randomization assumption from Definition 11.4.1, then

$$P(\mathbf{W} = \mathbf{w} | \mathbf{Y}(0), \mathbf{Y}(1)) = \prod_{j=1}^{n} P(W_j = w_j).$$

Let's now assume that the $\{W_j, Y_j(0), Y_j(1)\}$ are i.i.d., not just independent. Our estimand is $\mathrm{E}[\tau_1]$, which of course equals $\mathrm{E}[\tau_j]$ for any j — we are not trying to single out patient 1. We will first show

how powerful and elegant the randomization assumption is for enabling causal inference. By linearity,

$$\mathrm{E}[\tau_1] = \mathrm{E}[Y_1(1)] - \mathrm{E}[Y_1(0)],$$

and then we can try to estimate the two terms. A naive approach would be estimate $\mathrm{E}[Y_1(1)]$ using the sample mean for the treated group and $\mathrm{E}[Y_1(0)]$ using the sample mean for the control group. Without the assumption of randomized assignments, the naive approach could be severely biased: $\mathrm{E}[Y_1(1)]$ is an average over the whole population (including counterfactual outcomes for people who actually will not get the treatment), but the people assigned to treatment in the sample may tend to be different (e.g., in age or health status) than the people in the population overall. With the assumption of randomized assignments, matters become dramatically simpler: we then have

$$\theta_1 = \mathrm{E}[Y_1(1)] = \mathrm{E}[Y_1(1)|W_1 = 1] = \mathrm{E}[Y_1|W_1 = 1],$$

$$\theta_0 = \mathrm{E}[Y_1(0)] = \mathrm{E}[Y_1(0)|W_1 = 0] = \mathrm{E}[Y_1|W_1 = 0],$$

and it is straightforward to devise an estimator of $\mathrm{E}[Y_1|W_1 = 1]$ using the data from the treatment group, and $\mathrm{E}[Y_1|W_1 = 0]$ using the data from the control group.

11.5.1 Likelihood-based inference for population estimand

As above, suppose that the $\{W_j, Y_j(0), Y_j(1)\}$ are i.i.d. and that the potential outcomes are binary. As in predictive regression, where we condition on the predictors, we will condition on the assignments. Here the MLE of the population quantity $\mathrm{E}[\tau_1]$, given the $W_1, \ldots, W_n$, will be derived. With notation as before, the probability of the observed outcome being a 1 given the assignments is

$$\theta_0 = P(Y_1 = 1|W_1 = 0), \quad \text{and} \quad \theta_1 = P(Y_1 = 1|W_1 = 1).$$

In this subsection we will use likelihood methods to estimate θ_0, θ_1 efficiently, and combine them to estimate the population causal quantity

$$\mathrm{E}[\tau_1] = \theta_1 - \theta_0.$$

Factor

$$P(Y_1 = y_1, W_1 = w_1; \theta_0, \theta_1) = P(Y_1 = y_1|W_1 = w_1; \theta_0, \theta_1)P(W_1 = w_1).$$

We will make likelihood inference *conditional* through $P(Y_1 = y_1|W_1 = w_1, \theta_0, \theta_1)$, as $P(W_1 = w_1)$ has no direct information about θ_0, θ_1. So we will use the conditional log-likelihood

$$\begin{aligned}
\log L(\theta_0, \theta_1) &= \log P(\mathbf{Y} = \mathbf{y}|\mathbf{W} = \mathbf{w}; \theta_0, \theta_1) \\
&= \sum_{j=1}^{n} \log P(Y_j = y_j|W_j = w_j; \theta_0, \theta_1) \\
&= \sum_{j:w_j=0} \log \theta_0^{y_j}(1-\theta_0)^{1-y_j} + \sum_{j:w_j=1} \log \theta_1^{y_j}(1-\theta_1)^{1-y_j} \\
&= \sum_{j=1}^{n}(1-w_j) \log \theta_0^{y_j}(1-\theta_0)^{1-y_j} + \sum_{j=1}^{n} w_j \log \theta_1^{y_j}(1-\theta_1)^{1-y_j} \\
&= \log L_0(\theta_0) + \log L_1(\theta_1),
\end{aligned}$$

where

$$\log L_0(\theta_0) = \sum_{j=1}^{n}(1-w_j)\Big(y_j \log(\theta_0) + (1-y_j)\log(1-\theta_0)\Big),$$

$$\log L_1(\theta_1) = \sum_{j=1}^{n} w_j\Big(y_j \log(\theta_1) + (1-y_j)\log(1-\theta_1)\Big).$$

The result is two log-likelihoods for Bernoulli experiments, with probabilities of success θ_0 and θ_1, and sample sizes $\sum_{j=1}^{n}(1-w_j)$ (the number of control units) and $\sum_{j=1}^{n} w_j$ (the number of treated units), respectively. This simple structure has popped out without any questionable modeling assumptions; it follows entirely from conditioning on the assignments and the binary outcomes and assignments.

From results we have seen earlier about maximum likelihood with Bernoullis, the MLEs are

$$\hat{\theta}_0 = \frac{\sum_{j=1}^{n} Y_j(1-w_j)}{\sum_{j=1}^{n}(1-w_j)}, \quad \hat{\theta}_1 = \frac{\sum_{j=1}^{n} Y_j w_j}{\sum_{j=1}^{n} w_j},$$

respectively, assuming, of course, that $0 < \sum_{j=1}^{n} w_j < n$. Note that $\hat{\theta}_1$ is the fraction of outcomes which are 1 among the people who are treated, while $\hat{\theta}_0$ is the fraction of outcomes which are 1 among those in the control. Also, $\hat{\theta}_1$ is the sample mean of the outcomes in the treatment group, and $\hat{\theta}_0$ is the sample mean of the outcomes in the control group.

It follows that the MLE of $\mathrm{E}[\tau_1]$ is

$$\widehat{\mathrm{E}[\tau_1]} = \hat{\theta}_1 - \hat{\theta}_0,$$

which is simply the difference in sample means between the two groups. Some results about the MLEs of θ_0, θ_1, and $\mathrm{E}[\tau_1]$ are recorded below.

Theorem 11.5.1. *With the above assumptions and notation, we have*

$$\mathrm{E}[\hat{\theta}_0|\mathbf{W} = \mathbf{w}] = \theta_0, \quad \mathrm{E}[\hat{\theta}_1|\mathbf{W} = \mathbf{w}] = \theta_1,$$

$$\mathrm{Var}(\hat{\theta}_0|\mathbf{W}=\mathbf{w}) = \frac{\theta_0(1-\theta_0)}{\sum_{j=1}^n (1-w_j)}, \quad \mathrm{Var}(\hat{\theta}_1|\mathbf{W}=\mathbf{w}) = \frac{\theta_1(1-\theta_1)}{\sum_{j=1}^n w_j},$$

and the Fisher information in the sample is

$$\mathcal{I}_{\mathbf{Y}|\mathbf{W}=\mathbf{w}}(\theta_0,\theta_1) = \begin{pmatrix} \mathcal{I}_{\mathbf{Y}|\mathbf{W}=\mathbf{w}}(\theta_0) & 0 \\ 0 & \mathcal{I}_{\mathbf{Y}|\mathbf{W}=\mathbf{w}}(\theta_1) \end{pmatrix} = \begin{pmatrix} \frac{\sum_{j=1}^n (1-w_j)}{\theta_0(1-\theta_0)} & 0 \\ 0 & \frac{\sum_{j=1}^n w_j}{\theta_1(1-\theta_1)} \end{pmatrix}.$$

For the MLE of $\mathrm{E}[\tau_1]$*, we have*

$$\mathrm{E}[\hat{\theta}_1 - \hat{\theta}_0|\mathbf{W}=\mathbf{w}] = \mathrm{E}[\tau_1], \quad \mathrm{Var}(\hat{\theta}_1 - \hat{\theta}_0|\mathbf{W}=\mathbf{w}) = \frac{\theta_1(1-\theta_1)}{\sum_{j=1}^n w_j} + \frac{\theta_0(1-\theta_0)}{\sum_{j=1}^n (1-w_j)}.$$

Proof. All the likelihood calculations are standard from chapter 6, as is the variance result at the end. The final result for $\mathrm{E}[\tau_1]$ follows from noting that $\mathrm{E}[\tau_1] = \theta_1 - \theta_0$, and $\hat{\theta}_1, \hat{\theta}_0$ are conditionally independent given $\mathbf{W}$. ∎

Hence the maximum likelihood estimators of θ_0, θ_1 achieve the Cramér-Rao lower bound and can be combined to estimate $\mathrm{E}[\tau_1]$ efficiently and unbiasedly, where all of these statements are conditional on $\mathbf{W}$.

In words, in the context of our drug trial, this estimator is the fraction of the treated whose outcomes were a health improvement minus the fraction of the controls whose outcomes were a health improvement.

Approximate inference on $\widehat{\mathrm{E}[\tau_1]}$ can be based on the usual approximate pivot

$$\frac{\widehat{\mathrm{E}[\tau_1]} - \mathrm{E}[\tau_1]}{\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{\sum_{j=1}^n w_j} + \frac{\hat{\theta}_0(1-\hat{\theta}_0)}{\sum_{j=1}^n (1-w_j)}}} | (\mathbf{W}=\mathbf{w}) \dot{\sim} \mathcal{N}(0,1),$$

where for the asymptotics we suppose that both $\sum_{j=1}^n w_j$ and $\sum_{j=1}^n (1-w_j)$ get large.

The approximate pivot can be used to perform population testing about the value of $\mathrm{E}[\tau_1]$ or obtain population confidence intervals for $\mathrm{E}[\tau_1]$.

Example 11.5.2 (Continuing Example 11.4.3)**.** Then $\sum_{j=1}^n y_j w_j = 12$, $\sum_{j=1}^n w_j = 15$, which implies that $\hat{\theta}_1 = 12/15 = 0.80$, while

$$\mathrm{SE}(\hat{\theta}_1) = \sqrt{0.8 \times 0.2/15} \approx 0.103.$$

Likewise, $\hat{\theta}_0 = 3/17 \approx 0.176$, while $\mathrm{SE}(\hat{\theta}_1) \approx 0.092$. Taking this together,

$$\widehat{\mathrm{E}[\tau_1]} = 12/15 - 3/17 \approx 0.624,$$

	$Y_1(1) = 0$	$Y_1(1) = 1$
$Y_1(0) = 0$	p_{00}	p_{01}
$Y_1(0) = 1$	p_{10}	p_{11}

Table 11.3: Joint distribution of the potential outcomes.

with a standard error of roughly $\sqrt{0.103^2 + 0.092^2} \approx 0.138$. The estimate is roughly 4.52 standard errors above 0. A nominal 95% confidence interval for $\mathrm{E}[\tau_1]$ can be based on the above approximate pivot, giving $0.624 \pm 1.96 \times 0.138$, which simplifies to the interval

$$0.624 \pm 0.271.$$

The RCT suggests that if all the 32 patients were given the magnesium injection compared to the placebo then it would improve the feeling of well-being of between 35% to 89% of patients, which is statistically significant at any reasonable level. Of course there may be other concerns to consider, such as side effects, time taken by the procedure, pain of the procedure, other competitor treatments, etc.

☣ **11.5.3.** To think more deeply about this structure, let's take a step back. Suppose that the potential outcomes are binary (e.g., indicators for whether a patient recovers from an illness within a week). The population behavior is controlled by the parameters

$$p_{ik} = P(Y_1(0) = i, Y_1(1) = k), \quad \text{for} \quad i, k \in \{0, 1\},$$

which are displayed in Table 11.3. This table specifies the joint PMF of the potential outcomes $Y_1(0)$ and $Y_1(1)$, implying $\theta_1 = \mathrm{E}[Y_1(1)] = p_{01} + p_{11}$ and $\theta_0 = \mathrm{E}[Y_1(0)] = p_{10} + p_{11}$. By linearity,

$$\begin{aligned} \mathrm{E}[\tau_1] &= \mathrm{E}[Y_1(1) - Y_1(0)] = \theta_1 - \theta_0 = (p_{01} + p_{11}) - (p_{10} + p_{11}) \\ &= p_{01} - p_{10}. \end{aligned}$$

Note that this is the difference of the off-diagonal parameters p_{01}, p_{10}; the main-diagonal parameters p_{00}, p_{11} do not appear in the final expression for $\mathrm{E}[\tau_1]$. The same is true of

$$\mathrm{Var}(\tau_1) = \mathrm{E}[\tau_1^2] - (\mathrm{E}[\tau_1])^2 = (p_{01} + p_{10}) - (p_{01} - p_{10})^2.$$

There are two parameters $(\theta_0, \theta_1) \in [0, 1]^2$ which we can learn from the data, but three embedded in $p_{00}, p_{01}, p_{10}, p_{11}$, as these must sum up to 1. The two parameters $\theta_0 = (p_{10} + p_{11}), \theta_1 = (p_{01} + p_{11})$

are not enough to pin down all the $p_{00}, p_{01}, p_{10}, p_{11}$. We can consistently estimate the mean of the treatment effect,

$$\mathrm{E}[\tau_1] = p_{01} - p_{10} = (p_{01} + p_{11}) - (p_{10} + p_{11}) = \theta_1 - \theta_0,$$

but not the entire joint distribution of the potential outcomes — whatever statistical procedure we employ. Solving for p_{00}, p_{10}, p_{01} in terms of θ_0, θ_1,

$$\begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = \begin{pmatrix} 1 + p_{11} - \theta_1 - \theta_0 & \theta_1 - p_{11} \\ \theta_0 - p_{11} & p_{11} \end{pmatrix}, \quad \text{where } 0 \le p_{11} \le \min(\theta_0, \theta_1).$$

This is a statistical impossibility result! Not everything is estimable from data — there are too many unobservable potential outcomes to allow us to estimate all the parameters just from seeing the outcomes and the assignments. This is important, for interesting quantities like

$$\begin{aligned} \mathrm{Var}(\tau_1) &= (p_{01} + p_{10}) - (p_{01} - p_{10})^2 = (\theta_1 + \theta_0 - 2p_{11}) - (\theta_1 - \theta_0)^2 \\ &= \theta_0(1 - \theta_0) + \theta_1(1 - \theta_1) + \theta_0\theta_1 - 2p_{11}. \end{aligned}$$

cannot be precisely learned from the data. In statistics, we say that $p_{00}, p_{01}, p_{10}, p_{11}$ are *not identified* from the data, but $p_{01} - p_{10}$ is *identified*.

In 2000, A. Philip Dawid sharply criticized the use of potential outcomes in causal studies in the Journal of the American Statistical Association, principally due to this lack of identification. His article, plus the published discussion, provides a good entry into some of the more contentious issues in the causal literature. Particularly thought-provoking is the note by Judea Pearl who advocates an approach that focuses on using directed graphs, where there is a node for each variable and an edge (depicted as an arrow) from the node for variable X to the node for variable Y if X is hypothesized to have a causal effect on Y.

11.5.2 Bayesian inference for population estimand

With assumptions and notation as above, we now consider a Bayesian approach instead of the MLE. Suppose that $(\theta_0 \perp\!\!\!\perp \theta_1)|\mathbf{w}$ and we have conjugate Beta priors, given the assignments, $\theta_0|\mathbf{w} \sim \mathrm{Beta}(\alpha_0, \beta_0)$ and $\theta_1|\mathbf{w} \sim \mathrm{Beta}(\alpha_1, \beta_1)$. Then using the Bernoulli likelihoods $\log L_0(\theta_0)$ and $\log L_1(\theta_1)$ above we find that $(\theta_0 \perp\!\!\!\perp \theta_1)|\mathbf{y}, \mathbf{w}$, with

$$\theta_0|\mathbf{y}, \mathbf{w} \sim \mathrm{Beta}\Big(\alpha_0 + \sum_{j=1}^{n} y_j(1 - w_j), \beta_0 + \sum_{j=1}^{n}(1 - y_j)(1 - w_j)\Big),$$

and

$$\theta_1|\mathbf{y}, \mathbf{w} \sim \mathrm{Beta}\Big(\alpha_1 + \sum_{j=1}^{n} y_j w_j, \beta_1 + \sum_{j=1}^{n}(1 - y_j) w_j\Big).$$

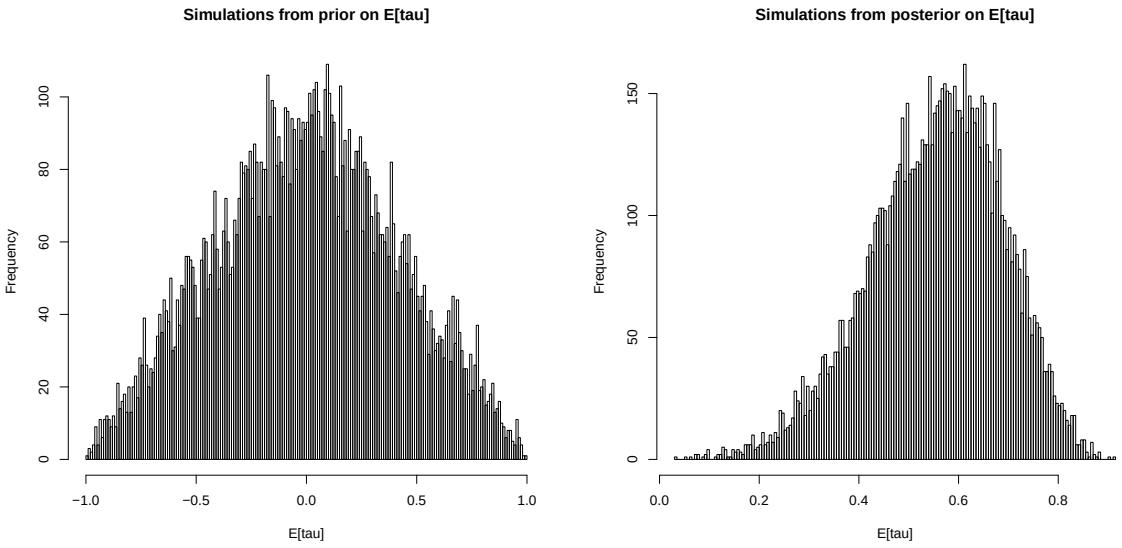


Figure 11.1: Histogram of simulations from prior (LHS) and posterior (RHS) of $\mathrm{E}(\tau_1)$.

Then

$$\mathrm{E}(\theta_1|\mathbf{y},\mathbf{w}) - \mathrm{E}(\theta_0|\mathbf{y},\mathbf{w}) = \frac{\alpha_1 + \sum_{j=1}^n y_j w_j}{\alpha_1 + \beta_1 + \sum_{j=1}^n w_j} - \frac{\alpha_0 + \sum_{j=1}^n y_j(1-w_j)}{\alpha_0 + \beta_0 + \sum_{j=1}^n 1 - w_j}.$$

We can simulate from the exact posterior $\mathrm{E}[\tau_1]|\mathbf{y},\mathbf{w}$ by simulating from the two independent posterior Beta distributions for θ_1 and θ_0 and calculating $\mathrm{E}[\tau_1] = \theta_1 - \theta_0$.

Example 11.5.4 (Continued from Example 11.5.2)**.** Take $\alpha_0 = \alpha_1 = 1$ and $\beta_0 = \beta_1 = 1$, and use $B = 10^6$ draws in the simulation. Then

$$\mathrm{E}[\mathrm{E}[\tau_1]|\mathbf{y},\mathbf{w}] \approx 0.554,$$

and a 95% credible interval is

$$[Q_{\mathrm{E}[\tau_1]|\mathbf{y},\mathbf{w}}(0.025), Q_{\mathrm{E}[\tau_1]|\mathbf{y},\mathbf{w}}(0.975)] \approx [0.264, 0.789].$$

The histograms of samples from the prior and posterior are given in Figure 11.1, showing the posterior is skewed, with a long left tail.

11.5.3 Testing of population estimand

Next we will consider hypothesis testing. Suppose the population null hypothesis is $H_0 : \mathrm{E}[\tau_1] = 0$, which would mean that the treatment has no population causal effect on the outcome compared to the control. We will test that against the composite alternative hypothesis $H_1 : \mathrm{E}[\tau_1] \neq 0$, that the treatment does have a population causal effect (it could be worse or better).

For the population null hypothesis of no causality, $\mathrm{E}[\tau_1] = 0$, form the test statistic

$$T = \frac{\widehat{\mathrm{E}[\tau_1]}}{\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{\sum_{j=1}^n w_j} + \frac{\hat{\theta}_0(1-\hat{\theta}_0)}{\sum_{j=1}^n (1-w_j)}}} | (\mathbf{W} = \mathbf{w}) \dot{\sim} \mathcal{N}(0,1)$$

where the approximate distribution statement is under the null. Then the null of no population causal effect for a nominal α-sized test is rejected if $|T| > Q_{\mathcal{N}(0,1)}(1-\alpha/2)$, while for a one-sided test $H_0 : \mathrm{E}[\tau_1] \leq 0$ against $H_1 : \mathrm{E}[\tau_1] > 0$, the null is rejected if $T > Q_{\mathcal{N}(0,1)}(1-\alpha)$. An alternative method is to use a likelihood ratio test.

Example 11.5.5 (Continuing Example 11.5.4)**.** The statistic T is approximately 4.52 for these data, which gives a nominal p-value of

$$1 - F_{\mathcal{N}(0,1)}(4.52) \approx 3 \times 10^{-6},$$

for a one-sided test.

11.6 A finite sample approach for experiments

Again, the focus is on the outcomes, the assignments, and the potential outcomes for the n units, which we collect together as $\mathbf{Y}, \mathbf{W}, \mathbf{Y}(1), \mathbf{Y}(0)$. We will now make a big switch, carrying out inference conditional on the potential outcomes

$$\mathbf{Y}(1) = \mathbf{y}(1), \quad \mathbf{Y}(0) = \mathbf{y}(0),$$

where $\mathbf{y}(1) = (y_1(1), \ldots, y_n(1))$ and $\mathbf{y}(0) = (y_1(0), \ldots, y_n(0))$. That is, we are viewing the potential outcomes as *fixed*, even though half of them will never be observed. Once we have conditioned, we move to the using the *finite sample approach* and make no assumptions about how the potential outcomes are generated. This is the same as the design-based framework from Chapter 10, where we make no assumptions about the finite population from which are sampling; instead, inference is driven by the sampling mechanism.

Typically the inferential focus in this approach is the *finite sample average treatment effect*

$$\bar{\tau} = n^{-1} \sum_{j=1}^n \{y_j(1) - y_j(0)\}.$$

Of course, other estimands may also be of interest, such as the median treatment effect in the sample or the proportion of patients for whom the treatment is beneficial.

Once we have decided to condition on $\mathbf{Y}(1), \mathbf{Y}(0)$, the only things left to write down probabilities for are $\mathbf{Y}$ and $\mathbf{W}$. But once we also condition on $\mathbf{W}$, the $\mathbf{Y}$ are determined by the switching equation

$$Y_j = W_j Y_j(1) + (1 - W_j) Y_j(0).$$

Thus

$$P(\mathbf{Y} = \mathbf{y}, \mathbf{W} = \mathbf{w} | \mathbf{Y}(1) = \mathbf{y}(1), \mathbf{Y}(0) = \mathbf{y}(0)) = P(\mathbf{W} = \mathbf{w} | \mathbf{Y}(1) = \mathbf{y}(1), \mathbf{Y}(0) = \mathbf{y}(0)).$$

Assuming that the assignments are randomized, this further simplifies to $P(\mathbf{W} = \mathbf{w})$. So in this approach we can focus on the randomization rather than having the burden of trying to model the Y_j or the potential outcomes. In the next subsection, we will derive a MoM estimator in this finite sample setting.

11.6.1 Method of moments based inference for finite sample estimand

Again we will focus on the core case where the assignments are randomized independently over the units:

$$P(\mathbf{W} = \mathbf{w}) = \prod_{j=1}^{n} P(W_j = w_j).$$

Using the above structure, we will deploy a method of moments estimator for $\bar{\tau}$. Following the MoM strategy, we start by deriving a moment equation (this time involving W_1, Y_1) and then use it to help come up with an estimator.

Theorem 11.6.1 (Finite sample approach)**.** *Assume the randomization of assignments and define*

$$G_1 = \frac{W_1 Y_1}{\mathrm{E}[W_1]} - \frac{(1 - W_1) Y_1}{\mathrm{E}[1 - W_1]}.$$

Then

$$\mathrm{E}[G_1 | \{Y_1(0) = y_1(0), Y_1(1) = y_1(1)\}] = y_1(1) - y_1(0),$$

and

$$\mathrm{Var}(G_1 | \{Y_1(0) = y_1(0), Y_1(1) = y_1(1)\}) = \frac{y_1^2(1)}{\mathrm{E}[W_1]} + \frac{y_1^2(0)}{1 - \mathrm{E}[W_1]} - \{y_1(1) - y_1(0)\}^2.$$

Proof. Note that

$$W_1 Y_1 = W_1 Y_1(1), \text{ and } (1 - W_1) Y_1 = (1 - W_1) Y_1(0).$$

These results follow from the key equation

$$Y_1 = W_1 Y_1(1) + (1 - W_1) Y_1(0),$$

by multiplying both sides by W_1 or by $1 - W_1$, and using the facts $W_1^2 = W_1, W_1(1 - W_1) = 0$ (which follow from the fact that W_1 is binary). We can also see these results directly. For example, $W_1 Y_1 = W_1 Y_1(1)$ is true since if $W_1 = 0$ then it just says $0 = 0$, while if $W_1 = 1$ then person 1 was assigned to the treatment group, so their outcome is $Y_1 = Y_1(1)$.

Conditioning on the potential outcomes,

$$\begin{aligned}&\mathrm{E}[W_1 Y_1 | \{Y_1(0) = y_1(0), Y_1(1) = y_1(1)\}] = \mathrm{E}[W_1 Y_1(1) | \{Y_1(0) = y_1(0), Y_1(1) = y_1(1)\}] \\ &= y_1(1)\mathrm{E}[W_1 | \{Y_1(0) = y_1(0), Y_1(1) = y_1(1)\}] = y_1(1)\mathrm{E}[W_1].\end{aligned}$$

Similarly,

$$\mathrm{E}[(1 - W_1)Y_1 | \{Y_1(0) = y_1(0), Y_1(1) = y_1(1)\}] = y_1(0)\mathrm{E}[1 - W_1].$$

Combining these results gives the stated result for the conditional mean of G_1. For the conditional variance, note that

$$G_1^2 = \frac{W_1 Y_1^2}{(\mathrm{E}[W_1])^2} + \frac{(1 - W_1)Y_1^2}{(\mathrm{E}[1 - W_1])^2},$$

as the cross term is zero (since $W_1(1 - W_1) = 0$). Thus,

$$\mathrm{E}[G_1^2 | \{Y_1(0) = y_1(0), Y_1(1) = y_1(1)\}] = y_1^2(1)\frac{\mathrm{E}[W_1]}{(\mathrm{E}[W_1])^2} + y_1^2(0)\frac{\mathrm{E}[1 - W_1]}{(\mathrm{E}[1 - W_1])^2} = \frac{y_1^2(1)}{\mathrm{E}[W_1]} + \frac{y_1^2(0)}{1 - \mathrm{E}[W_1]}.$$

∎

This probabilistic result suggests the method of moments estimator

$$\hat{\tau}_{\mathrm{MoM}}(\mathbf{W}) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{W_j Y_j}{\mathrm{E}[W_j]} - \frac{(1 - W_j)Y_j}{\mathrm{E}[1 - W_j]}\right),$$

for the estimand $\bar{\tau}$, which is conditionally unbiased given the potential outcomes:

$$\mathrm{E}[\hat{\tau}_{\mathrm{MoM}}(\mathbf{W}) | \{\mathbf{Y}(1) = \mathbf{y}(1), \mathbf{Y}(0) = \mathbf{y}(0)\}] = \frac{1}{n}\sum_{j=1}^{n}\{y_j(1) - y_j(0)\} = \bar{\tau}.$$

Note that $\mathrm{E}[W_j] = P(W_j = 1), \mathrm{E}[1 - W_j] = P(W_j = 0)$ are *known* probabilities (since we control how the random assignments are generated) so they are allowed to appear in the estimator. Also, note that in the simple case where fair coin flips are used for the randomized assignments,

$$\hat{\tau}_{\mathrm{MoM}}(\mathbf{W}) = \frac{1}{n/2}\sum_{j=1}^{n} W_j Y_j - \frac{1}{n/2}\sum_{j=1}^{n}(1 - W_j)Y_j.$$

More broadly, if n is large, then by LLN this is very likely to be close to the difference in sample means between the two groups,

$$\frac{1}{\sum_{j=1}^{n} W_j}\sum_{j=1}^{n} W_j Y_j - \frac{1}{\sum_{j=1}^{n}(1 - W_j)}\sum_{j=1}^{n}(1 - W_j)Y_j.$$

The method of moments estimator has conditional variance

$$\mathrm{Var}(\hat{\tau}_{\mathrm{MoM}}(\mathbf{W})|\{\mathbf{Y}(1)=\mathbf{y}(1),\mathbf{Y}(0)=\mathbf{y}(0)\}) = \frac{1}{n^2}\sum_{j=1}^{n}\left(\frac{y_j^2(1)}{\mathrm{E}[W_j]}+\frac{y_j^2(0)}{1-\mathrm{E}[W_j]}-\{y_j(1)-y_j(0)\}^2\right),$$

which will typically become small as n gets large.

☣ **11.6.2.** The conditional variance involves the quantities

$$n^{-1}\sum_{j=1}^{n}\frac{y_j^2(1)}{\mathrm{E}[W_j]}, \quad n^{-1}\sum_{j=1}^{n}\frac{y_j^2(0)}{1-\mathrm{E}[W_j]}, \quad \text{and } n^{-1}\sum_{j=1}^{n}y_j(0)y_j(1).$$

The first two terms can be unbiasedly estimated by

$$\frac{1}{n}\sum_{j=1}^{n}\frac{W_jY_j^2}{(\mathrm{E}[W_j])^2}, \quad \frac{1}{n}\sum_{j=1}^{n}\frac{(1-W_j)Y_j^2}{(1-\mathrm{E}[W_j])^2},$$

respectively, but the third is not possible to estimate well. This is a major problem. Instead of estimating the conditional variance of $\hat{\tau}_{\mathrm{MoM}}(\mathbf{W})$, many researchers resort to using an upper bound on the variance,

$$\lambda^2 = \frac{1}{n^2}\sum_{j=1}^{n}\left[\frac{y_j^2(1)}{\mathrm{E}[W_j]}+\frac{y_j^2(0)}{1-\mathrm{E}[W_j]}\right], \quad \text{estimated by} \quad \hat{\lambda}^2 = \frac{1}{n^2}\sum_{j=1}^{n}\left[\frac{W_jY_j^2}{(\mathrm{E}[W_j])^2}+\frac{(1-W_j)Y_j^2}{(1-\mathrm{E}[W_j])^2}\right].$$

Using the estimated upper bound for inference leads to conservative inference, i.e., the estimators are *at least* as precise as the reported standard errors say. We will see this approach in more detail in Section 11.6.4.

Example 11.6.3 (Continuing from Example 11.5.5)**.** Then $\hat{\tau}_{\mathrm{MoM}}(\mathbf{w}) \approx 0.562$. The square root of the conservative variance estimate is about 0.242.

11.6.2 Finite sample testing in randomized control trials

In the finite sample setting, there are two widely-used choices of null hypothesis. The first is the *Fisher null*, which states that the treatment effect is zero for every individual in the sample:

$$H_0 : \tau_j = 0, \ j = 1, \ldots, n,$$

where $\tau_j = y_j(1) - y_j(0)$. In our drug testing case, this says that the new drug has no improvement on the health outcome of any patient in the sample compared to the control. This is tested against the composite alternative $H_1 : \sum_{j=1}^{n}|\tau_j| > 0$, i.e., there is a causal effect on at least one patient.

The second finite sample null is the *Neyman null*, which states that the *average* causal effect across the sample is zero:

$$H_0 : \bar{\tau} = 0,$$

which is tested against the alternative $H_1 : \bar{\tau} \neq 0$. The Neyman null looks like a finite sample version of the population null hypothesis that the average causal effect in the population is 0. Note that the Fisher null implies the Neyman null, but not conversely.

The Neyman null allows some individual causal effects to be positive, and some to be negative, but they must exactly balance out over the sample. Exact balancing over a finite sample of units seems rather magical — there seems no scientific reason ever to expect that unless they are all individually 0, which would reduce to the Fisher null. To see even more clearly how fragile the Neyman null is, note that if the Neyman null holds and one patient drops out of the study or a new patient comes along and joins the study, then the Neyman null no longer holds unless the causal effect for the patient who dropped out or joined is 0.

11.6.3 Randomization test of Fisher null for the finite sample

The Fisher null says that the jth treatment effect $\tau_j = Y_j(1) - Y_j(0)$ is zero for every value of $j = 1, \ldots, n$. This implies that $Y_j(1) = Y_j(0) = Y_j$, revealing all potential outcomes to us as soon as we observe the Y_j. We will use the absolute value of the estimate

$$\hat{\tau}_{\text{MoM}}(\mathbf{w}) = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{w_j y_j}{\text{E}[W_j]} - \frac{(1 - w_j) y_j}{1 - \text{E}[W_j]} \right)$$

as test statistic. In determining the critical value of the test, the potential outcomes are held fixed, so the only source of randomness is the assignments given the potential outcomes: $\{\mathbf{W} | \mathbf{Y}(0) = \mathbf{y}(0), \mathbf{Y}(1) = \mathbf{y}(1)\}$. Furthermore, the randomization assumption means that $\mathbf{W}$ is independent of the potential outcomes. Then rejecting the Fisher null if

$$|\hat{\tau}_{\text{MoM}}(\mathbf{w})| > Q_{|\hat{\tau}_{\text{MoM}}(\mathbf{W})|}(1 - \alpha)$$

gives an α-sized test, which is called the *randomization test*.

The randomization test can be implemented by simulation, following much of the material from Chapter 10. Intuitively, we can imagine re-randomizing the units to treatment and control groups many times, each time following the same assignment mechanism as was actually used in the study. Drawing i.i.d. $\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(B)}$ from the assignment mechanism delivers the B i.i.d. draws

$$|\hat{\tau}_{\text{MoM}}(\mathbf{W}^{(1)})|, \ldots, |\hat{\tau}_{\text{MoM}}(\mathbf{W}^{(B)})|,$$

from which we can estimate the required quantile to set the critical value for the α-sized test.

It is important to take a step back and see what has been accomplished here. Up to simulation error (which we can measure and control the size of), the Fisher null of no finite sample causality can be

tested, without any need for asymptotics or parametric assumptions. The elegance and objectiveness of this statement has made this randomization test one of the most celebrated tests in all of statistics.

☣ **11.6.4.** The randomization test is conceptually very similar to a permutation test (see Chapter 10), and some authors use the terms interchangeably. However, in our terminology there is a difference in how permutation tests and randomization tests are used and interpreted. For both, the data are fixed and then shuffled in some way (randomly permuting which observation fall in which group for a permutation test, and re-running the random assignments for the randomization test). But the null hypothesis for the permutation test was that the population distributions for the two group were equal, whereas for the randomization test the null hypothesis is the Fisher null, which is a finite sample hypothesis.

It should be clear from Chapter 8 that we are somewhat skeptical of the use of p-values. However, if randomization-based causal p-values are needed, they can easily be estimated as

$$\hat{p} = \frac{1}{B}\sum_{l=1}^{B} I(|\hat{\tau}_{\text{MoM}}(\mathbf{W}^{(l)})| \geq |\hat{\tau}_{\text{MoM}}(\mathbf{w})|),$$

the fraction of the simulated statistics which are at least as large as the observed version $\hat{\tau}_{\text{MoM}}(\mathbf{w})$, as with permutation tests.

Example 11.6.5. (Continuing Example 11.4.3) Here we carry out a randomization test of the Fisher null of no causal impact of intra-muscular magnesium on fatigue compare to a placebo. It is based on $B = 10,000$ replications. Figure 11.2 shows the distribution of the test statistic under the null, while the horizontal red line shows the estimate value in the experiment. The estimated value seems exceptional. The simulation-based estimate of the exact p-value for the null is around

$$\hat{p} \approx 0.0387.$$

11.6.4 Asymptotic tests of Neyman null for the finite sample*

The Neyman null is more complicated to handle than the Fisher null. The approach is to construct an approximate pivot

$$T = \frac{\hat{\tau}_{\text{MoM}}(\mathbf{w})}{\sqrt{\lambda^2 - n^{-2}\sum_{j=1}^{n}\tau_j^2}} \dot{\sim} \mathcal{N}(0,1), \quad \text{where } \lambda^2 = \frac{1}{n^2}\sum_{j=1}^{n}\left[\frac{y_j^2(1)}{\text{E}[W_j]} + \frac{y_j^2(0)}{1-\text{E}[W_j]}\right].$$

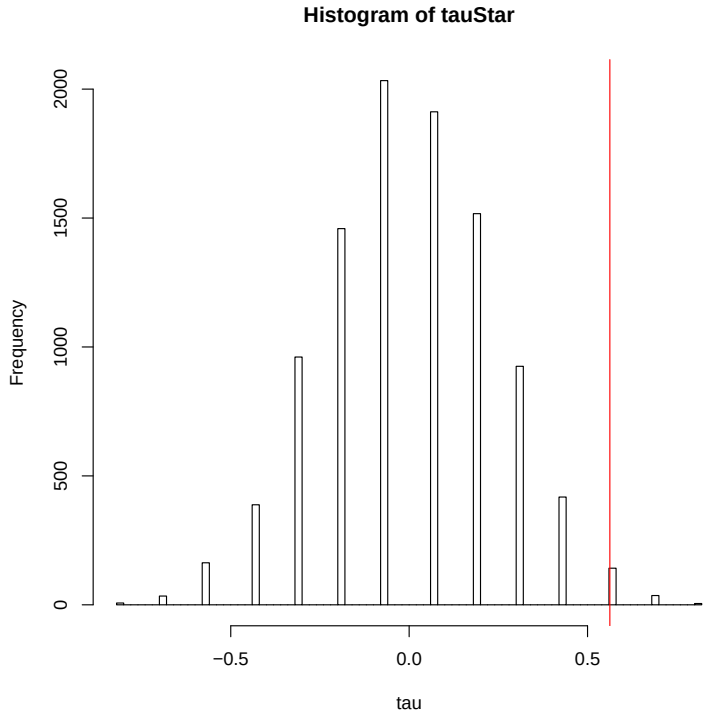


Figure 11.2: Null distribution of causal estimator and the estimated value (vertical red line)

Recall from Biohazard 11.6.2 that the variance in this denominator is not possible to estimate well, so we replace it by an estimated upper bound. This delivers the test statistic

$$\hat{\tau}_{\mathrm{MoM}}(\mathbf{w})/\hat{\lambda}, \text{ with } \hat{\lambda}^2 = \frac{1}{n^2}\sum_{j=1}^{n}\left[\frac{W_j Y_j^2}{(\mathrm{E}[W_j])^2} + \frac{(1-W_j)Y_j^2}{(1-\mathrm{E}[W_j])^2}\right],$$

which is compared to a $\mathcal{N}(0,1)$ distribution for testing. As we have inflated the variance, such a test will typically have nominal size which is less than α even if n is quite large. Hence this is a conservative asymptotic test. However, it is simple to compute.

11.7 Observational studies

Many researchers, particularly in the social sciences and public health, would like to make causal conclusions but have only data for which the assignments of the individuals were outside of the researcher's control. In statistics, such datasets are called *observational studies*. Most datasets in statistics are observational. Observational data are often used to provide descriptions or summaries, as well as generate predictions through predictive regressions. When we seek causal conclusions, the task is much harder.

Example 11.7.1. Let W_j be 1 if the jth person in a study went to Harvard before the age of 21 and 0 if they do not. Then let $Y_j(1)$ be their earnings at age 40 if they went to Harvard by 21 and $Y_j(0)$ be their earnings at age 40 if they did not go to Harvard by 21. Suppose we are able to find out W_j and Y_j for n individuals, where n is very large. Then

$$\tilde{\tau} = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{W_j Y_j}{\mathrm{E}[W_1]} - \frac{(1-W_j)Y_j}{1-\mathrm{E}[W_1]}\right)$$

would estimate the causal impact of going to Harvard on earnings if admissions were randomized. But admissions are not randomized! Parental characteristics, mathematical level reached in high school, geography, and many other factors impact both W_j and the potential earnings $Y_j(0), Y_j(1)$.

In RCTs, we used assignments, potential outcomes, and non-interference to define causal terms and randomization to drive causal inference. In observational studies we can still phrase causal statements using potential outcomes and non-interference, but it is fairly rare that we see useful forms of randomization appearing in nature or through societal actions.

11.7.1 Natural experiments

An influential counterexample was published in the American Economic Review in 1990 by Joshua Angrist, who used the random lottery draft numbers in the Vietnam war, to study the causal impact of veteran status on lifetime labor income. The lottery numbers, as the name suggested, were independent of any feature of the young men who were available for the draft — hence they form a randomization. This kind of secondary use of such randomization (i.e., a randomization exists but was generated by nature, not the researcher) is called a *natural experiment*. Many papers have been written in recent times which tell stories about possible natural experiments (e.g., tax thresholds, state lottery winnings, or geographical boundaries); some are convincing and some are threadbare. The 2021 Nobel Prize in Economics was awarded for this type of work to Angrist, David Card, and Guido Imbens.

Example 11.7.2 (Continued from Example 11.7.1)**.** If the Harvard admissions office were to calculate the average 40 year olds' earnings of the group of students just above the admissions cut-off and compare them to the average earnings of those just below the threshold, then that study would be close to a natural experiment viewing these marginal decisions as roughly random — and thus a form of administratively driven randomization. Then the use of RCT methods may well be appropriate — where the causal quantity being estimated is the impact on earnings of the marginal student being admitted to Harvard.

One approach to this overall research problem is to limit the scope of the social sciences and public health, only attempting to answer questions which can be answered by RCTs. But such RCTs are sometimes hard to carry out for some of the more important research questions which arise in those fields.

11.7.2 Conditioning on covariates

Instead of focusing on natural experiments, we follow a different rather more general route. This approach to observational studies is to condition on some *covariates*, and then assume that given those covariates the assignments are independent of the potential outcomes. Yet again we see that:

Conditioning is the soul of statistics.

We will see that causal inference *is* possible under these assumptions, and we can use statistical techniques analogous to those we deployed to analyze the RCTs. However, the conditional independence assumption is often hard to justify or verify. Write the covariates as

$$\mathbf{X} = (X_1, \ldots, X_n).$$

☣ **11.7.3.** The covariates must be *pretreatment variables*. This means that their values are determined before the assignments are made.

As with RCTs, we will take a potential outcomes approach and assume non-interference. But now we will generalize the assignment mechanism to

$$P(\mathbf{W}|\mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}), \mathbf{X}),$$

so the assignment probabilities can depend on the covariates.

11.7.3 Statistical model of observational studies

To analyze this setup we need to move from randomized assignments to *unconfounded assignments* given the covariates. It is this crucial assumption that will allow for causal inference for observational studies.

Assumption 11.7.4 (Unconfoundedness)**.** The assignments are called *unconfounded* or *ignorable* if they are conditionally independent of the potential outcomes, given the covariates:

$$[\mathbf{W} \perp\!\!\!\perp \{\mathbf{Y}(0), \mathbf{Y}(1)\}]|\mathbf{X}.$$

This important definition, allowing the dependence on covariates, is associated with the work of Donald B. Rubin from the 1970s and 1980s, who influentially advocated the use of potential outcomes in observational studies.

The right hand side of Figure 11.3 is graphical representation of unconfoundedness, with arrows going out from $\mathbf{X}$. The lack of a link between $\mathbf{W}$ and $\mathbf{Y}(0), \mathbf{Y}(1)$ denotes their conditional independence, given $\mathbf{X}$. The left-hand side of Figure 11.3 shows the same setup, but imposes randomization, which makes the assignments $\mathbf{W}$ independent of everything, which is depicted by it sitting by itself in the corner.

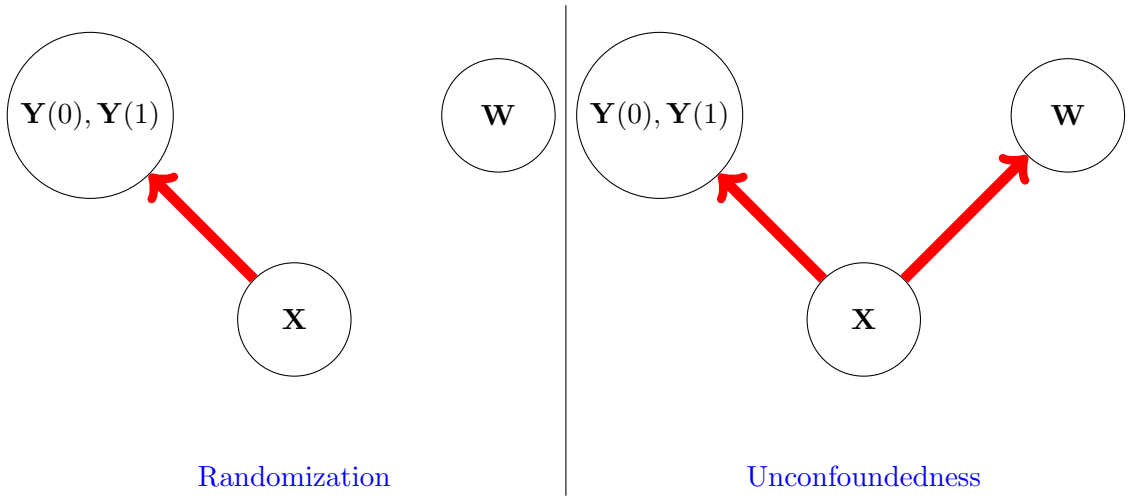


Figure 11.3: Graphical representation of two assignment mechanisms: randomization and unconfoundedness.

Section 2.5 in the Stat 110 book shows that conditional independence does not imply independence, nor does independence imply conditional independence, and warns that:

> *It is easy to make terrible blunders stemming from confusing independence and conditional independence.*

Example 11.7.5. Suppose that X, A, B, C are random variables which are all independent and

$$Y_1(1) = I(X + A > 0), \quad Y_1(0) = I(X + B > 0), \quad W_1 = I(X + C > 0).$$

Then $[W_1 \perp\!\!\!\perp \{Y_1(0), Y_1(1)\}]|X_1$, but W_1 and $\{Y_1(0), Y_1(1)\}$ are not independent as they share a common X_1.

Again, there will be both finite sample and population based approaches to causal inference, as outlined in Table 11.4.

Random: $\mathbf{W}, \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{Y}$ Data: $\mathbf{W}, \mathbf{X}, \mathbf{Y}$ Unconfounded: $[\mathbf{W} \perp\!\!\!\perp \{\mathbf{Y}(0), \mathbf{Y}(1)\}] \vert \mathbf{X}$		
Statistical strategy	Estimand	Inferential framework
Finite sample condition on $\mathbf{Y}(0) = \mathbf{y}(0), \mathbf{Y}(1) = \mathbf{y}(1)$	$\overline{\tau} = \frac{1}{n} \sum_{j=1}^{n} \{y_j(1) - y_j(0)\}$	$\mathbf{W} \vert \{\mathbf{Y}(0) = \mathbf{y}(0), \mathbf{Y}(1) = \mathbf{y}(1)\}, \mathbf{X}$
Population unconditional	$\mathrm{E}[\tau_1] = \mathrm{E}[Y_1(1)] - \mathrm{E}[Y_1(0)]$	Unconditional or $\mathbf{Y} \vert (\mathbf{W}, \mathbf{X})$

Table 11.4: Statistical strategies, estimands, and inferential frameworks for observational studies.

Definition 11.7.6 (Propensity score)**.** The *propensity score* $\lambda(x)$ is defined by

$$\lambda(x) = P(W_j = 1 | X_j = x) = \mathrm{E}[W_j | X_j = x].$$

So the propensity score is a binary predictive regression, giving the conditional probability of being in the treatment group. We will assume that $0 < \lambda(x) < 1$ for every x; this assumption is called *overlap*. If the unconfoundedness and overlap assumptions hold, the treatment assignments are said to be *strongly ignorable* given $\mathbf{X}$.

Assume unconfoundedness and that the $\mathbf{W}_j$ are conditionally independent given $\mathbf{X}$. Then

$$\begin{aligned} & P(\mathbf{W} = \mathbf{w} | \mathbf{Y}(\mathbf{0}), \mathbf{Y}(\mathbf{1}), \mathbf{X} = \mathbf{x}) \\ & = P(\mathbf{W} = \mathbf{w} | \mathbf{X} = \mathbf{x}), \quad \text{unconfounded} \\ & = \prod_{j=1}^{n} P(W_j = w_j | X_j = x_j) = \prod_{j=1}^{n} \lambda(x_j)^{w_j} \{1 - \lambda(x_j)\}^{1 - w_j}. \end{aligned}$$

☣ **11.7.7.** In randomized experiments, we typically know $P(W_j = 1)$ as we did the randomization (or the experimenter will tell us how they did it). In observational studies, the propensity score $\lambda(x)$ is typically not known, so we face a double challenge: unconfoundedness has to be justified by the researcher (and it cannot be established statistically or empirically) and $\lambda(x)$ has to be learned from the data.

For randomized experiments, the MLE and MoM estimators relied on the randomization assumption to drive two sets of probabilistic calculations:

- population estimation:

$$P(Y_1 = 1 | W = 1) = P(Y_1(1) = 1),$$

$$P(Y_1 = 0|W = 0) = P(Y_1(0) = 0).$$

- finite sample estimation:

$$\mathrm{E}[W_1 Y_1 | \{Y_1(1) = y_1(1), Y_1(0) = y_1(0)\}] = \mathrm{E}[W_1] y_1(1),$$

$$\mathrm{E}[(1 - W_1) Y_1 | \{Y_1(1) = y_1(1), Y_1(0) = y_1(0)\}] = \mathrm{E}[1 - W_1] y_1(0).$$

Neither of these results holds once randomization is dropped because conditional independence of $\mathbf{W}$ and $\{\mathbf{Y}(0), \mathbf{Y}(1)\}$ given $\mathbf{X}$ does not imply independence of $\mathbf{W}$ and $\{\mathbf{Y}(0), \mathbf{Y}(1)\}$. Hence both the population-based estimator and the finite sample estimator may well be biased and even inconsistent. We need some new thinking to recover from dropping the randomization assumption.

11.7.4 A population-based statistical model for observational studies

Consider again a population-based approach, with (W_1, Y_1, X_1), ..., (W_n, Y_n, X_n) i.i.d. By Adam's law,

$$\mathrm{E}[\tau_1] = \mathrm{E}[\mathrm{E}[\tau_1 | X]] = \mathrm{E}[\theta_1(X)] - \mathrm{E}[\theta_0(X)],$$

where

$$\theta_1(x) = \mathrm{E}[Y_1 | W_1 = 1, X_1 = x] = \mathrm{E}[Y_1(1) | W_1 = 1, X_1 = x], \quad \text{by unconfoundedness,}$$

and

$$\theta_0(x) = \mathrm{E}[Y_1 | W_1 = 0, X_1 = x] = \mathrm{E}[Y_1(0) | W_1 = 0, X_1 = x].$$

As an integral, Adam's law says that

$$\mathrm{E}[\tau_1] = \int_{-\infty}^{\infty} \{\theta_1(x) - \theta_0(x)\} f_X(x) \mathrm{d}x,$$

averaging over the random covariate X, which here is assumed to have the density function f_X.

To simplify the exposition, we will assume that outcomes are binary. Then

$$\theta_1(x) = P(Y_1 = 1 | W_1 = 1, X_1 = x)$$

and

$$\theta_0(x) = P(Y_1 = 1 | W_1 = 0, X_1 = x)$$

are binary prediction problems.

Likelihood-based inference for population estimand

Now let's go back to predictive regression and condition on the covariates:

$$P(Y_1 = y_1, W_1 = w_1 | X_1 = x) = P(Y_1 = y_1 | W_1 = w_1, X_1 = x) P(W_1 = w_1 | X_1 = x).$$

With assumptions and notation as in the previous subsection, the conditional probability mass function $P(Y_1 = y_1 | W_1 = w_1, X_1 = x)$ is determined by $\theta_0(x)$ and $\theta_1(x)$. Here we will estimate these binary predictive regressions, by parameterizing them as $\theta_0(x|\boldsymbol{\psi}_0), \theta_1(x|\boldsymbol{\psi}_1)$ and using likelihood methods to learn $\boldsymbol{\psi}_0$ and $\boldsymbol{\psi}_1$ from the data $(w_1, y_1, x_1), \ldots, (w_n, y_n, x_n)$.

Then the conditional log-likelihood for the observed outcomes given the assignments is

$$\begin{aligned}
\log L(\boldsymbol{\psi}_0, \boldsymbol{\psi}_1) &= \log P(\mathbf{Y} = \mathbf{y} | \mathbf{W} = \mathbf{w}, \mathbf{X} = \mathbf{x}, \boldsymbol{\psi}_0, \boldsymbol{\psi}_1) \\
&= \sum_{j=1}^{n} \log P(Y_j = y_j | W_j = w_j, X_j = x_j, \boldsymbol{\psi}_0, \boldsymbol{\psi}_1) \\
&= \sum_{j=1}^{n} (1 - w_j) \log P(Y_j = y_j | W_j = 0, X_j = x_j, \boldsymbol{\psi}_0, \boldsymbol{\psi}_1) \\
&+ \sum_{j=1}^{n} w_j \log P(Y_j = y_j | W_j = 1, X_j = x_j, \boldsymbol{\psi}_0, \boldsymbol{\psi}_1) \quad \text{(since } W_j \text{ is binary)} \\
&= \log L_0(\boldsymbol{\psi}_0) + \log L_1(\boldsymbol{\psi}_1),
\end{aligned}$$

where

$$\log L_0(\boldsymbol{\psi}_0) = \sum_{j=1}^{n} (1 - w_j) \Big(y_j \log\{\theta_0(x|\boldsymbol{\psi}_0)\} + (1 - y_j) \log\{1 - \theta_0(x|\boldsymbol{\psi}_0)\} \Big)$$

$$\log L_1(\boldsymbol{\psi}_1) = \sum_{j=1}^{n} w_j \Big(y_j \log\{\theta_1(x|\boldsymbol{\psi}_1)\} + (1 - y_j) \log\{1 - \theta_1(x|\boldsymbol{\psi}_1)\} \Big).$$

Then

$$\hat{\boldsymbol{\psi}}_0 = \underset{\boldsymbol{\psi}_0}{\text{argmax}} \ \log L_0(\boldsymbol{\psi}_0), \quad \hat{\boldsymbol{\psi}}_1 = \underset{\boldsymbol{\psi}_1}{\text{argmax}} \ \log L_1(\boldsymbol{\psi}_1)$$

which are the MLEs of the parameters which index binary predictive regressions on the subset of the data which is in the control group and the subset of the data which is in the treatment group, respectively. Chapter 6 reports the details of the properties of the maximum likelihood estimator of the binary predictive regression in the logistic case.

Then we estimate the population average treatment effect through

$$\widetilde{\mathrm{E}[\tau_1]} = \int \{\theta_1(x|\hat{\boldsymbol{\psi}}_1) - \theta_0(x|\hat{\boldsymbol{\psi}}_0)\} f_X(x) \mathrm{d}x,$$

where f_X is the density function of the covariates. Given the data on the covariates $x_1, \ldots, x_n$, a simple estimator in practice replaces the expectation with respect to X with the sample average

$$\widehat{\mathrm{E}[\tau_1]} = \frac{1}{n}\sum_{j=1}^{n}\{\theta_1(x_j|\hat{\boldsymbol{\psi}}_1) - \theta_0(x_j|\hat{\boldsymbol{\psi}}_0)\}.$$

11.7.5 Method of moments estimator of finite sample estimand

In the RCT case, we conditioned on the potential outcomes and developed a method of moments estimator of $\bar{\tau}$. To extend to the observational case, we additionally condition on the covariates. Recalling that $W_1Y_1 = W_1Y_1(1)$ and using the unconfoundedness assumption,

$$\mathrm{E}[W_1Y_1|Y_1(0) = y_1(0), Y_1(1) = y_1(1), X_1 = x_1] = y_1(1)\mathrm{E}[W_1|X_1 = x_1] = y_1(1)\lambda(x_1),$$

so

$$\mathrm{E}\left[\frac{W_1Y_1}{\lambda(X_1)}|Y_1(0) = y_1(0), Y_1(1) = y_1(1), X_1 = x_1\right] = y_1(1).$$

Similarly, using $(1 - W_1)Y_1 = (1 - W_1)Y_1(0)$ and unconfoundedness gives

$$\mathrm{E}\left[\frac{(1 - W_1)Y_1}{1 - \lambda(X_1)}|Y_1(0) = y_1(0), Y_1(1) = y_1(1), X_1 = x_1\right] = y_1(0).$$

If $\lambda(x)$ is known, we have the following MoM estimator for $\bar{\tau}$:

$$\hat{\tau}_{\mathrm{MoM}}(\mathbf{W}) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{W_jY_j}{\lambda(x_j)} - \frac{(1 - W_j)Y_j}{1 - \lambda(x_j)}\right).$$

Note that $\hat{\tau}_{\mathrm{MoM}}(\mathbf{W})$ weights terms by the inverse of the probability of assignment (this is a version of the Horvitz-Thompson estimator, which is discussed in Chapter 10 in the sampling context). It has the conditional unbiasedness property

$$\mathrm{E}[\hat{\tau}_{\mathrm{MoM}}(\mathbf{W})|\mathbf{Y}(0) = \mathbf{y}(0), \mathbf{Y}(1) = \mathbf{y}(1), \mathbf{X} = \mathbf{x}] = \frac{1}{n}\sum_{j=1}^{n}\{y_j(1) - y_j(0)\} = \bar{\tau},$$

elegantly mimicking the RCT version. The conditional variance is

$$\mathrm{Var}(\hat{\tau}_{\mathrm{MoM}}(\mathbf{W})|\mathbf{Y}(\mathbf{0}) = \mathbf{y}(0), \mathbf{Y}(1), \mathbf{X} = \mathbf{x}) = \frac{1}{n^2}\sum_{j=1}^{n}\sigma^2_{\tau,j},$$

where

$$\sigma^2_{\tau,j} = \frac{y_j^2(1)}{\lambda(x_j)} + \frac{y_j^2(0)}{1 - \lambda(x_j)} - \{y_j(1) - y_j(0)\}^2.$$

In the vast majority of observational studies $\lambda(x_j)$ is unknown, in which case it is usually parameterized, say as $\lambda(x_j|\boldsymbol{\psi})$. The method of moments estimator becomes

$$\tilde{\tau}_{\mathrm{MoM}}(\mathbf{W};\boldsymbol{\psi}) = \frac{1}{n}\sum_{j=1}^{n}\left(\frac{W_jY_j}{\lambda(x_j|\boldsymbol{\psi})} - \frac{(1 - W_j)Y_j}{1 - \lambda(x_j|\boldsymbol{\psi})}\right),$$

and then $\boldsymbol{\psi}$ is estimated, e.g., using maximum likelihood. To use maximum likelihood, let

$$\hat{\boldsymbol{\psi}}(\mathbf{W}) = \underset{\boldsymbol{\psi}}{\operatorname{argmax}} \log L(\boldsymbol{\psi}; \mathbf{W}), \text{ where } \log L(\boldsymbol{\psi}; \mathbf{W}) = \log P(\mathbf{W} = \mathbf{w} | \mathbf{X} = \mathbf{x}, \boldsymbol{\psi}).$$

Using independence over units,

$$\log L(\boldsymbol{\psi}; \mathbf{W}) = \sum_{j=1}^{n} W_j \log \lambda(x_j|\boldsymbol{\psi}) + (1 - W_j) \log\{1 - \lambda(x_j|\boldsymbol{\psi})\}.$$

We can then obtain the average treatment effect estimator

$$\tilde{\tau}_{\text{MoM}}(\mathbf{W}; \hat{\boldsymbol{\psi}}(\mathbf{W})).$$

Overall, these calculations are remarkable, as we can directly draw a line from the RCTs to observational studies, with very similar results. However, plugging in the estimator $\lambda(x_j|\hat{\boldsymbol{\psi}}(\mathbf{W}))$ means that $\tilde{\tau}_{\text{MoM}}(\mathbf{W}; \hat{\boldsymbol{\psi}}(\mathbf{W}))$ affects the bias and variance.

11.8 Recap

Causality is about how moving from control to treatment impacts outcomes. We formalized finite sample and population-based treatment effects through the potential outcomes framework, and introduced the crucial role of randomization. As always in statistics, it is important to think carefully about what the *estimand* is and what to *condition* on.

The main ideas and notation for the chapter are listed in Table 11.5.

In the population case, we focused on a maximum likelihood approach for binary outcomes, based upon a model which follows immediately from the structure of the problem. In the finite sample case, we focused on a method of moments approach. For hypothesis testing in the finite sample case, we saw that under the Fisher null hypothesis testing can be elegantly implemented without any need for asymptotics.

Social scientists and public health researchers often wish to tackle important research problems using observational studies. Sometimes natural experiments allow a secondary use of randomization to carry out causal inference. More common is the use of conditioning on covariates and then extending the idea of randomized assignments, if an unconfoundedness assumption can be justified. This conditioning argument again allows inference on both the finite sample treatment effects and the population-based treatment effects.

11.9 R, experiments, and causality

We have already covered in earlier chapters the computational techniques needed for this chapter, so this section just contains the code for the empirical illustrations in this chapter.

Code and output for Examples 11.4.3, 11.5.4 and 11.6.5

```
# data

n = 32; W = numeric(n); Y = numeric(n)
W[1:17] = 0; W[18:n] = 1
Y[1:14] = 0; Y[15:17]= 1
Y[18:20]= 0; Y[21:n] = 1
table(Y,W)

   W
Y    0  1
  0 14  3
  1  3 12

#------------
# MLE
theta1 = sum(W*Y)/sum(W)
se1    = sqrt(theta1*(1.0-theta1)/sum(W))
theta0 = sum((1-W)*Y)/sum(1-W)
se0    = sqrt(theta0*(1.0-theta0)/sum(1-W))

print(c(theta1,se1,theta0,se0),digits=3)
[1] 0.8000 0.1033 0.1765 0.0925

tauhat    = theta1-theta0
se.tau  = sqrt(se1^2 + se0^2)

print(c(tauhat,se.tau,tau/se.tau),digits=3)
[1] 0.624 0.139 4.498

#----------------
# Bayesian inference
set.seed(111); B=10^6

theta1 = rbeta(B,1.0+sum(W*Y),1.0+sum(W*(1.0-Y)))
theta0 = rbeta(B,1.0+sum((1.0-W)*Y),1.0+sum((1.0-W)*(1.0-Y)))
tau = theta1 - theta0

print(c(mean(tau),quantile(tau,probs=c(0.025,0.975))))
[1] 0.5541442 0.2642224 0.7894752

hist(rbeta(B,1,1)-rbeta(B,1,1),breaks=100,xlab="tau",
```

```
                    main="Simulation from prior on E[tau]")
hist(tau,breaks=100,xlab="tau",main="Simulation from
                    posterior on E[tau]")


#-----------------
#Fisher null p-value
set.seed(111);  B=10^4
tauMoM = mean(W*Y/0.5) - mean((1.0-W)*Y/0.5)

Wstar   = replicate(B,rbinom(n,1,0.5));  # randomize assignments
tauStar = apply(Wstar*Y/0.5,2,mean) - apply((1.0-Wstar)*Y/0.5,2,mean)

pdf("mag.pdf")
  hist(tauStar,breaks=100,xlab="tau")
  abline(v=tauMoM,col="red",lwd=1)
dev.off()
print(mean(abs(tauStar) >= abs(tauMoM)))
[1] 0.0387
```

Formula or idea	Description or name
causality	move assignment, impact on outcome
patient, individual, or unit	could be person, company, webpage, etc.
W	assignment
$Y(0), Y(1)$	potential outcomes
non-interference	outcome of jth unit only caused by jth assignment
$Y = Y(W)$	outcome
$\tau = Y(1) - Y(0)$	treatment effect (individual)
$\bar{\tau} = n^{-1} \sum_{j=1}^{n} \tau_j$	average treatment effect (finite sample)
$\mathrm{E}[\tau_1]$	treatment effect (population)
ethics of experiments	no intended harm, informed consent, safety strategy
randomized control trial (RCT)	$\mathbf{W} \perp\!\!\!\perp \mathbf{Y}(0), \mathbf{Y}(1)$ — randomized assignment
inference based on $\mathbf{W}, \mathbf{Y} \vert \mathbf{Y}(0), \mathbf{Y}(1)$	finite sample approach, estimand $\bar{\tau}$
$\mathrm{E}\left(\frac{W_1 Y_1}{E[W_1]} - \frac{(1-W_1)Y_1}{\mathrm{E}[1-W_1]} \vert Y_1(0) = y_1(0), Y_1(1) = y_1(1)\right)$ $= y_1(1) - y_1(0)$	method of moments
inference based on $\mathbf{Y} \vert \mathbf{W}$	population approach, estimand $\mathrm{E}[\tau_1]$
$\mathrm{E}[\tau_1] = \mathrm{E}[Y_1 \vert W_1 = 1] - \mathrm{E}[Y_1 \vert W_1 = 0]$	estimate directly
Fisher null or $\mathrm{E}[\tau_1] = 0$	hypothesis of no causal effect
randomization test	test of no causal effect
observational studies	
natural experiments	randomized by nature rather than researcher
$(\mathbf{W} \perp\!\!\!\perp \mathbf{Y}(0), \mathbf{Y}(1)) \vert \mathbf{X}$	unconfoundedness
$\lambda_j = P(W_j = 1 \vert X_j = x_j)$	propensity score
inference based on $\mathbf{W}, \mathbf{Y} \vert \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}$	finite sample approach, estimand $\bar{\tau}$
$\mathrm{E}\left(\frac{W_1 Y_1}{\lambda_1} - \frac{(1-W_1)Y_1}{1-\lambda_1} \vert Y_1(0) = y_1(0), Y_1(1) = y_1(1), X_1\right)$ $= y_1(1) - y_1(0)$	method of moments
inference based on $\mathbf{Y} \vert \mathbf{W}, \mathbf{X}$	population approach, estimand $\mathrm{E}[\tau_1]$
$\mathrm{E}[\tau_1 \vert X_1] = \mathrm{E}[Y_1 \vert W_1 = 1, X_1] - \mathrm{E}[Y_1 \vert W_1 = 0, X_1]$	estimate directly, then average out X

Table 11.5: Main ideas and notation in Chapter 11.

Bibliography

Atkinson, A. B. (2017). Pareto and the upper tail of the income distribution in the UK: 1799 to the present. *Economica 84*, 129–156.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.

Cameron, C. A. and P. K. Trivedi (2005). *Microeconometrics*. Cambridge: Cambridge University Press.

Christensen, R., W. Johnson, A. Branscum, and T. E. Hanson (2011). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*. Chapman & Hall/CRC Press.

Davison, A. C. (2003). *Statistical Models*. Cambridge: Cambridge University Press.

De Luca, G., J. R. Magnus, and F. Peracchi (2021). Posterior moments and quantiles for the normal location problem with laplace prior. *Communications in Statistics - Theory and Methods 50*, 4039–4049.

Diez, D. M., M. Cetinkaya-Rundel, and C. D. Barr (2020). *OpenIntro Statistics* (4 ed.). LeanPub.

Efron, B. (2012). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press.

Efron, B. (2013). A 250-year argument: Belief, behavior, and the bootstrap. *Bulletin of the American Mathematical Society 50*, 129–146.

Efron, B. and C. Morris (1977). Stein's paradox in statistics. *Scientific Americian 236*, 119–127.

Kahan, B. C., T. P. Morris, I. R. White, J. Carpenter, and S. Cro (2021). Estimands in published protocols of randomised trials: urgent improvement needed. *Trials 22*, 686.

Koenker, R. (2005). *Quantile Regression*. Cambridge: Cambridge University Press.

Mitchell, A. F. S. (1994). A note on posterior moments for a normal mean with double-exponential prior. *Journal of the Royal Statistical Society, Series B 56*, 605–610.

Peixoto, J. L. (1990). A property of well-formulated polynomial regression models. *American Statistician 44*, 26–30.

Peracchi, F. and A. F. M. Smith (1992). Exact and approximate posterior moments for a normal location parameter. *Journal of the Royal Statistical Society, Series B 54*, 793–804.

Wilkinson, L. (1999). *The Grammar of Graphics*. Springer.