Lecture 0: Course Introduction

We begin the course with an overview of the ideas and themes to be discussed this quarter.

Probability and Statistics

It's helpful to contrast *statistics* and *statistical inference* with *probability*.

- Probability: The distribution of the data is given to us, and we want to calculate something (e.g., probabilities) related to possible outcomes/events.
- Statistics (statistical inference): The distribution of the data is unknown, and we want to infer properties of the distribution (e.g., the value of some parameter, like the mean) from the observed data.

The mathematical framework of probability is what allows us to do statistical inference.

Parametric Modeling

A common approach to statistical inference begins by assuming the data comes from some reasonably simple model with one or more unknown parameters.

- Such models are typically approximations or simplifications of reality.
- They require assumptions about the data.

EXAMPLE 0.0.1: Suppose we model a collection of quantitative data $X_1, \ldots, X_n$ as being independent and identically distributed from a normal distribution with a mean μ and a variance σ^2. This requires us to make several assumptions:

- The observations are independent.
- The observations all have the same distribution.
- The distribution is normal.

The unknown parameters of interest here might be μ, σ^2, or both. $\diamond$

There are several different types of parametric inference that we could consider about an unknown parameter θ:

- We could estimate θ with a specific value.
- We could make a binary decision between two hypotheses about the value of θ.
- We could produce some set of values (like an interval) that we think contains θ.

We'll talk about each of these this quarter.

Frequentist and Bayesian Paradigms

There are two schools of thought about how probabilities can and can't be interpreted:

- Everyone agrees that probabilities can represent long-term frequencies of random events.
- The *frequentist* approach considers this to be the *only* meaningful interpretation of probability.
- The *Bayesian* approach believes that probabilities can also represent an individual's (possibly subjective) degree of belief about something unknown.

This difference in opinion leads to different interpretations of statistical inference:

- For frequentists, an unknown parameter is *fixed* (even though it's unknown).
- For Bayesians, an unknown parameter is a random variable that can be discussed in probabilistic terms.

EXAMPLE 0.0.2: Suppose θ is an unknown parameter, and we're interested in whether or not $\theta > 0$. Under the frequentist approach, we would say that θ either is or isn't greater than zero, and we simply don't know which. Hence, ideas such as "the probability that $\theta > 0$ based on the data" don't really make sense for frequentists. However, such probabilities are indeed considered meaningful under the Bayesian approach. ♢

Most classical statistical methods are frequentist. However, Bayesian methods have grown in popularity in recent decades. We'll discuss both this quarter.

Course Overview

We'll begin the quarter by discussing key concepts from probability, much of which may be review from your previous courses. Then we'll use these concepts to cover the three kinds of statistical inference we mentioned earlier.

Major Themes

There are several major themes that will come up throughout the course:

- We want to develop general approaches for solving statistical problems, as opposed to just a "grab bag" of methods for certain specific situations.
- We want to evaluate and compare different procedures to figure out if (or when) one is better than another.
- We'll often be interested in the *asymptotic* behavior of various statistical procedures, i.e., the behavior in the limit as the sample size tends to infinity. We'll see that such asymptotic behavior often obeys fairly general rules.

Lecture 1: Basic Probability

This lecture covers concepts from probability that will be needed later.

1.1 Random Variables and Distributions

For our purposes, random variables will be one of two types: discrete or continuous. (Also see the note immediately after Example 1.1.3.)

Discrete Random Variables

A random variable X is *discrete* if its set of possible values $\mathcal{X}$ is finite or countably infinite.

- The *probability mass function* (pmf) of a discrete random variable X is the nonnegative function $p(x) = P(X = x)$, where x denotes each possible value that X can take. It is always true that $\sum_{x \in \mathcal{X}} p(x) = 1$.
- The *cumulative distribution function* (cdf) of a random variable X is $F(x) = P(X \leq x)$. If X is discrete, then $F(x) = \sum_{\{t \in \mathcal{X}:\, t \leq x\}} p(t)$, and so the cdf consists of constant sections separated by jump discontinuities.

Specific types of discrete random variables include binomial, geometric, Poisson, and discrete uniform random variables.

Example 1.1.1: A fair coin is flipped three times. Let X be the total number of heads. The set of possible values of X is $\mathcal{X} = \{0, 1, 2, 3\}$, a finite set, so X is discrete. Its pmf is $p(0) = p(3) = 1/8$, $p(1) = p(2) = 3/8$. We can also recognize X as a binomial random variable. ♢

Example 1.1.2: A fair coin is flipped repeatedly until it comes up heads. Let X be the total number of flips needed to obtain heads. The set of possible values of X is $\mathcal{X} = \{1, 2, 3, \ldots\}$, a countably infinite set, so X is discrete. Its pmf is $p(x) = 2^{-x}$ for every $x \in \{1, 2, 3, \ldots\}$. We can also recognize X as a geometric random variable. ♢

Continuous Random Variables

A random variable X is *continuous* if its possible values form an uncountable set (e.g., some interval on $\mathbb{R}$) and the probability that X equals any such value *exactly* is zero.

- The *probability density function* (pdf) of a continuous random variable X is a nonnegative function $f(x)$ such that $\int_a^b f(x)\, dx = P(a \leq X \leq b)$ for any $a, b \in \mathbb{R}$. It is always true that $\int_{-\infty}^{\infty} f(t)\, dt = 1$.
- Again, the cdf of a random variable X is $F(x) = P(X \leq x)$. If X is continuous, then $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt$, and so the cdf is a continuous function.
- Note that the pdf can be obtained by differentiating the cdf.

Specific types of continuous random variables include normal, exponential, beta, gamma, chi-squared, Student's t, and continuous uniform random variables.

EXAMPLE 1.1.3: Let X be the amount of time in hours that an electrical component functions before breaking down. This random variable might have the pdf

$$f(x) = \lambda \exp(-\lambda x)\, I_{[0,\infty)}(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases}$$

which we recognize as an exponential distribution. The probability that the part functions for at least c hours is $P(X \geq 100) = \int_c^\infty f(x)\, dx = \exp(-c\lambda)$. ◊

> Note: The cdf is a more general description of a random variable than the pmf or pdf, since it has a single definition that applies for both discrete and continuous random variables. In fact, there is no difficulty in writing down the cdf of a "mixed" random variable that is neither wholly discrete nor wholly continuous. Such a cdf would simply include both jump discontinuities and regions where it is continuously increasing. However, such "mixed" random variables have neither a pmf nor a pdf in the senses considered here.

Clarification of Notation

We may sometimes need to clarify our notation for pmfs or pdfs in two ways:

- When dealing with more than one random variable, we may need to explicitly denote the random variable to which a pmf or pdf corresponds. If so, we will write $p^{(X)}(x)$ or $f^{(X)}(x)$ for the pmf or pdf (respectively) of X evaluated at x.
- A pmf or pdf often depends on one or more parameters. We may need to explicitly indicate the value of a parameter at which the pmf or pdf is calculated. If so, we will write $p_\theta(x)$ or $f_\theta(x)$ for the pmf or pdf evaluated at the parameter value θ.

Of course, we may write $p_\theta^{(X)}(x)$ or $f_\theta^{(X)}(x)$ when both types of clarification are needed.

Transformations of Random Variables

Let X be a random variable, and let $Y = g(X)$, where g is some strictly increasing function. The cdf of Y can be easily obtained from the cdf of X:

$$F^{(Y)}(y) = P(Y \leq y) = P[g(X) \leq y] = P\big[X \leq g^{-1}(y)\big] = F^{(X)}\big[g^{-1}(y)\big],$$

where g^{-1} denotes the inverse function of g.

- If X is discrete, then the pmf of Y is

$$p^{(Y)}(y) = P(Y = y) = P[g(X) = y] = \sum_{\{x \in \mathcal{X}:\, g(x) = y\}} P(X = x) = \sum_{\{x \in \mathcal{X}:\, g(x) = y\}} p^{(X)}(x).$$

- If X is continuous, then the pdf of Y can be obtained by differentiating the cdf of Y:

$$f^{(Y)}(y) = \frac{d}{dy} F^{(Y)}(y) = \Big\{\big[F^{(X)}\big]'\big[g^{-1}(y)\big]\Big\}\Big\{\big(g^{-1}\big)'(y)\Big\} = \frac{f^{(X)}[g^{-1}(y)]}{g'[g^{-1}(y)]}.$$

Multiple Random Variables

Let X and Y be discrete random variables that take values in $\mathcal{X}$ and $\mathcal{Y}$, respectively.

- The *joint* pmf of X and Y is $p^{(X,Y)}(x,y) = P(X = x, Y = y)$.

  Note: When we write $p(x,y)$ without clarification, we mean the joint pmf $p^{(X,Y)}(x,y)$.

- The *marginal* pmfs of X and Y are, respectively,

$$p^{(X)}(x) = P(X = x) = \sum_{y \in \mathcal{Y}} P(X = x, Y = y) = \sum_{y \in \mathcal{Y}} p^{(X,Y)}(x,y),$$
$$p^{(Y)}(y) = P(Y = y) = \sum_{x \in \mathcal{X}} P(X = x, Y = y) = \sum_{x \in \mathcal{X}} p^{(X,Y)}(x,y).$$

- The *conditional* pmfs of X given Y and of Y given X are, respectively,

$$p^{(X|Y)}(x \mid y) = P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p^{(X,Y)}(x,y)}{p^{(Y)}(y)},$$
$$p^{(Y|X)}(y \mid x) = P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{p^{(X,Y)}(x,y)}{p^{(X)}(x)}.$$

Now let X and Y be continuous random variables instead.

- The *joint* pdf of X and Y is a nonnegative function $f^{(X,Y)}(x,y)$ such that

$$\iint_A f^{(X,Y)}(x,y)\, dx\, dy = P[(X,Y) \in A] \qquad \text{for any set } A \subset \mathbb{R}^2 \ .$$

  Note: When we write $f(x,y)$ without clarification, we mean the joint pdf $f^{(X,Y)}(x,y)$.

- The *marginal* pdfs of X and Y are, respectively,

$$f^{(X)}(x) = \int_{-\infty}^{\infty} f^{(X,Y)}(x,y)\, dy, \qquad f^{(Y)}(y) = \int_{-\infty}^{\infty} f^{(X,Y)}(x,y)\, dx.$$

- The *conditional* pdfs of X given Y and of Y given X are, respectively,

$$f^{(X|Y)}(x \mid y) = \frac{f^{(X,Y)}(x,y)}{f^{(Y)}(y)}, \qquad f^{(Y|X)}(y \mid x) = \frac{f^{(X,Y)}(x,y)}{f^{(X)}(x)}.$$

  Note: It may seem intuitively reasonable to think of $f^{(X|Y)}(x \mid y)$ as "the pdf of X given that $Y = y$," so that $\int_A f^{(X|Y)}(x \mid y)\, dx = P(X \in A \mid Y = y)$ for any set $A \subset \mathbb{R}$. However, this is not technically correct. The quantity $P(X \in A \mid Y = y)$ cannot even be properly defined using our definition of conditional probability:

$$P(X \in A \mid Y = y) = \frac{P(X \in A, Y = y)}{P(Y = y)} = \frac{0}{0}$$

  since Y is a continuous random variable. See the note on page 146 of DeGroot and Schervish for additional explanation of what conditional pdfs actually represent.

Indepdenence

Random variables X and Y are called *independent* if $P(X \in A,\ Y \in B) = P(X \in A)\ P(Y \in B)$ for all sets A and B. Random variables X and Y are independent if and only if their joint pmf or pdf factorizes into their marginal pmfs or pdfs, i.e., $p^{(X,Y)}(x,y) = p^{(X)}(x)\ p^{(Y)}(y)$ or $f^{(X,Y)}(x,y) = f^{(X)}(x)\ f^{(Y)}(y)$ for all x and y.

1.2 Expectation and Related Concepts

There are several quantities that we can calculate to summarize random variables.

Expectation

For our purposes, the *expectation* or *expected value* $E(X)$ of a random variable X is defined as $E(X) = \sum_{x \in \mathcal{X}} x\, p(x)$ if X is discrete and $E(X) = \int_{-\infty}^{\infty} x\, f(x)\ dx$ if X is continuous. Similarly, the expectation of a function $g(X)$ of a random variable X can be computed as $E[g(X)] = \sum_{x \in \mathcal{X}} g(x)\, p(x)$ or $E[g(X)] = \int_{-\infty}^{\infty} g(x)\, f(x)\ dx$.

> Note: If you're familiar with more advanced notions of integration, you might recognize that a more general formula for all cases above is $E[g(X)] = \int_{-\infty}^{\infty} g(x)\ dF(x)$.

It is possible for an expectation to be ∞ or $-\infty$, or to fail to exist altogether.

> Note: The formal way to compute $E[g(X)]$ is to first compute the sum/integral separately over the positive and negative values of $g(X)$ and then add the results together. If the positive part yields ∞ and the negative part yields $-\infty$, then the final "answer" is $\infty - \infty$, which is undefined.

Example 1.2.1: Let X have a t distribution with one degree of freedom (also known as a Cauchy distribution), which has the pdf $f(x) = [\pi(1+x^2)]^{-1}$ for all $x \in \mathbb{R}$. Since the pdf is symmetric about zero, it might seem as though $E(X)$ should be zero. However, this is false:

$$E(X) = \int_{-\infty}^{\infty} \frac{x}{\pi(1+x^2)}\ dx = \underbrace{\int_{0}^{\infty} \frac{x}{\pi(1+x^2)}\ dx}_{\text{positive part}} + \underbrace{\int_{-\infty}^{0} \frac{x}{\pi(1+x^2)}\ dx}_{\text{negative part}} = \infty - \infty.$$

Thus, $E(X)$ does not exist. ♢

In the presence of multiple random variables, the sum or integral should be taken over the joint pdf, i.e., $E[g(X,Y)] = \iint_{\mathbb{R}^2} g(x,y)\, f(x,y)\ dx\ dy$ or $E[g(X,Y)] = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} g(x,y)\, p(x,y)$. For functions of X only or Y only, the sum or integral may equivalently be taken over the corresponding marginal pmf or pdf.

Clarification of Notation

The pmf or pdf of X often depends on one or more parameters. In general, the value of $E[g(X)]$ may also depend on these same parameters. To explicitly indicate this, we will write $E_\theta[g(X)]$ for the expectation of $g(X)$ computed with a parameter value θ.

> Note: You may have seen people write things like "$E_X[g(X)]$" for what we've called $E[g(X)]$ or $E_\theta[g(X)]$. The "$E_X[g(X)]$" notation is problematic for multiple reasons, and we will not use such notation in this course.

Variance

The *variance* $\text{Var}(X)$ of a random variable X is defined as $\text{Var}(X) = E\{[X - E(X)]^2\}$. An equivalent (an typically easier) formula is $\text{Var}(X) = E(X^2) - [E(X)]^2$. Similarly, the variance of a function $g(X)$ of a random variable X is $\text{Var}[g(X)] = E\{[g(X)]^2\} - \{E[g(X)]\}^2$.

> Note: If $E[g(X)]$ is infinite or does not exist, then $\text{Var}[g(X)]$ does not exist either. If $E[g(X)]$ is finite, then $\text{Var}[g(X)]$ is guaranteed to exist, although it may be ∞. It can be shown that if $E\{[g(X)]^2\}$ is finite, then $E[g(X)]$ is finite as well. Thus, if $\text{Var}[g(X)] < \infty$, then $E[g(X)]$ exists and is finite.

We will write $\text{Var}_\theta[g(X)]$ when necessary to explicitly indicate the dependence of the variance on a parameter value θ.

Covariance

The *covariance* $\text{Cov}(X, Y)$ of a random variable X and a random variable Y is defined as $\text{Cov}(X, Y) = E\{[X - E(X)][Y - E(Y)]\}$. An equivalent (and typically easier) formula is $\text{Cov}(X, Y) = E(XY) - E(X)\,E(Y)$. Similarly, the covariance of $g(X)$ and $h(Y)$ is $\text{Cov}[g(X), h(Y)] = E[g(X)\,h(Y)] - E[g(X)]\,E[h(Y)]$.

> Note: If either $E[g(X)]$ or $E[h(Y)]$ does not exist, then $\text{Cov}[g(X), h(Y)]$ does not exist either. However, if $\text{Var}[g(X)] < \infty$ and $\text{Var}[h(Y)] < \infty$, then $\text{Cov}[g(X), h(Y)]$ is guaranteed to exist and be finite (see the Cauchy-Schwarz inequality below).

The variance is simply a special case of the covariance:

$$\begin{aligned}\text{Var}[g(X)] = E\{[g(X)]^2\} - \{E[g(X)]\}^2 &= E[g(X)\,g(X)] - E[g(X)]\,E[g(X)] \\ &= \text{Cov}[g(X), g(X)].\end{aligned}$$

We will write $\text{Cov}_\theta[g(X), h(Y)]$ when necessary to explicitly indicate the dependence of the covariance on a parameter value θ.

Working with Expectations, Variances, and Covariances

Two key properties of expectation are as follows:

- If a and b are constants, then $E[a + b\,g(X)] = a + b\,E[g(X)]$.
- If X and Y are independent, then $E[g(X)\,h(Y)] = E[g(X)]\,E[h(Y)]$.

These properties can be used to derive other useful identities involving expectations, variances, and covariances. Suppose that a and b are constants.

- $\text{Var}[a + b\,g(X)] = b^2\,\text{Var}[g(X)]$.
- $\text{Cov}[a + b\,g(X),\ h(Y)] = \text{Cov}[g(X),\ a + b\,h(Y)] = b\,\text{Cov}[g(X), h(Y)]$.
- If X and Y are independent, then $\text{Cov}[g(X), h(Y)] = 0$. (The converse is false.)
- If X and Y are independent, then $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. A consequence of this result is that $\text{Var}(n^{-1} \sum_{i=1}^n X_i) = n^{-1}\,\text{Var}(X_1)$ if $X_1, \ldots, X_n$ are iid (independent and identically distributed).

Cauchy-Schwarz Inequality

If $\text{Var}[g(X)] < \infty$ and $\text{Var}[h(Y)] < \infty$, then $\text{Cov}[g(X), h(Y)]$ exists and is finite, and

$$\Big|\text{Cov}[g(X), h(Y)]\Big| \leq \sqrt{\text{Var}[g(X)]\ \text{Var}[h(Y)]},$$

with equality if and only if $g(X) = a + b\,h(Y)$ with probability 1 for some constants a and b.

> Note: The Cauchy-Schwarz inequality is actually a much more general result than just what is stated above.

Conditional Expectation and Variance

The expectation $E[g(X)]$ is computed using the pmf or pdf of X. We may also wish to consider the expectation of $g(X)$ conditional on the value of some other random variable Y. We call this the *conditional expectation* of $g(X)$ given $Y = y$ and compute it using the conditional pmf or pdf of X given $Y = y$:

$$E[g(X) \mid Y = y] = \sum_{x \in \mathcal{X}} g(x)\, p^{(X|Y)}(x \mid y) \quad \text{or} \quad E[g(X) \mid Y = y] = \int_{-\infty}^{\infty} g(x)\, f^{(X|Y)}(x \mid y)\, dx.$$

Notice that computing $E[g(X) \mid Y = y]$ yields a function of y, not a random variable. However, we can consider plugging the random variable Y into this function, which *does* yield a random variable. This random variable is what we mean when we write $E[g(X) \mid Y]$.

> Note: A formal treatment of conditional expectation is a bit more complicated than this, but the explanation above is good enough for our purposes.

Similarly, the *conditional variance* of $g(X)$ is

$$\text{Var}[g(X) \mid Y = y] = E\{[g(X)]^2 \mid Y = y\} - \{E[g(X) \mid Y = y]\}^2.$$

Again, we might consider either $\text{Var}[g(X) \mid Y = y]$, which is a function of y, or $\text{Var}[g(X) \mid Y]$, which is this same function evaluated at Y (yielding a random variable).

Iterated Expectation and Variance Formulas

Sometimes the expectation or variance of one random variable may be easier to compute when conditioned on another random variable. The marginal (unconditional) expectations and variances can be computed using the following formulas:

- $E[g(X)] = E\{E[g(X) \mid Y]\}$.
- $\text{Var}[g(X)] = E\{\text{Var}[g(X) \mid Y]\} + \text{Var}\{E[g(X) \mid Y]\}$.

These results are also sometimes called the *law of total expectation* and *law of total variance*, respectively.

EXAMPLE 1.2.2: Let X be the number of heads in Y independent flips of a fair coin, and let Y have a discrete uniform distribution on $\{1, \ldots, 5\}$. Then $p^{(X|Y)}(x \mid y)$ is the pmf of a $\text{Bin}(y, 1/2)$ distribution, so $E(X \mid Y = y) = y/2$ and $\text{Var}(X \mid Y = y) = y/4$ by standard results that we can look up. Also, $E(Y) = 3$ and $\text{Var}(Y) = 2$, again by standard results. Then

$$\begin{aligned} E(X) &= E[E(X \mid Y)] = E(Y/2) = E(Y)/2 = 3/2, \\ \text{Var}(X) &= E[\text{Var}(X \mid Y)] + \text{Var}[E(X \mid Y)] \\ &= E(Y/4) + \text{Var}(Y/2) = E(Y)/4 + \text{Var}(Y)/4 = 3/4 + 2/4 = 5/4. \end{aligned}$$

We could have obtained the same results without using these formulas, but the calculations would have been considerably more tedious. ♢

1.3 Convergence Concepts

We now consider the limiting behavior of sequences of random variables. In particular, we consider concepts and results related to *convergence* of such sequences.

Convergence of Real Numbers

Recall that a sequence of real numbers $\{a_n : n \geq 1\}$ *converges* to a ($a_n \to a$ as $n \to \infty$) if for every $\varepsilon > 0$, there exists $N \geq 1$ such that $|a_n - a| \leq \varepsilon$ for every $n \geq N$.

Convergence of Random Variables

For our purposes, there are two main notions of convergence for random variables. Let $\{X_n : n \geq 1\}$ be a sequence of random variables, and X be another random variable. Let F_n denote the cdf of X_n, and let F denote the cdf of X.

- We say that $\{X_n : n \geq 1\}$ *converges in probability* to X (written $X_n \to_P X$ as $n \to \infty$) if for every $\varepsilon > 0$, $P(|X_n - X| > \varepsilon) \to 0$.
- We say that $\{X_n : n \geq 1\}$ *converges in distribution* to X (written $X_n \to_D X$ as $n \to \infty$) if $F_n(x) \to F(x)$ at every point x where F is continuous.

Note that convergence in distribution is defined by convergence of cdfs, rather than the values of the actual random variables. For this reason, it is sometimes simply written as $F_n \to_D F$ or $X_n \to_D F$. We may also replace F with its "common name," e.g., $X_n \to_D N(0, 1)$.

Note: Convergence in distribution is also called *convergence in law* or *weak convergence*.

Theorem 1.3.1. *If $X_n \to_P X$, then $X_n \to_D X$.*

Thus, convergence in probability is stronger than convergence in distribution. However, in the case where the limiting random variable X is actually a constant, they are equivalent, as formalized in the following theorem.

Theorem 1.3.2. *Let $a \in \mathbb{R}$ be a constant. Then $X_n \to_P a$ if and only if $X_n \to_D a$.*

The next theorems provide additional useful results about convergence of random variables.

Theorem 1.3.3 (Continuous Mapping Theorem)**.** *Let $X_n \to_P a$ for some constant $a \in \mathbb{R}$, and let $g : \mathbb{R} \to \mathbb{R}$ be continuous at a. Then $g(X_n) \to_P g(a)$.*

Theorem 1.3.4 (Slutsky's Theorem)**.** *If $X_n \to_D X$, $Y_n \to_P b$, and $Z_n \to_P a$, where $a, b \in \mathbb{R}$ are constants, then $X_n Y_n + Z_n \to_D bX + a$.*

Weak Law of Large Numbers and Central Limit Theorem

We now state two extremely important asymptotic results: the weak law of large numbers and the central limit theorem.

Theorem 1.3.5 (Weak Law of Large Numbers, or WLLN)**.** *Let $\{X_n : n \geq 1\}$ be a sequence of iid random variables with $E(|X_1|) < \infty$. Let $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$. Then $\overline{X}_n \to_P E(X_1)$ as $n \to \infty$.*

Thus, the WLLN formalizes the intuitive notion that the expectation of a random variable may be interpreted as its long-run average.

> Note: Yes, there also exists a *strong* law of large numbers, which is similar but deals with a stronger form of convergence called *almost sure convergence* or *convergence with probability* 1. In more sophisticated versions of these theorems, the iid assumption can be relaxed much more for the weak law than for the strong law.

Theorem 1.3.6 (Central Limit Theorem, or CLT)**.** *Let $\{X_n : n \geq 1\}$ be a sequence of iid random variables with $\mathrm{Var}(X_1) = \sigma^2 < \infty$ and $E(X_1) = \mu$. Let $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$. Then $\sqrt{n}\big(\overline{X}_n - \mu\big) \to_D N(0, \sigma^2)$ as $n \to \infty$.*

Informally, the central limit theorem states that for large n, $\overline{X}_n$ is approximately normal with mean μ and variance σ^2/n. Notice that the WLLN and CLT yield different results because the CLT scales the quantity $\overline{X}_n - \mu$ by an extra factor of $\sqrt{n}$:

- The WLLN says that $\overline{X}_n - \mu \to_P 0$.
- The CLT says that $\sqrt{n}\big(\overline{X}_n - \mu\big) \to_D N(0, \sigma^2)$.

Delta Method

Let $\{Y_n : n \geq 1\}$ be a sequence of random variables such that $\sqrt{n}(Y_n - a) \to_D Z$ for some random variable Z and some constant $a \in \mathbb{R}$. Let $g : \mathbb{R} \to \mathbb{R}$ be a function. What can we say about the asymptotic behavior of $g(Y_n)$?

- First, note that since $1/\sqrt{n} \to 0$,

  $$Y_n - a = \big(1/\sqrt{n}\big)\big[\sqrt{n}(Y_n - a)\big] \to_D 0 \cdot Z = 0$$

  by Slutsky's theorem. Thus, $Y_n \to_D a$, and hence $Y_n \to_P a$.
- If g is continuous at a, then $g(Y_n) \to_P g(a)$ by continuous mapping theorem.

However, we can do better than this. Suppose g is differentiable at a, so that we may write $g(Y_n) - g(a) \approx g'(a)\,(Y_n - a)$ (a first-order Taylor expansion). Then by Slutsky's theorem,

$$\sqrt{n}[g(Y_n) - g(a)] \approx g'(a)\sqrt{n}(Y_n - a) \to_D g'(a)\,Z.$$

This is the basic idea of a technique called the *delta method*.

Theorem 1.3.7 (Delta Method)**.** *Let $\{Y_n : n \geq 1\}$ be a sequence of random variables such that $\sqrt{n}(Y_n - a) \to_D Z$ for some random variable Z and some constant $a \in \mathbb{R}$. Let $g : \mathbb{R} \to \mathbb{R}$ be continuously differentiable at a. Then $\sqrt{n}[g(Y_n) - g(a)] \to_D g'(a)\,Z$.*

Proof. Formally, $\sqrt{n}[g(Y_n) - g(a)] = g'(Y_n^\star)\sqrt{n}(Y_n - a)$ for some $Y_n^\star$ between Y_n and a. Note that for any $\varepsilon > 0$, $P(|Y_n^\star - a| > \varepsilon) \leq P(|Y_n - a| > \varepsilon)$, and $P(|Y_n - a| > \varepsilon) \to 0$ since $Y_n \to_P a$. Then $Y_n^\star \to_P a$, so $g'(Y_n^\star) \to_P g'(a)$ by the continuous mapping theorem. Since $\sqrt{n}(Y_n - a) \to_D Z$, the result follows by Slutsky's theorem. □

The following special case is by far the most common use of the delta method.

Corollary 1.3.8 (Delta Method, Normal Case)**.** *Let $\{Y_n : n \geq 1\}$ be a sequence of random variables such that $\sqrt{n}(Y_n - a) \to_D N(0, \tau^2)$ for some constants $a \in \mathbb{R}$ and $\tau^2 > 0$. Let $g : \mathbb{R} \to \mathbb{R}$ be continuously differentiable at a. Then $\sqrt{n}[g(Y_n) - g(a)] \to_D N(0,\, \tau^2[g'(a)]^2)$.*

Proof. Take $Z \sim N(0, \tau^2)$ in Theorem 1.3.7. □

Example 1.3.9: Suppose $X_1, X_2, \ldots$ are iid from the continuous uniform distribution on $[0, 60]$, and we want to find the asymptotic distribution of $(\overline{X}_n)^{-1}$. We have $E(X_1) = 30$ and $\mathrm{Var}(X_1) = 300$, so $\sqrt{n}(\overline{X}_n - 30) \to_D N(0, 300)$ by the CLT. Our function is $g(t) = t^{-1}$, and $g(30) = 1/30$. Its derivative is $g'(t) = -t^{-2}$, which is continuous at 30, and $g'(30) = -1/900$. Then by the Delta Method,

$$\sqrt{n}\left[\left(\overline{X}_n\right)^{-1} - \frac{1}{30}\right] \to_D N(0,\, 1/2700),$$

noting that $300(-1/900)^2 = 1/2700$. Thus, for large n, $(\overline{X}_n)^{-1}$ is approximately normal, with mean $1/30$ and variance $1/(2700\,n)$. ♢

Lecture 2: Basic Statistical Concepts

A *statistic* is any random variable that is calculated as a function of the data. In this lecture, we cover some basic statistical concepts that should be addressed before we proceed further.

- When sampling from a normal distribution, the sample mean $\overline{X}$ and the sample variance S^2 are two particularly important statistics. We will discuss the distribution of these statistics and another related statistic.
- Some statistics possess a property called *sufficiency*. We will define this property and explain its importance.
- Many common distributions belong to a group called the *exponential family*. We will define this group, provide examples of distributions that do and do not belong to it, and discuss its relevance to statistical inference.

2.1 Sampling from Normal Distributions

Let $X_1, \ldots, X_n \sim \text{iid}\, N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. A wide variety of statistical procedures are based on this simple setup, so it is important to study it in detail.

Expectation, Variance, and Covariance for Random Vectors

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random vector. Then $E(\boldsymbol{X})$ denotes a vector of length p with ith component $E(X_i)$, and $\text{Var}(\boldsymbol{X})$ denotes a $p \times p$ matrix with (i, j)th element $\text{Cov}(X_i, X_j)$.

> Note: $\text{Var}(\boldsymbol{X})$ is typically called the variance-covariance matrix of $\boldsymbol{X}$ since its ith diagonal element is $\text{Var}(X_i)$.

The various properties of univariate expectations and variances have multivariate extensions. Suppose that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are random vectors of length p. Also suppose that $\boldsymbol{a} \in \mathbb{R}^p$ and $\boldsymbol{B}$ is a $p \times p$ matrix.

- $E(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}) = \boldsymbol{a} + \boldsymbol{B}\, E(\boldsymbol{X})$, and $E(\boldsymbol{X} + \boldsymbol{Y}) = E(\boldsymbol{X}) + E(\boldsymbol{Y})$.
- $\text{Var}(\boldsymbol{a} + \boldsymbol{B}\boldsymbol{X}) = \boldsymbol{B}\, \text{Var}(\boldsymbol{X}) \boldsymbol{B}^T$.

Multivariate Normal Distribution

Let $\boldsymbol{Z}$ be a random vector, with $\boldsymbol{\theta} = E(\boldsymbol{Z})$ and $\boldsymbol{V} = \text{Var}(\boldsymbol{Z})$. The distribution of $\boldsymbol{Z}$ is called *multivariate normal*, which we write as $\boldsymbol{Z} \sim N_p(\boldsymbol{\theta}, \boldsymbol{V})$, if and only if $\boldsymbol{a}^T \boldsymbol{Z}$ has a (univariate) normal distribution for all $\boldsymbol{a} \in \mathbb{R}^p$. The following properties hold for $\boldsymbol{Z} \sim N_p(\boldsymbol{\theta}, \boldsymbol{V})$:

- If $\boldsymbol{V}$ is nonsingular, then the elements $Z_1, \ldots, Z_p$ of $\boldsymbol{Z}$ have joint pdf

$$f(z_1, \ldots, z_p) = \frac{1}{(2\pi)^{p/2}\, \det \boldsymbol{V}^{1/2}}\, \exp\left[-\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\theta})^T \boldsymbol{V}^{-1} (\boldsymbol{z} - \boldsymbol{\theta})\right].$$

- Z_i and Z_j are independent if and only if $V_{ij} = \text{Cov}(Z_i, Z_j) = 0$.

Let $\mathbf{0}_p$ denote a zeros vector of length p, and let $\boldsymbol{I}_p$ denote the $p \times p$ identity matrix. The distribution $N_p(\mathbf{0}_p, \boldsymbol{I}_p)$ is called the *p-variate standard normal* distribution and has a useful property stated in the following lemma.

Lemma 2.1.1. *Let $\boldsymbol{A}$ be a $p \times p$ matrix that is orthogonal ($\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}_p$), and let $\boldsymbol{Z} \sim N_p(\mathbf{0}_p, \boldsymbol{I}_p)$. Then $\boldsymbol{A}\boldsymbol{Z} \sim N_p(\mathbf{0}_p, \boldsymbol{I}_p)$.*

Proof. For any vector $\boldsymbol{b} \in \mathbb{R}^p$, the random vector $\boldsymbol{b}^T\boldsymbol{A}\boldsymbol{Z} = (\boldsymbol{A}^T\boldsymbol{b})^T\boldsymbol{Z}$ has a (univariate) normal distribution since $\boldsymbol{Z}$ is multivariate normal. Then $\boldsymbol{A}\boldsymbol{Z}$ is multivariate normal. Now simply note that $E(\boldsymbol{A}\boldsymbol{Z}) = \boldsymbol{A}\,E(\boldsymbol{Z}) = \mathbf{0}_p$ and $\mathrm{Var}(\boldsymbol{A}\boldsymbol{Z}) = \boldsymbol{A}\boldsymbol{I}_p\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}_p$. □

Chi-Squared Distribution

Let $\boldsymbol{Z} \sim N_p(\mathbf{0}_p, \boldsymbol{I}_p)$. The distribution of $\boldsymbol{Z}^T\boldsymbol{Z} = \sum_{i=1}^p Z_i^2$ is called a *chi-squared distribution with p degrees of freedom*, which we write as χ^2_p.

Lemma 2.1.2. *The χ^2_p distribution is the* Gamma$(p/2, 1/2)$ *distribution.*

Proof. First, note that the pdf of each Z_i^2 is

$$f^{(Z_i^2)}(u) = \frac{2\,f^{(Z_i)}\big(\sqrt{u}\big)}{2\sqrt{u}} = \frac{1}{\sqrt{2\pi\,u}}\,\exp\!\left(-\frac{u}{2}\right) = \frac{(1/2)^{1/2}}{\Gamma(1/2)}\,u^{-1/2}\exp\!\left(-\frac{u}{2}\right),$$

for $u > 0$ and zero otherwise, which is the pdf of a Gamma$(1/2, 1/2)$ distribution. Then since $Z_1^2, \ldots, Z_p^2 \sim \text{iid}\,\text{Gamma}(1/2, 1/2)$, their sum is $\sum_{i=1}^p Z_i^2 \sim \text{Gamma}(p/2, 1/2)$. (This result for the gamma distribution is stated and proven as Theorem 5.7.7 of DeGroot & Schervish.) □

The χ^2_p distribution has expectation p and variance $2p$.

Joint Distribution of the Sample Mean and Sample Variance

Let $X_1, \ldots, X_n \sim \text{iid}\,N(\mu, \sigma^2)$. Two commonly calculated summary statistics are

$$\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^n \big(X_i - \overline{X}\big)^2 = \frac{1}{n-1}\left[\sum_{i=1}^n X_i^2 - \big(\overline{X}\big)^2\right], \tag{2.1.1}$$

called (respectively) the *sample mean* and *sample variance*. By basic results on the normal distribution, it is clear that $\overline{X} \sim N(\mu, \sigma^2/n)$. The distribution of S^2, as well as the joint distribution of $\overline{X}$ and S^2, is provided by the following theorem.

Theorem 2.1.3. *Let $X_1, \ldots, X_n \sim \text{iid}\,N(\mu, \sigma^2)$, and let $\overline{X}$ and S^2 be defined as in (2.1.1). Then $\overline{X} \sim N(\mu, \sigma^2/n)$, and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Moreover, $\overline{X}$ and S^2 are independent.*

Proof. It suffices to prove the result for $\mu = 0$ and $\sigma^2 = 1$. Let $\boldsymbol{X} = (X_1, \ldots, X_n) \sim N_n(\mathbf{0}_n, \boldsymbol{I}_n)$. Now let $\boldsymbol{A}$ be an orthogonal $p \times p$ matrix for which all elements in the first row are $n^{-1/2}$. (Such a matrix can always be constructed, e.g., by the Gram-Schmidt process.) Then let $\boldsymbol{Y} = (Y_1, \ldots, Y_n) = \boldsymbol{A}\boldsymbol{X}$. Observe that $\boldsymbol{Y} \sim N(\mathbf{0}_n, \boldsymbol{I}_n)$ by Lemma 2.1.1, so the sum of the

squares of its last $n-1$ elements is $\sum_{i=2}^n Y_i^2 \sim \chi^2_{n-1}$. Now note that the first element is $Y_1 = n^{1/2}\overline{X}$, so we may write

$$\sum_{i=2}^n Y_i^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 = \boldsymbol{Y}^T\boldsymbol{Y} - Y_1^2 = \boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{X} - n\big(\overline{X}\big)^2 = \boldsymbol{X}^T\boldsymbol{X} - n\big(\overline{X}\big)^2 = \sum_{i=1}^n X_i^2 - n\big(\overline{X}\big)^2 = (n-1)S^2.$$

Finally, note that $Y_1, \ldots, Y_n$ are all independent, so Y_1 and and $\sum_{i=2}^n Y_i^2$ are independent. □

Without the normality assumption, some parts of Theorem 2.1.3 still hold, but others do not. Suppose $X_1, \ldots, X_n$ are iid with $E(X_1) = \mu$ and $\mathrm{Var}(X_1) = \sigma^2$, but suppose their distribution is not necessarily normal.

- We still have $E(\overline{X}) = \mu$ and $\mathrm{Var}(\overline{X}) = \sigma^2/n$. Also, we still have $E(S^2) = \sigma^2$, which agrees with Theorem 2.1.3.
- However, the distribution of $\overline{X}$ is not necessarily normal (though it is approximately normal for large n by the CLT), and the distribution of $(n-1)S^2/\sigma^2$ is not necessarily chi-squared. Moreover, $\overline{X}$ and S^2 are not necessarily independent.

Student's t Distribution

Let $Z \sim N(0,1)$ and $U \sim \chi^2_p$ be independent random variables. The distribution of the random variable

$$\frac{Z}{\sqrt{U/p}}$$

is called *Student's t distribution with p degrees of freedom*, which we write as t_p.

Lemma 2.1.4. *The pdf of the t_p distribution is, for all $t \in \mathbb{R}$,*

$$f(t) = \frac{\Gamma[(p+1)/2]}{\sqrt{\pi p}\,\Gamma(p/2)}\left(1 + \frac{t^2}{p}\right)^{-(p+1)/2}.$$

Proof. See pages 483–484 of DeGroot & Schervish. □

Various statistical procedures (many of which we will see later this quarter) involve the random variable

$$T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}} \tag{2.1.2}$$

Its distribution is of considerable importance and is given by the following theorem.

Theorem 2.1.5. *Let $X_1, \ldots, X_n \sim iid\, N(\mu, \sigma^2)$, and let T be defined as in (2.1.1) and (2.1.2). Then $T \sim t_{n-1}$.*

Proof. Let $Z = (\overline{X} - \mu)/\sqrt{\sigma^2/n}$ and $U = (n-1)S^2/\sigma^2$, By Theorem 2.1.3, Z and U are independent with $Z \sim N(0,1)$ and $U \sim \chi^2_{n-1}$. The result follows since $T = Z/\sqrt{U/(n-1)}$. □

2.2 Sufficient Statistics

Suppose we have a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ from some distribution with an unknown parameter θ. It is often the case that the "information" about θ that is contained in the sample $\boldsymbol{X}$ can be entirely summarized by some statistic $\boldsymbol{Y} = r(\boldsymbol{X})$. This idea is called *sufficiency*.

Definition of Sufficiency

A statistic $\boldsymbol{Y} = r(\boldsymbol{X})$ is said to be *sufficient* for an unknown parameter θ if and only if the conditional distribution of the sample $\boldsymbol{X}$ given the value of $\boldsymbol{Y}$ does not depend on θ.

EXAMPLE 2.2.1: Let $X_1, \ldots, X_n \sim \text{iid}\,\text{Bin}(1, \theta)$, and let $Y = \sum_{i=1}^n X_i$. Then the conditional pmf of $\boldsymbol{X} = (X_1, \ldots, X_n)$ given Y is

$$p_\theta^{(\boldsymbol{X}|Y)}(\boldsymbol{x} \mid y) = \frac{P_\theta(X_1 = x_1, \ldots, X_n = x_n, Y = y)}{P_\theta(Y = y)}$$

$$= \frac{P_\theta(X_1 = x_1, \ldots, X_n = x_n)}{P_\theta(\sum_{i=1}^n X_i = y)} = \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{y}\theta^y(1-\theta)^{n-y}} = \frac{\theta^y(1-\theta)^{n-y}}{\binom{n}{y}\theta^y(1-\theta)^{n-y}} = \frac{1}{\binom{n}{y}}.$$

if $y = \sum_{i=1}^n x_i$ (and zero if $y \neq \sum_{i=1}^n x_i$). This does not depend on θ, so Y is sufficient for θ. ◊

Factorization Theorem

The following result usually provides an easier way to show sufficiency of a statistic. It is also useful for identifying the sufficient statistic in the first place.

Theorem 2.2.2 (Factorization Theorem)**.** *Let θ be an unknown parameter, and let $\boldsymbol{X}$ be a sample with joint pmf $p_\theta(\boldsymbol{x})$ (if $\boldsymbol{X}$ is discrete) or joint pdf $f_\theta(\boldsymbol{x})$ (if $\boldsymbol{X}$ is continuous). A statistic $r(\boldsymbol{X})$ is sufficient for θ if and only if there exist functions $g(u, \theta)$ and $h(\boldsymbol{x})$ such that $p_\theta(\boldsymbol{x}) = g[r(\boldsymbol{x}), \theta]\, h(\boldsymbol{x})$ (if $\boldsymbol{X}$ is discrete) or $f_\theta(\boldsymbol{x}) = g[r(\boldsymbol{x}), \theta]\, h(\boldsymbol{x})$ (if $\boldsymbol{X}$ is continuous).*

Proof. See the proof of Theorem 7.7.1 of DeGroot & Schervish for a proof of the discrete case. The proof of the continuous case is beyond the scope of this course. □

The following result is in some sense already implied by the factorization theorem, but we state it separately for clarity.

Theorem 2.2.3. *Let $r(\boldsymbol{X})$ be sufficient for θ, and let q be an injective (one-to-one) function. Then $q[r(\boldsymbol{X})]$ is sufficient for θ.*

Proof. Let $g(u, \theta)$ and $h(\boldsymbol{x})$ be the functions that exist by the factorization theorem as applied to $r(\boldsymbol{X})$. Without loss of generality, let the codomain of q be its range, so that q is bijective and hence has an inverse function q^{-1}. Now simply apply the factorization theorem to $q[r(\boldsymbol{x})]$ with $\tilde{g}(u, \theta) = g[q^{-1}(u), \theta]$. □

EXAMPLE 2.2.4: For the situation of Example 2.2.1, we could simply note that the joint pmf of the sample $\boldsymbol{X}$ is $p_\theta(\boldsymbol{x}) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^y(1-\theta)^{n-y}$, where $y = \sum_{i=1}^n x_i$. Then $Y = \sum_{i=1}^n X_i$ is sufficient for θ by the factorization theorem with $g(y,\theta) = \theta^y(1-\theta)^{n-y}$ and $h(\boldsymbol{x}) = 1$. Note that $n^{-1}Y$ is also sufficient for θ by Theorem 2.2.3. ◊

Now consider the case of a sample from a normal distribution. Here the sufficient statistic varies according to which parameters are unknown.

EXAMPLE 2.2.5: Let $X_1, \ldots, X_n \sim \text{iid}\, N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ is unknown but $\sigma^2 > 0$ is known. The joint pdf of the sample is

$$\begin{aligned} f_\mu(\boldsymbol{x}) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right] \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + \mu^2\right)\right] \\ &= \underbrace{\exp\left[-\frac{1}{2\sigma^2}\left(-2\mu\sum_{i=1}^n x_i + \mu^2\right)\right]}_{g(y,\mu)} \underbrace{\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2\right)}_{h(\boldsymbol{x})}, \end{aligned}$$

where $y = \sum_{i=1}^n x_i$. Thus, $\sum_{i=1}^n X_i$ is sufficient for μ by the factorization theorem. Note that $\overline{X}$ is also sufficient for θ by Theorem 2.2.3. ◊

EXAMPLE 2.2.6: Let $X_1, \ldots, X_n \sim \text{iid}\, N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are both unknown. The joint pdf of the sample is the same as in Example 2.2.5, but we now factor it differently:

$$f_\mu(\boldsymbol{x}) = \underbrace{\frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + \mu^2\right)\right]}_{g[(y_1,y_2),(\mu,\sigma^2)]},$$

where $y_1 = \sum_{i=1}^n x_i$ and $y_2 = \sum_{i=1}^n x_i^2$, and where $h(\boldsymbol{x}) = 1$. Thus, $(\sum_{i=1}^n X_i,\ \sum_{i=1}^n X_i^2)$ is sufficient for (μ, σ^2) by the factorization theorem. Note that $(\overline{X}, S^2)$ is also sufficient for θ by Theorem 2.2.3. ◊

Minimal Sufficiency

Note from the factorization theorem that if $r(\boldsymbol{X})$ is sufficient for θ, then $[r(\boldsymbol{X}), s(\boldsymbol{X})]$ is also a sufficient for θ. However, since $r(\boldsymbol{X})$ alone is sufficient for θ, the combined statistic $[r(\boldsymbol{X})], s(\boldsymbol{X})]$ fails to reduce the data as much as possible.

EXAMPLE 2.2.7: In Example 2.2.5, we saw that if $X_1, \ldots, X_n \sim \text{iid}\, N(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ unknown and $\sigma^2 > 0$ known, then $\overline{X}$ is sufficient for μ. However, $(\overline{X}, S^2)$ is also sufficient for μ, though it does not reduce the data as much as possible. ◊

Formally, a statistic is called *minimal sufficient* for θ if it is a function of every other sufficient statistic for θ. This may be restated as follows: If $r(\boldsymbol{X})$ is sufficient for θ but there exists a non-injective (non–one-to-one) function q such that $q[r(\boldsymbol{X})]$ is also sufficient for θ, then $r(\boldsymbol{X})$ is not minimal sufficient.

EXAMPLE 2.2.8: In Examples 2.2.5 and Example 2.2.7, we saw that $\overline{X}$ and $(\overline{X}, S^2)$ were both sufficient for μ. Note that $\overline{X}$ is a function of $(\overline{X}, S^2)$, but $(\overline{X}, S^2)$ is not a function of $\overline{X}$. Thus, $(\overline{X}, S^2)$ is not a minimal sufficient statistic for μ. ◊

You may notice that in Example 2.2.8, we did not actually demonstrate that $\overline{X}$ was a minimal sufficient statistic for μ, only that $(\overline{X}, S^2)$ was not minimal sufficient for μ. There exist methods to prove minimal sufficiency, but it is usually clear by inspection that a sufficient statistic is or is not minimal sufficient.

> Note: Since there exists a concept of *sufficient* statistics, it is not surprising to learn that there also exists a concept of *necessary* statistics. A statistic is said to be *necessary* for θ if it is a function of every sufficient statistic for θ. It can be seen from this definition that a minimal sufficient statistic for θ is simply a statistic that is both necessary and sufficient for θ.

Sufficiency Principle

A common principle in statistical inference is that samples that yield the same sufficient statistic value should yield the same inference about θ. Mathematically, if $r(\boldsymbol{X})$ is sufficient for θ and $r(\boldsymbol{x}_1) = r(\boldsymbol{x}_2)$, then the inference about θ should be the same regardless of whether we observe $\boldsymbol{X} = \boldsymbol{x}_1$ or $\boldsymbol{X} = \boldsymbol{x}_2$. This is called the *sufficiency principle*. Most statistical procedures that we will see this quarter obey the sufficiency principle.

2.3 Exponential Family

Many commonly used distributions can be written in a single general form.

Definition of the Exponential Family

A distribution with unknown parameter θ belongs to the *exponential family* if its pmf $p_\theta(x)$ or pdf $f_\theta(x)$ can be written as

$$\exp\left[\sum_{j=1}^{k} \eta_j(\theta)\, r_j(x) - \psi(\theta)\right] h(x) \tag{2.3.1}$$

for some $k \geq 1$ and some functions $\eta_1, \ldots, \eta_k, r_1, \ldots, r_k, \psi, h$.

EXAMPLE 2.3.1: Consider the $N(\mu, \sigma^2)$ distribution with $\mu \in \mathbb{R}$ unknown and $\sigma^2 > 0$ known. The pdf of this distribution is

$$f_\mu(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \exp\left(\frac{\mu}{2\sigma^2}\, x - \frac{\mu^2}{\sigma^2}\right) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right).$$

Then

$$k = 1, \qquad \eta_1(\mu) = \frac{\mu}{\sigma^2}, \qquad r_1(x) = x, \qquad \psi(\mu) - \frac{\mu^2}{2\sigma^2}, \qquad h(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right),$$

and hence this distribution belongs to the exponential family. ◊

EXAMPLE 2.3.2: Now consider the $N(\mu, \sigma^2)$ distribution with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ both unknown. The pdf is the same as in Example 2.3.1, but we now break it up as

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \exp\left[\frac{\mu}{\sigma^2}\, x - \frac{1}{2\sigma^2}\, x^2 - \left(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\sigma^2\right)\right] \frac{1}{\sqrt{2\pi}}.$$

Then

$$k = 2, \qquad \eta_1(\mu, \sigma^2) = \frac{\mu}{\sigma^2}, \qquad r_1(x) = x, \qquad \eta_2(\mu, \sigma^2) = -\frac{1}{2\sigma^2}, \qquad r_2(x) = x^2,$$
$$\psi(\mu, \sigma^2) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log\sigma^2, \qquad h(x) = \frac{1}{\sqrt{2\pi}},$$

and hence this distribution belongs to the exponential family. ◊

Members and Non-Members of the Exponential Family

Many common distributions belong to the exponential family, including

- the binomial distribution where the success probability θ is unknown,
- the Poisson distribution where the rate parameter λ is unknown,
- the negative binomial distribution where the success/failure probability θ is unknown (which includes the geometric distribution as a special case),
- the normal distribution where the mean μ and variance σ^2 are unknown,
- the gamma distribution where the shape parameter α and rate parameter β are unknown (which includes the exponential distribution as a special case), and
- the beta distribution where the shape parameters α and β are unknown.

However, some common distributions do *not* belong to the exponential family, such as

- the discrete uniform distribution,
- the hypergeometric distribution,
- the continuous uniform distribution,
- Student's t distribution, and
- Snedecor's F distribution.

The following rule automatically disqualifies some distributions from the exponential family: If the *support* of the distribution (the set of values where the pmf or pdf is nonzero) depends on an unknown parameter, then the distribution is *not* in the exponential family. (Note that this automatically disqualifies the discrete and continuous uniform distributions.)

Importance of the Exponential Family

When we have a random sample from a distribution in the exponential family, many aspects of statistical inference become easier.

- Many theorems later in the course will involve regularity conditions that are difficult to check. However, such regularity conditions are often automatically satisfied if the sample consists of iid observations from a distribution in the exponential family.
- Some calculations and results take simple, standard forms for distributions in the exponential family. An example of such a result is given below.

Sufficiency and the Exponential Family

An iid sample from a distribution in the exponential family can always be reduced to a relatively small number of sufficient statistics by the following theorem.

Theorem 2.3.3. *Let $X_1, \ldots, X_n$ be iid observations from a distribution in the exponential family with pmf or pdf as stated in (2.3.1). Then $[\sum_{i=1}^n r_1(X_i), \ldots, \sum_{i=1}^n r_k(X_i)]$ is a sufficient statistic for θ.*

Proof. The joint pmf or pdf is of the sample is

$$\prod_{i=1}^{n} \exp\left[\sum_{j=1}^{k} \eta_j(\theta)\, r_j(x_i) - \psi(\theta)\right] h(x_i) = \exp\left\{\sum_{j=1}^{k}\left[\eta_j(\theta) \sum_{i=1}^{n} r_j(x_i)\right] - n\,\psi(\theta)\right\} \prod_{i=1}^{n} h(x_i),$$

and we now simply apply the factorization theorem. □

Example 2.3.4: Let $X_1, \ldots, X_n \sim \text{iid}\, N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ is unknown but $\sigma^2 > 0$ is known. By applying Theorem 2.3.3 to the pdf as written in Example 2.3.1, it can be seen that $\sum_{i=1}^n r_1(X_i) = \sum_{i=1}^n X_i$ is sufficient for μ. This result agrees with Example 2.2.5. ♢

Example 2.3.5: Let $X_1, \ldots, X_n \sim \text{iid}\, N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are both unknown. By applying Theorem 2.3.3 to the pdf as written in Example 2.3.2, it can be seen that $[\sum_{i=1}^n r_1(X_i),\ \sum_{i=1}^n r_2(X_i)] = (\sum_{i=1}^n X_i,\ \sum_{i=1}^n X_i^2)$ is sufficient for (μ, σ^2). This result agrees with Example 2.2.6. ♢

Lecture 3: Frequentist Estimation

Suppose we have an unknown parameter θ and some data $\boldsymbol{X}$. We may want to use the data to *estimate* the value of θ.

- An *estimator* $\hat{\theta}$ of an unknown parameter θ is any function of the data that is intended to approximate θ in some sense. Although we typically just write $\hat{\theta}$, it is actually $\hat{\theta}(\boldsymbol{X})$, a random variable.
- An *estimate* is the value an estimator takes for a particular set of data values. Thus, the estimator $\hat{\theta}(\boldsymbol{X})$ would yield the estimate $\hat{\theta}(\boldsymbol{x})$ if we observe the data $\boldsymbol{X} = \boldsymbol{x}$.

Good and Bad Estimators

Any function of the data can be considered an estimator for any parameter. However, it may not be a *good* estimator. A good estimator will "usually" be "close" to the parameter it estimates, in a sense to be formalized later. It may or may not be obvious whether an estimator is good, or whether one estimator is better than another.

EXAMPLE 3.0.1: Suppose $X_1, \ldots, X_n \sim$ iid Exp(λ), and we wish to estimate $\lambda = 1/E(X)$.

- Estimators such as $\hat{\lambda} = \left(\overline{X}\right)^{-1}$ or $\hat{\lambda} = (\text{sample median})^{-1}$ might be good estimators.
- Estimators such as $\hat{\lambda} = 1 + \left(\overline{X}\right)^3$ or $\hat{\lambda} = 17$ are probably bad estimators.
- Anything involving λ itself is not an estimator.

Note that $\hat{\lambda}(x_1, \ldots, x_n) = 17$ might be fine as an *estimate* for some particular data set $X_1 = x_1, \ldots, X_n = x_n$. However, the *estimator* $\hat{\lambda}(X_1, \ldots, X_n) = 17$, which ignores the data and returns 17 no matter what, is probably a bad estimator. ◊

3.1 Likelihood Function

Many procedures in statistical inference involve a mathematical object called the likelihood function.

Definition of Likelihood

Let θ be an unknown parameter, and let $\boldsymbol{X}$ be a sample with joint pmf $p_\theta(\boldsymbol{x})$ (if $\boldsymbol{X}$ is discrete) or joint pdf $f_\theta(\boldsymbol{x})$ (if $\boldsymbol{X}$ is continuous). The *likelihood function* for a particular set of data values $\boldsymbol{x}$ is $L_{\boldsymbol{x}}(\theta) = p_\theta(\boldsymbol{x})$ (if $\boldsymbol{X}$ is discrete) or $L_{\boldsymbol{x}}(\theta) = f_\theta(\boldsymbol{x})$ (if $\boldsymbol{X}$ is continuous).

- The function $L_{\boldsymbol{x}}(\theta)$ is simply a function of θ (though it may be a different function for different values of $\boldsymbol{x}$). The function itself is not random.
- Sometimes we may also need to consider $L_{\boldsymbol{X}}(\theta)$, which is a *random* function of θ.

EXAMPLE 3.1.1: Let $X \sim \text{Bin}(n, \theta)$. If we observe $X = x$, then $L_x(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$. We might also consider the random function $L_X(\theta) = \binom{n}{X}\theta^X(1-\theta)^{n-X}$. ◊

Interpretation of Likelihood

The likelihood function is essentially the same mathematical object as the joint pmf or pdf, but its *interpretation* is different.

- For $p_\theta(\boldsymbol{x})$ or $f_\theta(\boldsymbol{x})$, we think about fixing a parameter value θ and allowing $\boldsymbol{x}$ to vary.
- For $L_{\boldsymbol{x}}(\theta)$, we think about fixing a collection of sample values $\boldsymbol{x}$ and allowing θ to vary.

Since the pmf or pdf is nonnegative, the likelihood must be nonnegative as well.

What Likelihood Is Not

The likelihood is *not* a "pdf (or pmf) of θ given the data." There are several things wrong with such an interpretation:

- In frequentist inference, the unknown parameter θ is not a random variable, so talking about its "distribution" makes no sense.
- Even in Bayesian inference, the likelihood is still the same mathematical object as the pmf or pdf of the data. Hence, it describes probabilities of observing data values given certain parameter values, not the other way around.
- The likelihood may not even sum or integrate to 1 when summing or integrating over θ. In fact, it may sum or integrate to ∞, in which case we cannot even scale it to make it a pdf (or pmf).

Likelihood of Independent Samples

If the sample $\boldsymbol{X}$ consists of iid observations from a common individual pmf $p_\theta(x)$ or pdf $f_\theta(x)$, then the likelihood of the sample $\boldsymbol{X}$ is simply the product of the likelihoods associated with the individual observations $X_1, \ldots, X_n$, in which case $L_{\boldsymbol{x}}(\theta) = \prod_{i=1}^n L_{x_i}(\theta) = \prod_{i=1}^n f_\theta(x_i)$.

Log-Likelihood

It is often more convenient to work with the logarithm of the likelihood, $\ell_{\boldsymbol{x}}(\theta) = \log L_{\boldsymbol{x}}(\theta)$. We adopt the convention that $\ell_{\boldsymbol{x}}(\theta) = -\infty$ when $L_{\boldsymbol{x}}(\theta) = 0$.

3.2 Maximum Likelihood Estimation

An estimate $\hat{\theta}(\boldsymbol{x})$ of θ (with allowed parameter space Θ) is called a *maximum likelihood estimate* (MLE) of θ if $\hat{\theta}(\boldsymbol{x})$ maximizes $L_{\boldsymbol{x}}(\theta)$, the likelihood of θ, over Θ. An estimator that takes the value of the maximum likelihood estimate for every possible sample $\boldsymbol{X} = \boldsymbol{x}$ is called a *maximum likelihood estimator* (also MLE) . If the MLE is unique, then we can write

$$\hat{\theta}^{\text{MLE}} = \arg\max_{\theta \in \Theta} L_{\boldsymbol{X}}(\theta) = \arg\max_{\theta \in \Theta} \ell_{\boldsymbol{X}}(\theta),$$

noting that maximizing the likelihood is equivalent to maximizing the log-likelihood. Also note that the MLE can only take values within the allowed parameter space Θ.

Existence and Uniqueness

The maximum likelihood estimator need not be unique or even exist. It may be the case that for certain possible samples $\boldsymbol{X} = \boldsymbol{x}$, the likelihood function $L_{\boldsymbol{x}}(\theta)$ has a non-unique maximum or fails to achieve its maximum altogether. Further discussion of these possibilities can be found in Examples 7.5.8, 7.5.9, and 7.5.10 of DeGroot & Schervish.

Finding the MLE

Maximizing $L_{\boldsymbol{x}}(\theta)$, or equivalently $\ell_{\boldsymbol{x}}(\theta)$, is just a calculus problem. Typically we find all points in Θ where the derivative $\partial L_{\boldsymbol{x}}/\partial\theta$ is zero or undefined. Then we identify the global maximum, considering all critical points and boundary points of Θ.

EXAMPLE 3.2.1: Suppose $X_1, \ldots, X_n \sim$ iid Poisson(λ), where $\lambda \geq 0$, and we want to find the MLE of λ. The log-likelihood based on the sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ is

$$\ell_{\boldsymbol{x}}(\lambda) = \sum_{i=1}^{n} \log\left[\frac{\lambda^{x_i}\exp(-\lambda)}{(x_i)!}\right] = -n\lambda + n\overline{x}\log\lambda - \sum_{i=1}^{n}\log[(x_i)!].$$

Then

$$\frac{\partial}{\partial\lambda}\ell_{\boldsymbol{x}}(\lambda) = -n + \frac{n\overline{x}}{\lambda} = 0 \iff \lambda = \overline{x}.$$

It can be seen from the form of $\ell_{\boldsymbol{x}}(\lambda)$ that this critical point is indeed the maximizer, i.e.,

$$\ell_{\boldsymbol{x}}(\overline{x}) = \max_{\lambda\geq 0}\ell_{\boldsymbol{x}}(\lambda) \qquad \text{for all } \boldsymbol{x} \in (\mathbb{N}_0)^n,$$

where $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$. Thus, the maximum likelihood estimator of λ is $\hat{\lambda}^{\text{MLE}} = \overline{X}$. ◊

MLE with Multiple Parameters

The definition of the maximum likelihood estimator still holds if the unknown parameter θ is really $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$, i.e., if there are multiple unknown parameters. We still find the MLE $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$ the same way, though the calculus problem may be more complicated.

EXAMPLE 3.2.2: Let $X_1, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are both unknown, and we want to find the MLE of both parameters. The likelihood and log-likelihood based on the sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ are

$$L_{\boldsymbol{x}}(\sigma^2) = \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right] = (2\pi\sigma^2)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right],$$

$$\ell_{\boldsymbol{x}}(\sigma^2) = \log L_{\boldsymbol{x}}(\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Differentiating with respect to each parameter yields

$$\frac{\partial}{\partial\mu}\ell_{\boldsymbol{x}}(\mu,\sigma^2) = \frac{1}{\sigma^2}\sum_{i=1}^{n}x_i - \frac{n\mu}{\sigma^2} = \frac{n}{\sigma^2}(\overline{x}-\mu), \qquad \frac{\partial}{\partial(\sigma^2)}\ell_{\boldsymbol{x}}(\mu,\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$= \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}\big[(x_i-\mu)^2 - \sigma^2\big].$$

We now set *both* partial derivatives equal to zero and solve. First, note that

$$\frac{\partial}{\partial \mu}\ell_{\boldsymbol{x}}(\mu, \sigma^2) = \frac{n}{\sigma^2}(\overline{x} - \mu) = 0 \iff \mu = \overline{x}.$$

We can now substitute $\mu = \overline{x}$ into the other partial derivative, set it equal to zero, yielding

$$\frac{\partial}{\partial(\sigma^2)}\ell_{\boldsymbol{x}}(\mu, \sigma^2) = \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}\big[(x_i - \overline{x})^2 - \sigma^2\big] = 0 \iff \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \left(\frac{n-1}{n}\right)S^2.$$

It can be seen from the form of $\ell_{\boldsymbol{x}}(\mu, \sigma^2)$ that this point is indeed the maximizer, i.e.,

$$\ell_{\boldsymbol{x}}\left[\overline{x}, \left(\frac{n-1}{n}\right)S^2\right] = \max_{\mu\in\mathbb{R},\, \sigma^2>0} \ell_{\boldsymbol{x}}(\mu, \sigma^2). \tag{3.2.1}$$

> Note: The result in (3.2.1) holds for all $\boldsymbol{x} \in \mathbb{R}^p$ such that $x_1, \ldots, x_n$ are not all equal. If instead $x_1, \ldots, x_n$ are all equal, then $(n-1)S^2/n = 0$, which is outside the allowed parameter space for σ^2. However, since $X_1, \ldots, X_n$ are continuous random variables, $P(X_i = X_j) = 0$ for $i \neq j$, meaning that this issue arises with probability zero. Thus, we usually go ahead and define the MLE as if this situation cannot occur, despite the fact that it is technically possible.

Thus, the maximum likelihood estimators of μ and σ^2 are, respectively, $\hat{\mu}^{\text{MLE}} = \overline{X}$ and $(\hat{\sigma}^2)^{\text{MLE}} = (n-1)S^2/n$. Notice that the MLE of the variance is smaller than the usual sample variance by a factor of $(n-1)/n$. ♢

Example 3.2.3: Let $\boldsymbol{X}$ be an $n \times p$ matrix of known *constants* (*not* random variables, despite the capital letter), and let $Y_1, \ldots, Y_n$ be independent random variables with

$$Y_i \sim N\left(\sum_{j=1}^{p}\beta_j x_{ij},\ \sigma^2\right) \qquad \text{for each } i \in \{1, \ldots, n\},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ and $\sigma^2 > 0$ are both unknown. Let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. It can be shown (see, e.g., Theorem 11.5.1 of DeGroot & Schervish) that if the matrix $\boldsymbol{X}$ has rank p, then the maximum likelihood estimators of $\boldsymbol{\beta}$ and σ^2 are $\hat{\boldsymbol{\beta}}^{\text{MLE}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ and $(\hat{\sigma}^2)^{\text{MLE}} = n^{-1}\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^{\text{MLE}}\|_2^2$, respectively, where $\|\boldsymbol{u}\|_2^2 = \sum_{i=1}^{k} u_i^2$ for any $\boldsymbol{u} \in \mathbb{R}^k$.

> Note: If the rank of $\boldsymbol{X}$ is strictly less than p (which is automatically the case if $n < p$), then the MLE of $\boldsymbol{\beta}$ still exists but is not unique. However, the MLE of σ^2 does not exist (although we could take it to be zero if we expand the parameter space to $\sigma^2 \geq 0$ and adopt the convention that a normal distribution with variance zero is simply a degenerate distribution).

Note that $\hat{\boldsymbol{\beta}}^{\text{MLE}}$ is just the ordinary least squares estimator of $\boldsymbol{\beta}$. ♢

Sufficiency and Maximum Likelihood Estimation

Maximum likelihood estimation obeys the sufficiency principle, as shown below.

Theorem 3.2.4. *Let $r(\boldsymbol{X})$ be a sufficient statistic for θ, and let $\hat{\theta}^{\text{MLE}}$ be the unique maximum likelihood estimator of θ. Then $\hat{\theta}^{\text{MLE}}$ depends on $\boldsymbol{X}$ only through $r(\boldsymbol{X})$.*

Proof. By the factorization theorem, we may write $L_{\boldsymbol{x}}(\theta) = g[r(\boldsymbol{x}), \theta]\, h(\boldsymbol{x})$ for some functions g and h. Now note that maximizing $L_{\boldsymbol{x}}(\theta)$ as a function of θ for each $\boldsymbol{x}$ is equivalent to maximizing $g[r(\boldsymbol{x}), \theta]$ as a function of θ for each $\boldsymbol{x}$. The result follows. □

Invariance to Reparametrization

The next theorem provides a very convenient property of the maximum likelihood estimator.

Theorem 3.2.5. *Let $\hat{\theta}^{\mathrm{MLE}}$ be a maximum likelihood estimator of θ, and let g be a function with domain Θ. Then $\hat{\xi}^{\mathrm{MLE}} = g(\hat{\theta}^{\mathrm{MLE}})$ is a maximum likelihood estimator of $\xi = g(\theta)$.*

Proof. See the proof of Theorem 7.6.2 in DeGroot & Schervish. □

Example 3.2.6: Suppose that in Example 3.2.2, we had taken the second unknown parameter to be the standard deviation σ instead of the variance σ^2. Then the maximum likelihood estimator of σ would have simply been

$$\hat{\sigma}^{\mathrm{MLE}} = \sqrt{(\hat{\sigma}^2)^{\mathrm{MLE}}} = \sqrt{\frac{(n-1)S^2}{n}}$$

by Theorem 3.2.5. ◊

Numerical Calculation of Maximum Likelihood Estimates

It is often the case that the maximum likelihood estimator $\hat{\theta}^{\mathrm{MLE}}(\boldsymbol{X})$ cannot be expressed in closed form as a function of $\boldsymbol{X}$. However, the maximum likelihood *estimate* $\hat{\theta}^{\mathrm{MLE}}(\boldsymbol{x})$ for a particular sample $\boldsymbol{X} = \boldsymbol{x}$ can usually still be found numerically.

Example 3.2.7: Let $x_1, \ldots, x_n$ be known constants, and let $Y_1, \ldots, Y_n$ be independent random variables with

$$Y_i \sim \mathrm{Bin}\left[1, \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}\right] \qquad \text{for each } i \in \{1, \ldots, n\},$$

where $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$ are both unknown. (This is often called *logistic regression*.) The log-likelihood based on the sample $\boldsymbol{y} = (y_1, \ldots, y_n)$ is

$$\begin{aligned}\ell_{\boldsymbol{y}}(\alpha, \beta) = \log L_{\boldsymbol{y}}(\alpha, \beta) &= \log \prod_{i=1}^{n} \left[\frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}\right]^{y_i} \left[\frac{1}{1 + \exp(\alpha + \beta x_i)}\right]^{1-y_i} \\ &= \sum_{i=1}^{n} y_i(\alpha + \beta x_i) - \sum_{i=1}^{n} \log[1 + \exp(\alpha + \beta x_i)].\end{aligned}$$

Differentiating with respect to α and β yields

$$\frac{\partial}{\partial \alpha}\ell_{\boldsymbol{y}}(\alpha, \beta) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{\exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}, \qquad \frac{\partial}{\partial \beta}\ell_{\boldsymbol{y}}(\alpha, \beta) = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \frac{x_i \exp(\alpha + \beta x_i)}{1 + \exp(\alpha + \beta x_i)}.$$

Setting both partial derivatives above equal to zero yields a system of equations that cannot be solved in closed form. However, we can find a solution numerically to obtain maximum likelihood estimates $\hat{\alpha}^{\mathrm{MLE}}$ and $\hat{\beta}^{\mathrm{MLE}}$ for most samples $\boldsymbol{y} \in \{0, 1\}^p$. (However, no matter what the true values of α and β are, there is a nonzero probability of obtaining a sample such that the maximum likelihood estimates do not exist.) ◊

3.3 Estimators that Optimize Other Functions

Sometimes we may want to find an estimator by maximizing or minimizing some real-valued function other than the likelihood . There are many reasons why we might want to do this:

- The likelihood itself may be difficult to work with.
- We may be unsure of some aspect of our model (e.g., we may not know if the observations are normally distributed).
- We may want to favor certain kinds of estimates over others.

An estimator that is found by maximizing or minimizing some real-valued function other than the likelihood is called an *M-estimator*. Note that maximum likelihood estimation is a special case of M-estimation.

Example 3.3.1: In the regression setup of Example 3.2.3, the least squares estimator

$$\hat{\boldsymbol{\beta}}^{\mathrm{LS}} = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$$

is an M-estimator of $\boldsymbol{\beta}$. However, $\hat{\boldsymbol{\beta}}^{\mathrm{LS}} = \hat{\boldsymbol{\beta}}^{\mathrm{MLE}}$, so this estimator coincides with the maximum likelihood estimator. ♢

Example 3.3.2: In the regression setup of Example 3.2.3, we could instead consider the *least absolute deviation* estimator

$$\hat{\boldsymbol{\beta}}^{\mathrm{LAD}} = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_1,$$

where $\|\boldsymbol{u}\|_1 = \sum_{i=1}^k |u_i|$ for any $\boldsymbol{u} \in \mathbb{R}^k$. Then $\hat{\boldsymbol{\beta}}^{\mathrm{LAD}}$ is another M-estimator. ♢

Example 3.3.3: In the regression setup of Example 3.2.3, we could instead consider the *lasso* estimator

$$\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}} = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p}\big(\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\big),$$

where $\lambda > 0$ is some fixed constant.

> Reference: Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, **58** 267–288.

The addition of the term $\lambda\|\boldsymbol{\beta}\|_1$ does several things:

- It typically results in an estimate for which some components are exactly zero.
- The nonzero components are typically smaller in absolute value than the corresponding components of $\hat{\boldsymbol{\beta}}^{\mathrm{MLE}}$ (provided the MLE exists and is unique).
- $\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}}$ is unique in many cases when $\hat{\boldsymbol{\beta}}^{\mathrm{MLE}}$ is not. For example, $\hat{\boldsymbol{\beta}}^{\mathrm{LASSO}}$ is usually still unique even when $p > n$.

Many variations and extensions of the lasso have been developed. Such methods are sometimes called *regularized* or *penalized* regression. ♢

Lecture 4: Bayesian Estimation

Before we can discuss the Bayesian approach to estimation, we must first motivate and introduce the Bayesian philosophy of statistical inference.

4.1 Bayesian Philosophy

Suppose we have a jar that contains 99 fair coins and a single unfair coin that is rigged to come up as heads with probability 3/4. We draw one coin at random from the jar, and we let U be the heads probability of the coin we draw. Then

$$f^{(U)}(u) = \begin{cases} 1/2 & \text{with probability } 99/100 \\ 3/4 & \text{with probability } \ 1/100, \end{cases}$$

and hence the pmf of U is

$$f^{(U)}(u) = \begin{cases} 99/100 & \text{if } u = 1/2, \\ \ 1/100 & \text{if } u = 3/4. \end{cases}$$

After selecting a coin, we flip it five times, and we let X count the number of heads in these flips. Then for $x \in \{0, 1, \ldots, 5\}$, the conditional pmf of X given U is

$$f^{(X|U)}(x \mid u) = \begin{cases} \dfrac{5!}{x!\,(5-x)!}\left(\dfrac{1}{2}\right)^x\left(\dfrac{1}{2}\right)^{5-x} & \text{if } u = \dfrac{1}{2}, \\ \dfrac{5!}{x!\,(5-x)!}\left(\dfrac{3}{4}\right)^x\left(\dfrac{1}{4}\right)^{5-x} & \text{if } u = \dfrac{3}{4}. \end{cases}$$

Now suppose we wish to determine the probability that we have drawn the unfair coin conditional on the fact that we observe that all five flips are heads. This is simply

$$f^{(U|X)}(3/4 \mid 5) = \frac{f^{(U,X)}(3/4,\, 5)}{f^{(X)}(5)} = \frac{f^{(X|U)}(5 \mid 3/4)\; f^{(U)}(3/4)}{f^{(X)}(5)} \approx 0.071. \tag{4.1.1}$$

This calculation, called *Bayes' rule*, follows immediately from the definition of conditional probability. Both frequentists and Bayesians would agree that this calculation is valid.

Unknown Parameter

Suppose instead that we do not know how many coins of each type are in the jar. Then we do not know $f^{(U)}(u)$, the pmf of U, so we can no longer perform the calculation of Bayes' rule as in (4.1.1). Now suppose also that we let θ (instead of U) denote the heads probability of our selected coin. Then when we start flipping the coin, θ is either 1/2 or 3/4, and we simply do not know which. This is exactly the scenario of frequentist statistical inference.

Probability as Degree of Belief

On the other hand, the Bayesian philosophy allows probability to also describe a subjective degree of belief. Even though we are not told how many coins of each type are in the jar, we may believe (for example) that only a small fraction of the coins are unfair. Then we can construct a "distribution" of θ based on these subjective beliefs, which allows us once again to calculate Bayes' rule as in (4.1.1). Thus, if we are willing to specify a subjective "distribution" of θ in advance, we can obtain a conditional distribution of θ given the data we observe. This is the fundamental idea of the Bayesian philosophy of statistical inference.

4.2 Prior and Posterior

We now introduce the standard terminology and notation of the Bayesian approach. Suppose we have an unknown parameter θ and some data $\boldsymbol{X} = \boldsymbol{x}$ with joint pdf $f_\theta(\boldsymbol{x})$ or pmf $p_\theta(\boldsymbol{x})$. Since we now treat θ as a random variable, we should actually be writing $f_\theta(\boldsymbol{x}) = f^{(\boldsymbol{X}|\theta)}(\boldsymbol{x} \mid \theta)$ or $p_\theta(\boldsymbol{x}) = p^{(\boldsymbol{X}|\theta)}(\boldsymbol{x} \mid \theta)$ if we want to be consistent with our earlier notation. Either way, we will still write the likelihood as $L_{\boldsymbol{x}}(\theta)$.

> Note: We are now slightly abusing notation by using θ to represent both a random variable and a value that can be taken by that random variable. However, this is the standard way that things are written in the Bayesian approach.

Since θ is now treated as a random variable, we will also make a slight adjustment when writing down a statistical model. Instead of simply writing something like $X \sim \text{Bin}(n, \theta)$, we will now write $X \mid \theta \sim \text{Bin}(n, \theta)$ to explicitly show the conditioning on θ.

Prior

The *prior* distribution on θ represents our subjective beliefs about θ before observing the data. If we want to be consistent with our earlier notation, we should write the prior as $f^{(\theta)}(\theta)$ or $p^{(\theta)}(\theta)$ according to whether θ is continuous or discrete. However, the standard notation for a prior on θ is $\pi(\theta)$.

> Note: It should always be clear from context whether π refers to a prior distribution or to the constant that relates a circle's circumference to its diameter.

We will discuss how to choose a prior later.

Posterior

Once we have the prior $\pi(\theta)$ and the likelihood $L_{\boldsymbol{x}}(\theta)$, we can use Bayes' rule to find the distribution of θ given $\boldsymbol{X} = \boldsymbol{x}$. Using our original notation for conditional pmfs or pdfs, we would write

$$f^{(\theta|\boldsymbol{X})}(\theta \mid \boldsymbol{x}) = \frac{f^{(\boldsymbol{X}|\theta)}(\boldsymbol{x} \mid \theta)\, f^{(\theta)}(\theta)}{f^{(\boldsymbol{X})}(\boldsymbol{x})} \qquad \text{or} \qquad p^{(\theta|\boldsymbol{X})}(\theta \mid \boldsymbol{x}) = \frac{p^{(\boldsymbol{X}|\theta)}(\boldsymbol{x} \mid \theta)\, p^{(\theta)}(\theta)}{p^{(\boldsymbol{X})}(\boldsymbol{x})}.$$

Using the standard terminology of the Bayesian approach, we write this as

$$\pi(\theta \mid \boldsymbol{x}) = \frac{L_{\boldsymbol{x}}(\theta)\, \pi(\theta)}{m(\boldsymbol{x})}, \tag{4.2.1}$$

where $m(\boldsymbol{x})$ is the marginal distribution of $\boldsymbol{X}$ evaluated at $\boldsymbol{x}$ and $\pi(\theta \mid \boldsymbol{x})$ is called the *posterior* distribution of θ conditional on $\boldsymbol{X} = \boldsymbol{x}$. The posterior is the object on which all Bayesian inference about θ is based.

> Note: If $\boldsymbol{X}$ and θ are both discrete, then (4.2.1) arises directly from the definition of conditional probability. If $\boldsymbol{X}$ and θ are both continuous, then (4.2.1) arises directly from the way we defined conditional distributions of continuous random variables. However, if $\boldsymbol{X}$ is discrete and θ is continuous (or vice versa), then we technically have a problem, since we have not defined conditional distributions in this situation. In reality, this is not really an issue, as definitions can be cleaned up using more sophisticated techniques to make (4.2.1) valid in all such scenarios.

Recall that the marginal distribution of $\boldsymbol{X}$ is found by integrating or summing out θ from the joint distribution of $\boldsymbol{X}$ and θ. The joint distribution of $\boldsymbol{X}$ and θ is simply $L_{\boldsymbol{x}}(\theta)\, \pi(\theta)$, the numerator of (4.2.1). Thus, the marginal distribution of $\boldsymbol{X}$ is simply

$$m(\boldsymbol{x}) = \int L_{\boldsymbol{x}}(\theta)\, \pi(\theta)\, d\theta \qquad \text{or} \qquad m(\boldsymbol{x}) = \sum_{\theta \in \Theta} L_{\boldsymbol{x}}(\theta)\, \pi(\theta).$$

Thus, it can be seen that $m(\boldsymbol{x})$ is simply the normalizing constant that is needed to make the function $L_{\boldsymbol{x}}(\theta)\, \pi(\theta)$ a valid probability distribution for θ.

Example 4.2.1: Suppose $X_1, \ldots, X_n \sim \text{iid } p \sim \text{Bin}(1, \theta)$, where $0 \le \theta \le 1$, and we observe $\boldsymbol{X} = \boldsymbol{x}$. Now take our prior distribution for θ to be $\text{Beta}(a, b)$, which has pdf

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}\, I_{[0,1](\theta)}.$$

The likelihood is

$$L_{\boldsymbol{x}}(\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.$$

The marginal distribution of $\boldsymbol{x}$ is

$$\begin{aligned} m(\boldsymbol{x}) = \int_0^1 L_{\boldsymbol{x}}(\theta)\, \pi(\theta)\, d\theta &= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)} \int_0^1 \theta^{\sum_{i=1}^n x_i + a - 1}(1-\theta)^{n - \sum_{i=1}^n x_i + b - 1}\, d\theta \\ &= \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\, \frac{\Gamma(a + \sum_{i=1}^n x_i)\,\Gamma(b + n - \sum_{i=1}^n x_i)}{\Gamma(a+b+n)}, \end{aligned}$$

where the integral is computed by noting that the integrand is an unnormalized beta pdf. Then the posterior distribution of θ is

$$\pi(\theta \mid \boldsymbol{x}) = \frac{L_{\boldsymbol{x}}(\theta)\, \pi(\theta)}{m(\boldsymbol{x})} = \frac{\Gamma(a + \sum_{i=1}^n x_i)\,\Gamma(b + n - \sum_{i=1}^n x_i)}{\Gamma(a+b+n)} \theta^{\sum_{i=1}^n x_i + a - 1}(1-\theta)^{n - \sum_{i=1}^n x_i + b - 1}\, I_{[0,1]}(\theta),$$

which we recognize as the pdf of a $\text{Beta}(a + \sum_{i=1}^n x_i,\ b + n - \sum_{i=1}^n x_i)$ distribution. Thus, the posterior distribution of θ is $\theta \mid \boldsymbol{x} \sim \text{Beta}(a + \sum_{i=1}^n x_i,\ b + n - \sum_{i=1}^n x_i)$. ◊

Shortcut for Finding Posteriors

There is sometimes no need to actually compute the marginal distribution $m(\boldsymbol{x})$. Suppose we write down $L_{\boldsymbol{x}}(\theta)\,\pi(\theta)$, and we recognize that this function of θ looks like some distribution of θ with an incorrect normalizing constant. Then we know that $m(\boldsymbol{x})$ must simply be whatever is needed to fix this normalizing constant. In fact, we really only need to write down parts of $L_{\boldsymbol{x}}(\theta)\,\pi(\theta)$ that depend on θ, since we know the normalizing constants must work out properly if we recognize the distribution based on the form of the function of θ.

Example 4.2.2: In Example 4.2.1, we could have simply noted that

$$L_{\boldsymbol{x}}(\theta)\,\pi(\theta) \propto \theta^{\sum_{i=1}^n x_i + a - 1}(1-\theta)^{n - \sum_{i=1}^n x_i + b - 1}\, I_{[0,1]}(\theta),$$

which we recognize as the unnormalized pdf of a $\text{Beta}(a+\sum_{i=1}^n x_i,\ b+n-\sum_{i=1}^n x_i)$ distribution. Thus, we could have concluded immediately that $\theta \mid \boldsymbol{x} \sim \text{Beta}(a + \sum_{i=1}^n x_i,\ b + n - \sum_{i=1}^n x_i)$, without needing to do any further computation. ♢

Conjugate Priors

In Example 4.2.1, the prior was a beta distribution, and the posterior was another beta distribution. A family of distributions is called *conjugate* for a particular likelihood function if choosing a prior from that family leads to a posterior that is also from that family. Conjugate priors are often used because they are very convenient. In particular, the shortcut described above is guaranteed to work if a conjugate prior is chosen.

Example 4.2.3: Let $X_1, \ldots, X_n \mid \mu \sim \text{iid } N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ is unknown but $\sigma^2 > 0$ is known. Let the prior on μ be $\mu \sim N(\xi, \tau^2)$, where $\xi \in \mathbb{R}$ and $\tau^2 > 0$ are known. To find the posterior of μ, we first try the shortcut described above. Ignoring anything that is not a function of μ, we have

$$\begin{aligned}
L_{\boldsymbol{x}}(\mu)\,\pi(\mu) &\propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right]\exp\left[-\frac{(\mu-\xi)^2}{2\tau^2}\right] \\
&\propto \exp\left(\frac{\mu}{\sigma^2}\sum_{i=1}^n x_i - \frac{n\mu^2}{2\sigma^2} - \frac{\mu^2}{2\tau^2} + \frac{\xi\mu}{\tau^2}\right) \\
&\propto \exp\left[-\frac{(n\tau^2+\sigma^2)\mu^2}{2\sigma^2\tau^2} + \frac{(n\overline{x}\tau^2 + \xi\sigma^2)\mu}{\sigma^2\tau^2}\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\frac{n\tau^2+\sigma^2}{\sigma^2\tau^2}\right)\left(\mu^2 - 2\mu\frac{n\tau^2\overline{x}+\sigma^2\xi}{n\tau^2+\sigma^2}\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\left(\frac{n\tau^2+\sigma^2}{\sigma^2\tau^2}\right)\left(\mu - \frac{n\tau^2\overline{x}+\sigma^2\xi}{n\tau^2+\sigma^2}\right)^2\right],
\end{aligned}$$

which we recognize as another normal distribution. Thus, the posterior distribution of μ given $\boldsymbol{X} = \boldsymbol{x}$ is

$$\mu \mid \boldsymbol{x} \sim N\left(\frac{n\tau^2\overline{x}+\sigma^2\xi}{n\tau^2+\sigma^2},\ \frac{n\tau^2+\sigma^2}{\sigma^2\tau^2}\right).$$

It is perhaps more insightful to rewrite this as

$$\mu \mid \boldsymbol{x} \sim N\left[\frac{(\sigma^2/n)^{-1}}{(\tau^2)^{-1}+(\sigma^2/n)^{-1}}\overline{x} + \frac{(\tau^2)^{-1}}{(\tau^2)^{-1}+(\sigma^2/n)^{-1}}\xi,\ \frac{1}{(\tau^2)^{-1}+(\sigma^2/n)^{-1}}\right].$$

Note that the mean of the posterior is simply a weighted average of the sample mean $\overline{x}$ and the prior mean ξ, with weights proportional to the inverses of the σ^2/n (the variance of $\overline{x}$) and the prior variance τ^2. ◊

Choosing Priors

For this course, you will always be told what prior to use. However, in real applications, we would need to determine a prior distribution for ourselves. In principle, we should try to make a subjective prior to actually represent our subjective beliefs about θ. This sounds straightforward, but it is often quite difficult to do in practice. It also opens up our analysis to criticism if someone disagrees with our subjective choice.

> Note: This is not unique to the Bayesian approach. All statistical inference inherently includes subjective choices by the statistician. (For example, the choice to model observations as normally distributed is often at least somewhat subjective.) Bayesian inference with a subjective prior simply makes the subjectivity much more obvious.

What statisticians often do instead is to try to choose a prior that tries to say as little as possible about θ, i.e., it represents a total lack of prior knowledge about the value of θ. Such priors are often called *flat* or *uninformative* priors.

Example 4.2.4: In Example 4.2.3, we could specify a somewhat flat prior by taking the prior variance τ^2 to be very large. Note that if we do this, then the mean of the posterior distribution will be approximately $\overline{x}$, since the weight associated with $\overline{x}$ will be close to 1 while the weight associated with ξ will be close to 0. ◊

4.3 Bayes Estimators

Once we have the posterior distribution of a parameter θ, we can find a *Bayes estimate* $\hat{\theta}^B$. A Bayes estimate is simply a summary to report some sort of "center" of the posterior.

Posterior Mean

By far the most common choice of Bayes estimate is the posterior mean $E(\theta \mid x)$. (In fact, many people simply call the posterior mean *the* Bayes estimate.)

Example 4.3.1: In Example 4.2.1, we found $\theta \mid \boldsymbol{x} \sim \text{Beta}(a + \sum_{i=1}^n x_i,\ b + n - \sum_{i=1}^n x_i)$. Then the posterior mean is simply

$$E(\theta \mid \boldsymbol{x}) = \frac{a + \sum_{i=1}^n x_i}{a + b + n}$$

by standard properties of the beta distribution. ◊

Posterior Median

We could also report the posterior median as our Bayes estimate. However, this is sometimes harder to compute in closed form.

EXAMPLE 4.3.2: In Example 4.2.1, we found $\theta \mid \boldsymbol{x} \sim \text{Beta}(a + \sum_{i=1}^n x_i,\ b + n - \sum_{i=1}^n x_i)$. This distribution has a median that cannot be found in closed form (in general). ♢

Posterior Mode

We could also report the posterior mode as our Bayes estimate. This is is sometimes called a *maximum a posteriori* (MAP) estimate.

EXAMPLE 4.3.3: In Example 4.2.1, we found $\theta \mid \boldsymbol{x} \sim \text{Beta}(a + \sum_{i=1}^n x_i,\ b + n - \sum_{i=1}^n x_i)$. Then the posterior mode is simply

$$\arg\max_{0 \le \theta \le 1} \pi(\theta \mid \boldsymbol{x}) = \frac{a + \sum_{i=1}^n x_i - 1}{a + b + n - 2}$$

by standard properties of the beta distribution. ♢

Observe that maximizing the posterior is equivalent to maximizing $L_{\boldsymbol{x}}(\theta)\,\pi(\theta)$. Thus, the maximum a posteriori estimate is actually just maximizing the likelihood times the prior.

Bayes Estimates and Bayes Estimator

The Bayes estimate is whatever the measure we report based on the observed values $\boldsymbol{X} = \boldsymbol{x}$. Thus, the Bayes estimate is a function of $\boldsymbol{x}$. The *Bayes estimator* is the random variable obtained by inserting the random variable $\boldsymbol{X}$ into this function.

EXAMPLE 4.3.4: In Example 4.2.1, we found that the posterior mean was

$$E(\theta \mid \boldsymbol{x}) = \frac{a + \sum_{i=1}^n x_i}{a + b + n}.$$

Then the Bayes estimator is simply

$$\hat{\theta}^B = \frac{a + \sum_{i=1}^n X_i}{a + b + n},$$

which is a random variable. ♢

Frequentist Use of Bayes Estimators

Although we use the Bayesian philosophy to *derive* Bayes estimators, we can still *use* Bayes estimators even if we do not actually agree with the Bayesian philosophy. In the end, a Bayes estimator is simply a function of the data, just like any other estimator, so we can calculate it or consider its properties without worrying about the philosophy under which its form was derived.

Lecture 5: Finite-Sample Properties of Estimators

Conceptually, a good estimator should "usually" be "close" to the parameter it estimates. We now consider how to formalize this idea.

5.1 Bias and Variance

An estimator is simply a random variable. We begin by considering properties related to the expectation and variance of this random variable.

Bias

The *bias* of an estimator $\hat{\theta}$ of a parameter θ is $\text{Bias}_\theta(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta$. The estimator $\hat{\theta}$ is *unbiased* if $\text{Bias}_\theta(\hat{\theta}) = 0$ for all θ in the parameter space Θ.

EXAMPLE 5.1.1: Let $X_1, \ldots, X_n$ be iid random variables such that $\mu = E_\mu(X_1)$ is finite, and let $\overline{X}$ be the usual sample mean. Consider $\overline{X}/2$ as an estimator of μ. Then

$$\text{Bias}_\mu\big(\overline{X}/2\big) = E_\mu\big(\overline{X}/2\big) - \mu = -\mu/2.$$

Note that the bias is zero if μ happens to be zero, but not if $\mu \neq 0$, so this estimator is biased (i.e., not unbiased). ◊

EXAMPLE 5.1.2: Let $X_1, \ldots, X_n$ be iid random variables such that both $\mu = E_{\mu,\sigma^2}(X_1)$ and $\sigma^2 = \text{Var}_{\mu,\sigma^2}(X_1)$ are finite, and suppose $n \geq 2$. Let $\overline{X}$ and S^2 be the usual sample mean and sample variance, respectively. Then

$$E_{\mu,\sigma^2}(S^2) = \frac{1}{n-1} E_{\mu,\sigma^2}\left(\sum_{i=1}^n X_i^2 - n\overline{X}^2\right) = \frac{1}{n-1}\left[n(\mu^2+\sigma^2) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right] = \frac{n-1}{n-1}\sigma^2 = \sigma^2.$$

Thus, $\text{Bias}_{\mu,\sigma^2}(S^2) = \sigma^2 - \sigma^2 = 0$ for all values of σ^2, so S^2 is an unbiased estimator of σ^2. Note that if the $(n-1)^{-1}$ is replaced with n^{-1} in the definition of S^2 above, then the resulting estimator has expectation $(n-1)\sigma^2/n$ and thus is no longer unbiased. ◊

Unbiasedness is not, by itself, enough to ensure that an estimator is good. Similarly, an unbiased estimator is not necessarily better than a biased one.

EXAMPLE 5.1.3: Return to the situation of Example 5.1.2. The estimator $(X_1 - X_2)^2/2$ has expectation

$$\begin{aligned} E_{\mu,\sigma^2}\big[(X_1 - X_2)^2/2\big] &= E_{\mu,\sigma^2}(X_1^2/2) + E_{\mu,\sigma^2}(X_2^2/2) - E_{\mu,\sigma^2}(X_1)\,E_{\mu,\sigma^2}(X_2) \\ &= (\mu^2+\sigma^2)/2 + (\mu^2+\sigma^2)/2 - \mu^2 = \sigma^2 \end{aligned}$$

and is hence an unbiased estimator of σ^2. However, this estimator involves only the first two observations and ignores the remaining $n-2$ observations, so we probably would not want to use this estimator. In contrast, as shown in Example 5.1.2, the estimator $(n-1)S^2/n$ has expectation $(n-1)\sigma^2/n$ and is hence a biased estimator of σ^2. However, if n is large, then the bias is small, in which case this estimator may not be bad. ◊

It is often the case that we can "trade" a small amount of bias in order to improve an estimator in other ways. This idea will be discussed more later.

Variance

It can also be useful to consider the variance $\mathrm{Var}_\theta(\hat{\theta})$ of an estimator $\hat{\theta}$ of a parameter θ.

EXAMPLE 5.1.4: Return to the situation of Examples 5.1.2 and 5.1.3, and suppose further that the distribution of $X_1, \ldots, X_n$ is normal. Then

$$\mathrm{Var}_{\mu,\sigma^2}(S^2) = \left(\frac{\sigma^2}{n-1}\right)^2 \mathrm{Var}_{\mu,\sigma^2}\left[\frac{(n-1)S^2}{\sigma^2}\right] = \left(\frac{\sigma^2}{n-1}\right)^2 [2(n-1)] = \frac{2(\sigma^2)^2}{n-1},$$

noting that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ since $X_1, \ldots, X_n \sim \text{iid } N(\mu, \sigma^2)$. It follows that

$$\mathrm{Var}_{\mu,\sigma^2}\left[\left(\frac{n-1}{n}\right)S^2\right] = \left(\frac{n-1}{n}\right)^2 \mathrm{Var}_{\mu,\sigma^2}(S^2),$$

which is smaller than the variance of S^2. The variance of the estimator $(X_1 - X_2)^2/2$ can be found by noting that it is simply the sample variance of the first two observations, and thus

$$\mathrm{Var}_{\mu,\sigma^2}\left[\frac{(X_1 - X_2)^2}{2}\right] = 2(\sigma^2)^2 = (n-1)\,\mathrm{Var}_{\mu,\sigma^2}(S^2).$$

Unless n is very small, this estimator has much larger variance than either of the other two estimators discussed above. ♢

A smaller variance is usually better, but this is not always true. For example, a constant estimator (e.g., $\hat{\theta} = 17$) has zero variance but is clearly not a good estimator.

Bias-Variance Trade-Off

When comparing sensible estimators, an estimator with larger bias often has smaller variance, and vice versa. Thus, it may not be immediately clear which of several sensible estimators is to be preferred.

EXAMPLE 5.1.5: . Return to the situation of Examples 5.1.2–5.1.4. The estimators S^2 and $(X_1 - X_2)^2/2$ are both unbiased, but S^2 has smaller variance. Thus, S^2 is a better estimator than $(X_1 - X_2)^2/2$. However, the comparison between S^2 and $(n-1)S^2/n$ is not so clear. One estimator has smaller bias, while the other estimator has smaller variance. ♢

5.2 Mean Squared Error

The *mean squared error* of an estimator $\hat{\theta}$ of a parameter θ is $\mathrm{MSE}_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2]$. It provides one way to evaluate the overall performance of an estimator. The following theorem provides a useful way both to calculate and to interpret the mean squared error.

Theorem 5.2.1. *Let $\hat{\theta}$ be an estimator of θ. Then* $\mathrm{MSE}_\theta(\hat{\theta}) = [\mathrm{Bias}_\theta(\hat{\theta})]^2 + \mathrm{Var}_\theta(\hat{\theta})$.

Proof. $\mathrm{MSE}_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2] = [E_\theta(\hat{\theta} - \theta)]^2 + \mathrm{Var}_\theta(\hat{\theta} - \theta) = [\mathrm{Bias}_\theta(\hat{\theta})]^2 + \mathrm{Var}_\theta(\hat{\theta})$. □

EXAMPLE 5.2.2: Return to the situation of Examples 5.1.2–5.1.5. The mean squared errors of the estimators S^2 and $(n-1)S^2/n$ are

$$\begin{aligned}
\mathrm{MSE}_{\mu,\sigma^2}(S^2) &= \big[\mathrm{Bias}_{\mu,\sigma^2}(S^2)\big]^2 + \mathrm{Var}_{\mu,\sigma^2}(S^2) = \frac{2(\sigma^2)^2}{n-1}, \\
\mathrm{MSE}_{\mu,\sigma^2}\left[\left(\frac{n-1}{n}\right)S^2\right] &= \left\{\mathrm{Bias}_{\mu,\sigma^2}\left[\left(\frac{n-1}{n}\right)S^2\right]\right\}^2 + \mathrm{Var}_{\mu,\sigma^2}\left[\left(\frac{n-1}{n}\right)S^2\right] \\
&= \left[\left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2\right]^2 + \frac{2(n-1)(\sigma^2)^2}{n^2} = \frac{(2n-1)(\sigma^2)^2}{n^2} < \mathrm{MSE}_{\mu,\sigma^2}(S^2)
\end{aligned}$$

for all $n \geq 2$. Thus, for all $\mu \in \mathbb{R}$ and all $\sigma^2 > 0$, the mean squared error of the estimator $(n-1)S^2/n$ (the MLE of σ^2 for a normal sample) is smaller than the mean squared error of the unbiased estimator S^2. ♢

Let $\hat{\theta}$ and $\tilde{\theta}$ be estimators of θ. If $\mathrm{MSE}_\theta(\hat{\theta}) \leq \mathrm{MSE}_\theta(\tilde{\theta})$ for all $\theta \in \Theta$ and $\mathrm{MSE}_\theta(\hat{\theta}) < \mathrm{MSE}_\theta(\tilde{\theta})$ for some $\theta \in \Theta$, then the estimator $\hat{\theta}$ *dominates* the estimator $\tilde{\theta}$. In principle, it seems that we should avoid using any estimator that is dominated by another estimator. However, in practice, this policy is not always followed.

EXAMPLE 5.2.3: We showed in Example 5.2.2 that the unbiased sample variance S^2 is dominated by the maximum likelihood estimator $(n-1)S^2/n$. However, the sample variance is still often used in practice. ♢

EXAMPLE 5.2.4: Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_p)$, where $\boldsymbol{\mu} \in \mathbb{R}^p$ is unknown and $\sigma^2 > 0$ is known. Then $\boldsymbol{X}$ itself is the maximum likelihood estimator of $\boldsymbol{\mu}$, and it is unbiased. Indeed, it may seem that $\boldsymbol{X}$ is the *only* sensible estimator of $\boldsymbol{\mu}$. However, it can be shown that if $p \geq 3$, then there exist estimators that dominate $\boldsymbol{X}$, such as the *James-Stein estimator*

$$\hat{\boldsymbol{\mu}}^{\mathrm{JS}} = \left[1 - \frac{(p-2)\sigma^2}{\sum_{i=1}^p X_i^2}\right]\boldsymbol{X}.$$

> REFERENCE: James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, **1** 361–379.

Again, if $p \geq 3$, then there even exist other estimators that in turn dominate the James-Stein estimator. The proofs of these results are beyond the scope of this course. ♢

More commonly, when comparing sensible estimators, it is often the case that one estimator has smaller mean squared error for some parameter values, while the other estimator has smaller mean squared error for other parameter values. In this case, it is not at all clear which estimator is better.

EXAMPLE 5.2.5: Suppose $X \sim \mathrm{Bin}(n, \theta)$, where θ is unknown and $0 \leq \theta \leq 1$. Recall that the maximum likelihood estimator of θ is $\hat{\theta}^{\mathrm{MLE}} = X/n$. Its bias and variance are

$$\mathrm{Bias}_\theta\big(\hat{\theta}^{\mathrm{MLE}}\big) = E_\theta\left(\frac{X}{n}\right) - \theta = 0, \qquad \mathrm{Var}_\theta\big(\hat{\theta}^{\mathrm{MLE}}\big) = \mathrm{Var}_\theta\left(\frac{X}{n}\right) = \frac{\theta(1-\theta)}{n},$$

so its mean squared error is

$$\mathrm{MSE}_\theta\big(\hat{\theta}^{\mathrm{MLE}}\big) = \big[\mathrm{Bias}_\theta\big(\hat{\theta}^{\mathrm{MLE}}\big)\big]^2 + \mathrm{Var}_\theta\big(\hat{\theta}^{\mathrm{MLE}}\big) = \frac{\theta(1-\theta)}{n}.$$

If we instead put a $\mathrm{Beta}(a,b)$ prior on θ and conduct a Bayesian analysis, we find that the posterior mean is $\hat{\theta}^B = (X+a)/(n+a+b)$. Its bias and variance are

$$\mathrm{Bias}_\theta\big(\hat{\theta}^B\big) = E_\theta\bigg(\frac{X+a}{n+a+b}\bigg) - \theta = \frac{n\theta+a}{n+a+b} - \theta = \frac{(1-\theta)a - \theta b}{n+a+b},$$
$$\mathrm{Var}_\theta\big(\hat{\theta}^B\big) = \mathrm{Var}_\theta\bigg(\frac{X+a}{n+a+b}\bigg) = \frac{n\theta(1-\theta)}{(n+a+b)^2},$$

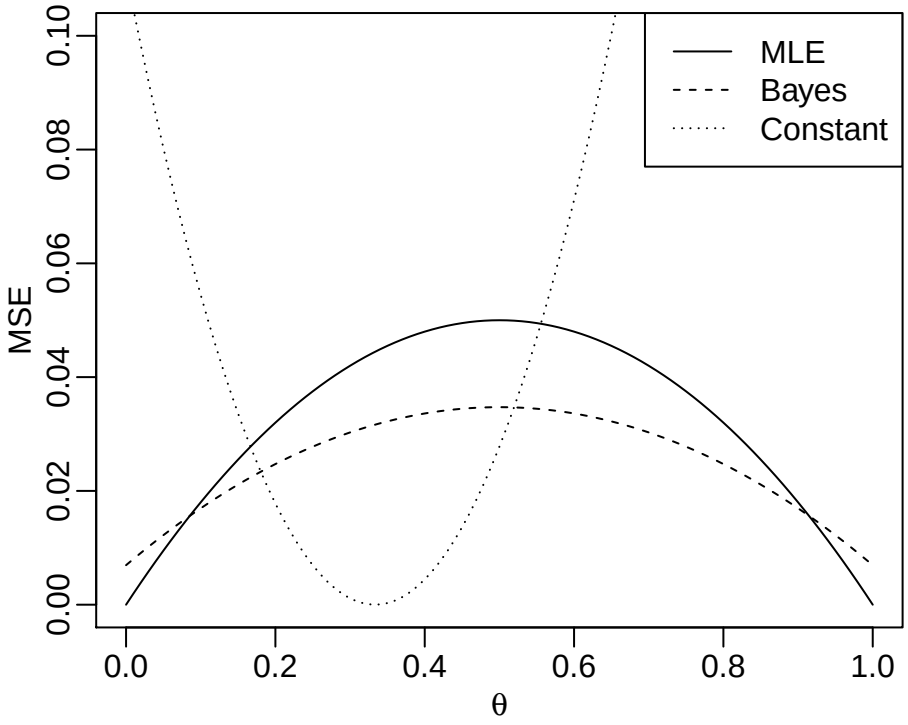
so its mean squared error is

$$\mathrm{MSE}_\theta\big(\hat{\theta}^B\big) = \big[\mathrm{Bias}_\theta\big(\hat{\theta}^B\big)\big]^2 + \mathrm{Var}_\theta\big(\hat{\theta}^B\big) = \frac{\big[(1-\theta)a - \theta b\big]^2 + n\theta(1-\theta)}{(n+a+b)^2}.$$

A rather stupid choice would be the constant estimator that ignores the data and just estimates c no matter what. This estimator has

$$\mathrm{Bias}_\theta(c) = E_\theta(c) - p = c - \theta, \quad \mathrm{Var}_\theta(c) = 0, \quad \mathrm{MSE}_\theta(c) = \big[\mathrm{Bias}_\theta(c)\big]^2 + \mathrm{Var}_\theta(c) = (c-\theta)^2.$$

The MSE of each estimator as a function of θ is plotted below in the case where $n = 5$. The Bayes estimator (posterior mean) uses $a = b = 1/2$. The constant estimator ignores the data and estimates $c = 1/3$ no matter what.



The plot shows the following:

- The Bayes estimator has smaller MSE than the maximum likelihood estimator unless the true parameter value θ is very close to 0 or 1. This illustrates the notion of a bias-variance trade-off. Although the Bayes estimator is biased (and the MLE is not), this bias allows a substantial reduction in variance.
- It is not difficult to see why the Bayes estimator is outperformed by the MLE when the true value of θ is close to 0 or 1. For $n = 5$ and $a = b = 1/2$ as shown in the plot, we are guaranteed to have $1/12 \le \hat{\theta}^B \le 11/12$ since $0 \le X \le 5$ no matter what.
- Note that the exact values of θ at which the curves for the MLE and Bayes estimator intersect are $(4 \pm \sqrt{11})/8$. It is not particularly surprising that these values are fairly close to 1/12 and 11/12.
- The constant estimator does very well if θ is actually near 1/3, but otherwise its performance can be very poor. ◊

We would observe similar results if we repeated this plot for other values of a, b, c, and n. ◊

Best Estimators

It is natural to ask whether we can find an estimator $\hat{\theta}$ of θ that has smaller mean squared error than every other estimator for $\theta \in \Theta$. However, no such estimator can exist. This conclusion is actually trivial, since the constant estimator $\hat{\theta} = c$ will always have smaller mean squared error than any other estimator if θ is actually equal to c. Thus, we must consider the idea of a "best" estimator in a narrower sense. There are two ways to do this:

- Take a weighted average of the MSE over all possible θ values, so that we can measure the performance of an estimator through a single number that takes into account all values of θ (rather than a function of θ). Then try to find the estimator that minimizes this "average MSE." It turns out that this is surprisingly easy, as we'll see.
- Restrict our attention to only estimators that meet a certain criterion, then try to find an estimator that is "best" (has lowest MSE for all θ) within this subset. The most common approach is to restrict our attention to unbiased estimators and try to find the *best unbiased estimator*.

The notion of average MSE optimality is addressed below, while the notion of best unbiased estimators is will be discussed later in the course.

Average MSE Optimality

Let $w(\theta)$ be a nonnegative weighting function that describes how much we want the various values of θ to "count" toward our weighted average MSE. Assume without loss of generality that $\int_\Theta w(\theta)\, d\theta = 1$ or $\sum_{\theta \in \Theta} w(\theta) = 1$ (whichever is appropriate). Then let

$$r_w\big(\hat{\theta}\big) = \int_\Theta \mathrm{MSE}_\theta\big(\hat{\theta}\big)\, w(\theta)\, d\theta \qquad \text{or} \qquad r_w\big(\hat{\theta}\big) = \sum_{\theta \in \Theta} \mathrm{MSE}_\theta\big(\hat{\theta}\big)\, w(\theta)$$

denote our weighted average MSE. The following theorem tells us how to minimize $r_w(\hat{\theta})$.

Theorem 5.2.6. *Let $\hat{\theta}^B$ denote the posterior mean of θ under the prior $\pi(\theta) = w(\theta)$. Then $r_w(\hat{\theta}^B) \le r_w(\hat{\theta})$ for any other estimator $\hat{\theta}$ of θ.*

Proof. We provide the proof for the case where the data and parameter are both continuous. (The proofs of the other cases are similar.) Let $f_\theta(\boldsymbol{x})$ be the joint pdf of the data, where $\boldsymbol{x} \in \mathbb{R}^n$, and let $\hat{\theta} = \hat{\theta}(\boldsymbol{x})$ be an estimator of θ other than $\hat{\theta}^B = \hat{\theta}^B(\boldsymbol{x})$. Then

$$\begin{aligned} r_w(\hat{\theta}) = \int_\Theta \mathrm{MSE}_\theta\big(\hat{\theta}\big)\, w(\theta)\, d\theta &= \int_\Theta E_\theta\Big[\big(\hat{\theta} - \theta\big)^2\Big]\, \pi(\theta)\, d\theta \\ &= \int_\Theta \left\{ \int_{\mathbb{R}^n} \big[\hat{\theta}(\boldsymbol{x}) - \theta\big]^2 f_\theta(\boldsymbol{x})\, d\boldsymbol{x} \right\} \pi(\theta)\, d\theta \\ &= \int_{\mathbb{R}^n} \left\{ \int_\Theta \big[\hat{\theta}(\boldsymbol{x}) - \theta\big]^2 \pi(\theta \mid \boldsymbol{x})\, d\theta \right\} m(\boldsymbol{x})\, d\boldsymbol{x}, \end{aligned}$$

noting that $f_\theta(\boldsymbol{x})\, \pi(\theta) = \pi(\theta \mid \boldsymbol{x})\, m(\boldsymbol{x})$. Now write the inner integral as

$$\begin{aligned} \int_\Theta \big[\hat{\theta}(\boldsymbol{x}) - \theta\big]^2 \pi(\theta \mid \boldsymbol{x})\, d\theta &= \int_\Theta \big[\hat{\theta}(\boldsymbol{x}) - \hat{\theta}^B(\boldsymbol{x}) + \hat{\theta}^B(\boldsymbol{x}) - \theta\big]^2 \pi(\theta \mid \boldsymbol{x})\, d\theta \\ &= \big[\hat{\theta}(\boldsymbol{x}) - \hat{\theta}^B(\boldsymbol{x})\big]^2 + 2\big[\hat{\theta}(\boldsymbol{x}) - \hat{\theta}^B(\boldsymbol{x})\big] \int_\Theta \big[\hat{\theta}^B(\boldsymbol{x}) - \theta\big] \pi(\theta \mid \boldsymbol{x})\, d\theta \\ &\quad + \int_\Theta \big[\hat{\theta}^B(\boldsymbol{x}) - \theta\big]^2 \pi(\theta \mid \boldsymbol{x})\, d\theta \\ &\ge \int_\Theta \big[\hat{\theta}^B(\boldsymbol{x}) - \theta\big]^2 \pi(\theta \mid \boldsymbol{x})\, d\theta, \end{aligned}$$

noting that $\int_\Theta [\hat{\theta}^B(\boldsymbol{x}) - \theta]\, \pi(\theta \mid \boldsymbol{x})\, d\theta = 0$. Then it follows that

$$r_w(\hat{\theta}) \ge \int_{\mathbb{R}^n} \left\{ \int_\Theta \big[\hat{\theta}^B(\boldsymbol{x}) - \theta\big]^2 \pi(\theta \mid \boldsymbol{x})\, d\theta \right\} m(\boldsymbol{x})\, d\boldsymbol{x} = \int_\Theta E_\theta\Big[\big(\hat{\theta} - \theta\big)^2\Big]\, \pi(\theta)\, d\theta = r_w(\hat{\theta}^B),$$

again noting that $f_\theta(\boldsymbol{x})\, \pi(\theta) = \pi(\theta \mid \boldsymbol{x})\, m(\boldsymbol{x})$. □

Note that although Theorem 5.2.6 involves the Bayes estimator and uses Bayesian notation in its proof, the result holds regardless of whether or not we actually believe in the Bayesian philosophy. Thus, we can still find Bayes estimators useful even if we are not willing to interpret them as a mean of some posterior distribution.

Example 5.2.7: In Example 5.2.5, suppose we want to find the estimator that minimizes a weighted average MSE with a weighting function of the form $w(\theta) = \theta^{c_1}(1-\theta)^{c_2}$, where $c_1 > -1$ and $c_2 > -1$ (to ensure that the integral of the weighting function is finite). Then $w(\theta)$, when multiplied by an appropriate constant, is the pdf of a Beta$(c_1 + 1,\, c_2 + 2)$ distribution. Then by Theorem 5.2.6, the estimator that minimizes the weighted average MSE under the weighting function $w(\theta)$ is simply the posterior mean of θ under a Beta$(c_1 + 1,\, c_2 + 1)$ prior, which is $\hat{\theta}^B = (X + c_1 + 1)/(n + c_1 + c_2 + 2)$. ◇

Lecture 6: Asymptotic Properties of Estimators

We now turn our attention to the limiting behavior of estimators as the sample size tends to infinity. This is referred to as the *asymptotic* behavior of an estimator.

6.1 Consistency

An estimator $\hat{\theta}_n$ of a parameter θ is *consistent* if $\hat{\theta}_n \to_P \theta$ for all θ in the parameter space Θ.

EXAMPLE 6.1.1: Suppose $\mu = E_\mu(X_1)$ is finite, and let $\overline{X}$ (more precisely, $\overline{X}_n$) be the usual sample mean of an iid sample $X_1, \ldots, X_n$.

- The estimator $\overline{X}_n$ is a consistent estimator of μ since $\overline{X}_n \to_P \mu$ (by the weak law of large numbers).
- The estimator $\overline{X}_n/2$ is not a consistent estimator of μ. since $\overline{X}_n/2 \to_P \mu/2$.
- The estimator $(n-1)\overline{X}_n/n$ is a consistent estimator of μ, despite the fact that its expectation is $E_\mu[(n-1)\overline{X}_n/n] = (n-1)\mu/n$ for each $n \geq 1$ (i.e., it is a biased estimator of μ for each $n \geq 1$).

In fact, if a_n is any sequence such that $a_n \to 1$, then $a_n\overline{X}_n$ is a consistent estimator of μ. $\diamond$

The following theorem can be helpful for showing consistency of an estimator.

Theorem 6.1.2. *If $E_\theta(\hat{\theta}_n) \to \theta$ and $\mathrm{Var}_\theta(\hat{\theta}_n) \to 0$ for all $\theta \in \Theta$, then $\hat{\theta}_n$ is a consistent estimator of θ.*

Proof. The proof uses Chebyshev's inequality (Theorem 6.2.2 of DeGroot & Schervish) and some simple manipulations involving the definition of convergence in probability. $\square$

Note that the conditions of Theorem 6.1.2 are *sufficient* conditions, not *necessary* conditions. Examples can be constructed in which an estimator is consistent despite the fact that the conditions of Theorem 6.1.2 fail.

Good Estimators Should Be Consistent

Consistency is perhaps the most basic property of a good estimator. Estimators derived using sensible statistical principles, such as maximum likelihood estimators and Bayes estimators, are usually consistent.

Summary of Regularity Conditions

A formal statement of the consistency of the maximum likelihood estimator requires the imposition of a variety of regularity conditions. A detailed list of one sufficient set of regularity conditions can be found in Section 6.4.

> Note: The conditions listed in Section 6.4 are intended as a single set of conditions to cover all results in this lecture. In fact, consistency of the MLE can be proved using slightly weaker conditions than those listed in Section 6.4.

For now, it will suffice to briefly summarize these regularity conditions as follows:

- The data $\boldsymbol{X} = (X_1, \ldots, X_n)$ is an iid sample with likelihood $L_{\boldsymbol{x}}(\theta) = \prod_{i=1}^{n} L_{x_i}(\theta)$.
- The parameter space Θ of the unknown parameter θ is an open subset (though not necessarily a proper subset) of the real line.
- The set $\mathcal{X} = \{x_1 \in \mathbb{R} : L_{x_1}(\theta) > 0\}$ does not depend on θ.
- If $L_{x_1}(\theta_1) = L_{x_1}(\theta_2)$ for almost all $x_1 \in \mathcal{X}$, then $\theta_1 = \theta_2$.
- The likelihood $L_{x_1}(\theta)$ must satisfy certain smoothness conditions as a function of θ.

Practically speaking, the most commonly encountered violation of these conditions is a situation where the set $\mathcal{X} = \{x_1 \in \mathbb{R} : L_{x_1}(\theta) > 0\}$ depends on θ (for example, if the distribution of the data is uniform over some interval that depends on θ). However, if all conditions are satisfied, then we have the following result.

Theorem 6.1.3. *Let $\hat{\theta}_n$ be the maximum likelihood estimator of θ based on the sample $\boldsymbol{X}_n = (X_1, \ldots, X_n)$. Then under the regularity conditions of Section 6.4, $\hat{\theta}_n$ is a consistent estimator of θ.*

Proof. The proof is beyond the scope of this course. □

> Note: The result of Theorem 6.1.3 actually follows as a corollary of Theorem 6.2.4, which will be presented in the next section. However, this logical implication is useless for actually *proving* Theorem 6.1.3 because the result of Theorem 6.1.3 will be used in the proof of Theorem 6.2.4.

6.2 Asymptotic Distribution of the MLE

Earlier in the course, we derived the asymptotic distribution of certain estimators by various ad hoc approaches. In particular, we have often exploited the central limit theorem and delta method in for estimators that can be expressed as some type of average from an iid sample. However, we now develop a much more general theory to address the asymptotic distribution of the maximum likelihood estimator.

Restrictions and Extensions

Some regularity conditions will still be required (see Section 6.4), so the results developed in this section will not apply to all maximum likelihood estimators. On the other hand, this basic approach is general enough to be extended to other M-estimators as well, though such extensions are beyond the scope of this course.

Derivatives of the Log-Likelihood

Throughout the lecture, we will write derivatives of the likelihood and log-likelihood as

$$L'_{\boldsymbol{X}}(\theta) = \frac{\partial}{\partial\theta} L_{\boldsymbol{X}}(\theta), \qquad L''_{\boldsymbol{X}}(\theta) = \frac{\partial^2}{\partial\theta^2} L_{\boldsymbol{X}}(\theta), \qquad \ell'_{\boldsymbol{X}}(\theta) = \frac{\partial}{\partial\theta} \ell_{\boldsymbol{X}}(\theta), \qquad \ell''_{\boldsymbol{X}}(\theta) = \frac{\partial^2}{\partial\theta^2} \ell_{\boldsymbol{X}}(\theta).$$

MLE Maximizes a Random Function

It is important to realize where the randomness in the distribution of the MLE comes from.

- The log-likelihood $\ell_{\boldsymbol{X}}(\theta)$ is a *random function* of θ. Different sample values $\boldsymbol{X} = \boldsymbol{x}$ lead to different log-likelihoods $\ell_{\boldsymbol{x}}(\theta)$, some of which are more probable than others.
- The MLE $\hat{\theta}$ is defined as the point at which this random function is maximized.
- Since the function is random, the point at which it is maximized is also random.

Log-Likelihoods for iid Samples are iid Sums

If $\boldsymbol{X} = (X_1, \ldots, X_n)$ is an iid sample, then the log-likelihood is $\ell_{\boldsymbol{X}}(\theta) = \sum_{i=1}^n \ell_{X_i}(\theta)$. Thus, for any particular θ, the log-likelihood itself is a sum of iid random variables. It follows that derivatives of the log-likelihood are also sums of iid random variables.

Score and Information

We now define two important functions based on the log-likelihood:

- The *score* or *score function* is simply $\ell'_{\boldsymbol{X}}(\theta)$, or equivalently, $\sum_{i=1}^n \ell'_{X_i}(\theta)$.
- The *information* or *Fisher information* is $I(\theta) = E_\theta\{[\ell'_{\boldsymbol{X}}(\theta)]^2\}$.

The following theorem provides some helpful results related to these quantities. It again requires some regularity conditions.

Lemma 6.2.1. *Under the regularity conditions of Section 6.4,* $E_\theta[\ell'_{\boldsymbol{X}}(\theta)] = 0$*, and*

$$I(\theta) = \mathrm{Var}_\theta[\ell'_{\boldsymbol{X}}(\theta)] = -E_\theta[\ell''_{\boldsymbol{X}}(\theta)] = -n\, E_\theta[\ell''_{X_1}(\theta)].$$

Proof. See the proof of Theorem 8.8.1 of DeGroot & Schervish for all but the last equality. For the last equality, simply note that $\ell''_{\boldsymbol{X}}(\theta) = \sum_{i=1}^n \ell''_{X_i}(\theta)$, and the terms in the sum are identically distributed. □

Lemma 6.2.1 provides the formula by which we usually calculate the Fisher information in practice, for two reasons:

- The second derivative may have fewer terms to consider than the first derivative.
- We only need to find a single expectation, as opposed to finding a variance or the expectation of a squared quantity.

The quantity $E_\theta[\ell''_{X_1}(\theta)]$ is often called $I_1(\theta)$, the *information per observation.*

Example 6.2.2: Let $X_1, \ldots, X_n \sim$ iid Poisson(λ), where $\lambda > 0$ is unknown. Suppose we want to find the Fisher information $I(\lambda)$. We find

$$\ell''_{X_1}(\lambda) = \frac{\partial^2}{\partial \lambda^2}[-\lambda + X_1 \log \lambda - \log(X_1!)] = \frac{\partial}{\partial \lambda}\left(-1 + \frac{X_1}{\lambda}\right) = -\frac{X_1}{\lambda^2},$$

so the information for the sample is $I(\lambda) = n\, I_1(\lambda) = -n\, E_\lambda(-X_1/\lambda^2) = n/\lambda$. ◊

EXAMPLE 6.2.3: Let $X_1, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ is unknown and $\sigma^2 > 0$ is known. Suppose we want to find the Fisher information $I(\mu)$. We find

$$\ell''_{X_1}(\mu) = \frac{\partial^2}{\partial \mu^2}\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(X_1-\mu)^2}{2\sigma^2}\right] = \frac{\partial}{\partial \mu}\left(\frac{X_1-\mu}{\sigma^2}\right) = -\frac{1}{\sigma^2},$$

so the information for the sample is simply $I(\mu) = n\, I_1(\mu) = -n\, E_\mu(-1/\sigma^2) = n/\sigma^2$. Note that in this example, the information $I(\mu)$ does not actually depend on μ. ◊

Main Result: Asymptotic Distribution of the MLE

The following result is among the most important in all of mathematical statistics. As with previous results in this lecture, it requires some regularity conditions.

Theorem 6.2.4. *Let $\hat{\theta}_n$ be the maximum likelihood estimator of θ based on the sample $\boldsymbol{X}_n = (X_1, \ldots, X_n)$. Then under the regularity conditions of Section 6.4,*

$$\sqrt{n}\big(\hat{\theta}_n - \theta\big) \to_D N\left[0,\, \frac{1}{I_1(\theta)}\right].$$

Proof. A fully rigorous proof is beyond the scope of this course, but we can still provide the basic idea. Begin with a Taylor expansion of $\ell'_{\boldsymbol{X}_n}(\hat{\theta}_n)$ around θ:

$$\ell'_{\boldsymbol{X}_n}(\hat{\theta}_n) = \ell_{\boldsymbol{X}_n}(\theta) + \big(\hat{\theta}_n - \theta\big)\ell''_{\boldsymbol{X}_n}(\theta) + \cdots,$$

where we can ignore the higher-order terms. (The justification of this claim is where most of the regularity conditions are used.) Now observe that the left-hand side is zero, so (neglecting the "$\cdots$" term entirely), we may rearrange and multiply by $\sqrt{n}$ to obtain

$$\sqrt{n}\big(\hat{\theta}_n - \theta\big) = -\sqrt{n}\left[\frac{\ell'_{\boldsymbol{X}_n}(\theta)}{\ell''_{\boldsymbol{X}_n}(\theta)}\right] = \frac{\sqrt{n}\left[\frac{1}{n}\ell'_{\boldsymbol{X}_n}(\theta) - 0\right]}{-\frac{1}{n}\ell''_{\boldsymbol{X}_n}(\theta)}.$$

Then by the central limit theorem,

$$\sqrt{n}\left[\frac{1}{n}\ell'_{\boldsymbol{X}_n}(\theta) - 0\right] = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\ell'_{X_i}(\theta) - 0\right] \to_D N[0,\, I_1(\theta)],$$

noting that $E_\theta[\ell'_{X_1}(\theta)] = 0$ and $\mathrm{Var}_\theta[\ell'_{X_1}(\theta)] = I_1(\theta)$ by Lemma 6.2.1. Next, observe that

$$-\frac{1}{n}\ell''_{\boldsymbol{X}_n}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\ell''_{X_i}(\theta) \to_P -E_\theta\big[\ell''_{X_i}(\theta)\big] = I_1(\theta)$$

by the weak law of large numbers. Then by Slutsky's theorem,

$$\sqrt{n}\big(\hat{\theta}_n - \theta\big) \to_D N\left[0,\, \frac{1}{I_1(\theta)}\right]$$

since the asymptotic variance is $I_1(\theta)/[I_1(\theta)]^{-2} = 1/I_1(\theta)$. □

The practical interpretation of Theorem 6.2.4 is that if n is large, then the distribution of the MLE $\hat{\theta}_n$ is approximately normal with mean θ (i.e., the true value) and variance $1/[n\,I_1(\theta)]$, or equivalently, $1/[I(\theta)]$.

> Note: It is intuitive to interpret asymptotic results as approximate results for large n, but such interpretations can occasionally produce incorrect statements when pushed too far. For instance, it is possible to construct a sequence of random variables Z_n such that $Z_n \to_D N(0,1)$ while $\mathrm{Var}(Z_n)$ does *not* converge to 1. In that case, it would be correct to say that the distribution of Z_n is approximately $N(0,1)$ when n is large, but it would be incorrect to say that $\mathrm{Var}(Z_n) \approx 1$ when n is large. Although theoretically possible, such seemingly self-contradictory situations are seldom encountered in practice, and the interpretation of asymptotic results as approximate results for large n is usually fairly safe.

As you might expect, we will exploit these asymptotic results heavily for inference procedures as the course continues.

EXAMPLE 6.2.5: Let $X_1, \ldots, X_n \sim$ iid Poisson(λ), where $\lambda > 0$ is unknown. We calculated in Example 6.2.2 that the Fisher information for the sample is $I(\lambda) = n/\lambda$. We also found in an earlier lecture that the MLE of λ is simply $\hat{\lambda}_n = \overline{X}_n$ (assuming that $\overline{X}_n > 0$, which is guaranteed to hold for large enough n).

> Note: This is where our assumption that the parameter space is an open interval becomes relevant. Since we have $\lambda > 0$ (strictly), we have
>
> $$\begin{aligned} P_\lambda\big(\overline{X}_n > 0 \text{ for some } n \geq 1\big) &= 1 - P_\lambda(X_n = 0 \text{ for all } n \geq 1) \\ &= 1 - \prod_{n=1}^{\infty} P_\lambda(X_i = 0) = 1 - \prod_{n=1}^{\infty} \exp(-\lambda) = 1 - 0 = 1, \end{aligned}$$
>
> so the MLE exists for sufficiently large n with probability 1.

Then by Theorem 6.2.4,

$$\sqrt{n}\big(\hat{\lambda}_n - \lambda\big) \to_D N\bigg(0, \frac{\lambda}{n}\bigg).$$

(Of course, we could have obtained the same result by the central limit theorem.) ◊

Asymptotic Distribution of Numerically Calculated MLEs

To see why Theorem 6.2.4 is useful, notice that it depends only on the regularity conditions and the fact that the MLE maximizes the likelihood. This fact means that Theorem 6.2.4 holds even in problems where we cannot find a closed-form solution for the MLE. Of course, if we cannot find a closed-form solution for the MLE, then it may also be difficult to calculate the Fisher information. There exist ways to get around this problem by using simpler quantities in place of the Fisher information, as we will see later in the course.

Asymptotic Distribution of Functions of MLEs

Recall that if $\hat{\theta}$ is an MLE of θ, then $g(\hat{\theta})$ is an MLE of $g(\theta)$. If g is continuously differentiable at the true value θ, then we can combine the delta method with Theorem 6.2.4 to obtain the asymptotic distribution of $g(\hat{\theta})$.

EXAMPLE 6.2.6: Let $X_1, \ldots, X_n \sim$ iid Poisson(λ), where $\lambda > 0$ is unknown. Note that each observation X_i has variance λ and hence standard deviation $\lambda^{1/2}$. The MLE of $\lambda^{1/2}$ is

$$\hat{\lambda}_n^{1/2} = \left(\overline{X}_n\right)^{1/2},$$

so we may apply the delta method to the result of Example 6.2.5 to obtain

$$\sqrt{n}\left(\hat{\lambda}_n^{1/2} - \lambda^{1/2}\right) \to_D N\left(0, \frac{1}{4n}\right),$$

noting that the derivative of the function $g(\lambda) = \lambda^{1/2}$ is $g'(\lambda) = \frac{1}{2}\lambda^{-1/2}$. ◊

Extension to Multiple Parameters

The result of Theorem 6.2.4 generalizes to the case of multiple unknown parameters. A full treatment of this topic is beyond the scope of the course, but we can still state the basic result. Suppose we write the unknown parameters as $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. First, the score function generalizes to a vector,

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\boldsymbol{X}}(\boldsymbol{\theta}) = \left(\frac{\partial}{\partial \theta_1} \ell_{\boldsymbol{X}}(\boldsymbol{\theta}) \quad \cdots \quad \frac{\partial}{\partial \theta_p} \ell_{\boldsymbol{X}}(\boldsymbol{\theta}) \right)^T.$$

Next, under certain generalized versions of the regularity conditions of Section 6.4, the Fisher information (now a $p \times p$ matrix) can be calculated as

$$I(\boldsymbol{\theta}) = \mathrm{Var}_{\boldsymbol{\theta}}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell_{\boldsymbol{X}}(\boldsymbol{\theta}) \right] = -E_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T} \ell_{\boldsymbol{X}}(\boldsymbol{\theta}) \right] = -n\, E_{\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T} \ell_{X_1}(\boldsymbol{\theta}) \right] = -n\, I_1(\boldsymbol{\theta}),$$

where

$$\frac{\partial^2}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^T} \ell_{\boldsymbol{X}}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial^2}{\partial \theta_1^2} \ell_{X_1}(\theta) & \cdots & \frac{\partial^2}{\partial \theta_1\, \partial \theta_p} \ell_{X_1}(\theta) \\ \vdots & & \vdots \\ \frac{\partial^2}{\partial \theta_1\, \partial \theta_p} \ell_{X_1}(\theta) & \cdots & \frac{\partial^2}{\partial \theta_p^2} \ell_{X_1}(\theta) \end{pmatrix}$$

is the matrix of second partial derivatives, sometimes called the Hessian. Then, again under suitable generalizations of the regularity conditions of Section 6.4, the MLE $\hat{\boldsymbol{\theta}}_n$ satisfies

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right) \to_D N_p\left\{\mathbf{0}_p,\ [I_1(\boldsymbol{\theta})]^{-1}\right\}.$$

> Note: You might notice that even if we are willing to overlook the regularity conditions, there is still a problem with the statement above: we have not defined convergence in distribution for random vectors. Thus, we will have to settle for understanding the result above at a more conceptual and imprecise level.

Again, the material here is presented only to provide some basic exposure to these more advanced multivariate concepts. We will not discuss them rigorously or in any further detail.

6.3 Asymptotic Efficiency

In some situations it may be possible to find the asymptotic distribution of estimators other than the MLE. Many sensible estimators $\tilde{\theta}_n$ of θ exhibit similar distributional convergence results to that of the MLE $\hat{\theta}_n$. Specifically, we often obtain a result of the form

$$\sqrt{n}\big(\tilde{\theta}_n - \theta\big) \to_D N[0,\, v(\theta)] \tag{6.3.1}$$

for some function $v(\theta)$, which can be called the *asymptotic variance* of $\tilde{\theta}_n$.

> Note: The factor of $\sqrt{n}$ in the convergence result above can introduce confusion over the term "asymptotic variance." Although we call $v(\theta)$ the asymptotic variance, the approximate distribution of $\tilde{\theta}_n$ for large n is normal with mean θ and variance $n^{-1}\, v(\theta)$.

The asymptotic variance provides another way to compare and evaluate estimators. Among estimators with convergence results of the form (6.3.1), we would usually prefer the estimator with the smallest asymptotic variance $v(\theta)$.

Asymptotic Relative Efficiency

We quantify this type of asymptotic variance comparison through a quantity called the *asymptotic relative efficiency* (ARE) of one estimator compared to another. If $\tilde{\theta}_n^{(1)}$ and $\tilde{\theta}_n^{(2)}$ are estimators of θ such that

$$\sqrt{n}\Big[\tilde{\theta}_n^{(1)} - \theta\Big] \to_D N\big[0,\, v^{(1)}(\theta)\big],$$
$$\sqrt{n}\Big[\tilde{\theta}_n^{(2)} - \theta\Big] \to_D N\big[0,\, v^{(2)}(\theta)\big],$$

then the asymptotic relative efficiency of $\tilde{\theta}_n^{(1)}$ compared to $\tilde{\theta}_n^{(2)}$ is

$$\mathrm{ARE}_\theta\Big[\tilde{\theta}_n^{(1)}, \tilde{\theta}_n^{(2)}\Big] = \frac{1/v^{(1)}(\theta)}{1/v^{(2)}(\theta)} = \frac{v^{(2)}(\theta)}{v^{(1)}(\theta)}.$$

Example 6.3.1: Let $X_1, \ldots, X_n \sim$ iid Poisson(λ), where $\lambda > 0$ is unknown. Suppose we plan to draw a new observation X_{new} sometime in the future, and we want to estimate $\zeta = P_\lambda(X_{\text{new}} = 0) = \exp(-\lambda)$. We know that the MLE of ζ is $\hat{\zeta}_n = \exp(-\hat{\lambda}_n) = \exp(-\overline{X}_n)$. By Theorem 6.2.4 and the delta method, we obtain

$$\sqrt{n}\big(\hat{\zeta}_n - \zeta\big) \to_D N[0,\, \lambda \exp(-2\lambda)].$$

However, another sensible estimator of ζ is the proportion of observations that are zero in the sample, i.e., $\tilde{\zeta}_n = n^{-1} \sum_{i=1}^n I_{\{0\}}(X_i)$. Note that $I_{\{0\}}(X_1), \ldots, I_{\{0\}}(X_n) \sim$ iid Bin$[1, \exp(-\lambda)]$, so

$$\sqrt{n}\big(\tilde{\zeta}_n - \zeta\big) \to_D N\{0,\, \exp(-\lambda)[1 - \exp(-\lambda)]\}.$$

Then the asymptotic relative efficiency of $\tilde{\zeta}_n$ compared to $\hat{\zeta}_n$ is

$$\mathrm{ARE}_\lambda\big(\tilde{\zeta}_n, \hat{\zeta}_n\big) = \frac{1/\{\exp(-\lambda)[1 - \exp(-\lambda)]\}}{1/[\lambda \exp(-2\lambda)]} = \frac{\lambda \exp(-2\lambda)}{\exp(-\lambda)[1 - \exp(-\lambda)]} = \frac{\lambda}{\exp(\lambda) - 1}.$$

Note that $\mathrm{ARE}_\lambda(\tilde{\zeta}_n, \hat{\zeta}_n) < 1$ for all $\lambda > 0$, meaning that $\tilde{\zeta}_n$ has larger asymptotic variance than the MLE $\hat{\zeta}_n$ for all parameter values. In fact, if λ is even moderately large, then the advantage of the MLE can be quite substantial, e.g., $\mathrm{ARE}_\lambda(\tilde{\zeta}_n, \hat{\zeta}_n) \approx 1/67$ if $\lambda = 5$. ⋄

Asymptotic relative efficiency can also be interpreted in terms of sample sizes. Suppose that $\text{ARE}_\theta[\tilde{\theta}_n^{(1)}, \tilde{\theta}_n^{(2)}] = 3$. Then the distribution of $\tilde{\theta}^{(1)}$ based on a sample of size n has approximately the same distribution as the distribution of $\tilde{\theta}^{(2)}$ based on a sample of size $3n$. In other words, an estimator that is three times as efficient as another (based on ARE) needs a sample size only a third as large as the other estimator in order to achieve approximately the same precision.

Asymptotic Efficiency

We might wish to go a step further than simply comparing two estimators using asymptotic relative efficiency. Specifically, we would like to know whether there exists an estimator that minimizes the asymptotic variance. The following theorem suggests an answer.

Theorem 6.3.2. *Let $\tilde{\theta}_n$ be an estimator such that (6.3.1) holds for some $v(\theta)$. Then under the regularity conditions of Section 6.4, $v(\theta) \geq [I_1(\theta)]^{-1}$.*

Proof. The proof is beyond the scope of this course. □

An estimator $\tilde{\theta}_n$ for which (6.3.1) holds with $v(\theta) = [I_1(\theta)]^{-1}$ (i.e., an estimator that attains the bound specified by Theorem 6.3.2) is called *asymptotically efficient*. Then the following result is immediately obvious by Theorem 6.2.4.

Corollary 6.3.3. *Let $\hat{\theta}_n$ be the maximum likelihood estimator of θ based on the sample $\boldsymbol{X}_n = (X_1, \ldots, X_n)$. Then under the regularity conditions of Section 6.4, $\hat{\theta}_n$ is asymptotically efficient.*

Estimators that are "close enough" to the MLE as $n \to \infty$ can also be asymptotically efficient. In particular, Bayes estimators are often asymptotically efficient as well.

EXAMPLE 6.3.4: Let $X_1, \ldots, X_n \sim$ iid Poisson(λ), where $\lambda > 0$ is unknown. It can be shown that the posterior mean of λ under a Gamma(a, b) prior is

$$\hat{\lambda}^B = \frac{a + \sum_{i=1}^n X_i}{b+n} = \left(\frac{n}{b+n}\right)\overline{X}_n + \left(\frac{b}{b+n}\right)\frac{a}{b}.$$

Now observe that

$$\begin{aligned}\sqrt{n}\big(\hat{\lambda}^B - \lambda\big) &= \sqrt{n}\left[\left(\frac{n}{b+n}\right)\overline{X}_n + \left(\frac{b}{b+n}\right)\frac{a}{b}\right] - \sqrt{n}\left[\left(\frac{n}{b+n}\right)\lambda + \left(\frac{b}{b+n}\right)\lambda\right] \\ &= \underbrace{\left(\frac{n}{b+n}\right)}_{\to 1}\underbrace{\sqrt{n}\big(\overline{X}_n - \lambda\big)}_{\to_D N\{0,[I_1(\lambda)]^{-1}\}} + \underbrace{\sqrt{n}\left(\frac{b}{b+n}\right)\left(\frac{a}{b} - \lambda\right)}_{\to 0} \to_D N\left[0, \frac{1}{I_1(\lambda)}\right]\end{aligned}$$

by Slutsky's theorem. Thus, $\hat{\lambda}^B$ is also asymptotically efficient. ◊

Uses of Asymptotically Inefficient Estimators

It might seem as though we should always use the MLE or another asymptotically efficient estimator, in which case the notion of asymptotic relative efficiency would be a bit pointless (since any two estimators of interest would have an ARE of 1). However, in practice, there may be compelling reasons to use (or at least to consider using) estimators that may not be asymptotically efficient.

- The MLE and other asymptotically efficient estimators might be too difficult to calculate, even numerically, while some other simpler estimator might exist.
- The MLE and other asymptotically efficient estimators might rely heavily on the assumptions underlying the model (e.g., that the observations are independent, or that their distribution is normal), and we may not trust these assumptions. We might instead prefer an estimator that performs reasonably well even if these assumptions are actually false. (We call such estimators *robust*.)

In these situations, the asymptotic relative efficiency quantifies how much we are "losing" or "gaining" in terms of efficiency by using one estimator instead of another if it turns out that all of the assumptions of the model actually are correct.

Example 6.3.5: Reconsider the estimators proposed in Example 6.3.1. Based on asymptotic relative efficiency alone, it might seem as though $\hat{\zeta}_n = \exp(-\overline{X}_n)$ is the superior choice. However, consider how this estimator performs if the distribution of $X_1, \ldots, X_n$ actually differs slightly from a Poisson(λ) distribution. Specifically, suppose that the actual distribution produces enormous values (e.g., $\lambda \cdot 10^6$) too often. These enormous values tend to cause $\overline{X}_n$ to also take enormous values, and thus $\hat{\zeta}_n = \exp(-\overline{X}_n)$ tends underestimate ζ quite badly. On the other hand, $\tilde{\zeta}_n$ still performs reasonable well as an estimator of ζ. The asymptotic relative efficiency as calculated in Example 6.3.1 tells us that the efficiency of $\tilde{\zeta}_n$ relative to $\hat{\zeta}_n$ drops off rapidly as λ increases. Thus, the larger the true value of λ is, the more a switch from $\hat{\zeta}_n$ to $\tilde{\zeta}_n$ hurts us if it turns out that all of the assumptions of the model are correct after all. ♢

6.4 Regularity Conditions for Earlier Results

The following regularity conditions are sufficient for

- Theorem 6.1.3 (consistency of the MLE),
- Lemma 6.2.1 (results on the score and information),
- Theorem 6.2.4 (asymptotic distribution of the MLE),
- Theorem 6.3.2 (lower bound for asymptotic variance), and
- Corollary 6.3.3 (asymptotic efficiency of the MLE).

Note: Some of the results above (especially consistency of the MLE) can actually be proven under slightly weaker conditions. The conditions below are intended as a single "universal" set of conditions that are sufficient for all of the results above to hold.

The conditions are as follows:

- The data $\boldsymbol{X} = (X_1, \ldots, X_n)$ is an iid sample with likelihood $L_{\boldsymbol{x}}(\theta) = \prod_{i=1}^n L_{x_i}(\theta)$.
- The parameter space Θ of the unknown parameter θ is an open subset (though not necessarily a proper subset) of the real line.
- The set $\mathcal{X} = \{x_1 \in \mathbb{R} : L_{x_1}(\theta) > 0\}$ does not depend on θ.
- If $L_{x_1}(\theta_1) = L_{x_1}(\theta_2)$ for almost all $x_1 \in \mathcal{X}$, then $\theta_1 = \theta_2$.

  Note: When we say a statement holds for "almost all $x_1 \in \mathcal{X}$," we mean that there exists a set $\mathcal{X}^\star \subseteq \mathcal{X}$ such that the statement holds for all $x_1 \in \mathcal{X}^\star$ and $P_\theta(X_1 \in \mathcal{X}^\star) = 1$ for all $\theta \in \Theta$.

- The likelihood $L_{\boldsymbol{x}}(\theta) = \prod_{i=1}^n L_{X_i}(\theta)$ is differentiable three times in θ with continuous third derivative, i.e., $L'''_{\boldsymbol{x}}(\theta)$ exists and is continuous for all $\theta \in \Theta$ and almost all $\boldsymbol{x} \in \mathcal{X}^n$.

  Note: This condition and the previous condition imply that the log-likelihood $\ell_{\boldsymbol{x}}(\theta) = \log L_{\boldsymbol{x}}(\theta)$ is also differentiable three times in θ with continuous third derivative.

- Depending on whether $\boldsymbol{X}$ is continuous or discrete, either

$$\int_{\mathcal{X}^n} L'''_{\boldsymbol{x}}(\theta)\, d\boldsymbol{x} = \frac{d^3}{d\theta^3} \int_{\mathcal{X}^n} L_{\boldsymbol{x}}(\theta)\, d\boldsymbol{x} \qquad \text{or} \qquad \sum_{\boldsymbol{x} \in \mathcal{X}^n} L'''_{\boldsymbol{x}}(\theta) = \frac{d^3}{d\theta^3} \sum_{\boldsymbol{x} \in \mathcal{X}^n} L_{\boldsymbol{x}}(\theta).$$

  Note that the right-hand side of either equation is equal to $(d^3/d\theta^3)(1) = 0$.

- For any $\theta_0 \in \Theta$, there exists $\delta_{\theta_0} > 0$ and a function $M_{\theta_0}(x)$ such that

$$\left|\ell'''_{x_1}(\theta)\right| \le M_{\theta_0}(x_1) \quad \text{for almost all } x_1 \in \mathcal{X} \text{ and all } \theta \in (\theta_0 - \delta_{\theta_0},\, \theta_0 + \delta_{\theta_0}),$$

  and $E_{\theta_0}[M_{\theta_0}(X_1)] < \infty$.

The conditions above are *sufficient* but not *necessary*. Many other sets of sufficient conditions can also be obtained.

Note: A quick survey of five different textbooks yields five slightly different sets of sufficient conditions for the results in this lecture. The conditions here are taken from Section 10.6.2 of *Statistical Inference* by George Casella and Roger L. Berger.

Lecture 7: More on Unbiased Estimators

Although the use of unbiased estimators is less emphasized in modern statistics than it was in the past, the property of unbiasedness leads to several interesting theoretical results for this class of estimators.

7.1 UMVUEs and the Cramér-Rao Inequality

Recall that in general there does not exist an estimator with smaller MSE than all other estimators for all values of the unknown parameter. However, it may be possible to find an estimator that is "best" (in the sense of smallest MSE) among some class of estimators. Note that if $\tilde{\theta}$ is an unbiased estimator of θ, then $\mathrm{MSE}_\theta(\tilde{\theta}) = [\mathrm{Bias}_\theta(\tilde{\theta})]^2 + \mathrm{Var}_\theta(\tilde{\theta}) = \mathrm{Var}_\theta(\tilde{\theta})$. Thus, finding an unbiased estimator with smallest MSE for all θ in the parameter space Θ is equivalent to finding an unbiased estimator with smallest variance for all $\theta \in \Theta$. If $\tilde{\theta}^\star$ is an unbiased estimator of θ such that $\mathrm{Var}_\theta(\tilde{\theta}^\star) \le \mathrm{Var}_\theta(\tilde{\theta})$ for all $\theta \in \Theta$ for all other unbiased estimators $\tilde{\theta}$ of θ, then $\tilde{\theta}^\star$ is called a *uniformly minimum–variance unbiased estimator* (UMVUE) of θ.

> Note: A UMVUE is an unbiased estimator for which the variance is uniformly minimum (among all unbiased estimators). Some people assume the uniform part to be implicitly understood and therefore use the term *minimum-variance unbiased estimator* (MVUE). Other people simply use the term *best unbiased estimator* (BUE), but such terminology is less descriptive since the notion of "best" could perhaps be interpreted in other ways.

It may not be immediately clear how to find a UMVUE or how to determine whether a particular unbiased estimator is a UMVUE. Some strategies will be discussed in this lecture, but there also exist other methods that are beyond the scope of this course.

Cramér-Rao Inequality

The following result provides a lower bound for the variance of any unbiased estimator of a function of a parameter. It requires some regularity conditions, and the conditions listed in Section 6.4 of Lecture 6 are again sufficient.

Theorem 7.1.1 (Cramér-Rao Inequality)**.** *Let $\tilde{\xi} = \tilde{\xi}(\boldsymbol{X})$ be an unbiased estimator of $\xi = g(\theta)$ based on the sample $\boldsymbol{X} = (X_1, \ldots, X_n)$, where $g : \Theta \to \mathbb{R}$ is continuously differentiable at θ with derivative $g'(\theta)$. Then under the regularity conditions of Section 6.4,*

$$\mathrm{Var}_\theta(\tilde{\xi}) \ge \frac{[g'(\theta)]^2}{I(\theta)} = \frac{[g'(\theta)]^2}{n\, I_1(\theta)} \quad \text{for all } \theta \in \Theta.$$

Proof. The covariance between the random variables $\tilde{\xi} = \tilde{\xi}(\boldsymbol{X})$ and $\ell'_{\boldsymbol{X}}(\theta)$ is

$$\mathrm{Cov}_\theta\big[\tilde{\xi},\, \ell'_{\boldsymbol{X}}(\theta)\big] = E_\theta\big[\tilde{\xi}\, \ell'_{\boldsymbol{X}}(\theta)\big] - E_\theta(\tilde{\xi})\, E[\ell'_{\boldsymbol{X}}(\theta)] = E_\theta\big[\tilde{\xi}\, \ell'_{\boldsymbol{X}}(\theta)\big],$$

noting that $E_\theta[\ell'_{\boldsymbol{X}}(\theta)] = 0$ by Lemma 6.2.1. Now observe that $\ell'_{\boldsymbol{X}}(\theta) = L'_{\boldsymbol{X}}(\theta)/L_{\boldsymbol{X}}(\theta)$, so

$$\mathrm{Cov}_\theta\big[\tilde{\xi},\, \ell'_{\boldsymbol{X}}(\theta)\big] = E_\theta\left[\tilde{\xi}\, \frac{L'_{\boldsymbol{X}}(\theta)}{L_{\boldsymbol{X}}(\theta)}\right] = \begin{cases} \int_{\mathbb{R}^n} \tilde{\xi}(\boldsymbol{x})\, L'_{\boldsymbol{x}}(\theta)\, d\boldsymbol{x} & \text{if } \boldsymbol{X} \text{ is continuous,} \\ \sum_{\boldsymbol{x} \in \mathcal{X}^n} \tilde{\xi}(\boldsymbol{x})\, L'_{\boldsymbol{x}}(\theta) & \text{if } \boldsymbol{X} \text{ is discrete,} \end{cases}$$

noting that $[L'_{\boldsymbol{x}}(\theta)/L_{\boldsymbol{x}}(\theta)]L_{\boldsymbol{x}}(\theta) = L'_{\boldsymbol{x}}(\theta)$. Then due to the regularity conditions, we have

$$\int_{\mathbb{R}^n} \tilde{\xi}(\boldsymbol{x})\, L'_{\boldsymbol{x}}(\theta)\, d\boldsymbol{x} = \int_{\mathbb{R}^n} \frac{\partial}{\partial\theta}\big[\tilde{\xi}(\boldsymbol{x})\, L_{\boldsymbol{x}}(\theta)\big]\, d\boldsymbol{x} = \frac{d}{d\theta}\int_{\mathbb{R}^n} \tilde{\xi}(\boldsymbol{x})\, L_{\boldsymbol{x}}(\theta)\, d\boldsymbol{x} = \frac{d}{d\theta}E_\theta(\tilde{\xi}) = g'(\theta),$$

$$\sum_{\boldsymbol{x}\in\mathcal{X}^n} \tilde{\xi}(\boldsymbol{x})\, L'_{\boldsymbol{x}}(\theta) = \sum_{\boldsymbol{x}\in\mathcal{X}^n} \frac{\partial}{\partial\theta}\big[\tilde{\xi}(\boldsymbol{x})\, L_{\boldsymbol{x}}(\theta)\big] = \frac{d}{d\theta}\sum_{\boldsymbol{x}\in\mathcal{X}^n} \tilde{\xi}(\boldsymbol{x})\, L_{\boldsymbol{x}}(\theta) = \frac{d}{d\theta}E_\theta(\tilde{\xi}) = g'(\theta).$$

Thus, we have $\mathrm{Cov}_\theta[\tilde{\xi},\, \ell'_{\boldsymbol{X}}(\theta)] = g'(\theta)$ in both the continuous and discrete cases. Then

$$\Big|g'(\theta)\Big| = \Big|\mathrm{Cov}_\theta\big[\tilde{\xi},\, \ell'_{\boldsymbol{X}}(\theta)\big]\Big| \le \sqrt{\mathrm{Var}_\theta(\tilde{\xi})\ \mathrm{Var}[\ell'_{\boldsymbol{X}}(\theta)]} = \sqrt{\mathrm{Var}_\theta(\tilde{\xi})\, I(\theta)},$$

where the inequality is the Cauchy-Schwarz inequality. The result follows immediately. □

> Note: The result of Theorem 7.1.1 is sometimes called the *information inequality*. Also, some people refer to the special case below as the Cramér-Rao inequality instead.

The following corollary addresses the special case of unbiased estimators of θ itself.

Corollary 7.1.2 (Cramér-Rao Inequality, Special Case)**.** *Let $\tilde{\theta}$ be an unbiased estimator of θ based on the sample $\boldsymbol{X} = (X_1, \ldots, X_n)$. Then under the regularity conditions of Section 6.4,*

$$\mathrm{Var}_\theta(\tilde{\theta}) \ge \frac{1}{I(\theta)} = \frac{1}{n\, I_1(\theta)} \qquad \textit{for all } \theta \in \Theta.$$

The Cramér-Rao inequality provides an obvious way of identifying a UMVUE: if the variance of an unbiased estimator attains the lower bound stated in the inequality (for all $\theta \in \Theta$), then the estimator is a UMVUE.

Example 7.1.3: Let $X_1, \ldots, X_n \sim$ iid Poisson(λ), where $\lambda > 0$ is unknown. We calculated in Example 6.2.2 of Lecture 6 that the Fisher information for the sample is $I(\lambda) = n/\lambda$. Now observe that $\hat{\lambda} = \overline{X}$ is clearly an unbiased estimator of λ, with variance

$$\mathrm{Var}_\lambda\big(\overline{X}\big) = \frac{\lambda}{n} = \frac{1}{I(\lambda)},$$

which attains the Cramér-Rao lower bound. Thus, $\hat{\lambda} = \overline{X}$ is a UMVUE of λ. ◊

The Cramér-Rao inequality may appear quite similar to Theorem 6.3.2 from Lecture 6, which states that if $\tilde{\theta}_n$ is a sequence of estimators such that $\sqrt{n}(\tilde{\theta}_n - \theta) \to_D N[0,\, v(\theta)]$, then $v(\theta) \ge [I(\theta)]^{-1}$. The difference is that the Cramér-Rao inequality holds for all n, i.e., it is *not* just an asymptotic result.

- Theorem 6.3.2 is "stronger" than the Cramér-Rao inequality in that Theorem 6.3.2 covers all estimators for which $\sqrt{n}(\tilde{\theta}_n - \theta)$ converges to a normal distribution centered at zero, including biased estimators.
- However, the Cramér-Rao inequality is "stronger" than Theorem 6.3.2 in that the bound provided by the Cramér-Rao inequality is an actual bound for the variance for any finite n, not just a bound on the asymptotic variance.

Thus, neither of the two results is stronger than the other.

Attainment of the Cramér-Rao Lower Bound

Depending on the situation, there may or may not exist an unbiased estimator that attains the Cramér-Rao lower bound. The following result provides a necessary and sufficient condition for an unbiased estimator to attain this bound.

Theorem 7.1.4. *Let $\tilde{\xi} = \tilde{\xi}(\boldsymbol{X})$ be an unbiased estimator of $\xi = g(\theta)$ based on the sample $\boldsymbol{X} = (X_1, \ldots, X_n)$, where $g : \Theta \to \mathbb{R}$ is continuously differentiable at θ with derivative $g'(\theta)$. Then under the regularity conditions of Section 6.4,*

$$\mathrm{Var}_\theta(\tilde{\xi}) = \frac{[g'(\theta)]^2}{I(\theta)} = \frac{[g'(\theta)]^2}{n\, I_1(\theta)} \quad \text{for all } \theta \in \Theta$$

if and only if there exists a function $b : \Theta \to \mathbb{R}$ such that $\tilde{\xi}(\boldsymbol{X}) = g(\theta) + b(\theta)\, \ell'_{\boldsymbol{X}}(\theta)$ with probability 1 for all $\theta \in \Theta$.

Proof. Note from the proof of Theorem 7.1.1 that $\mathrm{Var}_\theta(\tilde{\xi}) = [g'(\theta)]^2 / I(\theta)$ if and only if

$$\left| \mathrm{Cov}_\theta\big[\tilde{\xi},\, \ell'_{\boldsymbol{X}}(\theta)\big] \right| = \sqrt{\mathrm{Var}_\theta(\tilde{\xi})\ \mathrm{Var}[\ell'_{\boldsymbol{X}}(\theta)]}.$$

For any particular θ, this equality holds if and only if $\tilde{\xi}(\boldsymbol{X}) = a + b\, \ell'_{\boldsymbol{X}}(\theta)$ with probability 1 for some constants a and b. Hence, the equality holds for all $\theta \in \Theta$ if and only if there exist functions $a : \Theta \to \mathbb{R}$ and $b : \Theta \to \mathbb{R}$ such that $\tilde{\xi}(\boldsymbol{X}) = a(\theta) + b(\theta)\, \ell'_{\boldsymbol{X}}(\theta)$ with probability 1 for all $\theta \in \Theta$. Then since $\tilde{\xi}$ is unbiased,

$$g(\theta) = \xi = E_\theta(\tilde{\xi}) = a(\theta) + b(\theta)\, E_\theta[\ell'_{\boldsymbol{X}}(\theta)] = a(\theta)$$

by Lemma 6.2.1. Thus, $a(\theta) = g(\theta)$. □

Theorem 7.1.4 sometimes implies that no unbiased estimator of a particular parameter (or of a particular function of a parameter) attains the Cramér-Rao lower bound.

Example 7.1.5: Let $X_1, \ldots, X_n \sim$ iid Exp(λ), where $\lambda > 0$ is unknown. Then

$$\ell'_{\boldsymbol{X}}(\lambda) = \sum_{i=1}^n \ell'_{X_i}(\lambda) = \sum_{i=1}^n \left(\frac{1}{\lambda} - X_i \right) = \frac{n}{\lambda} - \sum_{i=1}^n X_i.$$

Now consider any unbiased estimator $\tilde{\lambda}$ of λ. The variance of $\tilde{\lambda}$ attains the Cramér-Rao lower bound if and only if there exists a function $b(\lambda)$ such that

$$\tilde{\lambda}(\boldsymbol{X}) = \lambda + b(\lambda) \left(\frac{n}{\lambda} - \sum_{i=1}^n X_i \right).$$

However, it is clear that no such function $b(\lambda)$ exists since $\tilde{\lambda}(\boldsymbol{X})$ cannot depend on λ. Thus, no unbiased estimator of λ attains the Cramér-Rao lower bound. ♢

UMVUEs and Non-Attainment of the Cramér-Rao Lower Bound

Note that attainment of the Cramér-Rao lower bound is a sufficient condition for an unbiased estimator to be a UMVUE. However, it is not a necessary condition.

- If an unbiased estimator attains the Cramér-Rao lower bound, then it is a UMVUE.
- However, in cases where no unbiased estimator attains the Cramér-Rao lower bound, a UMVUE can still exist, though we would need to use more sophisticated techniques to prove that its variance is indeed uniformly minimum.

EXAMPLE 7.1.6: In Example 7.1.5, even though no unbiased estimator of λ attains the Cramér-Rao lower bound, it can be shown using more sophisticated techniques that

$$\tilde{\lambda} = \frac{n-1}{\sum_{i=1}^{n} X_i} = \frac{n-1}{n\overline{X}}$$

is a UMVUE of λ. (In fact, it can also be shown that it is the unique UMVUE of λ.) ◊

7.2 Sufficiency and the Rao-Blackwell Theorem

Recall that the sufficiency principle states that statistical inference about a parameter θ should be based on a sufficient statistic for θ. Maximum likelihood estimators and Bayes estimators are always functions of a sufficient statistic, but such estimators may not be unbiased. It may not be immediately obvious how to construct a good unbiased estimator in certain situations.

EXAMPLE 7.2.1: Let $X_1, \ldots, X_n \sim$ iid Poisson(λ), where $\lambda > 0$ is unknown. Suppose we plan to draw a new observation X_{new} sometime in the future, and we want to estimate $\zeta = P_\lambda(X_{\text{new}} = 0) = \exp(-\lambda)$. We know that the MLE of ζ is $\hat{\zeta} = \exp(-\hat{\lambda}^{\text{MLE}}) = \exp(-\overline{X})$, but this is not an unbiased estimator of ζ. Alternatively, the estimator $\tilde{\zeta} = n^{-1}\sum_{i=1}^{n} I_{\{0\}}(X_i)$ (i.e., the sample proportion of observations that are zero) is clearly unbiased for ζ, but it is not clear if this is a UMVUE of ζ or if there exists some better unbiased estimator of ζ. (Intuition suggests that $\tilde{\zeta}$ is perhaps not a great estimator of ζ since it simply treats each observation as either "zero" or "not zero.") ◊

The following theorem can help with the construction of good unbiased estimators.

Theorem 7.2.2 (Rao-Blackwell)**.** *Let $\tilde{\xi} = \tilde{\xi}(\boldsymbol{X})$ be an unbiased estimator of $\xi = g(\theta)$, where the parameter space is Θ and $g : \Theta \to \mathbb{R}$ is a function. Let $\boldsymbol{Y} = r(\boldsymbol{X})$ be a sufficient statistic for θ, and let $\tilde{\xi}^\star = E(\tilde{\xi} \mid \boldsymbol{Y})$. Then $\tilde{\xi}^\star$ is an unbiased estimator of ξ, and $\text{Var}_\theta(\tilde{\xi}^\star) \le \text{Var}_\theta(\tilde{\xi})$ for all $\theta \in \Theta$.*

Proof. First note that $\tilde{\xi}^\star$ is unbiased since $E_\theta(\tilde{\xi}^\star) = E_\theta[E(\tilde{\xi} \mid \boldsymbol{Y})] = E_\theta(\tilde{\xi}) = \xi$, where the second equality is by the law of total expectation. Next, by the law of total variance,

$$\text{Var}_\theta(\tilde{\xi}) = E_\theta\big[\text{Var}(\tilde{\xi} \mid \boldsymbol{Y})\big] + \text{Var}_\theta\big[E(\tilde{\xi} \mid \boldsymbol{Y})\big] \ge \text{Var}_\theta\big[E(\tilde{\xi} \mid \boldsymbol{Y})\big] = \text{Var}_\theta(\tilde{\xi}^\star)$$

for all $\theta \in \Theta$. □

There are several things to note about the Rao-Blackwell theorem:

- No regularity conditions are required.
- It may appear as though sufficiency (for θ) of the statistic $\boldsymbol{Y}$ is not needed for the result to hold. However, if the statistic $\boldsymbol{Y}$ is not sufficient, then $E(\tilde{\xi} \mid \boldsymbol{Y})$ may depend on θ, in which case it is not an estimator at all. (This is also why we can skip writing the subscript θ on the conditional expectation.)
- In addition to having a (possibly) improved variance, the estimator $\tilde{\xi}^\star$ is also a function of the sufficient statistic $\boldsymbol{Y}$.
- If the original estimator $\tilde{\xi}$ is already a function of the sufficient statistic $\boldsymbol{Y}$, then the Rao-Blackwell theorem is not helpful since $\tilde{\xi}^\star$ will simply be the same estimator as $\tilde{\xi}$.

The Rao-Blackwell theorem can provide a way to find a good unbiased estimator in situations where it otherwise may not be clear how to construct one. If we can find *any* unbiased estimator of the unknown quantity of interest, then we can apply the Rao-Blackwell theorem to obtain a better unbiased estimator.

> Note: The process of applying the Rao-Blackwell theorem to a naïve unbiased estimator to obtain a better unbiased estimator is sometimes called *Rao-Blackwellization*.

Example 7.2.3: In Example 7.2.1, we stated that an unbiased estimator of $\zeta = \exp(-\lambda)$ is $\tilde{\zeta} = n^{-1}\sum_{i=1}^n I_{\{0\}}(X_i)$. However, consider the even simpler unbiased estimator defined by

$$\tilde{\zeta}_1 = I_{\{0\}}(X_1) = \begin{cases} 0 & \text{if } X_1 = 0, \\ 1 & \text{if } X_1 > 0. \end{cases}$$

Suppose we now apply the Rao-Blackwell theorem to the unbiased estimator $\tilde{\zeta}_1$ with the sufficient statistic $Y = \sum_{i=1}^n X_i$. Then

$$\begin{aligned} E\big(\tilde{\zeta}_1 \,\big|\, Y = y\big) &= 0 \cdot P\big(\tilde{\zeta}_1 = 0 \,\big|\, Y = y\big) + 1 \cdot P\big(\tilde{\zeta}_1 = 1 \,\big|\, Y = y\big) \\ &= P\big(\tilde{\zeta}_1 = 1 \,\big|\, Y = y\big) \\ &= P(X_1 > 0 \mid \textstyle\sum_{i=1}^n X_i = y) \\ &= 1 - P(X_1 = 0 \mid \textstyle\sum_{i=1}^n X_i = y) \\ &= 1 - \frac{P_\lambda(X_1 = 0,\ \sum_{i=1}^n X_i = y)}{P_\lambda(\sum_{i=1}^n X_i = y)} \\ &= 1 - \frac{P_\lambda(X_1 = 0,\ \sum_{i=2}^n X_i = y)}{P_\lambda(\sum_{i=1}^n X_i = y)} \\ &= 1 - \frac{P_\lambda(X_1 = 0)\ P_\lambda(\sum_{i=2}^n X_i = y)}{P_\lambda(\sum_{i=1}^n X_i = y)}. \end{aligned}$$

Now note that by basic properties of the Poisson distribution, $\sum_{i=1}^n X_i \sim \text{Poisson}(n\lambda)$ and $\sum_{i=2}^n X_i \sim \text{Poisson}[(n-1)\lambda]$. Then we have

$$E\big(\tilde{\zeta}_1 \,\big|\, Y = y\big) = 1 - \frac{\{\exp(-\lambda)\}\{[(n-1)\lambda]^y (y!)^{-1} \exp[-(n-1)\lambda]\}}{(n\lambda)^y (y!)^{-1} \exp(-n\lambda)} = \left(\frac{n-1}{n}\right)^y = \left(1 - \frac{1}{n}\right)^y.$$

Thus, a better unbiased estimator of $\zeta = \exp(-\lambda)$ is

$$\tilde{\zeta}^\star = \left(1 - \frac{1}{n}\right)^{\sum_{i=1}^n X_i}.$$

As an interesting observation, note that

$$\log \tilde{\zeta}^\star = \left(\sum_{i=1}^n X_i\right) \log\left(1 - \frac{1}{n}\right) = \overline{X}_n \log\left[\left(1 - \frac{1}{n}\right)^n\right] \to_P -\lambda$$

since $\overline{X}_n \to_P \lambda$ by the weak law of large numbers and $\log[(1 - \frac{1}{n})^n] \to \log(\frac{1}{e}) = -1$. Then $\tilde{\zeta}^\star \to_P \exp(-\lambda) = \zeta$, so $\tilde{\zeta}^\star$ is a consistent estimator of ζ as well. $\diamondsuit$

Lecture 8: Introduction to Hypothesis Testing

It is often the case that we wish to use data to make a binary decision about some unknown aspect of nature. For example, we may wish to decide whether or not it is plausible that a parameter takes some particular value. A frequentist approach to using data to make such decisions is *hypothesis testing*, also called *significance testing*.

> Note: There exist Bayesian counterparts of frequentist hypothesis tests, but the two philosophies differ more substantially for these types of binary decisions than for estimation problems.

8.1 Basic Structure of Hypothesis Tests

A hypothesis test consists of two hypothesis and a rejection region. The rejection region may be specified via a test statistic and a critical value. We define each of these terms below.

Hypotheses

A *hypothesis* is any statement about an unknown aspect of a distribution. In a hypothesis test, we have two hypotheses:

- H_0, the *null* hypothesis, and
- H_1, the *alternative* hypothesis.

Often a hypothesis is stated in terms of the value of one or more unknown parameters, in which case it is called a *parametric hypothesis*. Specifically, suppose we have an unknown parameter θ. Then parametric hypotheses about θ can be written in general as $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, where Θ_0 and Θ_1 are disjoint, i.e., $\Theta_0 \cap \Theta_1 = \varnothing$. We will typically assume hypotheses to be parametric unless clearly stated otherwise.

Example 8.1.1: Let $\mu \in \mathbb{R}$ be an unknown population mean. Parametric hypotheses about θ could be $H_0 : \mu \leq 2$ and $H_1 : \mu > 2$. A different set of parametric hypotheses could be $H_0 : \mu = 2$ and $H_1 : \mu \neq 2$. ◇

Hypotheses can be further classified as simple or composite.

- A hypothesis is *simple* if it fully specifies the distribution of the data (including all unknown parameter values). A parametric hypothesis is simple if it states specific values for *all* unknown parameters.
- A hypothesis is *composite* if it is not simple.

Note that taking both hypotheses to be simple is equivalent to allowing only two possible values for the unknown parameter θ, which is often unrealistic in practice. Thus, at least one hypothesis is typically composite, and sometimes both hypotheses are composite. (If only one hypothesis is composite, it is usually the alternative hypothesis H_1, for reasons that will become clear later.)

EXAMPLE 8.1.2: Let $X_1, \ldots, X_n \sim \text{iid } N(\mu, \sigma^2)$, and consider various sets of hypotheses:

- $H_0 : \mu = 40$ versus $H_1 : \mu = 45$, with σ^2 known. H_0 and H_1 are both simple.
- $H_0 : \mu = 40$ versus $H_1 : \mu \neq 40$, with σ^2 known. H_0 is simple, and H_1 is composite.
- $H_0 : \mu = 40$ versus $H_1 : \mu \neq 40$, with σ^2 unknown. H_0 and H_1 are both composite.
- $H_0 : \mu \leq 40$ versus $H_1 : \mu > 40$. H_0 and H_1 are both composite.
- $H_0 : (\mu, \sigma^2) = (40, 9)$ versus $H_1 : (\mu, \sigma^2) \neq (40, 9)$. H_0 is simple, and H_1 is composite.

Note that if σ^2 is unknown, any hypothesis that does not specify its value is composite. ◊

Rejection Region

A test of hypotheses H_0 and H_1 based on data $\boldsymbol{X}$ is a rule of the form

$$\text{Reject } H_0 \text{ (in favor of } H_1\text{) if and only if } \boldsymbol{X} \in R,$$

where R is a subset of the sample space $\mathcal{S}$. This set R is called the *rejection region*.

> Note: When we do not reject H_0, we typically simply say that we fail to reject H_0. Some people prefer to say instead that we accept H_0. However, the underlying theory is unaffected by which semantic interpretation we prefer.

EXAMPLE 8.1.3: Let $X \sim \text{Bin}(n, \theta)$, where $0 < \theta < 1$, and consider testing $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$. Perhaps the simplest nontrivial test of these hypotheses is to reject H_0 if and only if the trials are all successes or all failures, i.e., if and only if $X = 0$ or $X = n$. Then the rejection region is $R = \{0, n\}$. ◊

Essentially, a hypothesis test *is* its rejection region, in the sense that two tests of the same hypotheses based on the same data are identical tests if and only if they have the same rejection region.

Test Statistic

It is common to write the rejection region R in the form

$$R = \{\boldsymbol{x} \in \mathcal{S} : T(\boldsymbol{x}) \geq c\}, \tag{8.1.1}$$

where T is a real-valued function of the data and $c \in \mathbb{R}$.

- $T(\boldsymbol{X})$ is called the *test statistic*.
- c is called the *critical value*.

Different values of c yield different rejection regions, which we write as R_c.

> Note: Any rejection region R can be written in this form, since we can trivially take $T(\boldsymbol{x}) = I_R(\boldsymbol{x})$ and $c = 1$ (though we usually prefer to choose a test statistic that is less trivial). Thus, in particular, rejection regions of the form $\{\boldsymbol{x} \in \mathcal{S} : \tilde{T}(\boldsymbol{x}) > \tilde{c}\}$ can always be rewritten in the form of (8.1.1) for some T and c.

EXAMPLE 8.1.4: Let $X \sim \text{Bin}(n, \theta)$, where $0 < \theta < 1$, and consider testing $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$. A simple test of these hypotheses is to reject H_0 if and only if X/n is far enough from $1/2$. Then we could state the test statistic and rejection region as

$$T(X) = \left| \frac{X}{n} - \frac{1}{2} \right|, \qquad R_c = \{X \in \mathcal{S} : T(X) \geq c\},$$

where the sample space is $\mathcal{S} = \{0, 1, \ldots, n\}$ and $c \in \mathbb{R}$. ♢

Good and Bad Hypothesis Tests (and Non-Tests)

Every subset of the sample space can be a rejection region, and every rejection region corresponds to a hypothesis test. However, not all such hypothesis tests are actually sensible.

- A good hypothesis test should be more likely to reject H_0 if it is actually false than if it is actually true.
- Mathematically, the rejection region R corresponds to a sensible test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_1$ if $P_\theta(\boldsymbol{X} \in R)$ tends to be higher for $\theta \in \Theta_1$ than for $\theta \in \Theta_0$.
- A perfect hypothesis test would have $P_\theta(\boldsymbol{X} \in R)$ equal to 0 or 1 according to whether $\theta \in \Theta_0$ or $\theta \in \Theta_1$, respectively. However, this is typically impossible to achieve.

The probability $P_\theta(\boldsymbol{X} \in R)$, which is a function of θ, will be given a name in Section 8.2.

EXAMPLE 8.1.5: Let $X \sim \text{Bin}(n, \theta)$, where $0 < \theta < 1$, and consider testing $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$. Clearly the hypothesis tests proposed in Example 8.1.3 and Example 8.1.4 are good since X is more likely to fall in the rejection region if $\theta \neq 1/2$ than if $\theta = 1/2$. ♢

EXAMPLE 8.1.6: Let $X \sim \text{Bin}(n, \theta)$, where $0 < \theta < 1$, and consider testing $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$. A legal hypothesis test is simply to always reject H_0. The rejection region of this test is $\{0, 1, \ldots, n\}$, the entire sample space. Another legal hypothesis test is simply to never reject H_0. The rejection region of this test is $\varnothing$. However, these two hypothesis tests are obviously a waste of time. ♢

EXAMPLE 8.1.7: Let $X \sim \text{Bin}(n, \theta)$, where $0 < \theta < 1$, and consider testing $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$. Suppose we take the test statistic to be $T(X) = X$ and reject H_0 if and only if $X \geq c$. This is a legal hypothesis test. However, it is not a good test of these hypotheses since $P_\theta(X \geq c)$ is smaller for $\theta < 1/2$ than for $\theta = 1/2$. (Note, however, that it would be a good test of $H_0 : \theta = 1/2$ versus $H_1 : \theta > 1/2$.) ♢

EXAMPLE 8.1.8: Let $X \sim \text{Bin}(n, \theta)$, where $0 < \theta < 1$, and consider testing $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$. The seemingly perfect "test" that rejects H_0 if and only if $\theta \neq 1/2$ is not a hypothesis test at all, since it does not specify a rejection region as a subset of the sample space. (It specifies a rule in terms of the parameter value itself, which of course is impossible to apply in practice since the parameter value is unknown.) ♢

8.2 Properties of Hypothesis Tests

We now discuss basic properties of hypothesis tests in a probabilistic context. Remember that hypothesis tests as discussed here are a fundamentally frequentist concept, so probabilities discussed here are calculated as if the true parameter value is fixed but unknown.

Type I and Type II Errors

Since a perfect hypothesis test is typically impossible, there is some probability that our test will make the wrong decision.

- A *type I error* occurs if we reject H_0 when it is true, i.e., if $\theta \in \Theta_0$ and $\boldsymbol{X} \in R$.
- A *type II error* occurs if we fail to reject H_0 when it is false, i.e., if $\theta \in \Theta_1$ and $\boldsymbol{X} \notin R$.

The following table of possibilities may be helpful:

Truth	Data	Decision	Outcome
$H_0 : \theta \in \Theta_0$	$\boldsymbol{X} \notin R$	Fail to Reject H_0	Correct Decision
$H_0 : \theta \in \Theta_0$	$\boldsymbol{X} \in R$	Reject H_0	Type I Error
$H_1 : \theta \in \Theta_1$	$\boldsymbol{X} \notin R$	Fail to Reject H_0	Type II Error
$H_1 : \theta \in \Theta_1$	$\boldsymbol{X} \in R$	Reject H_0	Correct Decision

Of course, in reality we would not know whether a decision is correct or is an error. However, we can still consider the probability of each type of error.

- If $\theta \in \Theta_0$, then the probability of a type I error is $P_\theta(\boldsymbol{X} \in R)$.
- If $\theta \in \Theta_1$, then the probability of a type II error is $P_\theta(\boldsymbol{X} \notin R) = 1 - P_\theta(\boldsymbol{X} \in R)$.

The true value of θ is unknown, but these probabilities can be calculated for each possible θ.

Power Function

The *power function* of a hypothesis test with rejection region R is $\text{Power}(\theta) = P_\theta(\boldsymbol{X} \in R)$.

> Note: We will write $\text{Power}(\theta)$ to avoid any notational confusion, but be aware that this notation is nonstandard. Our textbook uses $\pi(\theta)$ for the power function, while another textbook uses $\beta(\theta)$. The latter choice is particularly confusing since many people instead use β to denote the probability of a type II error.

Notice that the power function provides the probabilities of both error types:

$$\text{Power}(\theta) = P_\theta(\boldsymbol{X} \in R) = \begin{cases} P_\theta(\text{type I error}) & \text{if } \theta \in \Theta_0, \\ 1 - P_\theta(\text{type II error}) & \text{if } \theta \in \Theta_1. \end{cases}$$

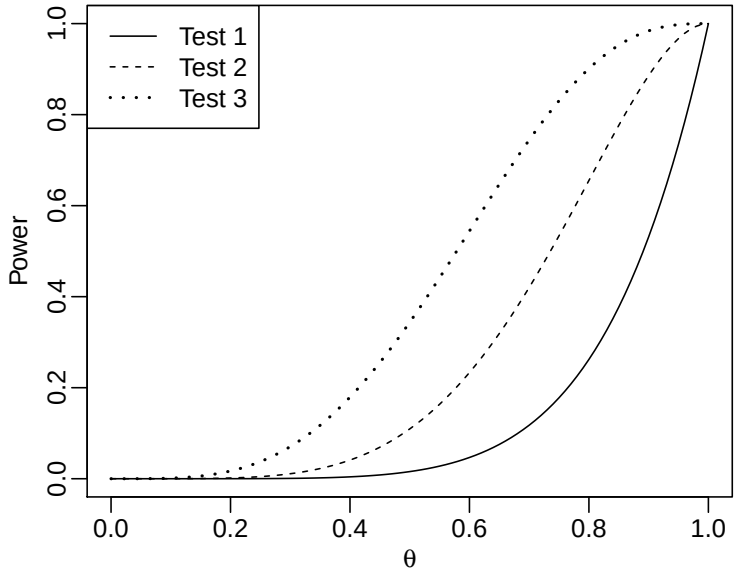
> Note: When people use the word "power" in the context of hypothesis tests, they usually mean $1 - P_\theta(\text{type II error})$, i.e., they mean the values of $\text{Power}(\theta)$ for $\theta \in \Theta_1$. The definition of the power function above is simply the logical extension to $\theta \in \Theta_0$ as well. Note, however, that it is actually *bad* if $\text{Power}(\theta)$ is large for $\theta \in \Theta_0$.

The "perfect" power function would be $\text{Power}(\theta) = I(\theta \in \Theta_1)$, but we know this is typically impossible since it corresponds to a "perfect" hypothesis test. More practically, we want $\text{Power}(\theta)$ to be small for $\theta \in \Theta_0$ and large for $\theta \in \Theta_1$.

EXAMPLE 8.2.1: Let $X \sim \text{Bin}(6, \theta)$, where $0 < \theta < 1$ and consider testing $H_0 : \theta \leq 1/2$ versus $H_1 : \theta > 1/2$ using one of the following three hypothesis tests:

- Test 1: Reject H_0 if and only if $X = 6$. The power function of this hypothesis test is $\text{Power}_1(\theta) = \theta^6$.
- Test 2: Reject H_0 if and only if $X \geq 5$. The power function of this hypothesis test is $\text{Power}_2(\theta) = \theta^6 + 6\theta^5(1-\theta) = \theta^5(6 - 5\theta)$.
- Test 3: Reject H_0 if and only if $X \geq 4$. The power function of this hypothesis test is $\text{Power}_3(\theta) = \theta^6 + 6\theta^5(1-\theta) + 15\,\theta^4(1-\theta)^2 = \theta^4(15 - 24\,\theta + 10\,\theta^2)$.

These functions are plotted below.



From the plot, it is easy to see the following:

- $\text{Power}_1(\theta)$ is very small for all $\theta \leq 1/2$, which is good. However, $\text{Power}_1(\theta)$ is still fairly small for most of the $\theta > 1/2$ region as well, which is not good.
- $\text{Power}_3(\theta)$ is fairly large for most of the $\theta > 1/2$ region, which is good. However, $\text{Power}_3(\theta)$ can be reasonably large even when $\theta \leq 1/2$, which is not so good.
- $\text{Power}_2(\theta)$ is in between $\text{Power}_1(\theta)$ and $\text{Power}_3(\theta)$.

Another way to think about this plot is as follows:

- Test 1 makes the fewest type I errors, while Test 3 makes the most (for all θ).
- Test 3 makes the fewest type II errors, while Test 1 makes the most (for all θ).

Which of these three tests is the "best" is a purely subjective question, which depends on the relative importance of type I and type II errors in the problem at hand. ♢

Error Trade-Off

In Example 8.2.1, all three hypothesis tests used the same test statistic and differed only in the choice of the critical value. When comparing a collection of tests of this form (i.e., when considering what to take as the critical value), there is always a trade-off of type I and type II errors.

- Increasing c tends to decrease $P_\theta(\boldsymbol{X} \in R_c) = P_\theta[T(\boldsymbol{X}) \geq c]$ for all θ. This decreases the probability of a type I error but increases the probability of a type II error.
- Decreasing c tends to increase $P_\theta(\boldsymbol{X} \in R_c) = P_\theta[T(\boldsymbol{X}) \geq c]$ for all θ. This decreases the probability of a type II error but increases the probability of a type I error.

However, when comparing hypothesis tests with *different* test statistics, it may be the case that one test outperforms the other in terms of both type I error and type II error.

Significance Levels and Sizes

The most common strategy (by far) is to fix some maximum probability of a type I error and to then try to find a test that has the smallest possible type II error probability subject to this constraint. This leads to the following terminology.

- A *level* (or *significance level*) of a test is any $\alpha \in \mathbb{R}$ such that $\text{Power}(\theta) \leq \alpha$ for all $\theta \in \Theta_0$. Thus, a level of a test is simply any upper bound for its type I error probability.
- The *size* of a test is $\sup_{\theta \in \Theta_0} \text{Power}(\theta)$. Thus, the size of a test is the smallest number that is a level of the test.

When possible, we usually try to report sizes and levels in such a way that the terms are interchangeable. In other words, when stating a level of the test, we usually state the size if it is known, even though larger values would also be levels. Similarly, when asked to find a test with a specified level α, we usually try to find a test with size α, even though tests with smaller sizes would also have level α.

Example 8.2.2: Consider again the three hypothesis tests of Example 8.2.1. Since the power function of each test is an increasing function of θ, we have

$$\sup_{0<\theta\leq 1/2} \text{Power}_j(\theta) = \text{Power}_j(1/2)$$

for each $j \in \{1, 2, 3\}$. Thus, we have the following:

- The size of Test 1 is $\sup_{\theta \in \Theta_0} \text{Power}(\theta) = \text{Power}_1(1/2) \approx 0.016$.
- The size of Test 2 is $\sup_{\theta \in \Theta_0} \text{Power}(\theta) = \text{Power}_2(1/2) \approx 0.109$.
- The size of Test 3 is $\sup_{\theta \in \Theta_0} \text{Power}(\theta) = \text{Power}_3(1/2) \approx 0.344$.

Note that in each case, the size of the test is also a level of the test. However, any number greater than the size is also a level of the test. ♢

Thus, fixing a maximum probability of a type I error is equivalent to specifying a level. In the next section, we will discuss how to actually construct hypothesis tests that have a specified level (either exactly or approximately).

8.3 Critical Values and Significance Levels

Suppose we have a test of the hypotheses $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$ that rejects H_0 if and only if $T(\boldsymbol{X}) \geq c$ for some test statistic $T(\boldsymbol{X})$ and some critical value c. We often wish to choose c so that the test will have a specified significance level α (such as $\alpha = 0.05$). The test has level α if

$$P_\theta[T(\boldsymbol{X}) \geq c] \leq \alpha \quad \text{for all } \theta \in \Theta_0,$$

so our goal is to find c such that this is the case.

Distribution of the Test Statistic

To work with $P_\theta[T(\boldsymbol{X}) \geq c]$, we need to know the distribution of the test statistic for every value of θ (or at least for every $\theta \in \Theta_0$). For this reason, we often choose a test statistic $T(\boldsymbol{X})$ that has some "standard" distribution (e.g., standard normal, Student's t, or chi-squared) when $\theta \in \Theta_0$.

EXAMPLE 8.3.1: Let $X_1, \ldots, X_n \sim \text{iid } N(\mu, \sigma^2)$, where $\sigma^2 > 0$ is known, and suppose we want to test $H_0 : \mu = 5$ versus $H_1 : \mu \neq 5$. We might take our test statistic to be

$$T(\boldsymbol{X}) = \frac{\left|\overline{X}_n - 5\right|}{\sqrt{\sigma^2/n}}$$

because this test statistic has the same distribution has the absolute value of a $N(0,1)$ random variable if H_0 is true, i.e., if $\mu = 5$. An equivalent test could be obtained by taking the test statistic to be any monotonically increasing function of the test statistic above, but such a test statistic might have a more complicated distribution. ♢

Suppose we know the distribution of $T(\boldsymbol{X})$ for each $\theta \in \Theta_0$. Let $F_\theta^{(T)}(t)$ denote the cdf of this distribution. Then our test has level α if and only if $P_\theta[T(\boldsymbol{X}) < c] \geq 1 - \alpha$ for all $\theta \in \Theta_0$, which holds if and only if

$$F_\theta^{(T)}(c) - P_\theta[T(\boldsymbol{X}) = c] \geq 1 - \alpha \quad \text{for all } \theta \in \Theta_0.$$

Simple Null Hypothesis

Suppose that our null hypothesis is $H_0 : \theta = \theta_0$ (i.e., suppose that $\Theta_0 = \{\theta_0\}$). Then our test has level α if and only if

$$F_{\theta_0}^{(T)}(c) - P_{\theta_0}[T(\boldsymbol{X}) = c] \geq 1 - \alpha.$$

Note that as long as $0 < \alpha \leq 1$, we can always find a value of c to satisfy this inequality since the left-hand side is a nondecreasing function of c that tends to 0 as $c \to -\infty$ and tends to 1 as $c \to \infty$.

Achieving a Specified Size

Now suppose that we wish to construct a test with *size* α (and suppose our null hypothesis is still $H_0 : \theta = \theta_0$. Our test has size α if and only if

$$F_{\theta_0}^{(T)}(c) - P_{\theta_0}[T(\boldsymbol{X}) = c] = 1 - \alpha.$$

It may or may not be possible to find such a test.

- If the distribution of $T(\boldsymbol{X})$ is continuous, then the equation above reduces to

  $$F_{\theta_0}^{(T)}(c) = 1 - \alpha.$$

  Since $T(\boldsymbol{X})$ is continuous, its cdf $F_{\theta_0}^{(T)}$ is continuous, and hence there always exists a value of c that satisfies this equation (as long as $0 < \alpha < 1$).

- If instead the distribution of $T(\boldsymbol{X})$ is discrete, then the expression

  $$F_{\theta_0}^{(T)}(c) - P_{\theta_0}[T(\boldsymbol{X}) = c]$$

  is no longer continuous as a function of c. Then there may or may not exist a value of c for which this expression is equal to $1 - \alpha$. If no such c exists, then there does not exist a test with size α based on the test statistic $T(\boldsymbol{X})$. In this case, we would typically try to find a test with size less than α (so that it still has *level* α) but as close to α as possible.

Example 8.3.2: In Example 8.3.1, we can obtain a test with size α by taking the critical value c to be the number such that $P(|Z| \geq c) = \alpha$ for a standard normal random variable Z. (For $\alpha = 0.05$, this is $c \approx 1.96$. For $\alpha = 0.10$, this is $c \approx 1.64$.) Any larger value of c would also yield a test with level α, but the size of such a test would be smaller than α. ♢

Composite Null Hypothesis

If the null hypothesis is composite, then it may not as easy to achieve a test with a specified level or size based on a particular test statistic $T(\boldsymbol{X})$. However, sometimes we find that

$$\sup_{\theta \in \Theta_0} P_\theta[T(\boldsymbol{X}) \geq c] = P_{\theta^\star}[T(\boldsymbol{X}) \geq c] \quad \text{for all } c \in \mathbb{R}$$

for some $\theta^\star \in \Theta_0 \cup \Theta_1$. (Often $\theta^\star$ is on the boundary of Θ_0.) Then we can proceed as if the set Θ_0 were instead simply $\{\theta^\star\}$, i.e., as if the null hypothesis were simply $H_0 : \theta = \theta^\star$.

Example 8.3.3: In Example 8.2.1 and Example 8.2.2,

$$\sup_{0 < \theta \leq 1/2} P_\theta(X \geq c) = P_{\theta=1/2}(X \geq c)$$

for all $c \in \mathbb{R}$ (which was why the sizes of the tests in Example 8.2.2 could be computed by evaluating the power function at $\theta = 1/2$). Then since the distribution of X is discrete, a test with size exactly α only exists for certain values of α. For example, there does not exist a test of this form with size 0.05. If we were asked to find a test with *level* 0.05, we could choose Test 1, which rejects H_0 if and only if $X = 6$. This test has *size* $1/64 \approx 0.016$, so 0.05 is indeed a level of this test. ♢

8.4 P-Values

The choice of the size or level of a test is typically subjective. This subjectivity can be somewhat unsatisfying, since two different people can reach opposite conclusions from the same data and the same test statistic simply because they chose to use different sizes or levels (and hence different critical values).

EXAMPLE 8.4.1: In Example 8.3.1 and Example 8.3.2, we considered a test that rejects H_0 if and only if the test statistic exceeds the number c such that $P(|Z| \geq c) = \alpha$ for a standard normal random variable Z. Suppose one person uses $\alpha = 0.05$ and $c \approx 1.96$, while another person uses $\alpha = 0.10$ and $c \approx 1.64$. Now suppose the observed test statistic value is 1.76. Then the first person will fail to reject H_0, while the second person will reject H_0. ◊

Thus, if we simply report whether or not we rejected H_0 at a certain level α, then we have somewhat oversimplified the conclusions that can be drawn from the data. A more informative way to report the conclusions of a hypothesis test is by stating a quantity called the p-value. Let $T(\boldsymbol{X})$ be a test statistic, and suppose we observe $\boldsymbol{X} = \boldsymbol{x}_{\text{obs}}$. Then the *p-value* of the test for the data $\boldsymbol{x}_{\text{obs}}$ is

$$p(\boldsymbol{x}_{\text{obs}}) = \sup_{\theta \in \Theta_0} P_\theta[T(\boldsymbol{X}) \geq T(\boldsymbol{x}_{\text{obs}})].$$

For a simple null hypothesis $H_0 : \theta = \theta_0$, the p-value reduces to

$$p(\boldsymbol{x}_{\text{obs}}) = P_{\theta_0}[T(\boldsymbol{X}) \geq T(\boldsymbol{x}_{\text{obs}})].$$

Thus, the p-value is the probability (under H_0) of observing a test statistic value at least as large as the one that actually was observed. The following theorem shows why the p-value is useful.

Theorem 8.4.2. *Let R_c be a rejection region of the form $R_c = \{\boldsymbol{x} : T(\boldsymbol{x}) \geq c\}$, where c is the smallest number such that the test associated with R_c has level α. Then $\boldsymbol{x}_{\text{obs}} \in R_c$ if and only if $p(\boldsymbol{x}_{\text{obs}}) \leq \alpha$.*

Proof. Suppose that $\boldsymbol{x}_{\text{obs}} \in R_c$. Then $T(\boldsymbol{x}_{\text{obs}}) \geq c$, so

$$p(\boldsymbol{x}_{\text{obs}}) = \sup_{\theta \in \Theta_0} P_\theta[T(\boldsymbol{X}) \geq T(\boldsymbol{x}_{\text{obs}})] \leq \sup_{\theta \in \Theta_0} P_\theta[T(\boldsymbol{X}) \geq c] \leq \alpha$$

since the test has level α. Now suppose instead that $\boldsymbol{x}_{\text{obs}} \notin R_c$. Then $T(\boldsymbol{x}_{\text{obs}}) < c$, so

$$p(\boldsymbol{x}_{\text{obs}}) = \sup_{\theta \in \Theta_0} P_\theta[T(\boldsymbol{X}) \geq T(\boldsymbol{x}_{\text{obs}})] > \alpha$$

since otherwise c would not be the smallest number such that the test associated with R_c has level α. □

Thus, Theorem 8.4.2 tells us that an equivalent way to make the final decision in a hypothesis test is to calculate the p-value $p(\boldsymbol{x}_{\text{obs}})$ for the observed data $\boldsymbol{x}_{\text{obs}}$ and reject H_0 at level α if and only if $p(\boldsymbol{x}_{\text{obs}}) \leq \alpha$. For this reason, the p-value is sometimes called the *observed significance level*.

EXAMPLE 8.4.3: In Example 8.4.1, the observed test statistic value 1.76 has p-value

$$p(1.76) = P(|Z| \geq 1.76) \approx 0.078,$$

where Z is a standard normal random variable. ◊

8.5 Logical Problems with Hypothesis Testing

Frequentist hypothesis testing has been an immensely popular tool of statistical inference for decades. However, there do exist scenarios in which hypothesis tests exhibit illogical behavior that some people (especially Bayesians) consider unacceptable.

Example 8.5.1: Suppose we wish to test whether a particular coin is fair or weighted in favor of heads. Then our hypotheses are $H_0 : \theta = 1/2$ and $H_1 : \theta > 1/2$, where θ denotes the probability that the coin yields heads on any given flip. Now suppose we are told that the following sequence of flips was observed (in order):

heads, heads, heads, heads, heads, tails.

There is some ambiguity here about how we should represent the data as a random variable.

- Perhaps the person flipping the coin decided to flip the coin repeatedly until obtaining tails. Let X be the number of times heads is observed for such an experiment before the first tails. Then $X \sim \text{Geometric}(\theta)$, and a sensible hypothesis test is to reject H_0 if and only if $X \geq c$ for some c. The observed value of X was $X = 5$, so the p-value is
$$p(5) = P_{\theta=1/2}(X \geq 5) = \frac{1}{32} \approx 0.031.$$
- Perhaps the person flipping the coin instead decided to flip the coin six times and record the results. Let X be the number of times heads is observed for such an experiment. Then $X \sim \text{Bin}(6, \theta)$, and a sensible hypothesis test is to reject H_0 if and only if $X \geq c$ for some c. The observed value of X was $X = 5$, so the p-value is
$$p(5) = P_{\theta=1/2}(X \geq 5) = \frac{7}{64} \approx 0.109.$$

Thus, the two different representations yield very different p-values and would therefore lead to opposite conclusions at both $\alpha = 0.05$ and $\alpha = 0.10$. This is troubling since there is no clear reason to prefer either representation over the other. Essentially, the result of our hypothesis test depends on knowing what the experimenter would have done under circumstances that are already known not to have occurred (e.g., whether the experimenter would have stopped flipping had tails occurred earlier than the sixth flip). ♢

Example 8.5.2: A researcher visits a lab and is allowed to use Machine A to conduct some measurements. These measurements are then used to perform a hypothesis test and reach a conclusion. However, the researcher later learns that the lab actually had two similar machines of this type (Machine A and Machine B), that another researcher also visited the lab the same day, and that the two machines were assigned to the two researchers randomly. Also, the machines are not identical: Machine A is a better model and hence provides more accurate measurements than Machine B. Although these new facts do not change the researcher's data or test statistic, they *do* change the distribution of that test statistic, which must instead be calculated as if there were probability 1/2 of using Machine A and probability 1/2 of using Machine B. Thus, the outcome of the hypothesis test can be altered even after the data has been collected by the mere existence of Machine B and the fact that it could have been used instead, even though it is already known that it was not used. ♢

EXAMPLE 8.5.3: Suppose a certain voltage θ is to be measured using a voltmeter for which the readings are iid $N(\theta, \sigma^2)$ random variables, where $\sigma^2 > 0$ is known. The sample mean is computed, and a hypothesis test is performed. However, it is later learned that the voltmeter had a maximum reading of 10 V, and any reading that otherwise would have been greater than 10 V would have instead been given as 10 V. This fact changes the distribution of the test statistic and could thus alter the outcome of the hypothesis test. Note that this change occurs even if all of the readings are less than 10 V, i.e., even if it is already known that the maximum did not actually matter. ♢

Source of the Issues

These types of examples seem to contradict common sense. The issue arises because the various probabilistic notions involved in hypothesis testing all involve summing or integrating over the entire sample space, i.e., over all possible data values that could have been observed. Thus, the results of the test can be affected by what would have happened for data values that did not actually occur. Note that this issue applies to frequentist inference in general, not just hypothesis testing. The same issues can also arise when calculating properties of estimators such as bias.

EXAMPLE 8.5.4: In Example 8.5.3, the existence of a maximum reading for the voltmeter would also affect the bias of the the sample mean. Note that the sample mean is still an unbiased estimator of the true mean of each reading on the voltmeter. However, the true mean of each reading on the voltmeter is now slightly less than the true voltage θ. ♢

These examples also highlight the differences between frequentist and Bayesian inference.

- Frequentist inference conditions on parameter values and integrates/sums over all possible data values that could be observed.
- Bayesian inference conditions on the observed data values and integrates/sums over all possible values of the parameter.

Thus, the issues that arise in the examples in this section do not arise in Bayesian inference. Since Bayesian methods are conditional on the data that is actually observed, they are unaffected by what could have happened for data values that did not actually occur.

Lecture 9: Likelihood Ratio Tests

Until now, we have considered problems where an appropriate test statistic can be chosen by common sense. However, we may encounter problems in which it is not clear what test statistic to use. Our first general method for finding tests is based on the likelihood function $L_{\boldsymbol{X}}(\theta)$ on the sets Θ_0 and Θ_1.

Likelihood Ratio Statistic

Let $\Theta = \Theta_0 \cup \Theta_1$. The *likelihood ratio statistic* $\Lambda(\boldsymbol{X})$ is defined as

$$\Lambda(\boldsymbol{X}) = \frac{\sup_{\theta \in \Theta_0} L_{\boldsymbol{X}}(\theta)}{\sup_{\theta \in \Theta} L_{\boldsymbol{X}}(\theta)},$$

and the *likelihood ratio test* rejects H_0 if and only if $[\Lambda(\boldsymbol{X})]^{-1} \geq c$, or equivalently, $\Lambda(\boldsymbol{X}) \leq k$, where $c \in (0,1)$ or $k = c^{-1} \in (0,1)$ is chosen to specify the level of the test.

> Note: You may wonder why we have written the likelihood ratio statistic $\Lambda(\boldsymbol{X})$ in such a way that we are forced to consider its inverse $[\Lambda(\boldsymbol{X})]^{-1}$ as the actual test statistic. In particular, it may seem as though we could simply reverse the numerator and denominator to avoid the problem altogether. Such an alteration to the definition would indeed work just fine. However, the definition of $\Lambda(\boldsymbol{X})$ above is fairly universal, so it would not be a good idea to stray too far from accepted conventions.

Note from the definition that $0 \leq \Lambda(\boldsymbol{X}) \leq 1$.

Simple Null Hypothesis

The likelihood ratio statistic can be written in a more convenient form if the following two conditions both hold:

- The null hypothesis is simple ($H_0 : \theta = \theta_0$).
- The maximum likelihood estimator $\hat{\theta}$ of θ on the parameter space $\Theta = \Theta_0 \cup \Theta_1$ exists.

Then

$$\Lambda(\boldsymbol{X}) = \frac{L_{\boldsymbol{X}}(\theta_0)}{L_{\boldsymbol{X}}(\hat{\theta})}.$$

Example 9.0.1: Let $X_1, \ldots, X_n \sim$ iid Exp(λ), where $\lambda > 0$, and consider testing $H_0 : \lambda = 2$ versus $H_1 : \lambda \neq 2$. The likelihood is (for $\lambda > 0$)

$$L_{\boldsymbol{X}}(\lambda) = \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} X_i\right).$$

Earlier in the course, we showed that the maximum likelihood estimator of λ is

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^{n} X_i} = \left(\overline{X}\right)^{-1}.$$

Then

$$L_{\boldsymbol{X}}(2) = 2^n \exp\left(-2\sum_{i=1}^n X_i\right) = \exp\left[-n\left(2\overline{X} - \log 2\right)\right],$$
$$L_{\boldsymbol{X}}(\hat{\lambda}) = \left(\frac{n}{\sum_{i=1}^n X_i}\right)^n \exp(-n) = \exp\left[-n\left(1 + \log \overline{X}\right)\right].$$

Then the likelihood ratio statistic is

$$\Lambda(\boldsymbol{X}) = \frac{L_{\boldsymbol{X}}(2)}{L_{\boldsymbol{X}}(\hat{\lambda})} = \frac{\exp\left[-n\left(2\overline{X} - \log 2\right)\right]}{\exp\left[-n\left(1 + \log \overline{X}\right)\right]} = \exp\left[n\left(1 + \log 2 + \log \overline{X} - 2\overline{X}\right)\right]$$
$$= \left[2\overline{X}\exp\left(1 - 2\overline{X}\right)\right]^n.$$

Thus, the likelihood ratio test of these hypotheses rejects H_0 if and only if

$$\left[2\overline{X}\exp\left(1 - 2\overline{X}\right)\right]^{-n} \geq c,$$

or equivalently,

$$\overline{X}\exp\left(-2\overline{X}\right) \leq c^\star,$$

where $c^\star = (2e)^{-1}c^{-1/n}$. Unfortunately, it is difficult to proceed any further in closed form. We need the distribution of the test statistic to specify a level via a critical value or to calculate a p-value, but both $[\Lambda(\boldsymbol{X})]^{-1}$ itself and the equivalent statistic $\overline{X}\exp(-2\overline{X})$ are difficult to work with. ♢

The situation at the end of Example 9.0.1 is not unusual when deriving the form of likelihood ratio tests. It is often the case that the distribution of the likelihood ratio statistic (or of some other equivalent statistic) is difficult to actually obtain. This is partially why people sometimes prefer other approaches to constructing tests. These other approaches are typically based on asymptotic properties (including asymptotic properties of the likelihood ratio test), as we will see later.

Composite Null Hypothesis

More work is required to find the likelihood ratio test when the null hypothesis is composite, particularly if there are one or more parameters with values unspecified by H_0. Finding the numerator of $\Lambda(\boldsymbol{X})$ typically requires first maximizing the likelihood function subject to the constraints of the null hypothesis, then evaluating the likelihood at this point.

Example 9.0.2: Let $X_1, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are both unknown, and consider testing $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ for some specified $\mu_0 \in \mathbb{R}$. The numerator of the likelihood ratio statistic is $\sup_{\sigma^2>0} L_{\boldsymbol{X}}(\mu_0, \sigma^2)$, and evaluating this requires first finding the value of σ^2 that maximizes $L_{\boldsymbol{X}}(\mu_0, \sigma^2)$, or equivalently, $\ell_{\boldsymbol{X}}(\mu_0, \sigma^2)$. Observe that

$$\frac{\partial}{\partial\sigma^2}\ell_{\boldsymbol{X}}(\mu_0, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n (X_i - \mu_0)^2 = 0 \iff \sigma^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_0)^2,$$

and it is clear that this value is indeed a maximum. Thus, the value of σ^2 that maximizes the expression in the numerator of $\Lambda(\boldsymbol{X})$ is

$$\tilde{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2.$$

Now recall that the (unconstrained) maximum likelihood estimators of μ and σ^2 are

$$\hat{\mu} = \overline{X}, \qquad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\big(X_i - \overline{X}\,\big)^2.$$

Then the likelihood ratio statistic is

$$\begin{aligned}\Lambda(\boldsymbol{X}) &= \frac{(2\pi\tilde{\sigma}_0^2)^{-n/2}\exp\big[-(2\tilde{\sigma}_0^2)^{-1}\sum_{i=1}^{n}(X_i-\mu_0)^2\big]}{(2\pi\hat{\sigma}^2)^{-n/2}\exp\big[-(2\hat{\sigma}^2)^{-1}\sum_{i=1}^{n}(X_i-\hat{\mu})^2\big]} \\ &= \frac{(\tilde{\sigma}_0^2)^{-n/2}\exp(-n/2)}{(\hat{\sigma}^2)^{-n/2}\exp(-n/2)} = \left(\frac{\hat{\sigma}^2}{\tilde{\sigma}_0^2}\right)^{n/2} = \left[\frac{\sum_{i=1}^{n}\big(X_i-\overline{X}\,\big)^2}{\sum_{i=1}^{n}(X_i-\mu_0)^2}\right]^{n/2}.\end{aligned}$$

Now observe that

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu_0)^2 = \sum_{i=1}^{n}\big(X_i-\overline{X}+\overline{X}-\mu_0\big)^2 &= \sum_{i=1}^{n}\big(X_i-\overline{X}\,\big)^2 + n\big(\overline{X}-\mu_0\big)^2 + 2\big(\overline{X}-\mu_0\big)\sum_{i=1}^{n}\big(X_i-\overline{X}\,\big) \\ &= \sum_{i=1}^{n}\big(X_i-\overline{X}\,\big)^2 + n\big(\overline{X}-\mu_0\big)^2.\end{aligned}$$

Then

$$\begin{aligned}[\Lambda(\boldsymbol{X})]^{-1} = \left[\frac{\sum_{i=1}^{n}\big(X_i-\overline{X}\,\big)^2 + n\big(\overline{X}-\mu_0\big)^2}{\sum_{i=1}^{n}\big(X_i-\overline{X}\,\big)^2}\right]^{n/2} &= \left[1+\frac{n\big(\overline{X}-\mu_0\big)^2}{\sum_{i=1}^{n}\big(X_i-\overline{X}\,\big)^2}\right]^{n/2} \\ &= \left[1+\frac{\big(\overline{X}-\mu_0\big)^2}{\hat{\sigma}^2}\right]^{n/2} = \left\{1+\frac{[T(\boldsymbol{X})]^2}{n-1}\right\}^{n/2},\end{aligned}$$

where

$$T(\boldsymbol{X}) = \frac{\big|\overline{X}-\mu_0\big|}{\sqrt{\hat{\sigma}^2/(n-1)}} = \frac{\big|\overline{X}-\mu_0\big|}{\sqrt{S^2/n}},$$

where $S^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i-\overline{X}\,)^2$ is the unbiased sample variance. Note that if H_0 is true (i.e., if $\mu = \mu_0$), then the distribution of $T(\boldsymbol{X})$ is the distribution of the absolute value of a Student's t random variable with $n-1$ degrees of freedom. Finally, observe that rejecting H_0 if and only if $[\Lambda(\boldsymbol{X})]^{-1} \geq c$ is equivalent to rejecting H_0 if and only if $T(\boldsymbol{X}) \geq c^\star$, where $c^\star = [(n-1)(c^{2/n}-1)]^{1/2}$. Thus, the likelihood ratio test is simply the conventional t test, and we can obtain a test with size α by rejecting H_0 if and only if $T(\boldsymbol{X})$ exceeds the appropriate quantile of a Student's t distribution with $n-1$ degrees of freedom. ♢

Lecture 10: Tests Based on Asymptotic Properties

A hypothesis test is only useful if we can determine a way to set its significance level, at least approximately. This task requires knowledge of the approximate distribution of the test statistic, which may be easier to find in the asymptotic limit. If we can choose a test statistic for which the asymptotic distribution is known, then we can use it to construct a hypothesis test.

10.1 Wald Tests

Recall that under certain regularity conditions, the asymptotic distribution of the maximum likelihood estimator is normal. More specifically, if $\hat{\theta}_n$ is the MLE of θ, then

$$\sqrt{n}\big(\hat{\theta}_n - \theta\big) \to_D N\bigg[0, \frac{1}{I_1(\theta)}\bigg],$$

where $I_1(\theta)$ denotes the Fisher information per observation. Then it follows that

$$\sqrt{n\, I_1(\theta)}\, \big(\hat{\theta}_n - \theta\big) = \sqrt{I(\theta)}\, \big(\hat{\theta}_n - \theta\big) \to_D N(0,1),$$

i.e., the quantity $[I(\theta)]^{1/2}\,(\hat{\theta}_n - \theta)$ has an approximate $N(0,1)$ distribution if n is large. Also note that under suitable regularity conditions, $I_1(\theta)$ is a continuous function of θ, and $\hat{\theta}_n \to_P \theta$. Then $I_1(\hat{\theta}_n) \to_P I_1(\theta)$, and so

$$\sqrt{I\big(\hat{\theta}_n\big)}\, \big(\hat{\theta}_n - \theta\big) = \frac{\sqrt{I_1\big(\hat{\theta}_n\big)}}{\sqrt{I_1(\theta)}}\, \sqrt{n\, I_1(\theta)}\, \big(\hat{\theta}_n - \theta\big) \to_D N(0,1)$$

by Slutsky's theorem. Thus, the quantity $[I(\hat{\theta}_n)]^{1/2}\,(\hat{\theta}_n - \theta)$ also has an approximate $N(0,1)$ distribution if n is large.

Definition of Wald Tests

The asymptotic results above suggest two similar hypothesis testing methods that should be useful for large n. One test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ with approximate size α is to reject H_0 if and only if

$$\sqrt{I(\theta_0)}\, \big|\hat{\theta}_n - \theta_0\big| \geq c, \tag{10.1.1}$$

where c is the number such that $P(|Z| \geq c) = \alpha$ for a standard normal random variable Z. Another test of the same hypotheses with approximate size α is to reject H_0 if and only if

$$\sqrt{I(\hat{\theta}_n)}\, \big|\hat{\theta}_n - \theta_0\big| \geq c, \tag{10.1.2}$$

where c is the same as above.

> Note: Recall that if $Z \sim N(0,1)$, then $W = Z^2 \sim \chi^2_1$. Thus, it is equivalent to state these tests as rejecting H_0 if and only if the squares of their respective test statistics are at least k, where k is the number such that $P(W \geq k) = \alpha$ for a χ^2_1 random variable W.

The two tests with rejection regions defined by (10.1.1) and (10.1.2) are called *Wald tests*. When people simply refer to "the Wald test" without further clarification, they usually mean the test associated with (10.1.2), which evaluates the Fisher information at the MLE.

> Note: The motivation for the use of (10.1.2) instead of (10.1.1) will be become clear later when we discuss confidence intervals based on the Wald test.

EXAMPLE 10.1.1: Let $X_1, \ldots, X_n \sim$ iid Exp(λ), where $\lambda > 0$, and consider testing $H_0 : \lambda = 2$ versus $H_1 : \lambda \neq 2$. Earlier in the course, we showed that the maximum likelihood estimator of λ is

$$\hat{\lambda}_n = \frac{n}{\sum_{i=1}^n X_i} = \left(\overline{X}_n\right)^{-1}.$$

Next, note that

$$\ell''_{\boldsymbol{X}}(\lambda) = \frac{\partial^2}{\partial \lambda^2}\left(n \log \lambda - \lambda \sum_{i=1}^n X_i\right) = -\frac{n}{\lambda^2}$$

so $I(\lambda) = -E_\lambda[\ell''_{\boldsymbol{X}}(\lambda)] = n/\lambda^2$. Then the Wald test statistic defined by (10.1.1) is

$$\sqrt{I(2)}\,\left|\hat{\lambda}_n - 2\right| = \frac{\sqrt{n}}{2}\,\left|\hat{\lambda}_n - 2\right| = \sqrt{n}\left|\frac{\hat{\lambda}_n}{2} - 1\right| = \sqrt{n}\left|1 - \frac{1}{2\overline{X}_n}\right|,$$

while the Wald test statistic defined by (10.1.2) is

$$\sqrt{I(\hat{\lambda}_n)}\,\left|\hat{\lambda}_n - 2\right| = \frac{\sqrt{n}}{\hat{\lambda}_n}\,\left|\hat{\lambda}_n - 2\right| = \sqrt{n}\left|1 - \frac{2}{\hat{\lambda}_n}\right| = \sqrt{n}\,\left|1 - 2\overline{X}_n\right|.$$

Each test rejects H_0 if and only if its test statistic is at least as large as some critical value c. (To obtain size $\alpha = 0.05$, we would take $c \approx 1.96$.) ◊

Observed Information

In practice, it may be the case that the maximum likelihood estimator $\hat{\theta}_n$ cannot be expressed in closed form but can be found numerically. In such situations, a closed-form expression for the Fisher information may be unavailable as well. For this reason, the Fisher information in the Wald test is often replaced by the quantity

$$J_{\boldsymbol{X}}\left(\hat{\theta}_n\right) = -\ell''_{\boldsymbol{X}}\left(\hat{\theta}_n\right),$$

which is called the *observed information*. If the MLE can be found numerically, then it is typically straightforward to compute the observed information numerically as well.

> Note: Actually, it can be argued that it is better to use the observed information than the Fisher information in the Wald test, regardless of whether or not the Fisher information can be written in closed form.

REFERENCE
Efron, B., and Hinkley, D. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, **65** 457–482.

The following result justifies the use of the observed information.

Lemma 10.1.2. *Let $\hat{\theta}_n$ be the maximum likelihood estimator of θ based on the sample $\boldsymbol{X}_n$. Then under the regularity conditions of Section 6.4 of Lecture 6,*

$$\frac{1}{n} J_{\boldsymbol{X}_n}(\hat{\theta}_n) \to_P I_1(\theta).$$

Proof. The proof is beyond the scope of this course. Note that correctly proving the result requires dealing with both the random function $J_{\boldsymbol{X}_n}$ and the random point $\hat{\theta}_n$ at which the function $J_{\boldsymbol{X}_n}$ is evaluated. □

A Wald test in which the Fisher information has been replaced by the observed information would typically still be called a Wald test.

10.2 Score Test

Recall that under certain regularity conditions, the asymptotic distribution of the score function $\ell'_{\boldsymbol{X}}(\theta)$ itself is normal. More specifically,

$$\sqrt{n}\left[\frac{1}{n}\ell'_{\boldsymbol{X}_n}(\theta) - 0\right] = \frac{1}{\sqrt{n}}\ell'_{\boldsymbol{X}_n}(\theta) \to_D N[0, I_1(\theta)].$$

Then it follows that

$$\frac{1}{\sqrt{n\, I_1(\theta)}}\ell'_{\boldsymbol{X}_n}(\theta) = \frac{1}{\sqrt{I(\theta)}}\ell'_{\boldsymbol{X}_n}(\theta) \to_D N(0,1),$$

i.e., the quantity $[I(\theta)]^{-1/2}\,\ell'_{\boldsymbol{X}}(\theta)$ has an approximate $N(0,1)$ distribution if n is large.

Definition of Score Test

The asymptotic result above suggests a hypothesis testing method that should be useful for large n. Specifically, a test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ with approximate size α is to reject H_0 if and only if

$$\frac{1}{\sqrt{I(\theta_0)}}|\ell'_{\boldsymbol{X}}(\theta_0)| \geq c, \tag{10.2.1}$$

where c is the number such that $P(|Z| \geq c) = \alpha$ for a standard normal random variable Z.

> Note: Recall that if $Z \sim N(0,1)$, then $W = Z^2 \sim \chi^2_1$. Thus, it is equivalent to state this test as rejecting H_0 if and only if the square of the test statistic above is at least k, where k is the number such that $P(W \geq k) = \alpha$ for a χ^2_1 random variable W.

A test with rejection region defined by (10.2.1) is called a *score test*.

EXAMPLE 10.2.1: Let $X_1, \ldots, X_n \sim$ iid Exp(λ), where $\lambda > 0$, and consider testing $H_0 : \lambda = 2$ versus $H_1 : \lambda \neq 2$. The score function is

$$\ell'_{\boldsymbol{X}}(\lambda) = \frac{\partial}{\partial\lambda}\left(n\log\lambda - \lambda\sum_{i=1}^{n} X_i\right) = \frac{n}{\lambda} - \sum_{i=1}^{n} X_i = n\left(\frac{1}{\lambda} - \overline{X}_n\right),$$

where $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i$. Next, note from Example 10.1.1 that $I(\lambda) = n/\lambda^2$. Then the score test statistic is

$$\frac{1}{\sqrt{I(2)}}\,|\ell'_{\boldsymbol{X}}(2)| = \frac{1}{\sqrt{n/4}}\left|n\left(\frac{1}{2} - \overline{X}_n\right)\right| = \sqrt{n}\,\left|1 - 2\overline{X}_n\right|,$$

and the score test rejects H_0 if and only if this test statistic is at least as large as some critical value c. (To obtain size $\alpha = 0.05$, we would take $c \approx 1.96$.) Note that in this particular example, the score test coincides with one of the Wald tests from Example 10.1.1. ◊

Unlike a Wald test, a score test does not involve the maximum likelihood estimator of the parameter of interest. Thus, even when no closed-form solution for the MLE exists, it may still be possible to express a score test in closed form.

EXAMPLE 10.2.2: Let $X_1, \ldots, X_n$ be iid continuous random variables with pdf

$$f_\theta(x) = \begin{cases} \dfrac{\theta}{(\theta + x)^2} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases}$$

where $\theta > 0$ is unknown. The score function is

$$\ell'_{\boldsymbol{X}}(\theta) = \frac{\partial}{\partial\theta}\left[n\log\theta - 2\sum_{i=1}^{n}\log(\theta + X_i)\right] = \frac{n}{\theta} - \sum_{i=1}^{n}\frac{2}{\theta + X_i}.$$

Note that no closed-form expression for the maximum likelihood estimator exists since solving the equation

$$\frac{n}{\theta} = \sum_{i=1}^{n}\frac{2}{\theta + X_i}$$

for θ symbolically is impossible if n is large. However, consider a score test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ for some specified $\theta_0 > 0$. Observe that

$$\ell''_{\boldsymbol{X}}(\theta) = \frac{\partial}{\partial\theta}\ell'_{\boldsymbol{X}}(\theta) = -\frac{n}{\theta^2} + \sum_{i=1}^{n}\frac{2}{(\theta + X_i)^2},$$

and it can be shown by simple calculus (which we omit for brevity) that

$$I(\theta) = -E_\theta[\ell''_{\boldsymbol{X}}(\theta)] = \frac{n}{\theta^2} - \frac{2n}{3\theta^2} = \frac{n}{3\theta^2}.$$

Then the score test statistic is

$$\frac{1}{\sqrt{I(\theta_0)}}\,|\ell'_{\boldsymbol{X}}(\theta_0)| = \frac{1}{\sqrt{n/(3\theta_0^2)}}\left|\frac{n}{\theta}_0 - \sum_{i=1}^{n}\frac{2}{\theta_0 + X_i}\right| = \sqrt{3n}\left|1 - \frac{1}{n}\sum_{i=1}^{n}\frac{2\theta_0}{\theta_0 + X_i}\right| = \sqrt{\frac{3}{n}}\left|\sum_{i=1}^{n}\frac{X_i - \theta_0}{X_i + \theta_0}\right|,$$

and the score test rejects H_0 if and only if this test statistic is at least as large as some critical value c. (To obtain size $\alpha = 0.05$, we would take $c \approx 1.96$.) Thus, we can express a score test for this example in closed form despite the fact that no closed-form solution exists for the maximum likelihood estimator. ◊

10.3 Asymptotics of Likelihood Ratio Tests

If we have already chosen to resort to asymptotic results to determine the distribution of a test statistic, it is logical to ask whether there exist such results for the likelihood ratio statistic itself. The following theorem provides precisely such a result.

Theorem 10.3.1. *Let $\Lambda(\boldsymbol{X}_n)$ be the likelihood ratio statistic for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ based on the sample $\boldsymbol{X}_n$. Then under the regularity conditions of Section 6.4 of Lecture 6,*

$$-2\log\Lambda(\boldsymbol{X}_n) \to_D \chi^2_1 \qquad \text{if } \theta = \theta_0,$$

where χ^2_1 denotes a chi-squared random variable with one degree of freedom.

Proof. Let $\hat{\theta}_n$ denote the MLE of θ. A Taylor expansion of $\ell_{\boldsymbol{X}_n}(\theta_0)$ around $\ell_{\boldsymbol{X}_n}(\hat{\theta}_n)$ yields

$$\begin{aligned}\ell_{\boldsymbol{X}_n}(\theta_0) &= \ell_{\boldsymbol{X}_n}\big(\hat{\theta}_n\big) + \ell'_{\boldsymbol{X}_n}\big(\hat{\theta}_n\big)\big(\theta_0 - \hat{\theta}_n\big) + \frac{1}{2}\ell''_{\boldsymbol{X}_n}\big(\hat{\theta}_n\big)\big(\theta_0 - \hat{\theta}_n\big)^2 + \cdots \\ &= \ell_{\boldsymbol{X}_n}\big(\hat{\theta}_n\big) + \frac{1}{2}\ell''_{\boldsymbol{X}_n}\big(\hat{\theta}_n\big)\big(\theta_0 - \hat{\theta}_n\big)^2 + \cdots\end{aligned}$$

since $\ell'_{\boldsymbol{X}_n}(\hat{\theta}_n) = 0$. (Also note that the purpose of the regularity conditions is to allow us to ignore the higher-order terms.) Now observe that

$$-2\log\Lambda(\boldsymbol{X}_n) = -2\log\left[\frac{L_{\boldsymbol{X}_n}(\theta_0)}{L_{\boldsymbol{X}_n}(\hat{\theta}_n)}\right] = -2\big[\ell_{\boldsymbol{X}_n}(\theta_0) - \ell_{\boldsymbol{X}_n}(\hat{\theta}_n)\big] \approx -\ell''_{\boldsymbol{X}_n}\big(\hat{\theta}_n\big)\big(\theta_0 - \hat{\theta}_n\big)^2$$

by the Taylor expansion. Then

$$\begin{aligned}-2\log\Lambda(\boldsymbol{X}_n) \approx -\ell''_{\boldsymbol{X}_n}\big(\hat{\theta}_n\big)\big(\theta_0 - \hat{\theta}_n\big)^2 &= J_{\boldsymbol{X}_n}\big(\hat{\theta}_n\big)\big(\hat{\theta}_n - \theta_0\big)^2 \\ &= \frac{n^{-1}J_{\boldsymbol{X}_n}\big(\hat{\theta}_n\big)}{I_1(\theta_0)}\Big[\sqrt{n\,I_1(\theta_0)}\,\big(\hat{\theta}_n - \theta_0\big)\Big]^2.\end{aligned}$$

Finally, $n^{-1}J_{\boldsymbol{X}_n}(\hat{\theta}_n)/I_1(\theta_0) \to_P 1$ by Lemma 10.1.2, and $[n\,I_1(\theta_0)]^{1/2}(\hat{\theta}_n - \theta_0) \to_D N(0,1)$ by Theorem 6.2.4 of Lecture 6. Then the result follows from Slutsky's theorem and the fact that the square of a $N(0,1)$ random variable is a χ^2_1 random variable. □

If n is large, Theorem 10.3.1 shows how to find a critical value that yields a likelihood ratio test with approximate size α. Specifically, a test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ with approximate size α is to reject H_0 if and only if

$$-2\log\Lambda(\boldsymbol{X}) \geq C$$

where C is the number such that $P(W \geq C) = \alpha$ for a χ^2_1 random variable W, or equivalently the number such that $P(|Z| \geq C^{1/2}) = \alpha$ for a $N(0,1)$ random variable Z.

> Note: This test is equivalent to taking the original critical value c for the likelihood ratio test as defined in Lecture 9 to be $c = \exp(C/2)$.

EXAMPLE 10.3.2: Let $X_1, \ldots, X_n \sim$ iid Exp(λ), where $\lambda > 0$, and consider testing $H_0 : \lambda = 2$ versus $H_1 : \lambda \neq 2$. From Example 9.0.1 of Lecture 9, the likelihood ratio statistic is

$$\Lambda(\boldsymbol{X}) = \left[2\overline{X}_n \exp\left(1 - 2\overline{X}_n\right)\right]^n.$$

Note that

$$-2\log\Lambda(\boldsymbol{X}) = -2n\left[1 + \log\left(2\overline{X}_n\right) - 2\overline{X}_n\right].$$

To obtain a likelihood ratio test with approximate size α, we should reject H_0 if and only if this test statistic is at least as large as some critical value C. (To obtain size $\alpha = 0.05$, we would take $C^{1/2} \approx 1.96$, and hence $C \approx 3.84$.) ♢

Extension to Multiple Parameters

Theorem 10.3.1 applies when there is a single unknown parameter θ. However, the result can be extended to the case of multiple unknown parameters. Note that the definition of the likelihood ratio statistic $\Lambda(\boldsymbol{X}_n)$ is unchanged even in the presence of multiple unknown parameters. Under certain regularity conditions, $-2\log\Lambda(\boldsymbol{X}_n)$ converges in distribution to a chi-squared random variable with ν degrees of freedom if H_0 is true, where ν denotes the number of parameters constrained by H_0 that are not constrained by the full parameter space. The details of such results are beyond the scope of this course.

Lecture 11: Confidence Intervals

Earlier in the course, we discussed methods to produce a single estimate of an unknown parameter θ. An alternative approach is to report a set of values of θ, usually an interval, that we believe contains θ, which we call a confidence set or confidence interval. Confidence sets/intervals are an inherently frequentist concept. (There exist Bayesian methods that serve a similar purpose, but they are typically referred to by a different name.)

11.1 Definition and Relationship to Hypothesis Testing

A confidence set is defined by a function C that maps a point $\boldsymbol{x}$ in the sample space to some subset of the parameter space. The confidence set $C(\boldsymbol{X})$ is the result of applying this function to the data $\boldsymbol{X}$.

Confidence Level

A *confidence level* of a confidence set $C(\boldsymbol{X})$ for a parameter θ is a number $\gamma \in [0,1]$ such that $P_\theta[\theta \in C(\boldsymbol{X})] \geq \gamma$ for all θ in the parameter space Θ.

- Confidence levels are typically expressed as the percentage $100\gamma\%$, and we typically refer to a confidence set for θ with confidence level γ as a $100\gamma\%$ confidence set (e.g., a 95% confidence set for θ when $\gamma = 0.95$).
- Note that the random part of the event $\theta \in C(\boldsymbol{X})$ is $C(\boldsymbol{X})$, not θ. For this reason, it is good to think and speak in terms of whether or not the confidence set $C(\boldsymbol{X})$ contains the parameter θ (rather than whether or not the parameter falls in the confidence set).

Relationship to Hypothesis Testing

The following theorem provides a method by which confidence intervals can be constructed.

Theorem 11.1.1. *For every $\theta_0 \in \Theta$, let R_{θ_0} be the rejection region of a hypothesis test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ with level α. Then $C(\boldsymbol{X}) = \{\theta_0 \in \Theta : \boldsymbol{X} \notin R_{\theta_0}\}$ is a $100(1-\alpha)\%$ confidence set.*

Proof. For every $\theta \in \Theta$, $P_\theta[\theta \in C(\boldsymbol{X})] = P_\theta(\boldsymbol{X} \notin R_\theta) = 1 - P_\theta(\boldsymbol{X} \in R_\theta) \geq 1 - \alpha$. □

Thus, Theorem 11.1.1 essentially states that confidence intervals are, in some sense, inverted hypothesis tests. A confidence interval with confidence level $1-\alpha$ is simply all the values of the parameter that would *not* be rejected as the null hypothesis value in a test with level α.

Example 11.1.2: Let $X_1, \ldots, X_n \sim \text{iid } N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ is unknown and $\sigma^2 > 0$ is known. A test of the hypotheses $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ with size α is to reject H_0 if and only if

$$\frac{\left|\overline{X} - \mu_0\right|}{\sqrt{\sigma^2/n}} \geq c,$$

where c is the number such that $P(|Z| \geq c) = \alpha$ for a standard normal random variable Z. Then we would *not* reject H_0 if and only if

$$\frac{|\overline{X} - \mu_0|}{\sqrt{\sigma^2/n}} < c \iff \overline{X} - c\sqrt{\frac{\sigma^2}{n}} < \mu_0 < \overline{X} + c\sqrt{\frac{\sigma^2}{n}}.$$

Thus,

$$\left(\overline{X} - c\sqrt{\frac{\sigma^2}{n}}\ ,\ \overline{X} + c\sqrt{\frac{\sigma^2}{n}}\right)$$

is a $100(1-\alpha)\%$ confidence interval for μ. ◊

Multiple Unknown Parameters

Although Theorem 11.1.1 is written in terms of a single unknown parameter θ, the same result essentially still holds in the case of multiple parameters, provided we make a slight notational adjustment. Let θ be the parameter for which we wish to construct a confidence interval, and let ψ denote the other unknown parameters. Let the parameter space be $\Theta \times \Psi$, where $\theta \in \Theta$ and $\psi \in \Psi$. Then Theorem 11.1.1 still holds, i.e., $C(\boldsymbol{X}) = \{\theta_0 \in \Theta : \boldsymbol{X} \notin R_{\theta_0}\}$ is still a $100(1-\alpha)\%$ confidence set for θ. The only difference is that each null hypothesis $H_0 : \theta = \theta_0$ is now a composite null hypothesis.

EXAMPLE 11.1.3: Let $X_1, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ is unknown and $\sigma^2 > 0$ is unknown. A test of the hypotheses $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ with size α is to reject H_0 if and only if

$$\frac{|\overline{X} - \mu_0|}{\sqrt{S^2/n}} \geq c,$$

where $c^\star$ is the number such that $P(|T| \geq c^\star) = \alpha$ for a random variable T that has Student's t distribution with $n-1$ degrees of freedom. Then

$$\left(\overline{X} - c^\star\sqrt{\frac{S^2}{n}}\ ,\ \overline{X} + c^\star\sqrt{\frac{S^2}{n}}\right)$$

is a $100(1-\alpha)\%$ confidence interval for μ. ◊

Note that the presence of additional unknown parameters merely complicates the process of finding a level-α test of $H_0 : \theta = \theta_0$, which is now composite. The process of inverting the test to find the confidence set is the same regardless of whether other parameters are unknown.

11.2 Asymptotic Confidence Intervals

If the sample size is large, asymptotic results can be used to approximate sizes of hypothesis tests. Confidence sets based on such asymptotically motivated tests may be called asymptotic confidence intervals. The stated confidence level of such sets is only approximate, with the quality of the approximation improving as the sample size increases.

Wald Intervals

The simplest asymptotic confidence intervals are those based on Wald tests. Recall that a Wald test of $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ rejects H_0 if and only if

$$\sqrt{I\big(\hat{\theta}_n\big)}\,\big|\hat{\theta}_n - \theta_0\big| \geq c,$$

where c is the number such that $P(|Z| \geq c) = \alpha$ for a standard normal random variable Z. Then we would *not* reject H_0 if and only if

$$\sqrt{I\big(\hat{\theta}_n\big)}\,\big|\hat{\theta}_n - \theta_0\big| < c \iff \hat{\theta}_n - \frac{c}{\sqrt{I\big(\hat{\theta}_n\big)}} < \theta_0 < \hat{\theta}_n + \frac{c}{\sqrt{I\big(\hat{\theta}_n\big)}}.$$

Thus,

$$\left(\hat{\theta}_n - \frac{c}{\sqrt{I\big(\hat{\theta}_n\big)}}\,,\ \hat{\theta}_n + \frac{c}{\sqrt{I\big(\hat{\theta}_n\big)}}\right)$$

is a $100(1-\alpha)\%$ confidence interval for θ.

> Note: The use of the Wald test for the creation of confidence intervals is one reason why the form of the test involving $I(\hat{\theta}_n)$ is more commonly used than the form of the test involving $I(\theta_0)$. If $I(\theta_0)$ is used instead, then θ_0 now appears in two places in the test statistic, and inverting the test is not as straightforward.

The usual naïve confidence intervals of the form

$$(\text{estimator}) \pm (\text{normal quantile})(\text{standard error of estimator})$$

are essentially just Wald intervals.

Example 11.2.1: Let $X \sim \text{Bin}(n, \theta)$, where $0 < \theta < 1$ is unknown. The Fisher information is $I(\theta) = n/[\theta(1-\theta)]$, and the maximum likelihood estimator of θ is $\hat{\theta}_n = X/n$ (provided that $0 < X < n$). Then a $100(1-\alpha)\%$ confidence interval for θ is

$$\left(\hat{\theta}_n - c\sqrt{\frac{\hat{\theta}_n\big(1-\hat{\theta}_n\big)}{n}}\,,\ \hat{\theta}_n + c\sqrt{\frac{\hat{\theta}_n\big(1-\hat{\theta}_n\big)}{n}}\right),$$

where c is the number such that $P(|Z| \geq c) = \alpha$ for a standard normal random variable Z. Note that the formula above yields $(0, 0)$ if $X = 0$ and $(1, 1)$ if $X = n$, but the MLE $\hat{\theta}_n$ does not exist in these cases anyway. ♢

Score Intervals

An alternative approach is to base asymptotic confidence intervals on score tests. Recall that a score test of $H_0 : \theta = \theta_0$ rejects H_0 if and only if

$$\frac{1}{\sqrt{I(\theta_0)}}\,|\ell'_{\boldsymbol{X}}(\theta_0)| \geq c,$$

where c is the number such that $P(|Z| \geq c) = \alpha$ for a standard normal random variable Z. Then we would *not* reject H_0 if and only if

$$\frac{1}{\sqrt{I(\theta_0)}} |\ell'_{\boldsymbol{X}}(\theta_0)| < c.$$

Rewriting this inequality in terms of θ_0 is not always straightforward since θ_0 appears in two places. Typically either approximations are used or solutions are computed numerically.

EXAMPLE 11.2.2: Let $X \sim \text{Bin}(n, \theta)$, where $0 < \theta < 1$ is unknown. The score function is

$$\ell'_X(\theta) = \frac{X}{\theta} - \frac{n - X}{1 - \theta} = \frac{X - n\theta}{\theta(1 - \theta)},$$

and the Fisher information is $I(\theta) = n/[\theta(1 - \theta)]$. Then the score interval for θ consists of those values of θ such that

$$\sqrt{\frac{\theta(1 - \theta)}{n}} \left| \frac{X - n\theta}{\theta(1 - \theta)} \right| = \frac{|X - n\theta|}{\sqrt{n\theta(1 - \theta)}} < c,$$

where c is the number such that $P(|Z| \geq c) = \alpha$ for a standard normal random variable Z. This inequality holds if and only if

$$(X - n\theta)^2 < cn\theta(1 - \theta) \iff (n^2 + cn)\theta^2 - 2n(X + c)\theta + X^2 < 0.$$

By the quadratic formula, this inequality holds if and only if

$$\frac{2n(X + c) - \sqrt{4n^2(X + c)^2 - 4(n^2 + cn)X^2}}{2(n^2 + cn)} < \theta < \frac{2n(X + c) + \sqrt{4n^2(X + c)^2 - 4(n^2 + cn)X^2}}{2(n^2 + cn)}.$$

The confidence interval implied by the above inequality is a bit of a mess. However, for the special case of a 95% confidence level, the interval above is closely approximated by

$$\left(\tilde{\theta}_n - 1.96\sqrt{\frac{\tilde{\theta}_n\big(1 - \tilde{\theta}_n\big)}{n}} \, , \, \tilde{\theta}_n + 1.96\sqrt{\frac{\tilde{\theta}_n\big(1 - \tilde{\theta}_n\big)}{n}} \right),$$

where $\tilde{\theta}_n = (X + 2)/(n + 4)$. Thus, the score interval can be approximated by computing the simpler Wald interval with $\hat{\theta}_n$ replaced by $\tilde{\theta}_n$. Note that $\tilde{\theta}_n$ may be interpreted as the "sample proportion" with an additional two "imaginary" successes and an additional two "imaginary" failures added to the "real" sample. ♢

Likelihood Ratio Intervals

Asymptotic confidence intervals based on the likelihood ratio test can also be constructed. However, the form of the likelihood ratio statistic is often more complicated than that of the Wald or score test statistics, so such likelihood ratio intervals often must be computed numerically.

Homework 1: Due at 11 a.m. on January 17

"DeGroot & Schervish $X.Y.Z$" means Exercise Z at the end of Section $X.Y$ in our text, *Probability and Statistics* (Fourth Edition) by Morris H. DeGroot and Mark J. Schervish.

1. DeGroot & Schervish 3.2.6.

2. DeGroot & Schervish 3.2.9.

3. DeGroot & Schervish 3.5.4.

4. DeGroot & Schervish 3.8.8.

5. DeGroot & Schervish 4.1.6. Also, find $\mathrm{Var}(1/X)$.

6. DeGroot & Schervish 4.2.4. Also, find the variance of the rectangle. *(The continuous uniform distribution on $[a, b]$ has mean $(a + b)/2$ and variance $(b - a)^2/12$. You may use these facts without proof.)*

7. DeGroot & Schervish 4.6.18.

8. Suppose we construct a random variable X as follows. Let $Y \sim \mathrm{Bin}(1, \theta)$, where $0 < \theta < 1$. If $Y = 1$, then $X = 0$. If instead $Y = 0$, then X has a Poisson(λ) distribution. Then the marginal distribution of X (*not* conditional on Y) is called a *zero-inflated Poisson* distribution. *(The mean and variance of the Poisson distribution are listed in Section 5.4 of DeGroot & Schervish. You may use these facts without proof.)*

   (a) Calculate $E(X)$ and $\mathrm{Var}(X)$, the marginal mean and variance of X.

   (b) Find the conditional pmf of Y given X.

9. The exponential distribution and the gamma distribution are related by the following property: Let $X_1, \ldots, X_n$ be iid Exp(β) random variables with pdf given in Definition 5.7.3 of DeGroot & Schervish. Then $Y_n = \sum_{i=1}^{n} X_i$ has a Gamma(α, β) distribution with pdf given in Definition 5.7.2 of DeGroot & Schervish. *(You may use this fact without proof.)*

   (a) Find sequences of constants k_n and m_n such that $k_n(Y_n - m_n) \to_D N(0, 1)$ as $n \to \infty$.

   (b) Find sequences of constants $k_n^\star$ and $m_n^\star$ such that $k_n^\star(Y_n^{-1} - m_n^\star) \to_D N(0, 1)$ as $n \to \infty$.

Solutions to Homework 1

"DeGroot & Schervish $X.Y.Z$" means Exercise Z at the end of Section $X.Y$ in our text, *Probability and Statistics* (Fourth Edition) by Morris H. DeGroot and Mark J. Schervish.

1. DeGroot & Schervish 3.2.6.

   ▷ Solution: First, note that

   $$Y = \begin{cases} 0 & \text{if} \quad 0 \le X < 1/2, \\ 1 & \text{if } 1/2 < X < 3/2, \\ 2 & \text{if } 3/2 < X < 5/2, \\ 3 & \text{if } 5/2 < X < 7/2, \\ 4 & \text{if } 7/2 < X \le 4. \end{cases}$$

   The value of Y is not clear if $X \in \{1/2, 3/2, 5/2, 7/2\}$, but $P(X \in \{1/2, 3/2, 5/2, 7/2\}) = 0$, so this ambiguity is irrelevant. Now note that the cdf of X is

   $$F(x) = P(X \le x) = \begin{cases} 0 & \text{if } x < 0, \\ x^2/16 & \text{if } 0 \le x \le 4, \\ 1 & \text{if } x > 4, \end{cases}$$

   and thus the pmf of Y is

   $$\begin{aligned} p(0) &= P(Y = 0) = P(0 \le X < 1/2) = F(1/2) - F(0) = 1/64 - 0 = 1/64, \\ p(1) &= P(Y = 1) = P(1/2 < X < 3/2) = F(3/2) - F(1/2) = 9/64 - 1/64 = 1/8, \\ p(2) &= P(Y = 2) = P(3/2 < X < 5/2) = F(5/2) - F(3/2) = 25/64 - 9/64 = 1/4, \\ p(3) &= P(Y = 3) = P(5/2 < X < 7/2) = F(7/2) - F(5/2) = 49/64 - 25/64 = 3/8, \\ p(4) &= P(Y = 4) = P(7/2 < X < 4) = F(4) - F(7/2) = 1 - 49/64 = 15/64, \end{aligned}$$

   with $p(y) = 0$ for all $y \notin \{0, 1, 2, 3, 4\}$. ◁

2. DeGroot & Schervish 3.2.9.

   ▷ Solution: Clearly we must have $c \ge 0$ since a pdf must be nonnegative. Note that $c = 0$ yields $f(x) = 0$ for all $x \in \mathbb{R}$, and $\int_{\mathbb{R}} 0 \, dx = 0 \ne 1$. Thus, we must have $c > 0$. However, for any $c > 0$,

   $$\int_{-\infty}^{\infty} f(x) \, dx = \int_0^{\infty} \frac{c}{1+x} \, dx = c \log(1 + \infty) - c \log(1 + 0) = \infty \ne 1.$$

   Thus, there does not exist any $c \in \mathbb{R}$ such that $f(x)$ is a pdf. ◁

3. DeGroot & Schervish 3.5.4.

   ▷ Solution to (a): The marginal pdf of X is

   $$f^{(X)}(x) = \int_{\mathbb{R}} f(x,y)\,dy = \int_0^{1-x^2} \frac{15}{4}x^2\,dy = \frac{15}{4}x^2(1-x^2)$$

   if $-1 \le x \le 1$, with $f^{(X)}(x) = 0$ otherwise. To find the marginal pdf of Y, note that $\{(x,y) \in \mathbb{R}^2 : 0 \le y \le 1-x^2\}$ can be written as $\{(x,y) \in \mathbb{R}^2 : |x| \le \sqrt{1-y},\ 0 \le y \le 1\}$. Then

   $$f^{(Y)}(y) = \int_{\mathbb{R}} f(x,y)\,dx = \int_{-\sqrt{1-y}}^{\sqrt{1-y}} \frac{15}{4}x^2\,dx = \frac{5}{2}(1-y)^{3/2},$$

   if $0 \le y \le 1$, with $f^{(Y)}(y) = 0$ otherwise. ◁

   ▷ Solution to (b): No, X and Y are not independent since $f^{(X)}(x)\,f^{(Y)}(y) \ne f(x,y)$. ◁

4. DeGroot & Schervish 3.8.8.

   ▷ Solution: The cdf of X is

   $$F^{(X)}(x) = \begin{cases} 0 & \text{if } x \le 0, \\ 1-\exp(-x) & \text{if } x > 0. \end{cases}$$

   Then the cdf of Y is $F^{(Y)}(y) = P(Y \le y) = P(X \le y^2) = F^{(X)}(y^2) = 1-\exp(-y^2)$ if $y > 0$, with $F^{(Y)}(y) = 0$ if $y \le 0$. We then differentiate to find that the pdf of Y is $f^{(Y)}(y) = 2y\exp(-y^2)$ if $y > 0$, with $f^{(Y)}(y) = 0$ if $y \le 0$. ◁

5. DeGroot & Schervish 4.1.6. Also, find $\mathrm{Var}(1/X)$.

   ▷ Solution: First, $E(1/X) = \int_{\mathbb{R}}(1/x)\,f(x)\,dx = \int_0^1 (1/x)\,2x\,dx = 2$. To find $\mathrm{Var}(1/X)$, we first find $E[(1/X)^2] = \int_{\mathbb{R}}(1/x)^2\,f(x)\,dx = \int_0^1 (1/x)^2\,2x\,dx = 2\int_0^1 \log x\,dx = 2\cdot\infty = \infty$. Thus, $\mathrm{Var}(X) = E[(1/X)^2] - [E(1/X)]^2 = \infty - 2^2 = \infty$. ◁

6. DeGroot & Schervish 4.2.4. Also, find the variance of the rectangle. *(The continuous uniform distribution on $[a,b]$ has mean $(a+b)/2$ and variance $(b-a)^2/12$. You may use these facts without proof.)*

   ▷ Solution: The area of the rectangle is XY, and $E(XY) = E(X)\,E(Y) = (1/2)(7) = 7/2$ since X and Y are independent. To find $\mathrm{Var}(XY)$, we first find

   $$\begin{aligned} E[(XY)^2] = E(X^2Y^2) = E(X^2)\,E(Y^2) &= \big\{[E(X)]^2 + \mathrm{Var}(X)\big\}\big\{[E(Y)]^2 + \mathrm{Var}(Y)\big\} \\ &= \big[(1/2)^2 + 1/12\big]\big[7^2 + 4/3\big] \\ &= (1/3)(151/3) = 151/9. \end{aligned}$$

   Then $\mathrm{Var}(XY) = E[(XY)^2] - [E(XY)]^2 = 151/9 - (7/2)^2 = 163/36$. ◁

7. DeGroot & Schervish 4.6.18.

   ▷ Solution: First, we compute the required expectations:

$$E(X) = \iint_{\mathbb{R}^2} x\, f(x,y)\, dx\, dy = \int_0^1 \int_0^1 (x^2 + xy)\, dx\, dy = \int_0^1 \left(\frac{1}{3} + \frac{y}{2}\right) dy = \frac{1}{3} + \frac{1}{4} = \frac{7}{12},$$
$$E(Y) = \iint_{\mathbb{R}^2} y\, f(x,y)\, dx\, dy = \int_0^1 \int_0^1 (y^2 + xy)\, dy\, dx = \int_0^1 \left(\frac{1}{3} + \frac{x}{2}\right) dx = \frac{1}{3} + \frac{1}{4} = \frac{7}{12},$$
$$E(XY) = \iint_{\mathbb{R}^2} xy\, f(x,y)\, dx\, dy = \int_0^1 \int_0^1 (x^2 y + xy^2)\, dx\, dy = \int_0^1 \left(\frac{y}{3} + \frac{y^2}{2}\right) dy = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}.$$

   Then $\mathrm{Cov}(X,Y) = E(XY) - E(X)\,E(Y) = (1/3) - (7/12)^2 = -1/144$. ◁

8. Suppose we construct a random variable X as follows. Let $Y \sim \mathrm{Bin}(1,\theta)$, where $0 < \theta < 1$. If $Y = 1$, then $X = 0$. If instead $Y = 0$, then X has a Poisson(λ) distribution. Then the marginal distribution of X (*not* conditional on Y) is called a *zero-inflated Poisson* distribution. *(The mean and variance of the Poisson distribution are listed in Section 5.4 of DeGroot & Schervish. You may use these facts without proof.)*

   (a) Calculate $E(X)$ and $\mathrm{Var}(X)$, the marginal mean and variance of X.

   (b) Find the conditional pmf of Y given X.

   ▷ Solution to (a): By the law of total expectation,

$$\begin{aligned} E(X) = E[E(X \mid Y)] &= \sum_{y=0}^{1} E(X \mid Y = y)\, P(Y = y) \\ &= E(X \mid Y = 0)\, P(Y = 0) + E(X \mid Y = 1)\, P(Y = 1) \\ &= \lambda(1-\theta) + 0 \cdot \theta = \lambda(1-\theta). \end{aligned}$$

   By the law of total variance,

$$\begin{aligned} \mathrm{Var}(X) &= E[\mathrm{Var}(X \mid Y)] + \mathrm{Var}[E(X \mid Y)] \\ &= E[\mathrm{Var}(X \mid Y)] + E\big\{[E(X \mid Y)]^2\big\} - \{E[E(X \mid Y)]\}^2 \\ &= \sum_{y=0}^{1} \mathrm{Var}(X \mid Y = y)\, P(Y = y) + \sum_{y=0}^{1} [E(X \mid Y = y)]^2\, P(Y = y) - [E(X)]^2 \\ &= \mathrm{Var}(X \mid Y = 0)\, P(Y = 0) + \mathrm{Var}(X \mid Y = 1)\, P(Y = 1) \\ &\qquad + [E(X \mid Y = 0)]^2\, P(Y = 0) + [E(X \mid Y = 1)]^2\, P(Y = 1) - [E(X)]^2 \\ &= \lambda(1-\theta) + 0 \cdot \theta + \lambda^2(1-\theta) + 0 \cdot \theta - [\lambda(1-\theta)]^2 = \lambda(1-\theta)(1+\lambda\theta). \end{aligned}$$

   An alternative approach is to calculate the marginal pmf of X and use it to find the expectation and variance directly. ◁

▷ Solution to (b): First,

$$\begin{aligned} p^{(Y|X)}(0 \mid 0) = P(Y = 0 \mid X = 0) &= \frac{P(X = 0,\ Y = 0)}{P(X = 0)} \\ &= \frac{P(X = 0,\ Y = 0)}{P(X = 0,\ Y = 0) + P(X = 0,\ Y = 1)} \\ &= \frac{P(X = 0 \mid Y = 0)\ P(Y = 0)}{P(X = 0 \mid Y = 0)\ P(Y = 0) + P(X = 0 \mid Y = 1)\ P(Y = 1)} \\ &= \frac{\exp(-\lambda)(1-\theta)}{\exp(-\lambda)(1-\theta) + 1 \cdot \theta} = \frac{(1-\theta)\exp(-\lambda)}{(1-\theta)\exp(-\lambda) + \theta}. \end{aligned}$$

It follows that

$$p^{(Y|X)}(1 \mid 0) = 1 - \frac{(1-\theta)\exp(-\lambda)}{(1-\theta)\exp(-\lambda) + \theta} = \frac{\theta}{(1-\theta)\exp(-\lambda) + \theta}.$$

Next, note that $p^{(X,Y)}(x, 1) = P(X = x,\ Y = 1) = 0$ for any $x > 0$. It follows immediately that $p^{(Y|X)}(1 \mid x) = P(Y = 1 \mid X = x) = 0$ for every $x > 0$, and hence $p^{(Y|X)}(0 \mid x) = 1$ for every $x > 0$. ◁

9. The exponential distribution and the gamma distribution are related by the following property: Let $X_1, \ldots, X_n$ be iid $\text{Exp}(\beta)$ random variables with pdf given in Definition 5.7.3 of DeGroot & Schervish. Then $Y_n = \sum_{i=1}^n X_i$ has a $\text{Gamma}(n, \beta)$ distribution with pdf given in Definition 5.7.2 of DeGroot & Schervish. *(You may use this fact without proof.)*

   (a) Find sequences of constants k_n and m_n such that $k_n(Y_n - m_n) \to_D N(0, 1)$ as $n \to \infty$.

   (b) Find sequences of constants $k_n^\star$ and $m_n^\star$ such that $k_n^\star(Y_n^{-1} - m_n^\star) \to_D N(0, 1)$ as $n \to \infty$.

▷ Solution to (a): Note that $E(X_1) = 1/\beta$ and $\text{Var}(X_1) = 1/\beta^2$. Then by the central limit theorem,

$$\sqrt{n}\left(\frac{1}{n}Y_n - \frac{1}{\beta}\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{\beta}\right) \to_D N\left(0,\ \frac{1}{\beta^2}\right)$$

as $n \to \infty$, which we may rewrite as

$$\frac{\beta}{\sqrt{n}}\left(Y_n - \frac{n}{\beta}\right) \to_D N(0, 1)$$

as $n \to \infty$. Thus, $k_n = \beta/\sqrt{n}$ and $m_n = n/\beta$. ◁

▷ Solution to (b): Let $g(y) = y^{-1}$, so that $g'(y) = -y^{-2}$ and $[g'(y)]^2 = y^{-4}$. Next, note that $(1/\beta^2)[g'(1/\beta)]^2 = (1/\beta^2)(1/\beta)^{-4} = \beta^2$. Then by the delta method,

$$\sqrt{n}\left[g\left(\frac{1}{n}Y_n\right) - g\left(\frac{1}{\beta}\right)\right] = \sqrt{n}\left(nY_n^{-1} - \beta\right) = n^{3/2}\left(Y_n^{-1} - \frac{\beta}{n}\right) \to_D N\left(0, \beta^2\right),$$

as $n \to \infty$, which may be rewritten as

$$\frac{n^{3/2}}{\beta}\left(Y_n^{-1} - \frac{\beta}{n}\right) \to_D N(0, 1)$$

as $n \to \infty$. Thus, $k_n^\star = n^{3/2}/\beta$ and $m_n^\star = \beta/n$. ◁

Homework 2: Due at 11 a.m. on January 29

"DeGroot & Schervish $X.Y.Z$" means Exercise Z at the end of Section $X.Y$ in our text, *Probability and Statistics* (Fourth Edition) by Morris H. DeGroot and Mark J. Schervish.

1. DeGroot & Schervish 8.2.10.

2. Let T have a Student's t distribution with ν degrees of freedom, where $\nu > 2$. It can be shown that $E(T)$ and $\mathrm{Var}(T)$ both exist. (You do not need to show this fact.)

   (a) Show that $E(T) = 0$. *Hint: Look at the pdf of a Student's t distribution. You should not need to do any calculus.*

   (b) Use the fact below (which you do not need to prove) to show that $\mathrm{Var}(T) = \nu/(\nu-2)$.

   Fact: If $W \sim \mathrm{Gamma}(\alpha, \beta)$, then $E(1/W) = \beta/(\alpha - 1)$.

   Hint: Recall how a Student's t random variable is constructed from a normal random variable and a chi-squared random variable. You should not need to do any calculus.

3. DeGroot & Schervish 7.7.5.

4. DeGroot & Schervish 7.7.9.

5. DeGroot & Schervish 7.8.2.

6. Let α and β be unknown parameters. Show that the $\mathrm{Beta}(\alpha, \beta)$ distribution and the $\mathrm{Gamma}(\alpha, \beta)$ both belong to the exponential family.

7. DeGroot & Schervish 7.5.8.

8. Let $X_1, \ldots, X_n \sim \text{iid } N(0, \sigma^2)$, where $\sigma^2 > 0$ is unknown. Find the maximum likelihood estimator of σ^2.

9. Let $X_n \sim \mathrm{Bin}(n, \theta)$, where θ is unknown.

   (a) Show that the maximum likelihood estimator of θ is $\hat{\theta}_n = X_n/n$.

   (b) Show that $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution as $n \to \infty$, and find the limiting distribution. *Hint: Consider X_n as a sum of independent random variables.*

   (c) Let $\xi = \arcsin\left(\sqrt{\theta}\right)$. Find (or state) the maximum likelihood estimator $\hat{\xi}_n$ of ξ.

   (d) Show that $\sqrt{n}(\hat{\xi}_n - \xi)$ converges in distribution as $n \to \infty$, and find the limiting distribution. What do you notice about the variance of the limiting distribution? *Note: The derivative of the arcsin function is* $\frac{d}{dt} \arcsin t = (1 - t^2)^{-1/2}$.

10. Let Y be a single observation of a $\mathrm{Geometric}(\theta)$ random variable with pmf $p_\theta(y) = (1-\theta)^y \theta$ for all integers $y \geq 0$ (and zero otherwise). *Note: Under this setup, Y counts the number of failures before the first success occurs in a sequence of iid trials.*

    (a) Find the maximum likelihood estimator of θ.

    (b) Explain the connection between the estimator in part (a) of this problem and the maximum likelihood estimator in part (a) of problem 9. *Hint: Recall the note about successes and failures.*

Solutions to Homework 2

"DeGroot & Schervish $X.Y.Z$" means Exercise Z at the end of Section $X.Y$ in our text, *Probability and Statistics* (Fourth Edition) by Morris H. DeGroot and Mark J. Schervish.

1. DeGroot & Schervish 8.2.10.

   ▷ Solution: Let $U_1 = X_1 + X_2 + X_3$ and $U_2 = X_4 + X_5 + X_6$. Since $X_1, \ldots, X_6 \sim \text{iid}\, N(0,1)$, we have $U_1, U_2 \sim \text{iid}\, N(0,3)$. Then $U_1/\sqrt{3},\ U_2/\sqrt{3} \sim \text{iid}\, N(0,1)$, so

   $$\frac{Y}{3} = \left(\frac{U_1}{\sqrt{3}}\right)^2 + \left(\frac{U_2}{\sqrt{3}}\right)^2 \sim \chi^2_2.$$

   Thus, $c = 1/3$. ◁

2. Let T have a Student's t distribution with ν degrees of freedom, where $\nu > 2$. It can be shown that $E(T)$ and $\text{Var}(T)$ both exist. (You do not need to show this fact.)

   (a) Show that $E(T) = 0$. *Hint: Look at the pdf of a Student's t distribution. You should not need to do any calculus.*

   (b) Use the fact below (which you do not need to prove) to show that $\text{Var}(T) = \nu/(\nu-2)$.

   Fact: If $W \sim \text{Gamma}(\alpha, \beta)$, then $E(1/W) = \beta/(\alpha - 1)$.

   Hint: Recall how a Student's t random variable is constructed from a normal random variable and a chi-squared random variable. You should not need to do any calculus.

   ▷ Solution to (a): The pdf of the Student's t distribution is symmetric, so since $E(T)$ exists we must have $E(T) = 0$. ◁

   ▷ Solution to (b): Note that $\text{Var}(T) = E(T^2) - [E(T)]^2 = E(T^2)$. Now let Z and U be independent random variables with distributions $Z \sim N(0,1)$ and $U \sim \chi^2_\nu$, or equivalently, $U \sim \text{Gamma}(\nu/2,\, 1/2)$. Then

   $$\text{Var}(T) = E(T^2) = E\left[\left(\frac{Z}{\sqrt{U/\nu}}\right)^2\right] = \nu\, E\left(\frac{Z^2}{U}\right) = \nu\, E(Z^2)\, E\left(\frac{1}{U}\right) = \nu\left[\frac{1/2}{(\nu/2) - 1}\right] = \frac{\nu}{\nu - 2},$$

   noting that $E(Z^2) = 1$. ◁

3. DeGroot & Schervish 7.7.5.

   ▷ Solution: Let $X_1, \ldots, X_n \sim \text{iid Gamma}(\alpha, \beta)$, where α is known and $\beta > 0$ is unknown. Then the joint pdf of $\boldsymbol{X} = (X_1, \ldots, X_n)$ is

   $$\begin{aligned} f_\beta(\boldsymbol{x}) &= \prod_{i=1}^n \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} \exp(-\beta x_i)\, I_{(0,\infty)}(x_i) \\ &= \underbrace{\beta^{n\alpha} \exp\left(-n\beta\, \overline{X}_n\right)}_{g(\overline{X}_n, \beta)} \underbrace{\left[\Gamma(\alpha)\right]^{-n} \left(\prod_{i=1}^n x_i\right)^{\alpha-1} I_{(0,\infty)}\left(\min_{i=1}^n x_i\right)}_{h(\boldsymbol{x})}. \end{aligned}$$

   Thus, $\overline{X}_n$ is sufficient for β by the factorization theorem. ◁

4. DeGroot & Schervish 7.7.9.

   ▷ SOLUTION: Let $X_1, \ldots, X_n \sim$ iid Unif(a, b), where a is known and $b > a$ is unknown. Then the joint pdf of $\boldsymbol{X} = (X_1, \ldots, X_n)$ is

$$f_b(\boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{b-a} I_{(a,b)}(x_i) = \underbrace{(b-a)^{-n} I_{(-\infty,b)}\Big(\max_{1\le i\le n} x_i\Big)}_{g(\max_{1\le i\le n} x_i,\, b)} \underbrace{I_{(a,\infty)}\Big(\min_{1\le i\le n} x_i\Big)}_{h(\boldsymbol{x})}.$$

   Thus, $\max_{1\le i\le n} X_i$ is sufficient for b by the factorization theorem. ◁

5. DeGroot & Schervish 7.8.2.

   ▷ SOLUTION: Let $X_1, \ldots, X_n \sim$ iid Beta(α, β), where $\alpha > 0$ and $\beta > 0$ are both unknown. Then the joint pdf of $\boldsymbol{X} = (X_1, \ldots, X_n)$ is

$$\begin{aligned} f_{\alpha,\beta}(\boldsymbol{x}) &= \prod_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} x_i^{\alpha-1}(1-x_i)^{\beta-1} I_{(0,1)}(x_i) \\ &= \underbrace{\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\right]^n \left(\prod_{i=1}^{n} x_i\right)^{\alpha-1} \left[\prod_{i=1}^{n}(1-x_i)\right]^{\beta-1}}_{g[(t_1,t_2),(\alpha,\beta)]} \underbrace{I_{(0,1)}\Big(\min_{1\le i\le n} x_i\Big) I_{(0,1)}\Big(\max_{1\le i\le n} x_i\Big)}_{h(\boldsymbol{x})}, \end{aligned}$$

   where $t_1 = \prod_{i=1}^n x_i$ and $t_2 = \prod_{i=1}^n (1-x_i)$. Thus, $(T_1, T_2) = [\prod_{i=1}^n X_i,\ \prod_{i=1}^n (1-X_i)]$ is sufficient for (α, β). ◁

6. Let α and β be unknown parameters. Show that the Beta(α, β) distribution and the Gamma(α, β) both belong to the exponential family.

   ▷ SOLUTION: First, the beta distribution has pdf

$$\begin{aligned} f(x) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{(0,1)}(x) \\ &= \exp[\alpha \log x + \beta \log(1-x) + \log\Gamma(\alpha+\beta) - \log\Gamma(\alpha) - \log\Gamma(\beta)]\, \frac{1}{x(1-x)} I_{(0,1)}(x). \end{aligned}$$

   Then

$$\begin{gathered} k = 2, \qquad \eta_1(\alpha,\beta) = \alpha, \qquad r_1(x) = x, \qquad \eta_2(\alpha,\beta) = \beta, \qquad r_2(x) = 1-x, \\ \psi(\alpha,\beta) = \log\Gamma(\alpha) + \log\Gamma(\beta) - \log\Gamma(\alpha+\beta), \qquad h(x) = \frac{1}{x(1-x)} I_{(0,1)}(x), \end{gathered}$$

   so this distribution belongs to the exponential family. The gamma distribution has pdf

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)\, I_{(0,\infty)}(x) = \exp[\alpha \log x - \beta x - \log\Gamma(\alpha) + \alpha \log\beta]\, \frac{1}{x} I_{(0,\infty)}(x).$$

   Then

$$\begin{gathered} k = 2, \qquad \eta_1(\alpha,\beta) = \alpha, \qquad r_1(x) = \log x, \qquad \eta_2(\alpha,\beta) = \beta, \qquad r_2(x) = -x, \\ \psi(\alpha,\beta) = \log\Gamma(\alpha) - \alpha \log\beta, \qquad h(x) = \frac{1}{x} I_{(0,1)}(x), \end{gathered}$$

   so this distribution also belongs to the exponential family. ◁

7. DeGroot & Schervish 7.5.8.

   ▷ Solution to (a): Let $m = \min_{1\le i\le n} x_i$. The likelihood is

   $$L_{\boldsymbol{x}}(\theta) = \prod_{i=1}^{n} \exp(\theta - x_i)\, I_{(-\infty,x_i)}(\theta) = \exp\left(-\sum_{i=1}^{n} x_i\right) \exp(n\theta)\, I_{(-\infty,m)}(\theta).$$

   Observe that the likelihood $L_{\boldsymbol{x}}(\theta)$ is strictly positive and strictly increasing for $\theta < m$, while $L_{\boldsymbol{x}}(\theta) = 0$ for all $\theta \ge m$. Now note that $\lim_{\theta\uparrow m} L_{\boldsymbol{x}}(\theta) = 1$. However, evaluating the likelihood at m itself yields $L_{\boldsymbol{x}}(m) = 0$. Thus, there is no value of θ that maximizes $L_{\boldsymbol{x}}(\theta)$, and so the maximum likelihood estimator does not exist. ◁

   ▷ Solution to (b): We could instead simply take the pdf to be

   $$f_\theta(x) = \begin{cases} \exp(\theta - x) & \text{if } x \ge \theta, \\ 0 & \text{if } x < \theta. \end{cases}$$

   Then $L_{\boldsymbol{x}}(m) = 1$, so the likelihood attains its maximum at $\theta = m$. Thus, the maximum likelihood estimator of θ is $\hat{\theta} = m = \min_{1\le i\le n} X_i$. ◁

8. Let $X_1, \ldots, X_n \sim \text{iid}\, N(0, \sigma^2)$, where $\sigma^2 > 0$ is unknown. Find the maximum likelihood estimator of σ^2.

   ▷ Solution: The likelihood and log-likelihood are

   $$L_{\boldsymbol{x}}(\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x_i^2}{2\sigma^2}\right) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2\right),$$
   $$\ell_{\boldsymbol{x}}(\sigma^2) = \log L_{\boldsymbol{x}}(\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n} x_i^2.$$

   Then

   $$\frac{\partial}{\partial(\sigma^2)}\ell_{\boldsymbol{x}}(\sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n} x_i^2 = 0 \iff \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2,$$

   and it can be seen that this critical point is indeed the maximizer, i.e.,

   $$\ell_{\boldsymbol{x}}\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) = \max_{\sigma^2>0} \ell_{\boldsymbol{x}}(\sigma^2) \qquad \text{for all } \boldsymbol{x} \in \mathbb{R}^p.$$

   Thus, the maximum likelihood estimator of σ^2 is $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n} X_i^2$. ◁

9. Let $X_n \sim \text{Bin}(n, \theta)$, where θ is unknown.

   (a) Show that the maximum likelihood estimator of θ is $\hat{\theta}_n = X_n/n$.

   (b) Show that $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution as $n \to \infty$, and find the limiting distribution. *Hint: Consider X_n as a sum of independent random variables.*

   (c) Let $\xi = \arcsin\big(\sqrt{\theta}\,\big)$. Find (or state) the maximum likelihood estimator $\hat{\xi}_n$ of ξ.

   (d) Show that $\sqrt{n}(\hat{\xi}_n - \xi)$ converges in distribution as $n \to \infty$, and find the limiting distribution. What do you notice about the variance of the limiting distribution? *Note: The derivative of the arcsin function is $\frac{d}{dt}\arcsin t = (1-t^2)^{-1/2}$.*

▷ Solution to (a): The likelihood is $L_x(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$, and hence the log-likelihood is $\ell_x(\theta) = \log\binom{n}{x} + x\log\theta + (n-x)\log(1-\theta)$. Then

$$\frac{\partial}{\partial\theta}\ell_x(\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta} = 0 \iff \theta = \frac{x}{n},$$

and it can be seen that this critical point is indeed the maximizer, i.e.,

$$\ell_x\left(\frac{x}{n}\right) = \max_{0\le\theta\le1}\ell_x(\theta) \qquad \text{for all } x \in \{0,\dots,n\}.$$

Thus, the maximum likelihood estimator of θ is $\hat{\theta}_n = X_n/n$. ◁

▷ Solution to (b): Write X_n as $X_n = \sum_{i=1}^n Z_i$, where $Z_1,\dots,Z_n \sim$ iid $\text{Bin}(1,\theta)$, and note that $E(Z_1) = \theta$ and $\text{Var}(Z_1) = \theta(1-\theta)$. Then by the central limit theorem, as $n \to \infty$, $\sqrt{n}(\hat{\theta}_n - \theta) = \sqrt{n}(n^{-1}\sum_{i=1}^n Z_i - \theta) \to_D N[0, \theta(1-\theta)]$. ◁

▷ Solution to (c): The maximum likelihood estimator of $\xi = \arcsin\big(\sqrt{\theta}\,\big)$ is simply $\hat{\xi}_n = \arcsin\Big(\sqrt{\hat{\theta}_n}\,\Big) = \arcsin\Big(\sqrt{X_n/n}\,\Big)$. ◁

▷ Solution to (d): Let $g(t) = \arcsin\big(\sqrt{t}\,\big)$, which has derivative

$$g'(t) = \left[\frac{1}{\sqrt{1-(\sqrt{t}\,)^2}}\right]\left(\frac{1}{2\sqrt{t}}\right) = \frac{1}{2\sqrt{t(1-t)}}.$$

Now note that $[g'(\theta)]^2 = 1/[4\theta(1-\theta)]$. Then $\sqrt{n}(\hat{\xi}_n - \xi) = \sqrt{n}[g\big(\hat{\theta}_n\big) - g(\theta)] \to_D N(0, 1/4)$ by the delta method. Notice that the variance of the limiting distribution is the same for all values of θ (or ξ), unlike our result in part (b). ◁

10. Let Y be a single observation of a Geometric(θ) random variable with pmf $p_\theta(y) = (1-\theta)^y\theta$ for all integers $y \geq 0$ (and zero otherwise), where $0 < \theta \leq 1$. *Note: Under this setup, Y counts the number of failures before the first success occurs in a sequence of iid trials.*

    (a) Find the maximum likelihood estimator of θ.

    (b) Explain the connection between the estimator in part (a) of this problem and the maximum likelihood estimator in part (a) of problem 9. *Hint: Recall the note about successes and failures.*

    ▷ Solution to (a): The likelihood is $L_y(\theta) = (1-\theta)^y\theta$, and hence the log-likelihood is $\ell_y(\theta) = \log L_y(\theta) = y\log(1-\theta) + \log\theta$. Then

$$\frac{\partial}{\partial\theta}\ell_y(\theta) = \frac{1}{\theta} - \frac{y}{1-\theta} = 0 \iff \theta = \frac{1}{y+1},$$

    and it can be seen that this critical point is indeed the maximizer, i.e.,

$$\ell_y\left(\frac{1}{y+1}\right) = \max_{0<\theta\leq 1}\ell_y(\theta) \qquad \text{for all integers } y \geq 0.$$

    Thus, the maximum likelihood estimator of θ is $\hat{\theta} = 1/(Y+1)$. ◁

    ▷ Solution to (b): In part (a) of problem 9, the maximum likelihood estimator was the number of successes divided by the total number of trials. Now consider the geometric random variable Y in this problem. Recall that Y counts the number of failures before the first success occurs. Then once this first success occurs, the number of successes that have occurred is 1, while the total number of trials that have occurred is $Y+1$. Thus, the maximum likelihood estimator can again be interpreted as the number of successes divided by the total number of trials. ◁

Homework 3: Due at 11 a.m. on February 14

"DeGroot & Schervish $X.Y.Z$" means Exercise Z at the end of Section $X.Y$ in our text, *Probability and Statistics* (Fourth Edition) by Morris H. DeGroot and Mark J. Schervish.

1. DeGroot & Schervish 7.3.12. Also find the posterior mean and the posterior mode.

2. DeGroot & Schervish 7.4.6.

3. Let $X_1, \ldots, X_n$ be iid random variables with a continuous uniform distribution on $[0, \theta]$, where $\theta > 0$ is unknown. Suppose we assign to θ the prior pdf

$$\pi(\theta) = \begin{cases} \dfrac{q\,k^q}{\theta^{q+1}} & \text{if } \theta \geq k, \\ 0 & \text{if } \theta < k, \end{cases}$$

   where $k > 0$ and $q > 0$ are constants. *Note: This is called the* Pareto(k, q) *distribution, and its mean is* $kq/(q-1)$ *if* $q > 1$ *(and* ∞ *if* $q \leq 1$*). You may use these facts without proof.*

   (a) Find the posterior distribution of θ.

   (b) Find (or simply state) the posterior mean of θ.

   (c) Find (or simply state) the posterior mode. Then explain why choosing to report the posterior mode as our Bayes estimator here would partially defeat the purpose of using a Bayes estimator.

4. Let $X_1, \ldots, X_n$ be iid random variables with pdf or pmf of the form

$$\exp[\eta(\theta)\, r(x) - \psi(\theta)]\, h(x),$$

   where $\theta \in \Theta$ is unknown. Suppose we assign to θ the prior pdf

$$\pi(\theta) = c \exp[-a\,\psi(\theta) + b\,\eta(\theta)],$$

   where the constants a, b, and c satisfy $c^{-1} = \int_\Theta \exp[-a\,\psi(\theta) + b\,\eta(\theta)]\, d\theta < \infty$. Show that the computation of the posterior distribution of θ essentially reduces to "updating" the values of the constants a, b, and/or c from the prior distribution.

5. Let $X \sim \text{Bin}(n, \theta)$, where $0 < \theta < 1$ and θ is unknown. Let $\xi = 1/\theta$. Prove that no unbiased estimator of ξ exists. *Hint: Here, an estimator is fully specified by the value it takes for each* $x \in \{0, \ldots, n\}$*. Let* $\tilde{\xi}$ *be any arbitrary estimator of* ξ*, and let* t_x *be the value that* $\tilde{\xi}$ *takes when* $X = x$*. Then look at the form of* $E_\theta(\tilde{\xi})$ *when* $\tilde{\xi}$ *is specified in this way.*

6. Let X be drawn from a discrete uniform distribution on $\{1, \ldots, N\}$, where $N \geq 1$ is an unknown positive integer.

   (a) Show (as mentioned in class) that the maximum likelihood estimator of N is $\hat{N} = X$.

   (b) Find the bias, variance, and mean squared error of $\hat{N}$. *(The discrete uniform distribution on* $\{1, \ldots, N\}$ *has mean* $(N+1)/2$ *and variance* $(N^2-1)/12$*. You may use these facts without proof.)*

   (c) Find an estimator $\tilde{N}$ that is an unbiased estimator of N.

   (d) Find $\text{MSE}_N(\tilde{N})$, and show that $\text{MSE}_N(\tilde{N}) > \text{MSE}_N(\hat{N})$ for all $N \geq 1$.

Solutions to Homework 3

"DeGroot & Schervish $X.Y.Z$" means Exercise Z at the end of Section $X.Y$ in our text, *Probability and Statistics* (Fourth Edition) by Morris H. DeGroot and Mark J. Schervish.

1. DeGroot & Schervish 7.3.12. Also find the posterior mean and the posterior mode.

   ▷ Solution: Before substituting in the specific values given in the problem, we first find the general form of the posterior for an exponential likelihood and gamma prior. Based on what the problem tells us, we have $X_1, \ldots, X_n \sim$ iid Exp(θ) and a Gamma(a, b) prior for θ. Ignoring constants, the posterior of θ is (for $\theta > 0$)

   $$\pi(\theta \mid \boldsymbol{x}) \propto \left[\prod_{i=1}^{n} \theta \exp(-\theta x_i)\right] \theta^{a-1} \exp(-b\theta) = \theta^{a+n-1} \exp\left[-\left(b + \sum_{i=1}^{n} x_i\right)\theta\right],$$

   which we recognize as an unnormalized Gamma$(a+n,\ b+\sum_{i=1}^{n} x_i)$ pdf. Thus, the general form of the posterior is $\theta \mid \boldsymbol{x} \sim$ Gamma$(a+n,\ b+\sum_{i=1}^{n} x_i)$. We now substitute in the specific values given in the problem. Clearly we have $n = 20$, and since the problem tells us that $\overline{x} = 3.8$, it follows that $\sum_{i=1}^{n} x_i = n\overline{x} = 76$. To determine the values of a and b, note that the mean of a Gamma(a, b) distribution is a/b, while the standard deviation is $\sqrt{a/b^2} = \sqrt{a}\,/b$. The problem tells us that $a/b = 0.2$ and $\sqrt{a}\,/b = 1$, from which it follows that $a = 0.04$ and $b = 0.2$. Then the posterior distribution for this observed data set is $\theta \mid \boldsymbol{x} \sim$ Gamma$(20.04, 76.2)$. The posterior mean is then $E(\theta \mid \boldsymbol{x}) = 20.04/76.2 \approx 0.263$, while the posterior mode is $\arg\max_{\theta>0} \pi(\theta \mid \boldsymbol{x}) = (20.04 - 1)/76.2 \approx 0.250$. ◁

2. DeGroot & Schervish 7.4.6.

   ▷ Solution: Let $X_1, \ldots, X_n \sim$ iid Poisson(θ) conditional on θ, and let the prior on θ be Gamma(a, b), where we also define $\mu_o = a/b$. Ignoring constants, the posterior distribution of θ is (for $\theta > 0$)

   $$\pi(\theta \mid \boldsymbol{x}) \propto \left[\prod_{i=1}^{n} \theta^{x_i} \exp(-\theta)\right] \theta^{a-1} \exp(-b\theta) \propto \theta^{\sum_{i=1}^{n} x_i + a - 1} \exp[-\theta(n+b)],$$

   which we recognize as an unnormalized Gamma$(a+\sum_{i=1}^{n} x_i,\ b+n)$ pdf. Thus, the posterior is $\theta \mid \boldsymbol{x} \sim$ Gamma$(a + \sum_{i=1}^{n} x_i,\ b+n)$, and the posterior mean of θ is (when written as a random variable)

   $$E(\theta \mid \boldsymbol{X}) = \frac{a + \sum_{i=1}^{n} X_i}{b+n} = \left(\frac{n}{n+b}\right)\overline{X}_n + \left(\frac{b}{n+b}\right)\frac{a}{b} = \gamma_n \overline{X}_n + (1-\gamma_n)\mu_0,$$

   where $\gamma_n = n/(b+n)$. Clearly $\gamma_n = 1/(1+n^{-1}b) \to 1$ as $n \to \infty$. ◁

3. Let $X_1, \ldots, X_n$ be iid random variables with a continuous uniform distribution on $[0, \theta]$, where $\theta > 0$ is unknown. Suppose we assign to θ the prior pdf

$$\pi(\theta) = \begin{cases} \dfrac{q\, k^q}{\theta^{q+1}} & \text{if } \theta \geq k, \\ 0 & \text{if } \theta < k, \end{cases}$$

where $k > 0$ and $q > 0$ are constants. *Note: This is called the* Pareto(k, q) *distribution, and its mean is* $kq/(q-1)$ *if* $q > 1$ *(and* ∞ *if* $q \leq 1$*). You may use these facts without proof.*

(a) Find the posterior distribution of θ.

▷ Solution to (a): Ignoring constants, the posterior distribution of θ is

$$\pi(\theta \mid \boldsymbol{x}) \propto \left[\prod_{i=1}^{n} \frac{1}{\theta} I_{[0,\theta]}(x_i) \right] \frac{1}{\theta^{q+1}} I_{[k,\infty)}(\theta) = \frac{1}{\theta^{q+n+1}} I_{[\max\{k, x_1, \ldots, x_n\},\, \infty)}(\theta),$$

which we recognize as an unnormalized Pareto$(\max\{k, x_1, \ldots, x_n\},\, q+n)$ pdf. Thus, $\theta \mid \boldsymbol{x} \sim$ Pareto$(\max\{k, x_1, \ldots, x_n\},\, q+n)$. ◁

(b) Find (or simply state) the posterior mean of θ.

▷ Solution to (b): Based on the facts provided in the note, the posterior mean is $E(\theta \mid \boldsymbol{x}) = \max\{k, x_1, \ldots, x_n\}(q+n)/(q+n-1)$. ◁

(c) Find (or simply state) the posterior mode. Then explain why choosing to report the posterior mode as our Bayes estimator here would partially defeat the purpose of using a Bayes estimator.

▷ Solution to (c): It can be seen from the form of the Pareto(k, q) pdf that the mode of a Pareto(k, q) distribution is simply k. Then the posterior mode is simply $\max\{k, x_1, \ldots, x_n\}$. As long as at least one observation is larger than k, the posterior mode is exactly the same as the MLE. ◁

4. Let $X_1, \ldots, X_n$ be iid random variables with pdf or pmf of the form

$$\exp[\eta(\theta)\, r(x) - \psi(\theta)]\, h(x),$$

where $\theta \in \Theta$ is unknown. Suppose we assign to θ the prior pdf

$$\pi(\theta) = c \exp[-a\,\psi(\theta) + b\,\eta(\theta)],$$

where the constants a, b, and c satisfy $c^{-1} = \int_\Theta \exp[-a\,\psi(\theta) + b\,\eta(\theta)]\, d\theta < \infty$. Show that the computation of the posterior distribution of θ essentially reduces to "updating" the values of the constants a, b, and/or c from the prior distribution.

▷ Solution: Ignoring constants, the posterior is

$$\begin{aligned}\pi(\theta \mid \boldsymbol{x}) &\propto \left\{\prod_{i=1}^n \exp[\eta(\theta)\, r(x_i) - \psi(\theta)]\right\} \exp[-a\,\psi(\theta) + b\,\eta(\theta)] \\ &\propto \exp\left\{-(a+n)\,\psi(\theta) + \left[b + \sum_{i=1}^n r(x_i)\right]\eta(\theta)\right\},\end{aligned}$$

so the properly normalized posterior is

$$\pi(\theta \mid \boldsymbol{x}) = \frac{\exp\left\{-(a+n)\,\psi(\theta) + \left[b + \sum_{i=1}^n r(x_i)\right]\eta(\theta)\right\}}{\displaystyle\int_\Theta \exp\left\{-(a+n)\,\psi(\theta) + \left[b + \sum_{i=1}^n r(x_i)\right]\eta(\theta)\right\} d\theta}.$$

Thus, the posterior as the same form as the prior, with a replaced by $a+n$, with b replaced by $b + \sum_{i=1}^n r(x_i)$, and with c replaced by the inverse of the denominator above. ◁

5. Let $X \sim \mathrm{Bin}(n, \theta)$, where $0 < \theta < 1$ and θ is unknown. Let $\xi = 1/\theta$. Prove that no unbiased estimator of ξ exists. *Hint: Here, an estimator is fully specified by the value it takes for each $x \in \{0, \ldots, n\}$. Let $\tilde{\xi}$ be any arbitrary estimator of ξ, and let t_x be the value that $\tilde{\xi}$ takes when $X = x$. Then look at the form of $E_\theta(\tilde{\xi})$ when $\tilde{\xi}$ is specified in this way.*

▷ Solution: Let $\tilde{\xi}$ be any estimator of ξ. Using the hint, we can write $E_\theta(\tilde{\xi}\,)$ as

$$E_\theta\big(\tilde{\xi}\,\big) = \sum_{x=0}^n t_x\, P_\theta(X = x) = \sum_{x=0}^n t_x \binom{n}{x} \theta^x (1-\theta)^{n-x},$$

which is some polynomial function of θ. For $\tilde{\xi}$ to be unbiased, this polynomial function of θ must equal $1/\theta$ for all $\theta \in (0, 1)$. However, this is impossible. (If it is not immediately clear why this is impossible, note that as $\theta \downarrow 0$, the polynomial function tends to whatever finite value it takes at zero, whereas $1/\theta$ tends to ∞.) ◁

6. Let X be drawn from a discrete uniform distribution on $\{1, \ldots, N\}$, where $N \geq 1$ is an unknown positive integer.

   (a) Show (as mentioned in class) that the maximum likelihood estimator of N is $\hat{N} = X$.

   ▷ Solution to (a): The likelihood is

$$L_x(N) = \frac{1}{N} I_{\{1,\ldots,N\}}(x) = \frac{1}{N} I_{\{x,x+1,\ldots\}}(N) = \begin{cases} 0 & \text{if } N < x, \\ \dfrac{1}{N} & \text{if } N \geq x. \end{cases}$$

   Then clearly $L_x(N)$ is maximized at $N = x$, so the MLE of N is $\hat{N} = X$. ◁

   (b) Find the bias, variance, and mean squared error of $\hat{N}$. *(The discrete uniform distribution on $\{1, \ldots, N\}$ has mean $(N+1)/2$ and variance $(N^2-1)/12$. You may use these facts without proof.)*

   ▷ Solution to (b): We have

$$\begin{aligned} \text{Bias}_N(\hat{N}) &= E_N(\hat{N}) - N = E_N(X) - N = \frac{N+1}{2} - N = -\left(\frac{N-1}{2}\right), \\ \text{Var}_N(\hat{N}) &= \text{Var}(X) = \frac{N^2-1}{12}, \\ \text{MSE}_N(\hat{N}) &= \left[\text{Bias}_N(\hat{N})\right]^2 + \text{Var}_N(\hat{N}) \\ &= \left[-\left(\frac{N-1}{2}\right)\right]^2 + \frac{N^2-1}{12} = \frac{N-1}{12}[3(N-1) + (N+1)] = \frac{(N-1)(2N-1)}{6}, \end{aligned}$$

   using the facts provided. ◁

   (c) Find an estimator $\tilde{N}$ that is an unbiased estimator of N.

   ▷ Solution to (c): Let $\tilde{N} = 2X - 1$. Then $E_N(\tilde{N}) = 2\,E_N(X) - 1 = N$ by the facts provided, so $\tilde{N}$ is an unbiased estimator of N. ◁

   (d) Find $\text{MSE}_N(\tilde{N})$, and show that $\text{MSE}_N(\tilde{N}) > \text{MSE}_N(\hat{N})$ for all $N \geq 1$.

   The question contained a mistake, as the two MSEs are actually equal if $N = 1$.

   ▷ Solution to (d): Since $\text{Bias}_N(\tilde{N}) = 0$, we have

$$\text{MSE}_N(\tilde{N}) = \text{Var}_N(\tilde{N}) = \text{Var}_N(2X - 1) = 4\,\text{Var}(X) = \frac{N^2-1}{3},$$

   again using the facts provided. Then simply observe that

$$\text{MSE}_N(\tilde{N}) = \frac{(N-1)(N+1)}{3} = \frac{(N-1)(2N+2)}{6} > \frac{(N-1)(2N-1)}{6} = \text{MSE}_N(\hat{N})$$

   for all $N > 1$. (Note that if $N = 1$, then both estimators have an MSE of zero.) ◁

Homework 4: Due at 11 a.m. on February 21

"DeGroot & Schervish $X.Y.Z$" means Exercise Z at the end of Section $X.Y$ in our text, *Probability and Statistics* (Fourth Edition) by Morris H. DeGroot and Mark J. Schervish.

1. Let $X \sim \text{Bin}(1, \theta)$ (a single observation). Observe that any estimator $\tilde{\theta} = \tilde{\theta}(X)$ of θ can be expressed as
$$\tilde{\theta}(X) = \begin{cases} t_0 & \text{if } X = 0, \\ t_1 & \text{if } X = 1, \end{cases}$$
where $t_0, t_1 \in \mathbb{R}$.
   (a) Find a formula for the mean squared error of $\tilde{\theta}$, and show that it can be written as a quadratic function of θ with coefficients that depend on t_0 and/or t_1.
   (b) Compute the weighted average MSE of $\tilde{\theta}$ using the weighting function $w(\theta) = 1$. (Write your answer in terms of t_0 and t_1 only.)
   (c) Find the estimator that minimizes this weighted average MSE by finding the values of t_0 and t_1 that minimize your answer to part (b).
   (d) Now show how this particular "optimal" estimator can instead be found without doing any of the calculations from parts (a), (b), or (c).

2. DeGroot & Schervish 8.8.2.

3. DeGroot & Schervish 8.8.4.

4. DeGroot & Schervish 8.8.6.

5. DeGroot & Schervish 8.8.14.

6. Let $X_1, \ldots, X_n \sim \text{iid Bin}(1, \theta)$, where $0 < \theta < 1$ and θ is unknown.
   (a) Find the Fisher information $I(\theta)$ for the sample.
   (b) We have shown before that the maximum likelihood estimator of θ is $\hat{\theta} = n^{-1} \sum_{i=1}^{n} X_i$. Use your answer to part (a) to state the asymptotic distribution of $\hat{\theta}$. Does it agree with the result obtained by using the central limit theorem?

7. Let $X_1, \ldots, X_n \sim \text{iid Pareto}(k, \alpha)$, where the $\text{Pareto}(k, \alpha)$ distribution has pdf
$$f(x) = \begin{cases} \dfrac{\alpha k^\alpha}{x^{\alpha+1}} & \text{if } x \geq k, \\ 0 & \text{if } x < k. \end{cases}$$
Suppose that $k > 0$ is known and $\alpha > 0$ is unknown.
   (a) Find the maximum likelihood estimator $\hat{\alpha}$ of α.
   (b) Find the asymptotic distribution of $\hat{\alpha}$.

8. Again let $X_1, \ldots, X_n \sim \text{iid Pareto}(k, \alpha)$, but now suppose that $k > 0$ is unknown and $\alpha > 0$ is known. Compute $E_k[\ell'_{\boldsymbol{X}}(k)]$, and explain why your answer does not contradict Lemma 6.2.1 from the notes.

Homework 4: Due at 11 a.m. on February 21

"DeGroot & Schervish $X.Y.Z$" means Exercise Z at the end of Section $X.Y$ in our text, *Probability and Statistics* (Fourth Edition) by Morris H. DeGroot and Mark J. Schervish.

1. Let $X \sim \text{Bin}(1, \theta)$ (a single observation). Observe that any estimator $\tilde{\theta} = \tilde{\theta}(X)$ of θ can be expressed as

$$\tilde{\theta}(X) = \begin{cases} t_0 & \text{if } X = 0, \\ t_1 & \text{if } X = 1, \end{cases}$$

where $t_0, t_1 \in \mathbb{R}$.

(a) Find a formula for the mean squared error of $\tilde{\theta}$, and show that it can be written as a quadratic function of θ with coefficients that depend on t_0 and/or t_1.

▷ Solution to (a): The mean squared error of $\tilde{\theta}$ is

$$\begin{aligned} \text{MSE}_\theta(\tilde{\theta}) &= E_\theta[(\tilde{\theta} - \theta)^2] = (t_0 - \theta)^2(1 - \theta) + (t_1 - \theta)^2\theta \\ &= (t_0^2 - 2t_0\theta + \theta^2) - (t_0^2\theta - 2t_0\theta^2 + \theta^3) + (t_1^2\theta - 2t_1\theta^2 + \theta^3) \\ &= t_0^2 + (-2t_0 - t_0^2 + t_1^2)\theta + (1 + 2t_0 - 2t_1)\theta^2, \end{aligned}$$

which is a quadratic function of θ with coefficients that depend on t_0 and/or t_1. ◁

(b) Compute the weighted average MSE of $\tilde{\theta}$ using the weighting function $w(\theta) = 1$. (Write your answer in terms of t_0 and t_1 only.)

▷ Solution to (b): The weighted average MSE of $\tilde{\theta}$ with $w(\theta) = 1$ is

$$\begin{aligned} \int_0^1 \text{MSE}_\theta(\tilde{\theta})\, w(\theta)\, d\theta &= \int_0^1 [t_0^2 + (-2t_0 - t_0^2 + t_1^2)\theta + (1 + 2t_0 - 2t_1)\theta^2] d\theta \\ &= t_0^2 - t_0 - \frac{1}{2}t_0^2 + \frac{1}{2}t_1^2 + \frac{1}{3} + \frac{2}{3}t_0 - \frac{2}{3}t_1 \\ &= \frac{1}{2}t_0^2 - \frac{1}{3}t_0 + \frac{1}{2}t_1^2 - \frac{2}{3}t_1 + \frac{1}{3}, \end{aligned}$$

which is a function of t_0 and t_1 only. ◁

(c) Find the estimator that minimizes this weighted average MSE by finding the values of t_0 and t_1 that minimize your answer to part (b).

▷ Solution to (c): Setting the partial derivatives of our answer from part (b) equal to zero yields the equations

$$t_0 - \frac{1}{3} = 0, \qquad t_1 - \frac{2}{3} = 0,$$

so the the weighted average MSE is minimized by taking $t_0 = 1/3$ and $t_1 = 2/3$. ◁

(d) Now show how this particular "optimal" estimator can instead be found without doing any of the calculations from parts (a), (b), or (c).

▷ Solution to (d): By Theorem 5.2.6 from the notes, the weighted average MSE with weighting function $w(\theta)$ is minimized by the posterior mean that results from a Bayesian analysis with prior $\pi(\theta) = w(\theta)$. Now note that $w(\theta) = 1$ for $0 \le \theta \le 1$ is the pdf of a Beta$(1,1)$ distribution. Then we can simply reuse the result from Examples 4.2.1 and 4.3.1 of the notes, in which we found that the posterior mean of θ for a Bin(n,θ) likelihood and a Beta(a,b) prior on θ is

$$\hat{\theta}^B(X) = E(\theta \mid X) = \frac{x+a}{n+a+b} = \frac{x+1}{3} = \begin{cases} 1/3 & \text{if } X = 0, \\ 2/3 & \text{if } X = 1, \end{cases}$$

noting that $n = a = b = 1$. ◁

2. DeGroot & Schervish 8.8.2.

▷ Solution: The second derivative of the log-likelihood is

$$\ell_X''(p) = \frac{\partial^2}{\partial p^2}[\log p + X \log(1-p)] = \frac{\partial}{\partial p}\left(\frac{1}{p} - \frac{X}{1-p}\right) = -\frac{1}{p^2} - \frac{X}{(1-p)^2}.$$

Then the Fisher information is

$$I(p) = -E_p[\ell_X''(p)] = -E_p\left[-\frac{1}{p^2} - \frac{X}{(1-p)^2}\right] = \frac{1}{p^2} + \frac{1}{(1-p)^2}\left(\frac{1-p}{p}\right) = \frac{1}{p^2(1-p)},$$

noting that $I(p) = I_1(p)$ since there is only one observation. ◁

3. DeGroot & Schervish 8.8.4.

▷ Solution: The second derivative of the log-likelihood is

$$\ell_X''(\sigma) = \frac{\partial^2}{\partial \sigma^2}\left[-\frac{1}{2}\log(2\pi) - \log\sigma - \frac{X^2}{2\sigma^2}\right] = \frac{\partial}{\partial \sigma}\left(-\frac{1}{\sigma} + \frac{X^2}{\sigma^3}\right) = \frac{1}{\sigma^2} - \frac{3X^2}{\sigma^4}.$$

Then the Fisher information is

$$I(\sigma) = -E_\sigma[\ell_X''(\sigma)] = -E_\sigma\left(\frac{1}{\sigma^2} - \frac{3X^2}{\sigma^4}\right) = -\frac{1}{\sigma^2} + \frac{3}{\sigma^2} = \frac{2}{\sigma^2},$$

noting that $I(\sigma) = I_1(\sigma)$ since there is only one observation. ◁

4. DeGroot & Schervish 8.8.6.

▷ Solution: Let $\ell_X^{(0)}(\theta)$ denote the likelihood in terms of θ, and let $\ell_X^{(1)}(\mu)$ denote the likelihood in terms of μ. Both likelihoods are simply equal to the pdf or pmf of X, which has the same form regardless of what the parameter is considered to be. Then

$$\ell_X^{(1)}(\mu) = \ell_X^{(0)}(\theta) = \ell_X^{(0)}[\psi(\mu)].$$

It follows that

$$\frac{\partial}{\partial\mu}\ell_X^{(1)}(\mu) = \frac{\partial}{\partial\mu}\ell_X^{(0)}[\psi(\mu)].$$

Let $[\ell_X^{(1)}]'(\mu)$ denote the left-hand side of the equation above. To compute the right-hand side, recall that the chain rule for taking derivatives states that

$$\frac{d}{dt}g[h(t)] = g'[h(t)]\,h'(t).$$

Applying this result with $g = \ell_X^{(0)}$ and $h = \psi$ yields

$$\frac{\partial}{\partial\mu}\ell_X^{(0)}[\psi(\mu)] = \left[\ell_X^{(0)}\right]'[\psi(\mu)]\;\psi'(\mu),$$

and therefore

$$\left[\ell_X^{(1)}\right]'(\mu) = \psi'(\mu)\left[\ell_X^{(0)}\right]'[\psi(\mu)].$$

Now let $E_\theta^{(0)}$ denote expectation taken with a particular value of θ, and let $E_\mu^{(1)}$ denote expectation taken with a particular value of μ, so that

$$E_{\psi(\mu)}^{(0)}[g(X)] = E_\mu^{(1)}[g(X)]$$

for any random quantity $g(X)$. Then the Fisher information in terms of θ is

$$I_0(\theta) = E_\theta^{(0)}\left(\left\{\left[\ell_X^{(0)}\right]'(\theta)\right\}^2\right).$$

Finally, the Fisher information in terms of μ is

$$\begin{aligned}
I_1(\mu) = E_\mu^{(1)}\left(\left\{\left[\ell_X^{(1)}\right]'(\mu)\right\}^2\right) &= E_\mu^{(1)}\left(\left\{\psi'(\mu)\left[\ell_X^{(0)}\right]'[\psi(\mu)]\right\}^2\right) \\
&= [\psi'(\mu)]^2\;E_\mu^{(1)}\left(\left\{\left[\ell_X^{(0)}\right]'[\psi(\mu)]\right\}^2\right) \\
&= [\psi'(\mu)]^2\;E_{\psi(\mu)}^{(0)}\left(\left\{\left[\ell_X^{(0)}\right]'[\psi(\mu)]\right\}^2\right) = [\psi'(\mu)]^2\;I_0[\psi(\mu)],
\end{aligned}$$

where the final equality is obtained by observing that the expectation in the next-to-last expression is precisely the result of evaluating the function I_0 at the point $\psi(\mu)$. ◁

5. DeGroot & Schervish 8.8.14.

▷ Solution: We begin by finding $I_1(\alpha)$, the Fisher information per observation. The second derivative of the log-likelihood of a single observation is

$$\ell''_{X_1}(\alpha) = \frac{\partial^2}{\partial\alpha^2}[\alpha\log\beta - \log\Gamma(\alpha) + (\alpha-1)\log X_1 - \beta X_1] = \frac{\partial}{\partial\alpha}\left[\log\beta - \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \log X_1\right]$$
$$= -\frac{\Gamma(\alpha)\,\Gamma''(\alpha) - [\Gamma'(\alpha)]^2}{[\Gamma(\alpha)]^2},$$

and thus

$$I_1(\alpha) = -E_\alpha\big[\ell''_{X_1}(\alpha)\big] = -E_\alpha\left\{-\frac{\Gamma(\alpha)\,\Gamma''(\alpha) - [\Gamma'(\alpha)]^2}{[\Gamma(\alpha)]^2}\right\} = \frac{\Gamma(\alpha)\,\Gamma''(\alpha) - [\Gamma'(\alpha)]^2}{[\Gamma(\alpha)]^2}.$$

The desired result then follows immediately by Theorem 6.2.4 of the notes. ◁

6. Let $X_1, \ldots, X_n \sim \text{iid}\,\text{Bin}(1,\theta)$, where $0 < \theta < 1$ and θ is unknown.

(a) Find the Fisher information $I(\theta)$ for the sample.

▷ Solution to (a): We begin by finding $I_1(\theta)$, the Fisher information per observation. The second derivative of the log-likelihood of a single observation is

$$\ell''_{X_1}(\theta) = \frac{\partial^2}{\partial\theta^2}[X\log\theta + (1-X)\log(1-\theta)] = \frac{\partial}{\partial\theta}\left(\frac{X}{\theta} - \frac{1-X}{1-\theta}\right) = -\frac{X}{\theta^2} - \frac{1-X}{(1-\theta)^2},$$

and thus

$$I_1(\theta) = -E_\theta\big[\ell''_{X_1}(\theta)\big] = -E_\theta\left[-\frac{X}{\theta^2} - \frac{1-X}{(1-\theta)^2}\right] = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}.$$

Then the Fisher information for the entire sample is $I(\theta) = n\,I_1(\theta) = n/[\theta(1-\theta)]$. ◁

(b) We have shown before that the maximum likelihood estimator of θ is $\hat\theta = n^{-1}\sum_{i=1}^n X_i$. Use your answer to part (a) to state the asymptotic distribution of $\hat\theta$. Does it agree with the result obtained by using the central limit theorem?

▷ Solution to (b): By Theorem 6.2.4 of the notes,

$$\sqrt{n}\big(\hat\theta_n - \theta\big) \to_D N[0,\, \theta(1-\theta)],$$

which agrees with the result obtained by using the central limit theorem. ◁

7. Let $X_1, \ldots, X_n \sim \text{iid}\,\text{Pareto}(k, \alpha)$, where the $\text{Pareto}(k, \alpha)$ distribution has pdf

$$f(x) = \begin{cases} \dfrac{\alpha k^\alpha}{x^{\alpha+1}} & \text{if } x \geq k, \\ 0 & \text{if } x < k. \end{cases}$$

Suppose that $k > 0$ is known and $\alpha > 0$ is unknown.

(a) Find the maximum likelihood estimator $\hat{\alpha}$ of α.

▷ Solution to (a): The derivative of the log-likelihood (i.e., the score function) is

$$\begin{aligned} \ell'_{\boldsymbol{X}}(\alpha) = \frac{\partial}{\partial \alpha}\left[n \log \alpha + n\alpha \log k - (\alpha + 1) \sum_{i=1}^n \log X_i \right] &= \frac{n}{\alpha} + n \log k - \sum_{i=1}^n \log X_i \\ &= \frac{n}{\alpha} - \sum_{i=1}^n \log\left(\frac{X_i}{k}\right), \end{aligned}$$

which equal zero if and only if

$$\alpha = \left[\frac{1}{n} \sum_{i=1}^n \log\left(\frac{X_i}{k}\right) \right]^{-1}.$$

(Note that the right-hand side of the equation above is undefined if $X_1 = \cdots = X_n = k$, but we can ignore this possibility since it occurs with probability 0 for all $\alpha > 0$.) Then since this is the only critical point and the log-likelihood clearly tends to $-\infty$ as $\alpha \to 0$ and as $\alpha \to \infty$, it follows that this point is indeed the maximum. Thus,

$$\hat{\alpha}_n = \left[\frac{1}{n} \sum_{i=1}^n \log\left(\frac{X_i}{k}\right) \right]^{-1}$$

is the maximum likelihood estimator of α. ◁

(b) Find the asymptotic distribution of $\hat{\alpha}$.

▷ Solution to (b): We begin by finding $I_1(\alpha)$, the Fisher information per observation. The second derivative of the log-likelihood of a single observation is

$$\ell''_{X_1}(\alpha) = \frac{\partial}{\partial \alpha}\left(\frac{1}{\alpha} + \log k - \log X_1 \right) = -\frac{1}{\alpha^2},$$

and thus

$$I_1(\alpha) = -E_\theta\left[\ell''_{X_1}(\alpha)\right] = -E_\theta\left(-\frac{1}{\alpha^2}\right) = \frac{1}{\alpha^2}.$$

Then

$$\sqrt{n}(\hat{\alpha}_n - \alpha) \to_D N(0, \alpha^2)$$

by Theorem 6.2.4 of the notes. ◁

8. Again let $X_1, \ldots, X_n \sim \text{iid}\,\text{Pareto}(k, \alpha)$, but now suppose that $k > 0$ is unknown and $\alpha > 0$ is known. Compute $E_k[\ell'_{\boldsymbol{X}}(k)]$, and explain why your answer does not contradict Lemma 6.2.1 from the notes.

   ▷ Solution: The score function is now

$$\ell'_{\boldsymbol{X}}(k) = \frac{\partial}{\partial k}\left[n \log \alpha + n\alpha \log k - (\alpha + 1) \sum_{i=1}^{n} \log X_i\right] = \frac{n\alpha}{k},$$

   and thus

$$E_k[\ell'_{\boldsymbol{X}}(k)] = E_k\left(\frac{n\alpha}{k}\right) = \frac{n\alpha}{k} \neq 0.$$

   At first glance, it may appear that this result contradicts Lemma 6.2.1 of the notes, which states that the expectation of the score is zero. However, there is no contradiction since one of the regularity conditions of Section 6.4 is now violated, which means that Lemma 6.2.1 does not apply. Specifically, the set

$$\mathcal{X} = \{x \in \mathbb{R} : \ell_{X_1}(k) > 0\} = \{x \in \mathbb{R} : x \geq k\} = [k, \infty)$$

   depends on the unknown parameter k, which is not permitted by the regularity conditions of Section 6.4. (Note that there would be no violation here if instead k were known, as was the case in the previous question.) ◁

Homework 5: Due at 11 a.m. on March 10

"DeGroot & Schervish $X.Y.Z$" means Exercise Z at the end of Section $X.Y$ in our text, *Probability and Statistics* (Fourth Edition) by Morris H. DeGroot and Mark J. Schervish.

1. Let $X_1, \ldots, X_n$ be iid continuous random variables with a pdf $f_\theta(x)$ that is symmetric about θ, where $\theta \in \mathbb{R}$ is unknown. Suppose that $\mathrm{Var}_\theta(X_1) = \sigma^2 < \infty$ is known, which implies that $E_\theta(X_1) = \theta$. Then θ is both the true mean and true median of the pdf $f_\theta(x)$, so it seems plausible that both the sample mean $\overline{X}_n$ and the sample median, which we will call M_n, could be good estimators of θ.

   (a) Use the central limit theorem to state the asymptotic distribution of the sample mean $\overline{X}_n$.

   Suppose $f_\theta(\theta)$, the value of the pdf at the true mean (and median), satisfies $f_\theta(\theta) > 0$. Then it can be shown that the asymptotic distribution of the sample median M_n is

$$\sqrt{n}(M_n - \theta) \to_D N\left(0, \frac{1}{4[f_\theta(\theta)]^2}\right).$$

   (You do not need to show this.)

   (b) Suppose $f_\theta(x)$ is the pdf of a $N(\theta, \sigma^2)$ distribution, where $\sigma^2 > 0$ is known. Compute $\mathrm{ARE}(M_n, \overline{X}_n)$, and use it to state which estimator performs better asymptotically.

   (c) Suppose $f_\theta(x) = \frac{1}{2}\lambda \exp(-\lambda|x - \theta|)$, where $\lambda > 0$ is known. Compute $\mathrm{ARE}(M_n, \overline{X}_n)$, and use it to state which estimator performs better asymptotically. *Note: Under this pdf,* $\mathrm{Var}_\theta(X_1) = 2\lambda^{-2}$*, and you may use this fact without proof.*

   (d) Suppose that

$$f_\theta(x) = \frac{\Gamma[(p+1)/2]}{\sqrt{\pi p}\,\Gamma(p/2)}\left[1 + \frac{(x-\theta)^2}{p}\right]^{-(p+1)/2},$$

   where $p \geq 3$ is an integer. (This is the pdf of Student's t distribution with p degrees of freedom that has been "shifted" so that its mean is θ instead of zero.) Compute $\mathrm{ARE}(M_n, \overline{X}_n)$, and use it to find an integer k such that the sample median performs better asymptotically if and only if $p \leq k$. *Note: Under this pdf,* $\mathrm{Var}_\theta(X_1) = p/(p-2)$*. Also, the gamma function takes the particular values*

$$\Gamma(3/2) = \frac{\sqrt{\pi}}{2}, \qquad \Gamma(2) = 1, \qquad \Gamma(5/2) = \frac{3\sqrt{\pi}}{4}, \qquad \Gamma(3) = 2,$$

   and it satisfies the inequality

$$\frac{\Gamma[(p+1)/2]}{\Gamma(p/2)} \leq \sqrt{p/2}$$

   for all $p > 0$*. You may use any of these facts without proof.*

2. DeGroot & Schervish 8.8.10.

3. DeGroot & Schervish 8.9.14.

4. Let $X_1, \ldots, X_n \sim \text{Bin}(1, \theta)$, where $0 < \theta < 1$ and $n \geq 3$. Suppose we wish to estimate the quantity $\xi = \theta^3$.

   (a) Find an unbiased estimator of ξ that is a function of X_1, X_2, and X_3 only.

   (b) Find an unbiased estimator of ξ that is a function of the sufficient statistic $\sum_{i=1}^{n} X_i$ and has smaller mean squared error than the estimator in part (a).

5. DeGroot & Schervish 9.1.2.

6. Let X be a single observation of an $\text{Exp}(\lambda)$ random variable, which has pdf

$$f_\lambda(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

   Consider testing $H_0 : \lambda \geq \lambda_0$ versus $H_1 : \lambda < \lambda_0$.

   (a) Find the power function of the hypothesis test that rejects H_0 if and only if $X \geq c$.

   (b) Let $0 < \alpha < 1$. Find a value of c such that the test in part (a) has size α.

   (c) For what true values of λ is $P_\lambda(\text{type II error}) \geq 1/2$ for the test in part (a) with size α as in (b)?

7. Let $X_1, X_2 \sim \text{iid}\, \text{Bin}(1, \theta)$, and consider testing $H_0 : \theta = 1/3$ versus $H_1 : \theta < 1/3$.

   (a) Find a test that has size $2/9$ exactly. *Note: It does not have to be a sensible test.*

   (b) Find the power function of the test from part (a), and use it to explain why this test is not a good test of these hypotheses.

Solutions to Homework 5

"DeGroot & Schervish $X.Y.Z$" means Exercise Z at the end of Section $X.Y$ in our text, *Probability and Statistics* (Fourth Edition) by Morris H. DeGroot and Mark J. Schervish.

1. Let $X_1, \ldots, X_n$ be iid continuous random variables with a pdf $f_\theta(x)$ that is symmetric about θ, where $\theta \in \mathbb{R}$ is unknown. Suppose that $\mathrm{Var}_\theta(X_1) = \sigma^2 < \infty$ is known, which implies that $E_\theta(X_1) = \theta$. Then θ is both the true mean and true median of the pdf $f_\theta(x)$, so it seems plausible that both the sample mean $\overline{X}_n$ and the sample median, which we will call M_n, could be good estimators of θ.

   (a) Use the central limit theorem to state the asymptotic distribution of the sample mean $\overline{X}_n$.

   ▷ Solution: $\sqrt{n}(\overline{X}_n - \theta) \to_D N(0, \sigma^2)$. ◁

   Suppose $f_\theta(\theta)$, the value of the pdf at the true mean (and median), satisfies $f_\theta(\theta) > 0$. Then it can be shown that the asymptotic distribution of the sample median M_n is

$$\sqrt{n}(M_n - \theta) \to_D N\left(0, \frac{1}{4[f_\theta(\theta)]^2}\right).$$

   (You do not need to show this.)

   (b) Suppose $f_\theta(x)$ is the pdf of a $N(\theta, \sigma^2)$ distribution, where $\sigma^2 > 0$ is known. Compute $\mathrm{ARE}(M_n, \overline{X}_n)$, and use it to state which estimator performs better asymptotically.

   ▷ Solution: $\mathrm{ARE}(M_n, \overline{X}_n) = 4[f_\theta(\theta)]^2/(1/\sigma^2) = 4(2\pi\sigma^2)^{-1}\sigma^2 = 2/\pi \approx 0.64$, so the sample mean performs better asymptotically. ◁

   (c) Suppose $f_\theta(x) = \frac{1}{2}\lambda \exp(-\lambda|x - \theta|)$, where $\lambda > 0$ is known. Compute $\mathrm{ARE}(M_n, \overline{X}_n)$, and use it to state which estimator performs better asymptotically. *Note: Under this pdf,* $\mathrm{Var}_\theta(X_1) = 2\lambda^{-2}$*, and you may use this fact without proof.*

   ▷ Solution: $\mathrm{ARE}(M_n, \overline{X}_n) = 4[f_\theta(\theta)]^2/(1/\sigma^2) = 4(\frac{1}{2}\lambda)^2/(\frac{1}{2}\lambda^2) = 2$, so the sample median performs better asymptotically. ◁

(d) Suppose that

$$f_\theta(x) = \frac{\Gamma[(p+1)/2]}{\sqrt{\pi p}\,\Gamma(p/2)}\left[1 + \frac{(x-\theta)^2}{p}\right]^{-(p+1)/2},$$

where $p \geq 3$ is an integer. (This is the pdf of Student's t distribution with p degrees of freedom that has been "shifted" so that its mean is θ instead of zero.) Compute $\text{ARE}(M_n, \overline{X}_n)$, and use it to find an integer k such that the sample median performs better asymptotically if and only if $p \leq k$. *Note: Under this pdf,* $\text{Var}_\theta(X_1) = p/(p-2)$. *Also, the gamma function takes the particular values*

$$\Gamma(3/2) = \frac{\sqrt{\pi}}{2}, \qquad \Gamma(2) = 1, \qquad \Gamma(5/2) = \frac{3\sqrt{\pi}}{4}, \qquad \Gamma(3) = 2,$$

and it satisfies the inequality

$$\frac{\Gamma[(p+1)/2]}{\Gamma(p/2)} \leq \sqrt{p/2}$$

for all $p > 0$. *You may use any of these facts without proof.*

▷ Solution: First note that

$$\text{ARE}(M_n, \overline{X}_n) = \frac{4[f_\theta(\theta)]^2}{1/\sigma^2} = \frac{4}{\pi p}\left\{\frac{\Gamma[(p+1)/2]}{\Gamma(p/2)}\right\}^2 \frac{p}{p-2} = \frac{4}{\pi(p-2)}\left\{\frac{\Gamma[(p+1)/2]}{\Gamma(p/2)}\right\}^2.$$

Then for all $p \geq 6$,

$$\text{ARE}(M_n, \overline{X}_n) \leq \frac{2p}{\pi(p-2)} = \frac{2}{\pi}\left(1 - \frac{2}{p}\right)^{-1} \leq \frac{2}{\pi}\left(1 - \frac{2}{6}\right)^{-1} = \frac{3}{\pi} < 1.$$

For $p = 5$, we have

$$\text{ARE}(M_n, \overline{X}_n) = \frac{4}{\pi(5-2)}\left\{\frac{\Gamma[(5+1)/2]}{\Gamma(5/2)}\right\}^2 = \frac{4}{3\pi}\left(\frac{8}{3\sqrt{\pi}}\right)^2 = \frac{256}{27\,\pi^2} \approx 0.96.$$

For $p = 4$, we have

$$\text{ARE}(M_n, \overline{X}_n) = \frac{4}{\pi(4-2)}\left\{\frac{\Gamma[(4+1)/2]}{\Gamma(4/2)}\right\}^2 = \frac{2}{\pi}\left(\frac{3\sqrt{\pi}}{4}\right)^2 = \frac{9}{8}.$$

For $p = 3$, we have

$$\text{ARE}(M_n, \overline{X}_n) = \frac{4}{\pi(3-2)}\left\{\frac{\Gamma[(3+1)/2]}{\Gamma(3/2)}\right\}^2 = \frac{4}{\pi}\left(\frac{2}{\sqrt{\pi}}\right)^2 = \frac{16}{\pi^2} \approx 1.62.$$

Thus, the sample median performs better asymptotically if and only if $p \leq 4$. ◁

2. DeGroot & Schervish 8.8.10.

▷ Solution: We begin by finding the Fisher information per observation, taking the parameter to be σ^2. (The parameter can be taken to be σ instead, which should eventually yield the same final answer.) Then the second derivative of the log-likelihood is

$$\begin{aligned}\ell''_{X_1}(\sigma^2) = \frac{\partial^2}{\partial(\sigma^2)^2}\ell_{X_1}(\sigma^2) &= \frac{\partial^2}{\partial(\sigma^2)^2}\left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{X_1^2}{2\sigma^2}\right] \\ &= \frac{\partial}{\partial(\sigma^2)}\left[-\frac{1}{2\sigma^2} + \frac{X_1^2}{2(\sigma^2)^2}\right] = \frac{1}{2(\sigma^2)^2} - \frac{X_1^2}{(\sigma^2)^3}.\end{aligned}$$

Then

$$I_1(\sigma^2) = -E_{\sigma^2}\left[\ell''_{X_1}(\sigma^2)\right] = -\frac{1}{2(\sigma^2)^2} + \frac{E_{\sigma^2}(X_1^2)}{(\sigma^2)^3} = \frac{1}{2(\sigma^2)^2}.$$

Next, let $g(\theta) = \log(\theta^{1/2}) = \frac{1}{2}\log\theta$, so that $g(\sigma^2) = \log\sigma$. Then $g'(\theta) = 1/(2\theta)$, so

$$\frac{[g'(\sigma^2)]^2}{n\,I_1(\sigma^2)} = \frac{[1/(2\sigma^2)]^2}{n/[2(\sigma^2)^2]} = \frac{2(\sigma^2)^2}{4n(\sigma^2)^2} = \frac{1}{2n}.$$

Then by the Cramér-Rao inequality, the variance of any unbiased estimator of $\log\sigma$ is at least $1/(2n)$. ◁

3. DeGroot & Schervish 8.9.14.

▷ Solution to (a): Note that $Y \sim \text{Poisson}(n\theta)$, so

$$\begin{aligned}E_\theta[\exp(-cY)] = \sum_{y=0}^{\infty}\exp(-cy)\,\frac{(n\theta)^y\,\exp(-n\theta)}{y!} &= \exp(-n\theta)\sum_{y=0}^{\infty}\frac{[n\theta\,\exp(-c)]^y}{y!} \\ &= \exp(-n\theta)\,\exp[n\theta\,\exp(-c)] \\ &= \exp\{-n\theta\,[1-\exp(-c)]\}.\end{aligned}$$

Hence, $E_\theta[\exp(-cY)] = \exp(-\theta)$ if and only if $n[1-\exp(-c)] = 1$, which holds if and only if $c = \log[n/(n-1)]$. ◁

▷ Solution to (b): From Example 6.2.2 of the notes, the Fisher information for the sample is $I(\theta) = n/\theta$. Now let $g(\theta) = \exp(-\theta)$. Then $g'(\theta) = -\exp(-\theta)$, so

$$\frac{[g'(\theta)]^2}{I(\theta)} = \frac{[-\exp(-\theta)]^2}{n/\theta} = \frac{\theta\exp(-2\theta)}{n}.$$

Then by the Cramér-Rao inequality, a lower bound for the variance of the unbiased estimator found in part (a) is $n^{-1}\theta\exp(-2\theta)$. ◁

4. Let $X_1, \ldots, X_n \sim \text{Bin}(1, \theta)$, where $0 < \theta < 1$ and $n \geq 3$. Suppose we wish to estimate the quantity $\xi = \theta^3$. *Note: It was okay to take $X_1, \ldots, X_n$ to be independent, since otherwise you are not told enough to complete the problem.*

   (a) Find an unbiased estimator of ξ that is a function of X_1, X_2, and X_3 only.

   ▷ Solution: Take $\tilde{\xi} = X_1 X_2 X_3$. ◁

   (b) Find an unbiased estimator of ξ that is a function of the sufficient statistic $\sum_{i=1}^n X_i$ and has smaller mean squared error than the estimator in part (a).

   ▷ Solution: We apply the Rao-Blackwell theorem. Let $Y = \sum_{i=1}^n X_i$. Then

$$\begin{aligned} P\big(\tilde{\xi} = 1 \,\big|\, Y = y\big) &= P(X_1 = X_2 = X_3 = 1 \mid Y = y) \\ &= \frac{P_\theta(X_1 = X_2 = X_3 = 1,\ Y = y)}{P_\theta(Y = y)} \\ &= \frac{P_\theta\left(\sum_{i=1}^3 X_i = 3\right) P_\theta\left(\sum_{i=4}^n X_i = y - 3\right)}{P_\theta(Y = y)} \\ &= \frac{\theta^3 \, \dfrac{(n-3)!}{(y-3)!\,(n-y)!}\theta^{y-3}(1-\theta)^{n-y}}{\dfrac{n!}{y!\,(n-y)!}\theta^y(1-\theta)^{n-y}} = \frac{(n-3)!\,y!}{n!\,(y-3)!} = \frac{y(y-1)(y-2)}{n(n-1)(n-2)}. \end{aligned}$$

   Thus, $\tilde{\xi}^\star = Y(Y-1)(Y-2)/[n(n-1)(n-2)]$ is an unbiased estimator that is a function of the sufficient statistic $\sum_{i=1}^n X_i$ and has smaller mean squared error than the estimator in part (a). *Note: Technically we have not shown that the MSE of $\tilde{\xi}^\star$ is smaller than that of $\tilde{\xi}$, merely that the MSE of $\tilde{\xi}^\star$ is not larger than that of $\tilde{\xi}$. Indeed, if $n = 3$, then the estimators coincide. However, it is true that $\tilde{\xi}^\star$ has strictly smaller MSE than $\tilde{\xi}$ if $n > 3$.* ◁

5. DeGroot & Schervish 9.1.2.

   ▷ Solution to (a): The power function of the test is

$$\text{Power}(\theta) = P_\theta(Y_n \leq 1.5) = P_\theta\left(\max_{1 \leq i \leq n} X_i \leq 1.5\right) = \prod_{i=1}^n P_\theta(X_i \leq 1.5) = [P_\theta(X_1 \leq 1.5)]^n = \left(\frac{1.5}{\theta}\right)^n$$

   for $\theta \geq 1.5$, and $\text{Power}(\theta) = 1$ for $\theta < 1.5$. ◁

   ▷ Solution to (b): $\text{Power}(\theta)$ is a non-increasing function of θ, so

$$\sup_{\theta \geq 2} \text{Power}(\theta) = \text{Power}(2) = \left(\frac{1.5}{2}\right)^n = \left(\frac{3}{4}\right)^n.$$

   Thus, the size of the test is $(3/4)^n$. ◁

6. Let X be a single observation of an $\text{Exp}(\lambda)$ random variable, which has pdf

$$f_\lambda(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \geq 0, \\ 0 & \text{if } x < 0. \end{cases}$$

Consider testing $H_0 : \lambda \geq \lambda_0$ versus $H_1 : \lambda < \lambda_0$.

(a) Find the power function of the hypothesis test that rejects H_0 if and only if $X \geq c$.

▷ SOLUTION: $\text{Power}(\lambda) = P_\lambda(X \geq c) = \int_c^\infty f_\lambda(x)\, dx = \exp(-\lambda c)$. ◁

(b) Let $0 < \alpha < 1$. Find a value of c such that the test in part (a) has size α.

▷ SOLUTION: $\text{Power}(\lambda)$ is a non-increasing function of λ, so

$$\sup_{\lambda \geq \lambda_0} \text{Power}(\lambda) = \text{Power}(\lambda_0) = \exp(-\lambda_0 c).$$

Thus, the size of the test is $\exp(-\lambda_0 c)$. Then the test has size α if and only if

$$\exp(-\lambda_0 c) = \alpha \iff c = -\frac{\log \alpha}{\lambda_0},$$

noting that $\log \alpha$ is negative since $0 < \alpha < 1$. ◁

(c) For what true values of λ is $P_\lambda(\text{type II error}) \geq 1/2$ for the test in part (a) with size α as in (b)?

▷ SOLUTION: $P_\lambda(\text{type II error}) \geq 1/2$ if and only if both $\lambda < \lambda_0$ and $\text{Power}(\lambda) \leq 1/2$. The test in part (a) with size α as in (b) has power function

$$\text{Power}(\lambda) = \exp\left[-\lambda\left(-\frac{\log \alpha}{\lambda_0}\right)\right] = \alpha^{\lambda/\lambda_0},$$

and hence

$$\text{Power}(\lambda) \leq 1/2 \iff \lambda \geq -\frac{\lambda_0 \log 2}{\log \alpha},$$

again noting that $\log \alpha$ is negative. Thus, $P_\lambda(\text{type II error}) \geq 1/2$ if and only if

$$-\frac{\lambda_0 \log 2}{\log \alpha} \leq \lambda < \lambda_0.$$

(Note that if $\alpha \geq 1/2$, then there are no such values of λ.) ◁

7. Let $X_1, X_2 \sim \text{iid}\,\text{Bin}(1, \theta)$, and consider testing $H_0 : \theta = 1/3$ versus $H_1 : \theta < 1/3$.

   (a) Find a test that has size 2/9 exactly. *Note: It does not have to be a sensible test.*

   ▷ SOLUTION: Note that there are only four possible values of (X_1, X_2), i.e., the sample space consists of only four points. If $\theta = 1/3$, then

   $$(X_1, X_2) = \begin{cases} (0,0) & \text{with probability } 4/9, \\ (0,1) & \text{with probability } 2/9, \\ (1,0) & \text{with probability } 2/9, \\ (1,1) & \text{with probability } 1/9. \end{cases}$$

   Thus, the only tests with size 2/9 exactly are the test that rejects H_0 if and only if $(X_1, X_2) = (0,1)$ and the test that rejects H_0 if and only if $(X_1, X_2) = (1,0)$. ◁

   (b) Find the power function of the test from part (a), and use it to explain why this test is not a good test of these hypotheses.

   ▷ SOLUTION: $\text{Power}(\theta) = \theta(1-\theta)$ for both of the tests from part (a). Note that $\text{Power}(1/3) > \text{Power}(\theta)$ for all $\theta < 1/3$. Thus, these tests are more likely to reject H_0 if it is true than if it is false, which is exactly the opposite of what a good hypothesis test should do. ◁