STAT 512

MATHEMATICAL STATISTICS

Spring, 2011

Lecture Notes

Joshua M. Tebbs

Department of Statistics

University of South Carolina

Contents

6 Functions of Random Variables

Complementary reading: Chapter 6 (WMS).

PROBLEM: Suppose Y is a continuous random variable, and consider a function of Y, say, $U = g(Y)$, where $g : \mathcal{R} \to \mathcal{R}$. The function $U = g(Y)$ is itself a random variable, and, thus, it has its own distribution. The goal of this chapter is to find distributions of functions of random variables. When there are multiple random variables, we will be interested in functions of the form $U = g(Y_1, Y_2, ..., Y_n)$, where $g : \mathcal{R}^n \to \mathcal{R}$.

REMARK: Here are some examples where this exercise might be of interest:

- In a medical experiment, Y denotes the systolic blood pressure for a group of cancer patients. How is $U = g(Y) = \log Y$ distributed?
- A field trial is undertaken to study Y, the yield for an experimental wheat cultivar, measured in bushels/acre. How is $U = g(Y) = \sqrt{Y}$ distributed?
- An actuary is examining the distributions of claim amounts, Y_1 and Y_2, for two competing policies. What is the distribution of $U = g(Y_1, Y_2) = Y_1/(Y_1 + Y_2)$? Here, $g : \mathcal{R}^2 \to \mathcal{R}$.
- In an early-phase clinical trial, the time to death is recorded for a sample of n rats, yielding data $Y_1, Y_2, ..., Y_n$. Researchers would like to find distribution of
$$U = g(Y_1, Y_2, ..., Y_n) = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$
the average time for the sample. Here, $g : \mathcal{R}^n \to \mathcal{R}$.

PREVAILING THEME: This chapter deals with finding distributions of functions of random variables. We will investigate three techniques for doing this:

(1) **Method of distribution functions**

(2) **Method of transformations**

(3) **Method of moment generating functions**.

6.1 The method of distribution functions (or "cdf technique")

SETTING: Suppose Y is a continuous random variable with cumulative distribution function (cdf) $F_Y(y) \equiv P(Y \leq y)$. The cdf technique is especially useful when the cdf $F_Y(y)$ can be written out in closed form (although this is not a requirement). This method can also be used if Y is vector valued (see Examples 6.2 and 6.3 in WMS).

Method of distribution functions:

1. If possible, find a closed form expression for $F_Y(y) = P(Y \leq y)$.
2. Find the support of U.
3. Write $F_U(u) = P(U \leq u)$, the cdf of U, in terms of $F_Y(y)$, the cdf of Y.
4. Differentiate $F_U(u)$ to obtain the pdf of U, $f_U(u)$.

Example 6.1. Suppose that $Y \sim \mathcal{U}(0,1)$. Find the distribution of $U = g(Y) = -\ln Y$.

SOLUTION. The cdf of $Y \sim \mathcal{U}(0,1)$ is given by

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ y, & 0 < y < 1 \\ 1, & y \geq 1. \end{cases}$$

The support for $Y \sim \mathcal{U}(0,1)$ is $R_Y = \{y : 0 < y < 1\}$; thus, because $u = -\ln y > 0$ (sketch a graph of the log function), it follows that the support for U is $R_U = \{u : u > 0\}$. Using the method of distribution functions, we have

$$\begin{aligned} F_U(u) = P(U \leq u) &= P(-\ln Y \leq u) \\ &= P(\ln Y > -u) \\ &= P(Y > e^{-u}) = 1 - P(Y \leq e^{-u}) = 1 - F_Y(e^{-u}). \end{aligned}$$

Notice how we have written the cdf of U as a function of the cdf of Y. Because $F_Y(y) = y$ for $0 < y < 1$; i.e., for $u > 0$, we have

$$F_U(u) = 1 - F_Y(e^{-u}) = 1 - e^{-u}.$$

Taking derivatives, we get, for $u > 0$,

$$f_U(u) = \frac{d}{du}F_U(u) = \frac{d}{du}(1 - e^{-u}) = e^{-u}.$$

Summarizing,

$$f_U(u) = \begin{cases} e^{-u}, & u > 0 \\ 0, & \text{otherwise.} \end{cases}$$

This is an exponential pdf with mean $\beta = 1$; that is, $U \sim$ exponential(1). $\square$

Example 6.2. Suppose that $Y \sim \mathcal{U}(-\pi/2, \pi/2)$. Find the distribution of the random variable defined by $U = g(Y) = \tan(Y)$.

SOLUTION. The cdf of $Y \sim \mathcal{U}(-\pi/2, \pi/2)$ is given by

$$F_Y(y) = \begin{cases} 0, & y \leq -\pi/2 \\ \frac{y+\pi/2}{\pi}, & -\pi/2 < y < \pi/2 \\ 1, & y \geq \pi/2. \end{cases}$$

The support for Y is $R_Y = \{y : -\pi/2 < y < \pi/2\}$. Sketching a graph of the tangent function over the principal branch from $-\pi/2$ to $\pi/2$, we see that $-\infty < u < \infty$. Thus, $R_U = \{u : -\infty < u < \infty\} \equiv \mathcal{R}$, the set of all reals. Using the method of distribution functions (and recalling the inverse tangent function), we have

$$\begin{aligned} F_U(u) = P(U \leq u) &= P[\tan(Y) \leq u] \\ &= P[Y \leq \tan^{-1}(u)] = F_Y[\tan^{-1}(u)]. \end{aligned}$$

Notice how we have written the cdf of U as a function of the cdf of Y. Because $F_Y(y) = (y + \pi/2)/\pi$ for $-\pi/2 < y < \pi/2$; i.e., for $u \in \mathcal{R}$, we have

$$\begin{aligned} F_U(u) &= F_Y[\tan^{-1}(u)] \\ &= \frac{\tan^{-1}(u) + \pi/2}{\pi}. \end{aligned}$$

The pdf of U, for $u \in \mathcal{R}$, is given by

$$f_U(u) = \frac{d}{du}F_U(u) = \frac{d}{du}\left[\frac{\tan^{-1}(u) + \pi/2}{\pi}\right] = \frac{1}{\pi(1+u^2)}.$$

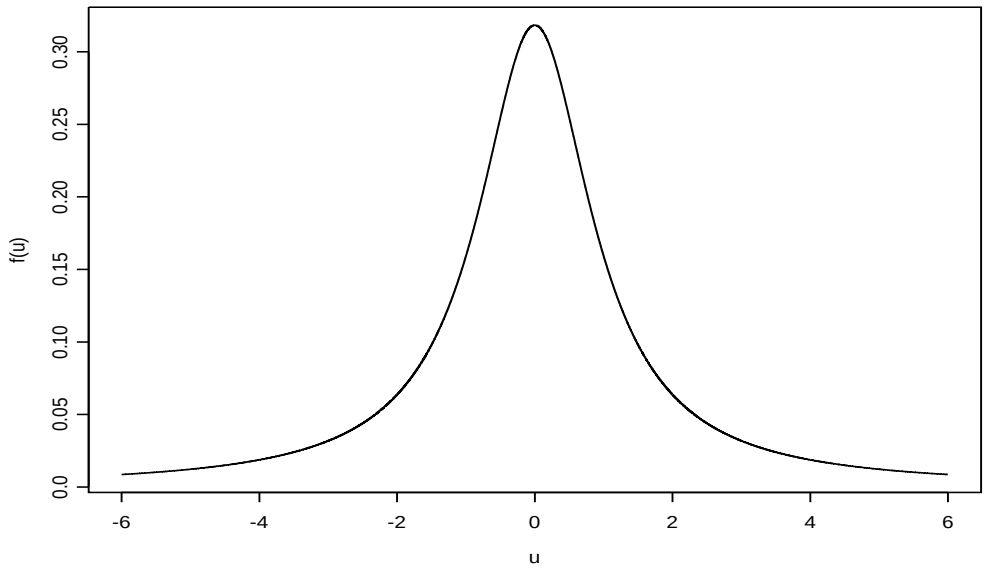

Figure 6.1: *The standard Cauchy probability density function.*

Summarizing,

$$f_U(u) = \begin{cases} \frac{1}{\pi(1+u^2)}, & -\infty < u < \infty \\ 0, & \text{otherwise.} \end{cases}$$

A random variable with this pdf is said to have a (standard) **Cauchy distribution**. One interesting fact about a Cauchy random variable is that none of its moments are finite! Thus, if U has a Cauchy distribution, $E(U)$, and all higher order moments, do not exist. EXERCISE: If U is standard Cauchy, show that $E(|U|) = +\infty$. $\square$

6.2 The method of transformations

SETTING: Suppose that Y is a continuous random variable with cdf $F_Y(y)$ and support R_Y, and let $U = g(Y)$, where $g : R_Y \to \mathcal{R}$ is a continuous, **one-to-one** function defined over R_Y. Examples of such functions include continuous (strictly) **increasing/decreasing** functions. Recall from calculus that if g is one-to-one, it has an unique inverse g^{-1}. Also recall that if g is increasing (decreasing), then so is g^{-1}.

METHOD OF TRANSFORMATIONS: Suppose that $g(y)$ is a strictly increasing function of y defined over R_Y. Then, it follows that $u = g(y) \Longleftrightarrow g^{-1}(u) = y$ and

$$\begin{aligned} F_U(u) = P(U \leq u) &= P[g(Y) \leq u] \\ &= P[Y \leq g^{-1}(u)] = F_Y[g^{-1}(u)]. \end{aligned}$$

Differentiating $F_U(u)$ with respect to u, we get

$$f_U(u) = \frac{d}{du}F_U(u) = \frac{d}{du}F_Y[g^{-1}(u)] = f_Y[g^{-1}(u)] \underbrace{\frac{d}{du}g^{-1}(u)}_{\text{chain rule}}.$$

Now as g is increasing, so is g^{-1}; thus, $\frac{d}{du}g^{-1}(u) > 0$. If $g(y)$ is strictly decreasing, then $F_U(u) = 1 - F_Y[g^{-1}(u)]$ and $\frac{d}{du}g^{-1}(u) < 0$ (verify!), which gives

$$f_U(u) = \frac{d}{du}F_U(u) = \frac{d}{du}\{1 - F_Y[g^{-1}(u)]\} = -f_Y[g^{-1}(u)]\frac{d}{du}g^{-1}(u).$$

Combining both cases, we have shown that the pdf of U, where nonzero, is given by

$$f_U(u) = f_Y[g^{-1}(u)]\left|\frac{d}{du}g^{-1}(u)\right|.$$

It is again important to keep track of the support for U. If R_Y denotes the support of Y, then R_U, the support for U, is given by $R_U = \{u : u = g(y);\ y \in R_Y\}$.

Method of transformations:

1. Verify that the transformation $u = g(y)$ is continuous and one-to-one over R_Y.
2. Find the support of U.
3. Find the inverse transformation $y = g^{-1}(u)$ and its derivative (with respect to u).
4. Use the formula above for $f_U(u)$.

Example 6.3. Suppose that $Y \sim$ exponential(β); i.e., the pdf of Y is

$$f_Y(y) = \begin{cases} \frac{1}{\beta}e^{-y/\beta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Let $U = g(Y) = \sqrt{Y}$. Use the method of transformations to find the pdf of U.

SOLUTION. First, we note that the transformation $g(y) = \sqrt{y}$ is a continuous strictly increasing function of y over $R_Y = \{y : y > 0\}$, and, thus, $g(y)$ is one-to-one. Next, we need to find the support of U. This is easy since $y > 0$ implies $u = \sqrt{y} > 0$ as well. Thus, $R_U = \{u : u > 0\}$. Now, we find the inverse transformation:

$$g(y) = u = \sqrt{y} \Longleftrightarrow \underbrace{y = g^{-1}(u) = u^2}_{\text{inverse transformation}}$$

and its derivative:

$$\frac{d}{du}g^{-1}(u) = \frac{d}{du}(u^2) = 2u.$$

Thus, for $u > 0$,

$$\begin{aligned} f_U(u) &= f_Y[g^{-1}(u)]\left|\frac{d}{du}g^{-1}(u)\right| \\ &= \frac{1}{\beta}e^{-u^2/\beta} \times |2u| = \frac{2u}{\beta}e^{-u^2/\beta}. \end{aligned}$$

Summarizing,

$$f_U(u) = \begin{cases} \frac{2u}{\beta}e^{-u^2/\beta}, & u > 0 \\ 0, & \text{otherwise.} \end{cases}$$

This is a **Weibull distribution** with parameters $m = 2$ and $\alpha = \beta$; see Exercise 6.26 in WMS. The Weibull family of distributions is common in engineering and actuarial science applications. □

Example 6.4. Suppose that $Y \sim \text{beta}(\alpha = 6, \beta = 2)$; i.e., the pdf of Y is given by

$$f_Y(y) = \begin{cases} 42y^5(1-y), & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

What is the distribution of $U = g(Y) = 1 - Y$?

SOLUTION. First, we note that the transformation $g(y) = 1-y$ is a continuous decreasing function of y over $R_Y = \{y : 0 < y < 1\}$, and, thus, $g(y)$ is one-to-one. Next, we need to find the support of U. This is easy since $0 < y < 1$ clearly implies $0 < u < 1$. Thus, $R_U = \{u : 0 < u < 1\}$. Now, we find the inverse transformation:

$$g(y) = u = 1 - y \Longleftrightarrow \underbrace{y = g^{-1}(u) = 1 - u}_{\text{inverse transformation}}$$

and its derivative:

$$\frac{d}{du}g^{-1}(u) = \frac{d}{du}(1-u) = -1.$$

Thus, for $0 < u < 1$,

$$\begin{aligned} f_U(u) &= f_Y[g^{-1}(u)]\left|\frac{d}{du}g^{-1}(u)\right| \\ &= 42(1-u)^5[1-(1-u)] \times |-1| = 42u(1-u)^5. \end{aligned}$$

Summarizing,

$$f_U(u) = \begin{cases} 42u(1-u)^5, & 0 < u < 1 \\ 0, & \text{otherwise.} \end{cases}$$

We recognize this is a beta distribution with parameters $\alpha = 2$ and $\beta = 6$. $\square$

QUESTION: What happens if $u = g(y)$ is not a one-to-one transformation? In this case, we can still use the method of transformations, but we have "break up" the transformation $g : R_Y \to R_U$ into disjoint regions where g is one-to-one.

RESULT: Suppose that Y is a continuous random variable with pdf $f_Y(y)$ and that $U = g(Y)$, not necessarily a one-to-one (but continuous) function of y over R_Y. Furthermore, suppose that we can partition R_Y into a finite collection of sets, say, $B_1, B_2, ..., B_k$, where $P(Y_i \in B_i) > 0$ for all i, and $f_Y(y)$ is continuous on each B_i. Furthermore, suppose that there exist functions $g_1(y), g_2(y), ..., g_k(y)$ such that $g_i(y)$ is defined on B_i, $i = 1, 2, ..., k$, and the $g_i(y)$ satisfy

(a) $g(y) = g_i(y)$ for all $y \in B_i$

(b) $g_i(y)$ is monotone on B_i, so that $g_i^{-1}(\cdot)$ exists uniquely on B_i.

Then, the pdf of U is given by

$$f_U(u) = \begin{cases} \sum_{i=1}^{k} f_Y[g_i^{-1}(u)]\left|\frac{d}{du}g_i^{-1}(u)\right|, & u \in R_U \\ 0, & \text{otherwise.} \end{cases}$$

That is, writing the pdf of U can be done by adding up the terms $f_Y[g_i^{-1}(u)]|\frac{d}{du}g_i^{-1}(u)|$ corresponding to each disjoint set B_i, for $i = 1, 2, ..., k$.

Example 6.5. Suppose that $Y \sim \mathcal{N}(0,1)$; that is, Y has a standard normal distribution; i.e.,

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Consider the transformation $U = g(Y) = Y^2$. This transformation is not one-to-one on $R_Y = \mathcal{R} = \{y : -\infty < y < \infty\}$, but it is one-to-one on $B_1 = (-\infty, 0)$ and $B_2 = [0, \infty)$ (separately) since $g(y) = y^2$ is decreasing on B_1 and increasing on B_2. Furthermore, note that B_1 and B_2 partitions R_Y. Summarizing,

Partition	Transformation	Inverse transformation
$B_1 = (-\infty, 0)$	$g_1(y) = y^2 = u$	$g_1^{-1}(u) = -\sqrt{u} = y$
$B_2 = [0, \infty)$	$g_2(y) = y^2 = u$	$g_2^{-1}(u) = \sqrt{u} = y$

And, on both sets B_1 and B_2,

$$\left| \frac{d}{du} g_i^{-1}(u) \right| = \frac{1}{2\sqrt{u}}.$$

Clearly, $u = y^2 > 0$; thus, $R_U = \{u : u > 0\}$, and the pdf of U is given by

$$f_U(u) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-(-\sqrt{u})^2/2} \left(\frac{1}{2\sqrt{u}} \right) + \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{u})^2/2} \left(\frac{1}{2\sqrt{u}} \right), & u > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Thus, for $u > 0$, and recalling that $\Gamma(1/2) = \sqrt{\pi}$, $f_U(u)$ collapses to

$$\begin{aligned} f_U(u) &= \frac{2}{\sqrt{2\pi}} e^{-u/2} \left(\frac{1}{2\sqrt{u}} \right) \\ &= \frac{1}{\sqrt{2\pi}} u^{\frac{1}{2}-1} e^{-u/2} = \frac{1}{\sqrt{\pi} 2^{1/2}} u^{\frac{1}{2}-1} e^{-u/2} = \frac{1}{\Gamma(1/2) 2^{1/2}} u^{\frac{1}{2}-1} e^{-u/2}. \end{aligned}$$

Summarizing, the pdf of U is

$$f_U(u) = \begin{cases} \frac{1}{\Gamma(1/2) 2^{1/2}} u^{\frac{1}{2}-1} e^{-u/2}, & u > 0 \\ 0, & \text{otherwise.} \end{cases}$$

That is, $U \sim \text{gamma}(1/2, 2)$. Recall that the gamma$(1/2, 2)$ distribution is the same as a χ^2 distribution with 1 degree of freedom; that is, $U \sim \chi^2(1)$. $\square$

6.3 Several independent random variables

RECALL: In STAT 511, we talked about the notion of independence when dealing with n-variate random vectors. Recall that $Y_1, Y_2, ..., Y_n$ are (mutually) **independent** random variables if and only if

$$F_{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{i=1}^{n} F_{Y_i}(y_i)$$

or, equivalently, if and only if

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \prod_{i=1}^{n} f_{Y_i}(y_i).$$

That is, the joint cdf $F_{\boldsymbol{Y}}(\boldsymbol{y})$ factors into the product of the marginal cdfs. Similarly, the joint pdf (pmf) $f_{\boldsymbol{Y}}(\boldsymbol{y})$ factors into the product of the marginal pdfs (pmfs).

NOTATION REMINDER: The random vector $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$. A realization of $\boldsymbol{Y}$ is $\boldsymbol{y} = (y_1, y_2, ..., y_n)$. $\boldsymbol{Y}$ is random; $\boldsymbol{y}$ is fixed.

MATHEMATICAL EXPECTATION: Suppose that $Y_1, Y_2, ..., Y_n$ are (mutually) **independent** random variables. For real valued functions $g_1, g_2, ..., g_n$,

$$E[g_1(Y_1)g_2(Y_2)\cdots g_n(Y_n)] = E[g_1(Y_1)]E[g_2(Y_2)]\cdots E[g_n(Y_n)],$$

provided that each expectation exists; that is, the expectation of the product is the product of the expectations. **This result only holds for independent random variables!**

Proof. We'll prove this for the continuous case (the discrete case follows analogously). Suppose that $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$ is a vector of (mutually) independent random variables with joint pdf $f_{\boldsymbol{Y}}(\boldsymbol{y})$. Then,

$$\begin{aligned}
E\left[\prod_{i=1}^{n} g_i(Y_i)\right] &= \int_{\mathcal{R}^n} [g_1(y_1)g_2(y_2)\cdots g_n(y_n)] f_{\boldsymbol{Y}}(\boldsymbol{y}) d\boldsymbol{y} \\
&= \int_{\mathcal{R}}\int_{\mathcal{R}}\cdots\int_{\mathcal{R}} [g_1(y_1)g_2(y_2)\cdots g_n(y_n)] f_{Y_1}(y_1) f_{Y_2}(y_2)\cdots f_{Y_n}(y_n) d\boldsymbol{y} \\
&= \int_{\mathcal{R}} g_1(y_1) f_{Y_1}(y_1) dy_1 \int_{\mathcal{R}} g_2(y_2) f_{Y_2}(y_2) dy_2 \cdots \int_{\mathcal{R}} g_n(y_n) f_{Y_n}(y_n) dy_n \\
&= E[g_1(Y_1)]E[g_2(Y_2)]\cdots E[g_n(Y_n)]. \quad \square
\end{aligned}$$

IMPORTANT: Suppose that $a_1, a_2, ..., a_n$ are constants and that $Y_1, Y_2, ..., Y_n$ are **independent** random variables, where Y_i has mgf $m_{Y_i}(t)$, for $i = 1, 2, ..., n$. Define the **linear combination**

$$U = \sum_{i=1}^{n} a_i Y_i = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n.$$

Then, the moment generating function of U is given by

$$m_U(t) = \prod_{i=1}^{n} m_{Y_i}(a_i t).$$

Proof. Using the definition, the moment generating function of U is

$$\begin{aligned} m_U(t) = E(e^{tU}) &= E\left[e^{t(a_1Y_1+a_2Y_2+\cdots+a_nY_n)}\right] \\ &= E\left(e^{a_1tY_1}e^{a_2tY_2}\cdots e^{a_ntY_n}\right) \\ &= E(e^{a_1tY_1})E(e^{a_2tY_2})\cdots E(e^{a_ntY_n}) \\ &= m_{Y_1}(a_1t)m_{Y_2}(a_2t)\cdots m_{Y_n}(a_nt) = \prod_{i=1}^{n} m_{Y_i}(a_it). \ \square \end{aligned}$$

COROLLARY: If $a_1 = a_2 = \cdots = a_n = 1$ in the last result, the linear combination $U = \sum_{i=1}^{n} Y_i$ and

$$m_U(t) = \prod_{i=1}^{n} m_{Y_i}(t).$$

That is, the mgf of the **sum** $U = \sum_{i=1}^{n} Y_i$ is the product of the marginal mgfs.

Example 6.6. Suppose that $Y_1, Y_2, ..., Y_n$ are independent $\mathcal{N}(\mu_i, \sigma_i^2)$ random variables for $i = 1, 2, ..., n$. Find the distribution of the linear combination

$$U = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n.$$

SOLUTION. Because $Y_1, Y_2, ..., Y_n$ are independent, we know from the last result that

$$\begin{aligned} m_U(t) &= \prod_{i=1}^{n} m_{Y_i}(a_it) \\ &= \prod_{i=1}^{n} \exp[\mu_i(a_it) + \sigma_i^2(a_it)^2/2] \\ &= \exp\left[\left(\sum_{i=1}^{n} a_i\mu_i\right)t + \left(\sum_{i=1}^{n} a_i^2\sigma_i^2\right)t^2/2\right]. \end{aligned}$$

We recognize this as the moment generating function of a normal random variable with mean $E(U) = \sum_{i=1}^{n} a_i\mu_i$ and variance $V(U) = \sum_{i=1}^{n} a_i^2\sigma_i^2$. Because mgfs are unique, we may conclude that

$$U \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i\mu_i, \sum_{i=1}^{n} a_i^2\sigma_i^2\right).$$

That is, the distribution of a linear combination of independent normal random variables is normally distributed. □

CONCEPTUALIZATION: In many statistical problems, a collection of random variables, say, $Y_1, Y_2, ..., Y_n$ can be viewed as independent observations from the same probability model. Statisticians like to call this common model the **population distribution** because, at least conceptually, we can envisage the observations $Y_1, Y_2, ..., Y_n$ as being randomly drawn from a population where $f_Y(y)$ describes the population; i.e., the pdf (pmf) $f_Y(y)$ describes how the observations $Y_1, Y_2, ..., Y_n$ are marginally distributed.

IID OBSERVATIONS: Suppose that $Y_1, Y_2, ..., Y_n$ are **independent** observations, where each Y_i has the common pdf (pmf) $f_Y(y)$. A succinct way to express this is to say that

"$Y_1, Y_2, ..., Y_n$ is an **iid** sample from $f_Y(y)$."

The collection $Y_1, Y_2, ..., Y_n$ is often called a **random sample**, and the model $f_Y(y)$ represents the population distribution. The acronym "iid" is read "**i**ndependent and **i**dentically **d**istributed."

REMARK: With an iid sample $Y_1, Y_2, ..., Y_n$ from $f_Y(y)$, there may be certain characteristics of $f_Y(y)$ that we would like investigate, especially if the exact form of $f_Y(y)$ is not known. For example, we might like to **estimate** the mean or variance of the distribution; i.e., we might like to estimate $E(Y) = \mu$ and/or $V(Y) = \sigma^2$. An obvious **estimator** for $E(Y) = \mu$ is the **sample mean**

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i;$$

i.e., the arithmetic average of the sample $Y_1, Y_2, ..., Y_n$. An estimator for $V(Y) = \sigma^2$ is

the **sample variance**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

Both $\overline{Y}$ and S^2 are values that are computed from the sample; i.e., they are computed from the observations (i.e., data) $Y_1, Y_2, ..., Y_n$, so they are called **statistics**. Note that

$$E(\overline{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \mu$$

and

$$V(\overline{Y}) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{\sigma^2}{n}.$$

That is, the mean of sample mean $\overline{Y}$ is the same as the underlying population mean μ. The variance of the sample mean $\overline{Y}$ equals the population variance σ^2 divided by n (the sample size).

Example 6.7. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y)$, where

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

That is, $Y_1, Y_2, ..., Y_n \sim$ iid $\mathcal{N}(\mu, \sigma^2)$. What is the distribution of the sample mean $\overline{Y}$?

SOLUTION. It is important to recognize that the sample mean $\overline{Y}$ is simply a linear combination of the observations $Y_1, Y_2, ..., Y_n$, with $a_1 = a_2 = \cdots = a_n = \frac{1}{n}$; i.e.,

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{Y_1}{n} + \frac{Y_2}{n} + \cdots + \frac{Y_n}{n}.$$

We know that $Y_1, Y_2, ..., Y_n$ are iid $\mathcal{N}(\mu, \sigma^2)$ so

$$\overline{Y} \sim \mathcal{N}\left(\sum_{i=1}^{n} a_i\mu, \sum_{i=1}^{n} a_i^2\sigma^2\right),$$

where $a_1 = a_2 = \cdots = a_n = \frac{1}{n}$; that is,

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

PUNCHLINE: If we have an iid sample of normal observations, the sample mean $\overline{Y}$ is also normally distributed. □

6.4 The method of moment generating functions

UNIQUENESS: Suppose that Z_1 and Z_2 are random variables with mgfs $m_{Z_1}(t)$ and $m_{Z_2}(t)$, respectively. If $m_{Z_1}(t) = m_{Z_2}(t)$ for all t, then Z_1 and Z_2 have the same distribution. This is called the **uniqueness property** of moment generating functions.

PUNCHLINE: The mgf completely determines the distribution! How can we use this result? Suppose that we have a transformation $U = g(Y)$ or $U = g(Y_1, Y_2, ..., Y_n)$. If we can compute $m_U(t)$, the mgf of U, and can recognize it as one we already know (e.g., Poisson, normal, gamma, binomial, etc.), then we can use the uniqueness property to conclude that U has that distribution (we've been doing this informally all along; see, e.g., Example 6.6).

REMARK: When $U = g(Y)$, using the mgf method requires us to know the mgf of Y up front. Thus, if you do not know $m_Y(t)$, it is best to try another method. This turns out to be true because, in executing the mgf technique, we must be able to express the mgf of U as a function of the mgf of Y (as we'll see in the examples which follow). Similarly, if $U = g(Y_1, Y_2, ..., Y_n)$, the mgf technique is not helpful unless you know the marginal mgfs $m_{Y_1}(t), m_{Y_2}(t), ..., m_{Y_n}(t)$.

Method of moment generating functions:

1. Derive the mgf of U, which is given by $m_U(t) = E(e^{tU})$.
2. Try to recognize $m_U(t)$ as a moment generating function that you already know.
3. Because mgfs are unique, U must have the same distribution as the one whose mgf you recognized.

Example 6.8. Suppose that $Y \sim \text{gamma}(\alpha, \beta)$. Use the method of mgfs to derive the distribution of $U = g(Y) = 2Y/\beta$.

SOLUTION. We know that the mgf of Y is

$$m_Y(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha},$$

for $t < 1/\beta$. Now, the mgf of U is given by

$$\begin{aligned} m_U(t) = E(e^{tU}) = E[e^{t(2Y/\beta)}] &= E[e^{(2t/\beta)Y}] \\ &= m_Y(2t/\beta) \\ &= \left[\frac{1}{1-\beta(2t/\beta)}\right]^{\alpha} = \left(\frac{1}{1-2t}\right)^{\alpha}, \end{aligned}$$

for $t < 1/2$. However, we recognize $m_U(t) = (1-2t)^{-\alpha}$ as the $\chi^2(2\alpha)$ mgf. Thus, by uniqueness, we can conclude that $U = 2Y/\beta \sim \chi^2(2\alpha)$. $\square$

MGF TECHNIQUE: The method of moment generating functions is very useful (and commonly applied) when we have **independent** random variables $Y_1, Y_2, ..., Y_n$ and interest lies in deriving the distribution of the sum

$$U = g(Y_1, Y_2, ..., Y_n) = Y_1 + Y_2 + \cdots + Y_n.$$

In particular, we know that

$$m_U(t) = \prod_{i=1}^{n} m_{Y_i}(t),$$

where $m_{Y_i}(t)$ denotes the marginal mgf of Y_i. Of course, if $Y_1, Y_2, ..., Y_n$ are **iid**, then not only are the random variables independent, they also all have the same distribution! Thus, because mgfs are unique, the mgfs must be the same too. Summarizing, if $Y_1, Y_2, ..., Y_n$ are **iid**, each with mgf $m_Y(t)$,

$$m_U(t) = \prod_{i=1}^{n} m_Y(t) = [m_Y(t)]^n.$$

Example 6.9. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from

$$p_Y(y) = \begin{cases} p^y(1-p)^{1-y}, & y = 0, 1 \\ 0, & \text{otherwise.} \end{cases}$$

That is, $Y_1, Y_2, ..., Y_n$ are iid Bernoulli(p) random variables. What is the distribution of the sum $U = Y_1 + Y_2 + \cdots + Y_n$?

SOLUTION. Recall that the Bernoulli mgf is given by $m_Y(t) = q + pe^t$, where $q = 1 - p$. Using the last result, we know that

$$m_U(t) = [m_Y(t)]^n = (q + pe^t)^n,$$

which we recognize as the mgf of a $b(n,p)$ random variable! Thus, by the uniqueness property of mgfs, we have that $U = Y_1 + Y_2 + \cdots + Y_n \sim b(n,p)$. $\square$

Example 6.10. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from

$$f_Y(y) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} y^{\alpha-1} e^{-y/\beta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

That is, $Y_1, Y_2, ..., Y_n$ are iid gamma(α, β) random variables. What is the distribution of the sum $U = Y_1 + Y_2 + \cdots + Y_n$?

SOLUTION. Recall that the gamma mgf is, for $t < 1/\beta$,

$$m_Y(t) = \left(\frac{1}{1-\beta t}\right)^\alpha.$$

Using the last result we know that, for $t < 1/\beta$,

$$m_U(t) = [m_Y(t)]^n = \left[\left(\frac{1}{1-\beta t}\right)^\alpha\right]^n = \left(\frac{1}{1-\beta t}\right)^{\alpha n},$$

which we recognize as the mgf of a gamma$(\alpha n, \beta)$ random variable. Thus, by the uniqueness property of mgfs, we have that $U = Y_1 + Y_2 + \cdots + Y_n \sim$ gamma$(\alpha n, \beta)$. $\square$

COROLLARY: If $Y_1, Y_2, ..., Y_n$ is an iid sample of exponential random variables with mean β, then $U = Y_1 + Y_2 + \cdots + Y_n \sim$ gamma(n, β). This follows from Example 6.10 by taking $\alpha = 1$. $\square$

Example 6.11. As another special case of Example 6.10, take $\alpha = 1/2$ and $\beta = 2$ so that $Y_1, Y_2, ..., Y_n$ are iid $\chi^2(1)$ random variables. The result in Example 6.10 says that $U = Y_1 + Y_2 + \cdots + Y_n \sim$ gamma$(n/2, 2)$ which is the same as the $\chi^2(n)$ distribution. Thus, the sum of independent $\chi^2(1)$ random variables follows a $\chi^2(n)$ distribution. $\square$

GENERALIZATION: If $Y_1, Y_2, ..., Y_n$ are independent (not iid) random variables where $Y_i \sim \chi^2(\nu_i)$, then $U = Y_1 + Y_2 + \cdots + Y_n \sim \chi^2(\nu)$, where $\nu = \sum_i \nu_i$.

Example 6.12. Suppose that $Y_1, Y_2, ..., Y_n$ are independent $\mathcal{N}(\mu_i, \sigma_i^2)$ random variables. Find the distribution of

$$U = \sum_{i=1}^n \left(\frac{Y_i - \mu_i}{\sigma_i}\right)^2.$$

SOLUTION. Define

$$Z_i = \frac{Y_i - \mu_i}{\sigma_i},$$

for each $i = 1, 2, ..., n$. Observe the following facts.

- $Z_1, Z_2, ..., Z_n$ are independent $\mathcal{N}(0,1)$ random variables. That $Z_i \sim \mathcal{N}(0,1)$ follows from standardization. That $Z_1, Z_2, ..., Z_n$ are independent follows because **functions of independent random variables are themselves independent**.
- From Example 6.5, we know that $Z_1^2, Z_2^2, ..., Z_n^2$ are independent $\chi^2(1)$ random variables. This is true because $Z_i \sim \mathcal{N}(0,1) \Longrightarrow Z_i^2 \sim \chi^2(1)$ and because $Z_1^2, Z_2^2, ..., Z_n^2$ are functions of $Z_1, Z_2, ..., Z_n$ (which are independent).
- Finally, from Example 6.11 we know that

$$U = \sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{\sigma_i} \right)^2 = \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n). \; \square$$

6.5 Bivariate transformations

REMARK: So far in this chapter, we have talked about transformations involving a single random variable Y. It is sometimes of interest to consider a **bivariate transformation** such as

$$\begin{aligned} U_1 &= g_1(Y_1, Y_2) \\ U_2 &= g_2(Y_1, Y_2). \end{aligned}$$

To discuss such transformations, we will assume that Y_1 and Y_2 are jointly **continuous** random variables. Furthermore, for the following methods to apply, the transformation needs to be one-to-one. We start with the joint distribution of $\boldsymbol{Y} = (Y_1, Y_2)$. Our first goal is to derive the joint distribution of $\boldsymbol{U} = (U_1, U_2)$.

BIVARIATE TRANSFORMATIONS: Suppose that $\boldsymbol{Y} = (Y_1, Y_2)$ is a continuous random vector with joint pdf $f_{Y_1,Y_2}(y_1, y_2)$. Let $g : \mathcal{R}^2 \to \mathcal{R}^2$ be a continuous one-to-one vector-valued mapping from R_{Y_1,Y_2} to R_{U_1,U_2}, where $U_1 = g_1(Y_1, Y_2)$ and $U_2 = g_2(Y_1, Y_2)$,

and where R_{Y_1,Y_2} and R_{U_1,U_2} denote the two-dimensional supports of $\boldsymbol{Y} = (Y_1, Y_2)$ and $\boldsymbol{U} = (U_1, U_2)$, respectively. If $g_1^{-1}(u_1, u_2)$ and $g_2^{-1}(u_1, u_2)$ have continuous partial derivatives with respect to both u_1 and u_2, and the Jacobian, J, where, with "det" denoting "determinant",

$$J = \det \begin{vmatrix} \frac{\partial g_1^{-1}(u_1,u_2)}{\partial u_1} & \frac{\partial g_1^{-1}(u_1,u_2)}{\partial u_2} \\ \frac{\partial g_2^{-1}(u_1,u_2)}{\partial u_1} & \frac{\partial g_2^{-1}(u_1,u_2)}{\partial u_2} \end{vmatrix} \neq 0,$$

then

$$f_{U_1,U_2}(u_1, u_2) = \begin{cases} f_{Y_1,Y_2}[g_1^{-1}(u_1, u_2), g_2^{-1}(u_1, u_2)]|J|, & (u_1, u_2) \in R_{U_1,U_2} \\ 0, & \text{otherwise,} \end{cases}$$

where $|J|$ denotes the absolute value of J.

RECALL: The determinant of a 2×2 matrix, e.g.,

$$\det \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

IMPORTANT: When performing a bivariate transformation, the function $g : \mathcal{R}^2 \to \mathcal{R}^2$ must be one-to-one. In addition, we need to keep track of what the transformation $U_1 = g_1(Y_1, Y_2), U_2 = g_2(Y_1, Y_2)$ "does" to the support R_{Y_1,Y_2}. Remember, g is a vector-valued function that maps points in R_{Y_1,Y_2} to R_{U_1,U_2}.

Steps to perform a bivariate transformation:

1. Find $f_{Y_1,Y_2}(y_1, y_2)$, the joint distribution of Y_1 and Y_2. This may be given in the problem. If Y_1 and Y_2 are independent, then $f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1)f_{Y_2}(y_2)$.
2. Find R_{U_1,U_2}, the support of $\boldsymbol{U} = (U_1, U_2)$.
3. Find the inverse transformations $y_1 = g_1^{-1}(u_1, u_2)$ and $y_2 = g_2^{-1}(u_1, u_2)$.
4. Find the Jacobian, J, of the inverse transformation.
5. Use the formula above to find $f_{U_1,U_2}(u_1, u_2)$, the joint distribution of U_1 and U_2.

NOTE: If desired, marginal distributions $f_{U_1}(u_1)$ and $f_{U_2}(u_2)$ can be found by integrating the joint distribution $f_{U_1,U_2}(u_1, u_2)$ as we learned in STAT 511.

Example 6.13. Suppose that $Y_1 \sim$ gamma$(\alpha, 1)$, $Y_2 \sim$ gamma$(\beta, 1)$, and that Y_1 and Y_2 are independent. Define the transformation

$$\begin{aligned} U_1 &= g_1(Y_1, Y_2) = Y_1 + Y_2 \\ U_2 &= g_2(Y_1, Y_2) = \frac{Y_1}{Y_1 + Y_2}. \end{aligned}$$

Find each of the following distributions:

(a) $f_{U_1,U_2}(u_1, u_2)$, the joint distribution of U_1 and U_2,

(b) $f_{U_1}(u_1)$, the marginal distribution of U_1, and

(c) $f_{U_2}(u_2)$, the marginal distribution of U_2.

SOLUTIONS. (a) Since Y_1 and Y_2 are independent, the joint distribution of Y_1 and Y_2 is

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2) &= f_{Y_1}(y_1) f_{Y_2}(y_2) \\ &= \frac{1}{\Gamma(\alpha)} y_1^{\alpha-1} e^{-y_1} \times \frac{1}{\Gamma(\beta)} y_2^{\beta-1} e^{-y_2} \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} y_1^{\alpha-1} y_2^{\beta-1} e^{-(y_1+y_2)}, \end{aligned}$$

for $y_1 > 0$, $y_2 > 0$, and 0, otherwise. Here, $R_{Y_1,Y_2} = \{(y_1, y_2) : y_1 > 0, y_2 > 0\}$. By inspection, we see that $u_1 = y_1 + y_2 > 0$, and $u_2 = \frac{y_1}{y_1+y_2}$ must fall between 0 and 1. Thus, the support of $\boldsymbol{U} = (U_1, U_2)$ is given by

$$R_{U_1,U_2} = \{(u_1, u_2) : u_1 > 0, 0 < u_2 < 1\}.$$

The next step is to derive the inverse transformation. It follows that

$$\begin{array}{c} u_1 = g_1(y_1, y_2) = y_1 + y_2 \\ u_2 = g_2(y_1, y_2) = \frac{y_1}{y_1+y_2} \end{array} \Longrightarrow \begin{array}{c} y_1 = g_1^{-1}(u_1, u_2) = u_1 u_2 \\ y_2 = g_2^{-1}(u_1, u_2) = u_1 - u_1 u_2 \end{array}.$$

The Jacobian is given by

$$J = \det \begin{vmatrix} \frac{\partial g_1^{-1}(u_1,u_2)}{\partial u_1} & \frac{\partial g_1^{-1}(u_1,u_2)}{\partial u_2} \\ \frac{\partial g_2^{-1}(u_1,u_2)}{\partial u_1} & \frac{\partial g_2^{-1}(u_1,u_2)}{\partial u_2} \end{vmatrix} = \det \begin{vmatrix} u_2 & u_1 \\ 1 - u_2 & -u_1 \end{vmatrix} = -u_2 u_1 - u_1(1 - u_2) = -u_1.$$

We now write the joint distribution for $\boldsymbol{U} = (U_1, U_2)$. For $u_1 > 0$ and $0 < u_2 < 1$, we have that

$$\begin{aligned} f_{U_1,U_2}(u_1,u_2) &= f_{Y_1,Y_2}[g_1^{-1}(u_1,u_2), g_2^{-1}(u_1,u_2)]|J| \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)}(u_1u_2)^{\alpha-1}(u_1-u_1u_2)^{\beta-1}e^{-[u_1u_2+(u_1-u_1u_2)]} \times |-u_1|. \end{aligned}$$

Rewriting this expression, we get

$$f_{U_1,U_2}(u_1,u_2) = \begin{cases} \frac{u_2^{\alpha-1}(1-u_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}u_1^{\alpha+\beta-1}e^{-u_1}, & u_1 > 0, 0 < u_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

ASIDE: We see that U_1 and U_2 are **independent** since the support $R_{U_1,U_2} = \{(u_1,u_2) : u_1 > 0, 0 < u_2 < 1\}$ does not constrain u_1 by u_2 or vice versa and since the nonzero part of $f_{U_1,U_2}(u_1,u_2)$ can be factored into the two expressions $h_1(u_1)$ and $h_2(u_2)$, where

$$h_1(u_1) = u_1^{\alpha+\beta-1}e^{-u_1}$$

and

$$h_2(u_2) = \frac{u_2^{\alpha-1}(1-u_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}.$$

(b) To obtain the marginal distribution of U_1, we integrate the joint pdf $f_{U_1,U_2}(u_1,u_2)$ over u_2. That is, for $u_1 > 0$,

$$\begin{aligned} f_{U_1}(u_1) &= \int_{u_2=0}^{1} f_{U_1,U_2}(u_1,u_2)du_2 \\ &= \int_{u_2=0}^{1} \frac{u_2^{\alpha-1}(1-u_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)}u_1^{\alpha+\beta-1}e^{-u_1}du_2 \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)}u_1^{\alpha+\beta-1}e^{-u_1}\int_{u_2=0}^{1}\underbrace{u_2^{\alpha-1}(1-u_2)^{\beta-1}}_{\text{beta}(\alpha,\beta)\text{kernel}}du_2 \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)}u_1^{\alpha+\beta-1}e^{-u_1} \times \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \\ &= \frac{1}{\Gamma(\alpha+\beta)}u_1^{\alpha+\beta-1}e^{-u_1}. \end{aligned}$$

Summarizing,

$$f_{U_1}(u_1) = \begin{cases} \frac{1}{\Gamma(\alpha+\beta)}u_1^{\alpha+\beta-1}e^{-u_1}, & u_1 > 0, \\ 0, & \text{otherwise.} \end{cases}$$

We recognize this as a gamma$(\alpha+\beta, 1)$ pdf; thus, marginally, $U_1 \sim$ gamma$(\alpha+\beta, 1)$.

(c) To obtain the marginal distribution of U_2, we integrate the joint pdf $f_{U_1,U_2}(u_1,u_2)$ over u_1. That is, for $0 < u_2 < 1$,

$$\begin{aligned} f_{U_2}(u_2) &= \int_{u_1=0}^{\infty} f_{U_1,U_2}(u_1,u_2)du_1 \\ &= \int_{u_1=0}^{\infty} \frac{u_2^{\alpha-1}(1-u_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} u_1^{\alpha+\beta-1}e^{-u_1}du_1 \\ &= \frac{u_2^{\alpha-1}(1-u_2)^{\beta-1}}{\Gamma(\alpha)\Gamma(\beta)} \underbrace{\int_{u_1=0}^{\infty} u_1^{\alpha+\beta-1}e^{-u_1}du_1}_{=\,\Gamma(\alpha+\beta)} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u_2^{\alpha-1}(1-u_2)^{\beta-1}. \end{aligned}$$

Summarizing,

$$f_{U_2}(u_2) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u_2^{\alpha-1}(1-u_2)^{\beta-1}, & 0 < u_2 < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Thus, marginally, $U_2 \sim \text{beta}(\alpha,\beta)$. $\square$

REMARK: Suppose that $\boldsymbol{Y} = (Y_1, Y_2)$ is a continuous random vector with joint pdf $f_{Y_1,Y_2}(y_1,y_2)$, and suppose that we would like to find the distribution of a single random variable

$$U_1 = g_1(Y_1, Y_2).$$

Even though there is no U_2 present here, the bivariate transformation technique can still be useful! In this case, we can define a "dummy variable" $U_2 = g_2(Y_1,Y_2)$ that is of no interest to us, perform the bivariate transformation to obtain $f_{U_1,U_2}(u_1,u_2)$, and then find the marginal distribution of U_1 by integrating $f_{U_1,U_2}(u_1,u_2)$ out over the dummy variable u_2. While the choice of U_2 is arbitrary, there are certainly bad choices. Stick with something easy; usually $U_2 = g_2(Y_1,Y_2) = Y_2$ does the trick.

EXERCISE: Suppose that Y_1 and Y_2 are random variables with joint pdf

$$f_{Y_1,Y_2}(y_1,y_2) = \begin{cases} 8y_1y_2, & 0 < y_1 < y_2 < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Find the pdf of $U_1 = Y_1/Y_2$.

REMARK: The transformation method can also be extended to handle n-variate transformations. Suppose that $Y_1, Y_2, ..., Y_n$ are continuous random variables with joint pdf $f_{\boldsymbol{Y}}(\boldsymbol{y})$ and define

$$\begin{aligned} U_1 &= g_1(Y_1, Y_2, ..., Y_n) \\ U_2 &= g_2(Y_1, Y_2, ..., Y_n) \\ &\vdots \\ U_n &= g_n(Y_1, Y_2, ..., Y_n). \end{aligned}$$

If this transformation is one-to-one, the procedure that we discussed for the bivariate case extends straightforwardly; see WMS, pp 330.

6.6 Order statistics

DEFINITION: Suppose that $Y_1, Y_2, ..., Y_n$ are iid observations from $f_Y(y)$. As we have discussed, the values $Y_1, Y_2, ..., Y_n$ can be envisioned as a **random sample** from a population where $f_Y(y)$ describes the behavior of individuals in this population. Define

$$\begin{aligned} Y_{(1)} &= \text{smallest of } Y_1, Y_2, ..., Y_n \\ Y_{(2)} &= \text{second smallest of } Y_1, Y_2, ..., Y_n \\ &\vdots \\ Y_{(n)} &= \text{largest of } Y_1, Y_2, ..., Y_n. \end{aligned}$$

The new random variables, $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ are called **order statistics**; they are simply the observations $Y_1, Y_2, ..., Y_n$ ordered from low to high.

GOALS: We are interested in understanding how a single order statistic is distributed (e.g., minimum, maximum, sample median, etc.). In addition, we might want to derive the distribution of a function of order statistics, say, $R = Y_{(n)} - Y_{(1)}$, the sample range. Throughout our discussion, we assume that the observations $Y_1, Y_2, ..., Y_n$ are continuous so that, theoretically, ties are not possible.

PDF OF $Y_{(1)}$: Suppose that $Y_1, Y_2, ..., Y_n$ are iid observations from the pdf $f_Y(y)$ or, equivalently, from the cdf $F_Y(y)$. To derive $f_{Y_{(1)}}(y)$, the marginal pdf of the **minimum** order statistic, we will use the distribution function technique. The cdf of $Y_{(1)}$ is

$$\begin{aligned} F_{Y_{(1)}}(y) &= P(Y_{(1)} \leq y) \\ &= 1 - P(Y_{(1)} > y) \\ &= 1 - P(\{Y_1 > y\} \cap \{Y_2 > y\} \cap \cdots \cap \{Y_n > y\}) \\ &= 1 - P(Y_1 > y)P(Y_2 > y) \cdots P(Y_n > y) \\ &= 1 - [P(Y_1 > y)]^n = 1 - [1 - F_Y(y)]^n. \end{aligned}$$

Thus, for values of y in the support of $Y_{(1)}$,

$$\begin{aligned} f_{Y_{(1)}}(y) &= \frac{d}{dy} F_{Y_{(1)}}(y) \\ &= \frac{d}{dy}\{1 - [1 - F_Y(y)]^n\} \\ &= -n[1 - F_Y(y)]^{n-1}[-f_Y(y)] = nf_Y(y)[1 - F_Y(y)]^{n-1}, \end{aligned}$$

and 0, otherwise. This is the marginal pdf of the minimum order statistic. □

Example 6.14. An engineering system consists of 5 components placed in series; that is, the system fails when the first component fails. Suppose that the $n = 5$ component lifetimes $Y_1, Y_2, ..., Y_5$ are assumed to be iid exponential observations with mean β. Since the system fails when the first component fails, system failures can be determined (at least, probabilistically) by deriving the pdf of $Y_{(1)}$, the minimum order statistic. Recall that for the exponential model, the pdf is

$$f_Y(y) = \begin{cases} \frac{1}{\beta} e^{-y/\beta}, & y > 0 \\ 0, & \text{otherwise} \end{cases}$$

and the cdf is given by

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ 1 - e^{-y/\beta}, & y > 0. \end{cases}$$

Using the formula for the pdf of the minimum order statistic, we see that, with $n = 5$

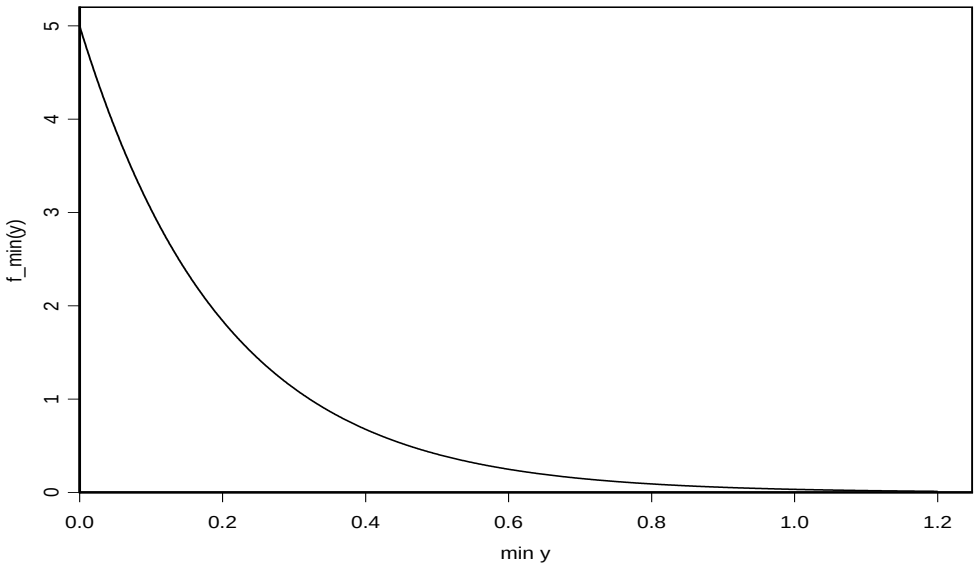

Figure 6.2: *The probability density function of $Y_{(1)}$, the minimum order statistic in Example* 6.14 *when $\beta = 1$ year. This represents the distribution of the lifetime of a series system, which is exponential with mean* $1/5$.

components, the distribution of the lifetime of the series system is given by

$$\begin{aligned} f_{Y_{(1)}}(y) &= nf_Y(y)[1 - F_Y(y)]^{n-1} \\ &= 5\left(\frac{1}{\beta}e^{-y/\beta}\right)\left[1 - \left(1 - e^{-y/\beta}\right)\right]^{5-1} \\ &= \frac{5}{\beta}e^{-y/\beta}\left(e^{-y/\beta}\right)^4 \\ &= \frac{5}{\beta}e^{-5y/\beta} = \frac{1}{(\beta/5)}e^{-y/(\beta/5)}, \end{aligned}$$

for $y > 0$. That is, the minimum order statistic $Y_{(1)}$, which measures the lifetime of the system, follows an exponential distribution with mean $E(Y_{(1)}) = \beta/5$. □

Example 6.15. Suppose that, in Example 6.14, the mean component lifetime is $\beta = 1$ year, and that an engineer is claiming the system with these settings will likely last at least 6 months (before repair is needed). Is there evidence to support his claim?

SOLUTION. We can compute the probability that the system lasts longer than 6 months,

which occurs when $Y_{(1)} > 0.5$. Using the pdf for $Y_{(1)}$ (see Figure 6.2), we have

$$P(Y_{(1)} > 0.5) = \int_{0.5}^{\infty} \frac{1}{1/5} e^{-y/(1/5)} dy = \int_{0.5}^{\infty} 5e^{-5y} dy \approx 0.082.$$

Thus, chances are that the system would not last longer than six months. There is not very much evidence to support the engineer's claim. □

PDF OF $Y_{(n)}$: Suppose that $Y_1, Y_2, ..., Y_n$ are iid observations from the pdf $f_Y(y)$ or, equivalently, from the cdf $F_Y(y)$. To derive $f_{Y_{(n)}}(y)$, the marginal pdf of the **maximum** order statistic, we will use the distribution function technique. The cdf of $Y_{(n)}$ is

$$\begin{aligned} F_{Y_{(n)}}(y) &= P(Y_{(n)} \leq y) \\ &= P(\{Y_1 \leq y\} \cap \{Y_2 \leq y\} \cap \cdots \cap \{Y_n \leq y\}) \\ &= P(Y_1 \leq y)P(Y_2 \leq y) \cdots P(Y_n \leq y) \\ &= [P(Y_1 \leq y)]^n = [F_Y(y)]^n. \end{aligned}$$

Thus, for values of y in the support of $Y_{(n)}$,

$$\begin{aligned} f_{Y_{(n)}}(y) &= \frac{d}{dy} F_{Y_{(n)}}(y) \\ &= \frac{d}{dy} \{[F_Y(y)]^n\} \\ &= n f_Y(y)[F_Y(y)]^{n-1}, \end{aligned}$$

and 0, otherwise. This is the marginal pdf of the maximum order statistic. □

Example 6.16. The proportion of rats that successfully complete a designed experiment (e.g., running through a maze) is of interest for psychologists. Denote by Y the proportion of rats that complete the experiment, and suppose that the experiment is replicated in 10 different rooms. Assume that $Y_1, Y_2, ..., Y_{10}$ are iid beta random variables with $\alpha = 2$ and $\beta = 1$. Recall that for this beta model, the pdf is

$$f_Y(y) = \begin{cases} 2y, & 0 < y < 1 \\ 0, & \text{otherwise.} \end{cases}$$

Find the pdf of $Y_{(10)}$, the largest order statistic. Also, calculate $P(Y_{(10)} > 0.90)$.

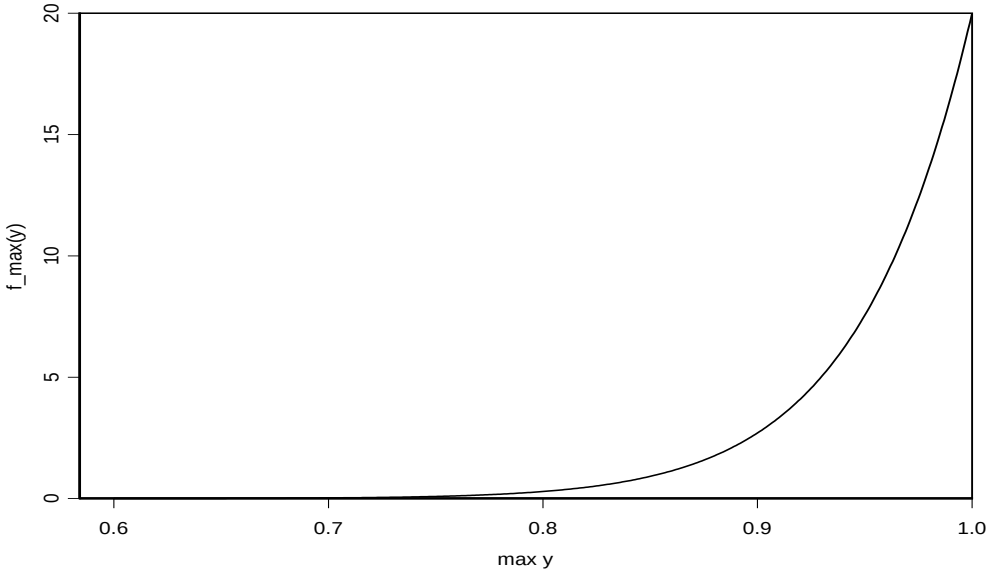

Figure 6.3: *The pdf for $Y_{(10)}$, the largest order statistic in Example* 6.16.

SOLUTION. Direct calculation shows that the cdf of Y is given by

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ y^2, & 0 < y < 1 \\ 1, & y \geq 1. \end{cases}$$

Using the formula for the pdf of the maximum order statistic, for $0 < y < 1$,

$$f_{Y_{(10)}}(y) = nf_Y(y)[F_Y(y)]^{n-1} = 10(2y)(y^2)^9 = 20y^{19}.$$

Thus, the pdf of $Y_{(10)}$ is given by

$$f_{Y_{(10)}}(y) = \begin{cases} 20y^{19}, & 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

and this probability density function is depicted in Figure 6.3. Note that this is the pdf of a beta($\alpha = 20, \beta = 1$) random variable; i.e., $Y_{(10)} \sim$ beta$(20, 1)$. Furthermore,

$$P(Y_{(10)} > 0.90) = \int_{0.90}^{1} 20y^{19}dy = y^{20}\Big|_{0.90}^{1} = 1 - (0.9)^{20} \approx 0.88. \ \square$$

PDF OF $Y_{(k)}$: Suppose that $Y_1, Y_2, ..., Y_n$ are iid observations from the pdf $f_Y(y)$ or, equivalently, from the cdf $F_Y(y)$. To derive $f_{Y_{(k)}}(y)$, the pdf of the k**th order statistic**, we appeal to a multinomial-type argument. Define

Class	Description	#Y_i's
1	the Y_i's less than y	$k-1$
2	the Y_i's equal to y	1
3	the Y_i's greater than y	$n-k$

Thus, since $Y_1, Y_2, ..., Y_n$ are independent, we have, by appeal to the multinomial model,

$$f_{Y_{(k)}}(y) = \frac{n!}{(k-1)!1!(n-k)!}[F_Y(y)]^{k-1}[f_Y(y)]^1[1-F_Y(y)]^{n-k},$$

where we interpret

$$\begin{aligned} F_Y(y) &= P(Y_i < y) \\ f_Y(y) &= P(Y_i = y) \\ 1 - F_Y(y) &= P(Y_i > y). \end{aligned}$$

Thus, the pdf of the kth order statistic $Y_{(k)}$ is given by

$$f_{Y_{(k)}}(y) = \frac{n!}{(k-1)!(n-k)!}[F_Y(y)]^{k-1}f_Y(y)[1-F_Y(y)]^{n-k},$$

for values of y in the support of $Y_{(k)}$, and 0, otherwise. □

Example 6.17. Suppose that $Y_1, Y_2, ..., Y_n$ are iid $\mathcal{U}(0,1)$ observations. What is the distribution of the kth order statistic $Y_{(k)}$?

SOLUTION. Recall that for this model, the pdf is

$$f_Y(y) = \begin{cases} 1, & 0 < y < 1 \\ 0, & \text{otherwise} \end{cases}$$

and the cdf of Y is

$$F_Y(y) = \begin{cases} 0, & y \leq 0 \\ y, & 0 < y < 1 \\ 1, & y \geq 1. \end{cases}$$

Using the formula for the pdf of the kth order statistic, we have, for $0 < y < 1$,

$$\begin{aligned} f_{Y_{(k)}}(y) &= \frac{n!}{(k-1)!(n-k)!}[F_Y(y)]^{k-1} f_Y(y)[1-F_Y(y)]^{n-k} \\ &= \frac{n!}{(k-1)!(n-k)!} y^{k-1}(1-y)^{n-k} \\ &= \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} y^{k-1}(1-y)^{(n-k+1)-1}. \end{aligned}$$

You should recognize this as a beta pdf with $\alpha = k$ and $\beta = n - k + 1$. That is, $Y_{(k)} \sim \text{beta}(k, n-k+1)$. $\square$

TWO ORDER STATISTICS: Suppose that $Y_1, Y_2, ..., Y_n$ are iid observations from the pdf $f_Y(y)$ or, equivalently, from the cdf $F_Y(y)$. For $j < k$, the joint distribution of $Y_{(j)}$ and $Y_{(k)}$ is

$$\begin{aligned} f_{Y_{(j)},Y_{(k)}}(y_j, y_k) &= \frac{n!}{(j-1)!(k-1-j)!(n-k)!}[F_Y(y_j)]^{j-1} \\ &\quad \times f_Y(y_j)[F_Y(y_k) - F_Y(y_j)]^{k-1-j} f_Y(y_k)[1-F_Y(y_k)]^{n-k}, \end{aligned}$$

for values of $y_j < y_k$ in the support of $Y_{(j)}$ and $Y_{(k)}$, and 0, otherwise. $\square$

REMARK: Informally, this result can again be derived using a multinomial-type argument, only this time, using the 5 classes

Class	Description	#Y_i's
1	the Y_i's less than y_j	$j-1$
2	the Y_i's equal to y_j	1
3	the Y_i's greater than y_j but less than y_k	$k-1-j$
4	the Y_i's equal to y_k	1
5	the Y_i's greater than y_k	$n-k$

EXERCISE: Suppose that $Y_1, Y_2, ..., Y_5$ is an iid sample of $n = 5$ exponential observations with mean $\beta = 1$.

(a) Find the joint distribution of $Y_{(1)}$ and $Y_{(5)}$.

(b) Find the probability that the sample range $R = Y_{(5)} - Y_{(1)}$ exceeds 2. That is, compute $P(R > 2) = P(Y_{(5)} - Y_{(1)} > 2)$. *Hint:* You have the joint distribution of $Y_{(1)}$ and $Y_{(5)}$ in part (a).

7 Sampling Distributions and the Central Limit Theorem

Complementary reading: Chapter 7 (WMS).

7.1 Introduction

REMARK: For the remainder of this course, we will often treat a collection of random variables $Y_1, Y_2, ..., Y_n$ as a **random sample**. This is understood to mean that

- the random variables $Y_1, Y_2, ..., Y_n$ are independent
- each Y_i has common pdf (pmf) $f_Y(y)$. This probability model $f_Y(y)$ can be discrete (e.g., Bernoulli, Poisson, geometric, etc.) or continuous (e.g., normal, gamma, uniform, etc.). It could also be a mixture of continuous and discrete parts.

REVIEW: In mathematical statistics, it is common to refer to a collection of random variables with these properties as an **iid sample**. The acronymn "iid" means "**i**ndependent and **i**dentically **d**istributed." The model $f_Y(y)$ is called the **population distribution**; it represents the distribution from which the sample values $Y_1, Y_2, ..., Y_n$ are drawn.

DEFINITION: A **statistic**, say T, is a function of the random variables $Y_1, Y_2, ..., Y_n$. A statistic can depend on known constants, but it can not depend on unknown parameters.

NOTE: To emphasize the dependence of T on $Y_1, Y_2, ..., Y_n$, we may write

$$T = T(Y_1, Y_2, ..., Y_n).$$

In addition, while it will often be the case that $Y_1, Y_2, ..., Y_n$ constitute a random sample (i.e., that they are iid), our definition of a statistic T holds in more general settings. In practice, it is common to view $Y_1, Y_2, ..., Y_n$ as **data** from an experiment or observational study and T as some summary measure (e.g., sample mean, sample variance, etc.).

Example 7.1. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y)$. For example, each of the following are statistics:

- $T(Y_1, Y_2, ..., Y_n) = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, the sample mean.
- $T(Y_1, Y_2, ..., Y_n) = \frac{1}{2}[Y_{(n/2)} + Y_{([n/2]+1)}]$, the sample median (if n is even).
- $T(Y_1, Y_2, ..., Y_n) = Y_{(1)}$, the minimum order statistic.
- $T(Y_1, Y_2, ..., Y_n) = Y_{(n)} - Y_{(1)}$, the sample range.
- $T(Y_1, Y_2, ..., Y_n) = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2$, the sample variance.

IMPORTANT: Since $Y_1, Y_2, ..., Y_n$ are random variables, any statistic $T = T(Y_1, Y_2, ..., Y_n)$, being a function of $Y_1, Y_2, ..., Y_n$, is also a random variable. Thus, T has, among other characteristics, its own mean, its own variance, and its own probability distribution!

DEFINITION: The probability distribution of a statistic T is called the **sampling distribution** of T. The sampling distribution of T describes mathematically how the values of T vary in repeated sampling from the population distribution $f_Y(y)$. Sampling distributions play a central role in statistics.

7.2 Sampling distributions related to the normal distribution

Example 7.2. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, and consider the statistic

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

the **sample mean**. From Example 6.7 (notes), we know that

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

Furthermore, the quantity

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1). \ \square$$

Example 7.3. In the interest of pollution control, an experimenter records Y, the amount of bacteria per unit volume of water (measured in mg/cm^3). The population distribution for Y is assumed to be normal with mean $\mu = 48$ and variance $\sigma^2 = 100$; that is, $Y \sim \mathcal{N}(48, 100)$. As usual, let Z denote a standard normal random variable.

(a) What is the probability that a single water specimen's bacteria amount will exceed 50 mg/cm^3?

SOLUTION. Here, we use the population distribution $\mathcal{N}(48, 100)$ to compute

$$\begin{aligned} P(Y > 50) &= P\left(Z > \frac{50 - 48}{10}\right) \\ &= P(Z > 0.2) = 0.4207. \end{aligned}$$

(b) Suppose that the experimenter takes a random sample of $n = 100$ water specimens, and denote the observations by $Y_1, Y_2, ..., Y_{100}$. What is the probability that the sample mean $\overline{Y}$ will exceed 50 mg/cm^3?

SOLUTION. Here, we need to use the sampling distribution of the sample mean $\overline{Y}$. Since the population distribution is $\mathcal{N}(48, 100)$, we know that

$$\overline{Y} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \sim \mathcal{N}\left(48, 1\right).$$

Thus,

$$\begin{aligned} P(\overline{Y} > 50) &= P\left(Z > \frac{50 - 48}{1}\right) \\ &= P(Z > 2) = 0.0228. \ \square \end{aligned}$$

EXERCISE: How large should the sample size n be so that $P(\overline{Y} > 50) < 0.01$?

RECALL: If $Y_1, Y_2, ..., Y_n$ are independent $\mathcal{N}(\mu_i, \sigma_i^2)$ random variables, then

$$\sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{\sigma_i}\right)^2 \sim \chi^2(n).$$

We proved this in the last chapter. See Example 6.12 (notes).

SPECIAL CASE: If $Y_1, Y_2, ..., Y_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, then

$$\sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma}\right)^2 \sim \chi^2(n).$$

NEW RESULT: If $Y_1, Y_2, ..., Y_n$ are iid $\mathcal{N}(\mu, \sigma^2)$, then

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{Y_i - \overline{Y}}{\sigma}\right)^2 \sim \chi^2(n-1).$$

In addition, $\overline{Y}$ and S^2 are **independent**.

REMARK: We will not prove the independence result, in general; this would be proven in a more advanced course, although WMS proves this for the $n = 2$ case. The statistics $\overline{Y}$ and S^2 are independent only if the observations $Y_1, Y_2, ..., Y_n$ are iid $\mathcal{N}(\mu, \sigma^2)$. If the normal model changes (or does not hold), then $\overline{Y}$ and S^2 are no longer independent.

Proof. We will prove that

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

First, we write

$$\begin{aligned}\underbrace{\sum_{i=1}^{n}\left(\frac{Y_i - \mu}{\sigma}\right)^2}_{W_1} &= \sum_{i=1}^{n}\left(\frac{Y_i - \overline{Y} + \overline{Y} - \mu}{\sigma}\right)^2 \\ &= \underbrace{\sum_{i=1}^{n}\left(\frac{Y_i - \overline{Y}}{\sigma}\right)^2}_{W_2} + \underbrace{\sum_{i=1}^{n}\left(\frac{\overline{Y} - \mu}{\sigma}\right)^2}_{W_3},\end{aligned}$$

since the cross product

$$2\sum_{i=1}^{n}\left(\frac{Y_i - \overline{Y}}{\sigma}\right)\left(\frac{\overline{Y} - \mu}{\sigma}\right) = 0.$$

Now, we know that $W_1 \sim \chi^2(n)$. Also, we can rewrite W_3 as

$$\begin{aligned}\sum_{i=1}^{n}\left(\frac{\overline{Y} - \mu}{\sigma}\right)^2 &= n\left(\frac{\overline{Y} - \mu}{\sigma}\right)^2 \\ &= \left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2(1),\end{aligned}$$

since

$$\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1),$$

and the square of a standard normal is distributed as $\chi^2(1)$. So, we have

$$\begin{aligned}W_1 &= W_2 + W_3 \\ &= \frac{(n-1)S^2}{\sigma^2} + W_3.\end{aligned}$$

Since W_2 is a function of S^2 and W_3 is a function of $\overline{Y}$, W_2 and W_3 are independent. Thus, the mgf of W_1 is given by

$$\begin{aligned} m_{W_1}(t) = E(e^{tW_1}) &= E\left\{e^{t[(n-1)S^2/\sigma^2+W_3]}\right\} \\ &= E\left\{e^{t[(n-1)S^2/\sigma^2]}e^{tW_3}\right\} \\ &= E\left\{e^{t[(n-1)S^2/\sigma^2]}\right\}E(e^{tW_3}). \end{aligned}$$

But, $m_{W_1}(t) = (1-2t)^{-n/2}$ since $W_1 \sim \chi^2(n)$ and $m_{W_3}(t) = (1-2t)^{-1/2}$ since $W_3 \sim \chi^2(1)$; both of these mgfs are valid for $t < 1/2$. Thus, it follows that

$$(1-2t)^{-n/2} = E\left\{e^{t[(n-1)S^2/\sigma^2]}\right\}(1-2t)^{-1/2}.$$

Hence, it must be the case that

$$E\left\{e^{t[(n-1)S^2/\sigma^2]}\right\} = E(e^{tW_2}) = m_{W_2}(t) = (1-2t)^{-(n-1)/2},$$

for values of $t < 1/2$. Thus, $W_2 \sim \chi^2(n-1)$ by the uniqueness property of mgfs. □

Example 7.4. In an ecological study examining the effects of Hurricane Katrina, researchers choose $n = 9$ plots and, for each plot, record Y, the amount of dead weight material (recorded in grams). Denote the nine dead weights by $Y_1, Y_2, ..., Y_9$, where Y_i represents the dead weight for plot i. The researchers model the data $Y_1, Y_2, ..., Y_9$ as an iid $\mathcal{N}(100, 32)$ sample. What is the probability that the sample variance S^2 of the nine dead weights is less than 20? That is, what is $P(S^2 < 20)$?

SOLUTION. We know that

$$\frac{(n-1)S^2}{\sigma^2} = \frac{8S^2}{32} \sim \chi^2(8).$$

Thus,

$$\begin{aligned} P(S^2 < 20) &= P\left[\frac{8S^2}{32} < \frac{8(20)}{32}\right] \\ &= P[\chi^2(8) < 5] \approx 0.24. \ \square \end{aligned}$$

Note that the table of χ^2 probabilities (Table 6, pp 794-5, WMS) offers little help in computing $P[\chi^2(8) < 5]$. I found this probability using the `pchisq(5,8)` command in R.

EXERCISE: How large should the sample size n be so that $P(S^2 < 20) < 0.01$?

7.2.1 The t distribution

THE t DISTRIBUTION: Suppose that $Z \sim \mathcal{N}(0,1)$ and that $W \sim \chi^2(\nu)$. If Z and W are independent, then the random variable

$$T = \frac{Z}{\sqrt{W/\nu}}$$

has a t **distribution** with ν degrees of freedom. This is denoted $T \sim t(\nu)$.

THE t PDF: Suppose that the random variable T has a t distribution with ν degrees of freedom. The pdf for T is given by

$$f_T(t) = \begin{cases} \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\ \Gamma(\nu/2)}(1+t^2/\nu)^{-(\nu+1)/2}, & -\infty < t < \infty \\ 0, & \text{otherwise.} \end{cases}$$

REMARK: It is possible to derive the t pdf using a bivariate transformation argument. The good news is that, in practice, we will never use the formula for the t pdf to find probabilities. Computing gives areas (probabilities) upon request; in addition, tabled values (giving limited probabilities) are readily available. See Table 5 (WMS).

FACTS ABOUT THE t DISTRIBUTION:

- continuous and **symmetric** about 0
- indexed by a parameter called the **degrees of freedom** (thus, there are infinitely many t distributions!)
- in practice, ν will usually be an integer (and is often related to the sample size)
- As $\nu \to \infty$, $t(\nu) \to \mathcal{N}(0,1)$; thus, when ν becomes larger, the $t(\nu)$ and the $\mathcal{N}(0,1)$ distributions look more alike
- $E(T) = 0$ and $V(T) = \frac{\nu}{\nu-2}$ for $\nu > 2$
- When compared to the standard normal distribution, the t distribution, in general, is less peaked, and has more mass in the tails. Note that $V(T) > 1$.

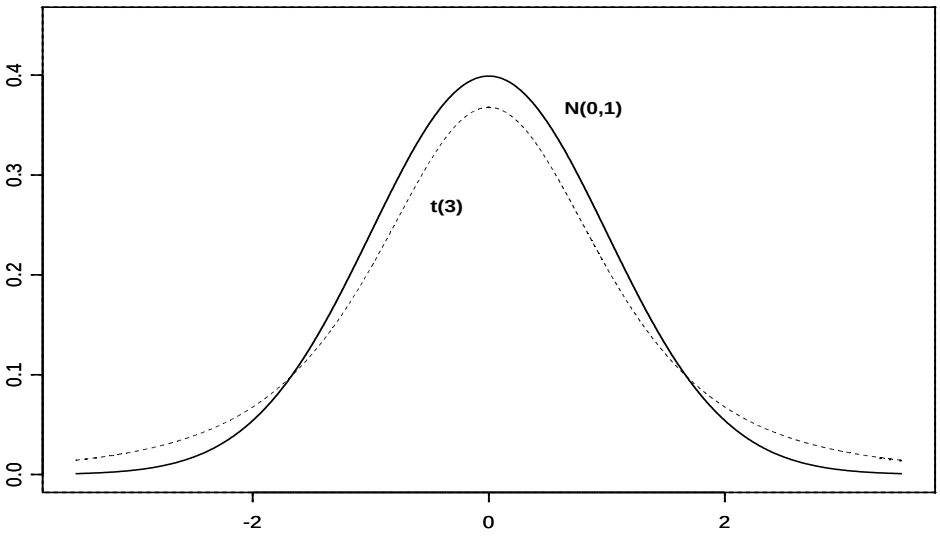

Figure 7.4: *The* $t(3)$ *distribution* (dotted) *and the* $\mathcal{N}(0,1)$ *distribution* (solid).

RELATIONSHIP WITH THE CAUCHY DISTRIBUTION: When $\nu = 1$, the t pdf reduces to

$$f_T(t) = \begin{cases} \frac{1}{\pi(1+t^2)}, & -\infty < t < \infty \\ 0, & \text{otherwise.} \end{cases}$$

which we recognize as the pdf of a Cauchy random variable. Recall that no moments are finite for the Cauchy distribution.

IMPORTANT RESULT: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\mu, \sigma^2)$ sample. From past results, we know that

$$\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \quad \text{and} \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

In addition, we know that $\overline{Y}$ and S^2 are independent, so the two quantities above (being functions of $\overline{Y}$ and S^2, respectively) are independent too. Thus,

$$t = \frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} \sim \frac{\text{“}\mathcal{N}(0,1)\text{”}}{\sqrt{\text{“}\chi^2(n-1)\text{”}/(n-1)}}$$

has a $t(n-1)$ distribution. But, simple algebra shows that

$$t = \frac{\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{\overline{Y}-\mu}{S/\sqrt{n}}.$$

This allows us to conclude that if $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\mu, \sigma^2)$ sample,

$$t = \frac{\overline{Y}-\mu}{S/\sqrt{n}} \sim t(n-1).$$

COMPARISON: You should see the effect of estimating σ, the population standard deviation, with S, the sample standard deviation. Recall that if $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\mu, \sigma^2)$ sample,

$$Z = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Thus, when we replace σ with its natural estimator S, we go from a standard normal sampling distribution to a t sampling distribution with $n-1$ degrees of freedom. Of course, if n is large, then we know that these sampling distributions will be "close" to each other.

DERIVATION OF THE t PDF: We know that $Z \sim \mathcal{N}(0,1)$, that $W \sim \chi^2(\nu)$, and that Z and W are independent. Thus, the joint pdf of (Z, W) is given by

$$f_{Z,W}(z,w) = \underbrace{\frac{1}{\sqrt{2\pi}} e^{-z^2/2}}_{\mathcal{N}(0,1) \text{ pdf}} \underbrace{\frac{1}{\Gamma(\nu/2)2^{\nu/2}} w^{\nu/2-1} e^{-w/2}}_{\chi^2(\nu) \text{ pdf}},$$

for $-\infty < z < \infty$ and $w > 0$. Consider the bivariate transformation

$$\begin{aligned} T = g_1(Z,W) &= \frac{Z}{\sqrt{W/\nu}} \\ U = g_2(Z,W) &= W. \end{aligned}$$

The support of (Z, W) is the set $R_{Z,W} = \{(z,w) : -\infty < z < \infty, w > 0\}$. The support of (T, U) is the image space of $R_{Z,W}$ under $g : \mathcal{R}^2 \to \mathcal{R}^2$, where g is defined as above; i.e., $R_{T,U} = \{(t,u) : -\infty < t < \infty, u > 0\}$. The (vector-valued) function g is one-to-one, so the inverse transformation exists and is given by

$$\begin{aligned} z = g_1^{-1}(t,u) &= t\sqrt{u/\nu} \\ w = g_2^{-1}(t,u) &= u. \end{aligned}$$

The Jacobian of the transformation is

$$J = \det \begin{vmatrix} \frac{\partial g_1^{-1}(t,u)}{\partial t} & \frac{\partial g_1^{-1}(t,u)}{\partial u} \\ \frac{\partial g_2^{-1}(t,u)}{\partial t} & \frac{\partial g_2^{-1}(t,u)}{\partial u} \end{vmatrix} = \det \begin{vmatrix} \sqrt{u/\nu} & t/2\sqrt{u\nu} \\ 0 & 1 \end{vmatrix} = \sqrt{u/\nu}.$$

We have the support of (T, U), the inverse transformation, and the Jacobian; we are now ready to write the joint pdf of (T, U). For all $-\infty < t < \infty$ and $u > 0$, this joint pdf is given by

$$\begin{aligned} f_{T,U}(t,u) &= f_{Z,W}[g_1^{-1}(t,u), g_2^{-1}(t,u)]|J| \\ &= \frac{1}{\sqrt{2\pi}} e^{-(t\sqrt{u/\nu})^2/2} \frac{1}{\Gamma(\nu/2)2^{\nu/2}} u^{\nu/2-1} e^{-u/2} \times |\sqrt{u/\nu}| \\ &= \frac{1}{\sqrt{2\pi\nu}\Gamma(\nu/2)2^{\nu/2}} u^{(\nu+1)/2-1} e^{-\frac{u}{2}\left(1+\frac{t^2}{\nu}\right)}. \end{aligned}$$

To find the marginal pdf of T, we simply integrate $f_{T,U}(t,u)$ with respect to u; that is,

$$\begin{aligned} f_T(t) &= \int_0^\infty f_{T,U}(t,u)du \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi\nu}\Gamma(\nu/2)2^{\nu/2}} u^{(\nu+1)/2-1} e^{-\frac{u}{2}\left(1+\frac{t^2}{\nu}\right)} du \\ &= \frac{1}{\sqrt{2\pi\nu}\Gamma(\nu/2)2^{\nu/2}} \int_0^\infty \underbrace{u^{(\nu+1)/2-1} e^{-\frac{u}{2}\left(1+\frac{t^2}{\nu}\right)}}_{\text{gamma}(a,b) \text{ kernel}} du, \end{aligned}$$

where $a = (\nu+1)/2$ and $b = 2\left(1+\frac{t^2}{\nu}\right)^{-1}$. The gamma kernel integral above equals

$$\Gamma(a)b^a = \Gamma[(\nu+1)/2]\left[2\left(1+\frac{t^2}{\nu}\right)^{-1}\right]^{(\nu+1)/2},$$

so that the pdf of T becomes

$$\begin{aligned} f_T(t) &= \frac{1}{\sqrt{2\pi\nu}\Gamma(\nu/2)2^{\nu/2}} \Gamma[(\nu+1)/2]\left[2\left(1+\frac{t^2}{\nu}\right)^{-1}\right]^{(\nu+1)/2} \\ &= \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\,\Gamma(\nu/2)}(1+t^2/\nu)^{-(\nu+1)/2}, \end{aligned}$$

for all $-\infty < t < \infty$. We recognize this as the pdf of a t random variable with ν degrees of freedom. $\square$

7.2.2 The F distribution

THE F DISTRIBUTION: Suppose that $W_1 \sim \chi^2(\nu_1)$ and that $W_2 \sim \chi^2(\nu_2)$. If W_1 and W_2 are independent, then the quantity

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

has an F **distribution** with ν_1 (numerator) and ν_2 (denominator) degrees of freedom. This is denoted $F(\nu_1, \nu_2)$.

REMARK: It is possible to derive the F pdf using a bivariate transformation (similar to the argument we just made in deriving the t pdf). If $W \sim F(\nu_1, \nu_2)$, the pdf of W, for all $w > 0$, is given by

$$f_W(w) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2}) \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} w^{(\nu_1-2)/2}}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2}) \left(1 + \frac{\nu_1 w}{\nu_2}\right)^{(\nu_1+\nu_2)/2}}.$$

Like the t pdf, we will never use the formula for the F pdf to find probabilities. Computing gives areas (probabilities) upon request; in addition, F tables (though limited in their use) are readily available. See Table 7 (WMS).

FACTS ABOUT THE F DISTRIBUTION:

- continuous and **skewed right**
- indexed by two **degrees of freedom** parameters ν_1 and ν_2; these are usually integers and are often related to sample sizes
- If $W \sim F(\nu_1, \nu_2)$, then $E(W) = \nu_2/(\nu_2 - 2)$, for $\nu_2 > 2$. A formula for $V(W)$ is given on pp 368 (WMS). Note that $E(W) \approx 1$ if ν_2 is large.

FUNCTIONS OF t AND F: The following results are useful. Each of the following facts can be proven using the method of transformations.

1. If $W \sim F(\nu_1, \nu_2)$, then $1/W \sim F(\nu_2, \nu_1)$.

2. If $T \sim t(\nu)$, then $T^2 \sim F(1, \nu)$.

3. If $W \sim F(\nu_1, \nu_2)$, then $(\nu_1/\nu_2)W/[1 + (\nu_1/\nu_2)W] \sim \text{beta}(\nu_1/2, \nu_2/2)$.

Example 7.5. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution. Recall that

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \quad \text{and} \quad T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

Now, write

$$\begin{aligned} T^2 = \left(\frac{\overline{Y} - \mu}{S/\sqrt{n}}\right)^2 &= \left(\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}\right)^2 \frac{\sigma^2}{S^2} \\ &= \frac{\left(\frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}\right)^2/1}{\frac{(n-1)S^2}{\sigma^2}/(n-1)} \\ &\sim \frac{\text{“}\chi^2(1)\text{”}/1}{\text{“}\chi^2(n-1)\text{”}/(n-1)} \sim F(1, n-1), \end{aligned}$$

since the numerator and denominator are independent; this follows since $\overline{Y}$ and S^2 are independent when the underlying population distribution is normal. We have informally established the second result (immediately above) for the case wherein ν is an integer greater than 1. $\square$

AN IMPORTANT APPLICATION: Suppose that we have two **independent** samples:

$$Y_{11}, Y_{12}, ..., Y_{1n_1} \sim \text{iid } \mathcal{N}(\mu_1, \sigma_1^2)$$
$$Y_{21}, Y_{22}, ..., Y_{2n_2} \sim \text{iid } \mathcal{N}(\mu_2, \sigma_2^2).$$

Define the statistics

$$\begin{aligned} \overline{Y}_{1+} = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j} &= \text{sample mean for sample 1} \\ \overline{Y}_{2+} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j} &= \text{sample mean for sample 2} \\ S_1^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n_1}(Y_{1j} - \overline{Y}_{1+})^2 &= \text{sample variance for sample 1} \\ S_2^2 = \frac{1}{n_2 - 1}\sum_{j=1}^{n_2}(Y_{2j} - \overline{Y}_{2+})^2 &= \text{sample variance for sample 2.} \end{aligned}$$

We know that

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1) \quad \text{and} \quad \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1).$$

Furthermore, as the samples are independent, $(n_1-1)S_1^2/\sigma_1^2$ and $(n_2-1)S_2^2/\sigma_2^2$ are as well. Thus, the quantity

$$F = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/(n_2-1)} \sim \frac{\text{“}\chi^2(n_1-1)\text{”}/(n_1-1)}{\text{“}\chi^2(n_2-1)\text{”}/(n_2-1)} \sim F(n_1-1, n_2-1).$$

But, algebraically,

$$F = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/(n_2-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}.$$

Thus, we conclude that

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1).$$

In addition, if the two population variances σ_1^2 and σ_2^2 are equal; i.e., $\sigma_1^2 = \sigma_2^2 = \sigma^2$, say, then

$$F = \frac{S_1^2/\sigma^2}{S_2^2/\sigma^2} = \frac{S_1^2}{S_2^2} \sim F(n_1-1, n_2-1).$$

7.3 The Central Limit Theorem

RECALL: If $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, then we know the sample mean $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$. This begs the question: *“What is the sampling distribution of $\overline{Y}$ if the observations (data) are not normally distributed?”*

CENTRAL LIMIT THEOREM: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution with mean $E(Y) = \mu$ and $V(Y) = \sigma^2 < \infty$. Let $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ denote the sample mean and define

$$U_n = \sqrt{n}\left(\frac{\overline{Y}-\mu}{\sigma}\right) = \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}}.$$

Then, as $n \to \infty$, the cumulative distribution function (cdf) of U_n converges pointwise to the cdf of a $\mathcal{N}(0,1)$ random variable.

NOTATION: We write $U_n \xrightarrow{d} \mathcal{N}(0,1)$. The symbol "$\xrightarrow{d}$" is read, "converges in distribution to." The mathematical statement that

$$U_n = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

implies that, for large n, $\overline{Y}$ has an **approximate** normal sampling distribution with mean μ and variance σ^2/n. Thus, it is common to write

$$\overline{Y} \sim \mathcal{AN}(\mu, \sigma^2/n).$$

REMARK: Note that this result is very powerful! The Central Limit Theorem (CLT) states that averages will be approximately normally distributed even if the underlying population distribution, say, $f_Y(y)$, is not! This is not an exact result; it is only an approximation.

HOW GOOD IS THE APPROXIMATION?: Since the CLT only offers an approximate sampling distribution for $\overline{Y}$, one might naturally wonder exactly how good the approximation is. In general, the goodness of the approximation jointly depends on

(a) *sample size*. The larger the sample size n, the better the approximation.

(b) *symmetry* in the underlying population distribution $f_Y(y)$. The more symmetric $f_Y(y)$ is, the better the approximation. If $f_Y(y)$ is highly skewed (e.g., exponential), we need a larger sample size for the CLT to "kick in." Recall from STAT 511 that

$$\xi = \frac{E[(Y-\mu)^3]}{\sigma^3}$$

the **skewness coefficient**, quantifies the skewness in the distribution of Y.

RESULT: Suppose U_n is a sequence of random variables; denote by $F_{U_n}(u)$ and $m_{U_n}(t)$ the corresponding sequence of cdfs and mgfs, respectively. Then, if $m_{U_n}(t) \longrightarrow m_U(t)$ pointwise for all t in an open neighborhood of 0, then there exists a cdf $F_U(u)$ where $F_{U_n}(u) \longrightarrow F_U(u)$ pointwise at all points where $F_U(u)$ is continuous. *That is, convergence of mgfs implies convergence of cdfs.* We say that the sequence of random variables U_n **converges in distribution** to U and write $U_n \xrightarrow{d} U$.

LEMMA: Recall from calculus that, for all $a \in \mathcal{R}$,

$$\lim_{n\to\infty}\left(1+\frac{a}{n}\right)^n = e^a.$$

A slight variant of this result states that if $a_n \to a$, as $n \to \infty$, then

$$\lim_{n\to\infty}\left(1+\frac{a_n}{n}\right)^n = e^a.$$

PROOF OF THE CLT: To prove the CLT, we will use the last result (and the lemma above) to show that the mgf of

$$U_n = \sqrt{n}\left(\frac{\overline{Y}-\mu}{\sigma}\right)$$

converges to $m_U(t) = e^{t^2/2}$, the mgf of a standard normal random variable. We will then be able to conclude that $U_n \xrightarrow{d} \mathcal{N}(0,1)$, thereby establishing the CLT. Let $m_Y(t)$ denote the common mgf of each $Y_1, Y_2, ..., Y_n$. We know that this mgf $m_Y(t)$ is finite for all $t \in (-h, h)$, for some $h > 0$. Define

$$X_i = \frac{Y_i - \mu}{\sigma},$$

and let $m_X(t)$ denote the common mgf of each $X_1, X_2, ..., X_n$ (the Y_i's are iid; so are the X_i's). This mgf $m_X(t)$ exists for all $t \in (-\sigma h, \sigma h)$. Simple algebra shows that

$$U_n = \sqrt{n}\left(\frac{\overline{Y}-\mu}{\sigma}\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i.$$

Thus, the mgf of U_n is given by

$$\begin{aligned} m_{U_n}(t) = E(e^{tU_n}) = E\left[e^{(t/\sqrt{n})\sum_{i=1}^n X_i}\right] &= E\left[e^{(t/\sqrt{n})X_1}e^{(t/\sqrt{n})X_2}\cdots e^{(t/\sqrt{n})X_n}\right] \\ &= E\left[e^{(t/\sqrt{n})X_1}\right]E\left[e^{(t/\sqrt{n})X_2}\right]\cdots E\left[e^{(t/\sqrt{n})X_n}\right] \\ &= \left[m_X(t/\sqrt{n})\right]^n. \end{aligned}$$

Now, consider the McLaurin series expansion (i.e., a Taylor series expansion about 0) of $m_X(t/\sqrt{n})$; we have

$$m_X(t/\sqrt{n}) = \sum_{k=0}^{\infty} m_X^{(k)}(0)\frac{(t/\sqrt{n})^k}{k!},$$

where $m_X^{(k)}(0) = (d^k/dt^k)m_X(t)|_{t=0}$. Recall that $m_X(t)$ exists for all $t \in (-\sigma h, \sigma h)$, so this power series expansion is valid for all $|t/\sqrt{n}| < \sigma h$; i.e., for all $|t| < \sqrt{n}\sigma h$. Because each X_i has mean 0 and variance 1 (verify!), it is easy to see that

$$\begin{aligned} m_X^{(0)}(0) &= 1 \\ m_X^{(1)}(0) &= 0 \\ m_X^{(2)}(0) &= 1. \end{aligned}$$

Thus, our series expansion above becomes

$$m_X(t/\sqrt{n}) = 1 + \frac{(t/\sqrt{n})^2}{2!} + R_X(t/\sqrt{n}),$$

where $R_X(t/\sqrt{n})$ is the remainder term in the expansion; i.e.,

$$R_X(t/\sqrt{n}) = \sum_{k=3}^{\infty} m_X^{(k)}(0)\frac{(t/\sqrt{n})^k}{k!}.$$

The key to finishing the proof is recognizing that

$$\lim_{n\to\infty} nR_X(t/\sqrt{n}) = 0.$$

This is not difficult to see since the $k = 3$ term in $R_X(t/\sqrt{n})$ contains an $n\sqrt{n}$ in its denominator; the $k = 4$ term contains an n^2 in its denominator, and so on, and since $m_X^{(k)}(0)/k!$ is finite for all k. The last statement also is true when $t = 0$ since $R_X(0/\sqrt{n}) = 0$. Thus, for any fixed t, we can write

$$\begin{aligned} \lim_{n\to\infty} m_{U_n}(t) &= \lim_{n\to\infty} \left[m_X(t/\sqrt{n})\right]^n \\ &= \lim_{n\to\infty} \left[1 + \frac{(t/\sqrt{n})^2}{2!} + R_X(t/\sqrt{n})\right]^n \\ &= \lim_{n\to\infty} \left\{1 + \frac{1}{n}\left[\frac{t^2}{2} + nR_X(t/\sqrt{n})\right]\right\}^n. \end{aligned}$$

Finally, let $a_n = \frac{t^2}{2} + nR_X(t/\sqrt{n})$. It is easy to see that $a_n \to t^2/2$, since $nR_X(t/\sqrt{n}) \to 0$. Thus, the last limit equals $e^{t^2/2}$. We have shown that

$$\lim_{n\to\infty} m_{U_n}(t) = e^{t^2/2},$$

the mgf of a standard normal distribution; this completes the proof. □

Example 7.6. A chemist is studying the degradation behavior of vitamin B_6 in a multivitamin. The chemist selects a random sample of $n = 36$ multivitamin tablets, and for each tablet, counts the number of days until the B_6 content falls below the FDA requirement. Let $Y_1, Y_2, ..., Y_{36}$ denote the measurements for the 36 tablets, and assume that $Y_1, Y_2, ..., Y_{36}$ is an iid sample from a Poisson distribution with mean 50.

(a) What is the approximate probability that the average number of days $\overline{Y}$ will exceed 52? That is, what is $P(\overline{Y} > 52)$?

SOLUTION. Recall that in the Poisson model, $\mu = \sigma^2 = 50$. The Central Limit Theorem says that

$$\overline{Y} \sim \mathcal{AN}\left(50, \frac{50}{36}\right).$$

Thus,

$$P(\overline{Y} > 52) \approx P\left(Z > \frac{52 - 50}{\sqrt{50/36}}\right) = P(Z > 1.70) = 0.0446.$$

(b) How many tablets does the researcher need to observe so that $P(\overline{Y} < 49.5) \approx 0.01$?

SOLUTION. We want to find the n such that

$$P(\overline{Y} < 49.5) \approx P\left(Z < \frac{49.5 - 50}{\sqrt{50/n}}\right) = P\left(Z < \frac{49.5 - 50}{\sqrt{50/n}}\right) \approx 0.01.$$

Thus, we need to solve

$$\frac{49.5 - 50}{\sqrt{50/n}} = -2.33$$

for n; note that $z = -2.33$ is the 1st percentile of the standard normal distribution. It follows that $n \approx 1086$. □

7.4 The normal approximation to the binomial

IMPORTANCE: An important application of the Central Limit Theorem deals with approximating the sampling distributions of functions of **count data**; such data are pervasive in statistical problems.

RECALL: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a Bernoulli(p) distribution; that is, $Y_i = 1$, if the ith trial is a "success," and $Y_i = 0$, otherwise. Recall that the

probability mass function (pmf) for the Bernoulli random variable is

$$p_Y(y) = \begin{cases} p^y(1-p)^{1-y}, & y = 0, 1 \\ 0, & \text{otherwise.} \end{cases}$$

That is, the sample $Y_1, Y_2, ..., Y_n$ is a random string of zeros and ones, where $P(Y_i = 1) = p$, for each i. Recall that in the Bernoulli model,

$$\mu = E(Y) = p \quad \text{and} \quad \sigma^2 = V(Y) = p(1-p).$$

From Example 6.9 (notes), we know that

$$X = \sum_{i=1}^{n} Y_i,$$

the number of "successes," has a binomial distribution with parameters n and p; that is, $X \sim b(n, p)$. Define the **sample proportion** $\widehat{p}$ as

$$\widehat{p} = \frac{X}{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Note that $\widehat{p}$ is an average of iid values of 0 and 1; thus, the CLT must apply! That is, for large n,

$$\widehat{p} \sim \mathcal{AN}\left[p, \frac{p(1-p)}{n}\right].$$

HOW GOOD IS THE APPROXIMATION?: Since we are sampling from a "binary" population (almost as discrete as one can get!), one might naturally wonder how well the normal distribution approximates the true sampling distribution of $\widehat{p}$. The approximation is **best** when

(a) n is large (the approximation improves as n increases), and

(b) p is close to 1/2. Recall that, for $Y \sim b(1, p)$,

$$\xi = \frac{E[(Y-\mu)^3]}{\sigma^3} = \frac{1-2p}{\sqrt{p(1-p)}}.$$

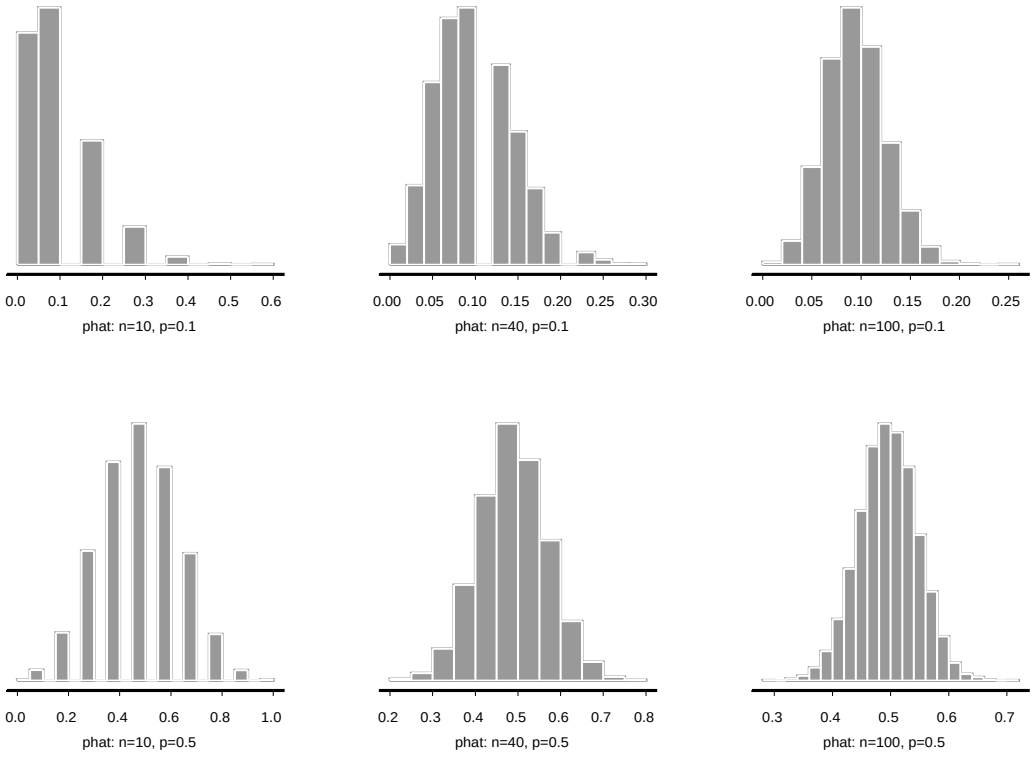

Figure 7.5: *The approximate sampling distributions for $\widehat{p}$ for different n and p.*

RULES OF THUMB: One can feel comfortable using the normal approximation as long as np and $n(1-p)$ are larger than 10. Other guidelines have been proposed in the literature. This is just a guideline.

Example 7.7. Figure 7.5 presents **Monte Carlo distributions** for 10,000 simulated values of $\widehat{p}$ for each of six select cases:

Case 1: $n = 10, p = 0.1$ Case 2: $n = 40, p = 0.1$ Case 3: $n = 100, p = 0.1$

Case 4: $n = 10, p = 0.5$ Case 5: $n = 40, p = 0.5$ Case 6: $n = 100, p = 0.5$

One can clearly see that the normal approximation is not good when $p = 0.1$, except when n is very large. On the other hand, when $p = 0.5$, the normal approximation is already pretty good when $n = 40$. $\square$

Example 7.8. Dimenhydrinate, also known by the trade names Dramamine and Gravol, is an over-the-counter drug used to prevent motion sickness. The drug's manufacturer claims that dimenhydrinate helps reduce motion sickness in 40 percent of the population. A random sample of $n = 200$ individuals is recruited in a study to test the manufacturer's claim. Define $Y_i = 1$, if the the ith subject responds to the drug, and $Y_i = 0$, otherwise, and assume that $Y_1, Y_2, ..., Y_{200}$ is an iid Bernoulli($p = 0.4$) sample; note that $p = 0.4$ corresponds to the company's claim. Let X count the number of subjects that respond to the drug; we then know that $X \sim b(200, 0.4)$. What is the probability that 60 or less respond to the drug? That is, what is $P(X \leq 60)$?

SOLUTION. We compute this probability in two ways; first, we compute $P(X \leq 60)$ **exactly** using the $b(200, 0.4)$ model; this is given by

$$P(X \leq 60) = \sum_{x=0}^{60} \binom{200}{x} (0.4)^x (1-0.4)^{200-x} = 0.0021.$$

I used the R command `pbinom(60,200,0.4)` to compute this probability. Alternatively, we can use the **CLT approximation** to the binomial to find this probability; note that the sample proportion

$$\widehat{p} = \frac{X}{n} \sim \mathcal{AN}\left[0.4, \frac{0.4(1-0.4)}{200}\right].$$

Thus,

$$\begin{aligned} P(X \leq 60) &= P(\widehat{p} \leq 0.3) \\ &\approx P\left[Z \leq \frac{0.3 - 0.4}{\sqrt{\frac{0.4(1-0.4)}{200}}}\right] \\ &= P(Z \leq -2.89) = 0.0019. \end{aligned}$$

As we can see, the CLT approximation is very close to the true (exact) probability. Here, $np = 200 \times 0.4 = 80$ and $n(1-p) = 200 \times 0.6 = 120$, both of which are large. Thus, we can feel comfortable with the normal approximation. □

QUESTION FOR THOUGHT: We have observed here that $P(X \leq 60)$ is very, very small under the assumption that $p = 0.4$, the probability of response for each subject, claimed by the manufacturer. If we, in fact, did observe this event $\{X \leq 60\}$, what might this suggest about the manufacturer's claim that $p = 0.4$?

8 Estimation

Complementary reading: Chapter 8 (WMS).

8.1 Introduction

REMARK: Up until now (i.e., in STAT 511 and the material so far in STAT 512), we have dealt with **probability models**. These models, as we know, can be generally divided up into two types: discrete and continuous. These models are used to describe populations of individuals.

- In a clinical trial with n patients, let p denote the probability of response to a new drug. A $b(1, p)$ model is assumed for each subject's response (e.g., respond/not).
- In an engineering application, the lifetime of an electrical circuit, Y, is under investigation. An exponential(β) model is assumed.
- In a public-health study, Y, the number of sexual partners in the past year, is recorded for a group of high-risk HIV patients. A Poisson(λ) model is assumed.
- In an ecological study, the amount of dead-weight (measured in g/plot), Y, is recorded. A $\mathcal{N}(\mu, \sigma^2)$ model is assumed.

Each of these situations employs a probabilistic model that is indexed by population parameters. *In real life, these parameters are unknown.* An important statistical problem, thus, involves **estimating** these parameters with a random sample $Y_1, Y_2, ..., Y_n$ (i.e., an iid sample) from the population. We can state this problem generally as follows.

GENERAL PROBLEM: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a population which is described by the model $f_Y(y; \theta)$. Here, $f_Y(y; \theta)$ is a pmf or pdf that describes the population of interest, and θ is a **parameter** that indexes the model. *The statistical problem of interest is to estimate θ with the observed data $Y_1, Y_2, ..., Y_n$.*

TERMINOLOGY: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y;\theta)$. A **point estimator** $\widehat{\theta}$ is a function of $Y_1, Y_2, ..., Y_n$ that estimates θ. Since $\widehat{\theta}$ is (in general) a function of $Y_1, Y_2, ..., Y_n$, it is a **statistic**. In practice, θ could be a scalar or vector.

Example 8.1. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson distribution with mean θ. We know that the probability mass function (pmf) for Y is given by

$$f_Y(y;\theta) = \begin{cases} \frac{\theta^y e^{-\theta}}{y!}, & y = 0, 1, 2, ... \\ 0, & \text{otherwise.} \end{cases}$$

Here, the parameter is $\theta = E(Y)$. What estimator should we use to estimate θ?

Example 8.2. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{U}(0, \theta)$ distribution. We know that the probability density function (pdf) for Y is given by

$$f_Y(y;\theta) = \begin{cases} \frac{1}{\theta}, & 0 < y < \theta \\ 0, & \text{otherwise.} \end{cases}$$

Here, the parameter is θ, the upper limit of the support of Y. What estimator should we use to estimate θ?

Example 8.3. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution. We know that the probability density function (pdf) for Y is given by

$$f_Y(y;\boldsymbol{\theta}) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}, & -\infty < y < \infty \\ 0, & \text{otherwise.} \end{cases}$$

Here, the parameter is $\boldsymbol{\theta} = (\mu, \sigma^2)$, a vector of two parameters (the mean and the variance). What estimator should we use to estimate $\boldsymbol{\theta}$? Or, equivalently, we might ask how to estimate μ and σ^2 separately.

"GOOD" ESTIMATORS: In general, a "good" estimator $\widehat{\theta}$ has the following properties:

(1) $\widehat{\theta}$ is **unbiased** for θ, and

(2) $\widehat{\theta}$ has **small variance**.

8.2 Bias and mean-squared error

TERMINOLOGY: An estimator $\widehat{\theta}$ is said to be **unbiased** for θ if

$$E(\widehat{\theta}) = \theta,$$

for all possible values of θ. If $\widehat{\theta}$ is not an unbiased estimator; i.e., if $E(\widehat{\theta}) \neq \theta$, then we say that $\widehat{\theta}$ is biased. In general, the **bias** of an estimator is

$$B(\widehat{\theta}) \equiv E(\widehat{\theta}) - \theta.$$

If $B(\widehat{\theta}) > 0$, then $\widehat{\theta}$ overestimates θ. If $B(\widehat{\theta}) < 0$, then $\widehat{\theta}$ underestimates θ. If $\widehat{\theta}$ is unbiased, then, of course, $B(\widehat{\theta}) = 0$.

Example 8.1 (revisited). Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a Poisson distribution with mean θ. Recall that, in general, the **sample mean** $\overline{Y}$ is an unbiased estimator for a population mean μ. For the Poisson model, the (population) mean is $\mu = E(Y) = \theta$. Thus, we know that

$$\widehat{\theta} = \overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is an unbiased estimator of θ. Recall also that the variance of the sample mean, $V(\overline{Y})$, is, in general, the population variance σ^2 divided by n. For the Poisson model, the (population) variance is $\sigma^2 = \theta$; thus, $V(\widehat{\theta}) = V(\overline{Y}) = \theta/n$. $\square$

Example 8.2 (revisited). Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{U}(0, \theta)$ distribution, and consider the point estimator $Y_{(n)}$. Intuitively, this seems like a reasonable estimator to use; the largest order statistic should be fairly close to θ, the upper endpoint of the support. To compute $E(Y_{(n)})$, we have to know how $Y_{(n)}$ is distributed, so we find its pdf. For $0 < y < \theta$, the pdf of $Y_{(n)}$ is

$$\begin{aligned} f_{Y_{(n)}}(y) &= n f_Y(y)[F_Y(y)]^{n-1} \\ &= n\left(\frac{1}{\theta}\right)\left(\frac{y}{\theta}\right)^{n-1} = n\theta^{-n}y^{n-1}, \end{aligned}$$

so that

$$E(Y_{(n)}) = \int_0^\theta y \times \underbrace{n\theta^{-n}y^{n-1}}_{=f_{Y_{(n)}}(y)} dy = n\theta^{-n}\left(\frac{1}{n+1}\right)y^{n+1}\Big|_0^\theta = \left(\frac{n}{n+1}\right)\theta.$$

We see that $Y_{(n)}$ is a **biased estimator** of θ (it underestimates θ on average). But,

$$\widehat{\theta} = \left(\frac{n+1}{n}\right)Y_{(n)}$$

is an unbiased estimator because

$$E(\widehat{\theta}) = E\left[\left(\frac{n+1}{n}\right)Y_{(n)}\right] = \left(\frac{n+1}{n}\right)E(Y_{(n)}) = \left(\frac{n+1}{n}\right)\left(\frac{n}{n+1}\right)\theta = \theta. \ \square$$

EXERCISE: Compute $V(\widehat{\theta})$.

Example 8.3 (revisited). Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution. To estimate μ, we know that a good estimator is $\overline{Y}$. The sample mean $\overline{Y}$ is unbiased; i.e., $E(\overline{Y}) = \mu$, and, furthermore, $V(\overline{Y}) = \sigma^2/n$ decreases as the sample size n increases. To estimate σ^2, we can use the **sample variance**; i.e.,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2.$$

Assuming the normal model, the sample variance is unbiased. To see this, recall that

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

so that

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

since the mean of a χ^2 random variable equals its degrees of freedom. Thus,

$$n-1 = E\left[\frac{(n-1)S^2}{\sigma^2}\right] = \left(\frac{n-1}{\sigma^2}\right)E(S^2) \implies E(S^2) = \sigma^2,$$

showing that S^2 is an unbiased estimator of the population variance σ^2. To compute the variance of S^2 as an estimator, recall that

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

since the variance of a χ^2 random variable equals twice its degrees of freedom. Therefore,

$$2(n-1) = V\left[\frac{(n-1)S^2}{\sigma^2}\right] = \left[\frac{(n-1)^2}{\sigma^4}\right] V(S^2)$$
$$\implies V(S^2) = \frac{2\sigma^4}{n-1}. \ \square$$

ESTIMATING FUNCTIONS OF PARAMETERS: In some problems, the goal is to estimate a function of θ, say, $\tau(\theta)$. The following example illustrates how we can find an unbiased estimator of a function of θ.

Example 8.4. Suppose that $Y_1, Y_2, ..., Y_n$ are iid exponential observations with mean θ. Derive an unbiased estimator for $\tau(\theta) = 1/\theta$.

SOLUTION. Since $E(\overline{Y}) = \theta$, one's intuition might suggest to try $1/\overline{Y}$ as an estimator for $1/\theta$. First, note that

$$E\left(\frac{1}{\overline{Y}}\right) = E\left(\frac{n}{\sum_{i=1}^n Y_i}\right) = nE\left(\frac{1}{T}\right),$$

where $T = \sum_{i=1}^n Y_i$. Recall that $Y_1, Y_2, ..., Y_n$ iid exponential$(\theta) \implies T \sim$ gamma(n, θ), so therefore

$$\begin{aligned} E\left(\frac{1}{\overline{Y}}\right) = nE\left(\frac{1}{T}\right) &= n\int_{t=0}^{\infty} \frac{1}{t} \underbrace{\frac{1}{\Gamma(n)\theta^n} t^{n-1}e^{-t/\theta}}_{\text{gamma}(n,\theta) \text{ pdf}} dt \\ &= \frac{n}{\Gamma(n)\theta^n} \underbrace{\int_{t=0}^{\infty} t^{(n-1)-1}e^{-t/\theta} dt}_{=\ \Gamma(n-1)\theta^{n-1}} \\ &= \frac{n\Gamma(n-1)\theta^{n-1}}{\Gamma(n)\theta^n} = \frac{n\Gamma(n-1)}{(n-1)\Gamma(n-1)\theta} = \left(\frac{n}{n-1}\right)\frac{1}{\theta}. \end{aligned}$$

This shows that $1/\overline{Y}$ is a biased estimator of $\tau(\theta) = 1/\theta$. However,

$$E\left(\frac{n-1}{n\overline{Y}}\right) = \left(\frac{n-1}{n}\right) E\left(\frac{1}{\overline{Y}}\right) = \left(\frac{n-1}{n}\right)\left(\frac{n}{n-1}\right)\frac{1}{\theta} = \frac{1}{\theta}.$$

This shows that

$$\widehat{\tau(\theta)} = \frac{n-1}{n\overline{Y}}$$

is an unbiased estimator of $\tau(\theta) = 1/\theta$. $\square$

TERMINOLOGY: The **mean-squared error** (MSE) of a point estimator $\widehat{\theta}$ is given by

$$\text{MSE}(\widehat{\theta}) \equiv E[(\widehat{\theta} - \theta)^2] = V(\widehat{\theta}) + [B(\widehat{\theta})]^2.$$

We see that the MSE combines the

- the precision (variance) of $\widehat{\theta}$ and
- accuracy (bias) of $\widehat{\theta}$.

Of course, if $\widehat{\theta}$ is unbiased for θ, then $\text{MSE}(\widehat{\theta}) = V(\widehat{\theta})$, since $B(\widehat{\theta}) = 0$.

INTUITIVELY: Suppose that we have two unbiased estimators, say, $\widehat{\theta}_1$ and $\widehat{\theta}_2$. Then we would prefer to use the one with the **smaller variance**. That is, if $V(\widehat{\theta}_1) < V(\widehat{\theta}_2)$, then we would prefer $\widehat{\theta}_1$ as an estimator. *Note that it only makes sense to choose an estimator on the basis of its variance when both estimators are unbiased.*

CURIOSITY: Suppose that we have two estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ and that both of them are not unbiased (e.g., one could be unbiased and other isn't, or possibly both are biased). On what grounds should we now choose between $\widehat{\theta}_1$ and $\widehat{\theta}_2$? In this situation, a reasonable approach is to choose the estimator with the **smaller mean-squared error**. That is, if $\text{MSE}(\widehat{\theta}_1) < \text{MSE}(\widehat{\theta}_2)$, then we would prefer $\widehat{\theta}_1$ as an estimator.

Example 8.5. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid Bernoulli(p) sample, where $0 < p < 1$. Define $X = Y_1 + Y_2 + \cdots + Y_n$ and the two estimators

$$\widehat{p}_1 = \frac{X}{n} \quad \text{and} \quad \widehat{p}_2 = \frac{X+2}{n+4}.$$

Which estimator should we use to estimate p?

SOLUTION. First, we should note that $X \sim b(n, p)$, since X is the sum of iid Bernoulli(p) observations. Thus,

$$E(\widehat{p}_1) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}(np) = p$$

(i.e., $\widehat{p}_1$ is unbiased) and

$$E(\widehat{p}_2) = E\left(\frac{X+2}{n+4}\right) = \frac{1}{n+4}E(X+2) = \frac{1}{n+4}[E(X)+2] = \frac{np+2}{n+4}.$$

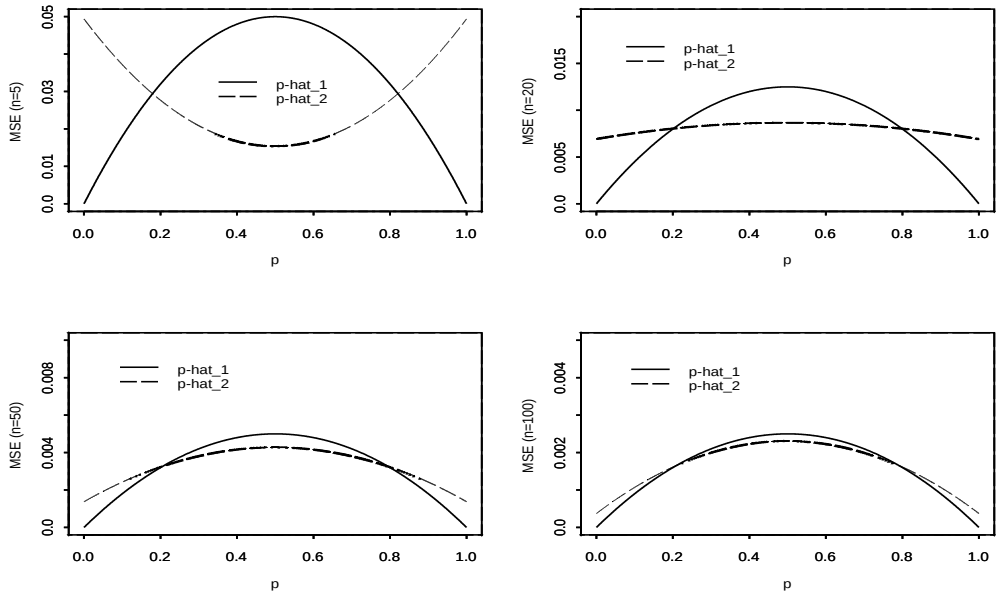

Figure 8.6: *Plots of* $\text{MSE}(\widehat{p}_1)$ *and* $\text{MSE}(\widehat{p}_2)$ *for different sample sizes in Example* 8.5.

Thus, to compare $\widehat{p}_1$ and $\widehat{p}_2$ as estimators, we should use the estimators' mean-squared errors (since $\widehat{p}_2$ is biased). The variances of $\widehat{p}_1$ and $\widehat{p}_2$ are, respectively,

$$V(\widehat{p}_1) = V\left(\frac{X}{n}\right) = \frac{1}{n^2}V(X) = \frac{1}{n^2}[np(1-p)] = \frac{p(1-p)}{n}$$

and

$$V(\widehat{p}_2) = V\left(\frac{X+2}{n+4}\right) = \frac{1}{(n+4)^2}V(X+2) = \frac{1}{(n+4)^2}V(X) = \frac{np(1-p)}{(n+4)^2}.$$

The mean-squared error of $\widehat{p}_1$ is

$$\begin{aligned}\text{MSE}(\widehat{p}_1) &= V(\widehat{p}_1) + [B(\widehat{p}_1)]^2 \\ &= \frac{p(1-p)}{n} + (p-p)^2 = \frac{p(1-p)}{n},\end{aligned}$$

which is equal to $V(\widehat{p}_1)$ since $\widehat{p}_1$ is unbiased. The mean-squared error of $\widehat{p}_2$ is

$$\begin{aligned}\text{MSE}(\widehat{p}_2) &= V(\widehat{p}_2) + [B(\widehat{p}_2)]^2 \\ &= \frac{np(1-p)}{(n+4)^2} + \left(\frac{np+2}{n+4} - p\right)^2.\end{aligned}$$

ANALYSIS: Figure 8.6 displays values of MSE($\widehat{p}_1$) and MSE($\widehat{p}_2$) graphically for $n = 5$, 20, 50, and 100. We can see that neither estimator is uniformly superior; i.e., neither estimator delivers a smaller MSE for all $0 < p < 1$. However, for smaller sample sizes, $\widehat{p}_2$ often beats $\widehat{p}_1$ (in terms of MSE) when p is in the vicinity of 0.5; otherwise, $\widehat{p}_1$ often provides smaller MSE.

8.3 The standard error of an estimator

TERMINOLOGY: The **standard error** of a point estimator $\widehat{\theta}$ is simply the standard deviation of the estimator. We denote the standard error of $\widehat{\theta}$ by

$$\sigma_{\widehat{\theta}} = \sqrt{V(\widehat{\theta})}.$$

Table 8.1 (WMS, pp 397) summarizes some common point estimators and their standard errors. We now review these.

8.3.1 One population mean

SITUATION: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample with mean μ and variance σ^2 and that interest lies in estimating the **population mean** μ.

POINT ESTIMATOR: To estimate the (population) mean μ, a natural point estimator to use is the **sample mean**; i.e.,

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

FACTS: We have shown that, in general,

$$\begin{aligned} E(\overline{Y}) &= \mu \\ V(\overline{Y}) &= \frac{\sigma^2}{n}. \end{aligned}$$

STANDARD ERROR: The **standard error** of $\overline{Y}$ is equal to

$$\sigma_{\overline{Y}} = \sqrt{V(\overline{Y})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

8.3.2 One population proportion

SITUATION: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid Bernoulli(p) sample, where $0 < p < 1$, and that interest lies in estimating the **population proportion** p. Recall that $X = \sum_{i=1}^{n} Y_i \sim b(n, p)$, since X is the sum of iid Bernoulli(p) observations.

POINT ESTIMATOR: To estimate the (population) proportion p, a natural point estimator to use is the **sample proportion**; i.e.,

$$\widehat{p} = \frac{X}{n} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

FACTS: It is easy to show (verify!) that

$$\begin{aligned} E(\widehat{p}) &= p \\ V(\widehat{p}) &= \frac{p(1-p)}{n}. \end{aligned}$$

STANDARD ERROR: The **standard error** of $\widehat{p}$ is equal to

$$\sigma_{\widehat{p}} = \sqrt{V(\widehat{p})} = \sqrt{\frac{p(1-p)}{n}}.$$

8.3.3 Difference of two population means

SITUATION: Suppose that we have two **independent** samples; i.e.,

Sample 1: $Y_{11}, Y_{12}, ..., Y_{1n_1}$ are iid with mean μ_1 and variance σ_1^2

Sample 2: $Y_{21}, Y_{22}, ..., Y_{2n_2}$ are iid with mean μ_2 and variance σ_2^2

and that interest lies in estimating the **population mean difference** $\theta \equiv \mu_1 - \mu_2$. As noted, we assume that the samples themselves are independent (i.e., observations from one sample are independent from observations in the other sample).

NEW NOTATION: Because we have two samples, we need to adjust our notation accordingly. Here, we use the conventional notation Y_{ij} to denote the jth observation from

sample i, for $i = 1, 2$ and $j = 1, 2, ..., n_i$. The symbol n_i denotes the sample size from sample i. It is not necessary that the sample sizes n_1 and n_2 are equal.

POINT ESTIMATOR: To estimate the population mean difference $\theta = \mu_1 - \mu_2$, a natural point estimator to use is the **difference of the sample means**; i.e.,

$$\widehat{\theta} \equiv \overline{Y}_{1+} - \overline{Y}_{2+},$$

where

$$\overline{Y}_{1+} = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j} \quad \text{and} \quad \overline{Y}_{2+} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j}.$$

This notation is also standard; the "+" symbol is understood to mean that the subscript it replaces has been "summed over."

FACTS: It is easy to show (verify!) that

$$\begin{aligned} E(\overline{Y}_{1+} - \overline{Y}_{2+}) &= \mu_1 - \mu_2 \\ V(\overline{Y}_{1+} - \overline{Y}_{2+}) &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \end{aligned}$$

STANDARD ERROR: The **standard error** of $\widehat{\theta} = \overline{Y}_{1+} - \overline{Y}_{2+}$ is equal to

$$\sigma_{\overline{Y}_{1+} - \overline{Y}_{2+}} = \sqrt{V(\overline{Y}_{1+} - \overline{Y}_{2+})} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

8.3.4 Difference of two population proportions

SITUATION: Suppose that we have two **independent** samples; i.e.,

$$\begin{aligned} &\text{Sample 1:} \quad Y_{11}, Y_{12}, ..., Y_{1n_1} \text{ are iid Bernoulli}(p_1) \\ &\text{Sample 2:} \quad Y_{21}, Y_{22}, ..., Y_{2n_2} \text{ are iid Bernoulli}(p_2) \end{aligned}$$

and that interest lies in estimating the **population proportion difference** $\theta \equiv p_1 - p_2$. Again, it is not necessary that the sample sizes n_1 and n_2 are equal. As noted, we assume that the samples themselves are independent (i.e., observations from one sample are independent from observations in the other sample). Define

$$X_1 = \sum_{j=1}^{n_1} Y_{1j} \quad \text{and} \quad X_2 = \sum_{j=1}^{n_2} Y_{2j}.$$

We know that $X_1 \sim b(n_1, p_1)$, $X_2 \sim b(n_2, p_2)$, and that X_1 and X_2 are independent (since the samples are). The **sample proportions** are

$$\widehat{p}_1 = \frac{X_1}{n_1} = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j} \quad \text{and} \quad \widehat{p}_2 = \frac{X_2}{n_2} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j}.$$

POINT ESTIMATOR: To estimate the population proportion difference $\theta = p_1 - p_2$, a natural point estimator to use is the **difference of the sample proportions**; i.e.,

$$\widehat{\theta} \equiv \widehat{p}_1 - \widehat{p}_2.$$

FACTS: It is easy to show (verify!) that

$$\begin{aligned} E(\widehat{p}_1 - \widehat{p}_2) &= p_1 - p_2 \\ V(\widehat{p}_1 - \widehat{p}_2) &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}. \end{aligned}$$

STANDARD ERROR: The **standard error** of $\widehat{\theta} = \widehat{p}_1 - \widehat{p}_2$ is equal to

$$\sigma_{\widehat{p}_1 - \widehat{p}_2} = \sqrt{V(\widehat{p}_1 - \widehat{p}_2)} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

8.4 Estimating the population variance

RECALL: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample with mean μ and variance σ^2. The **sample variance** is defined as

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

In Example 8.3 (notes), we showed that if $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\mu, \sigma^2)$ sample, then the sample variance S^2 is an **unbiased estimator** of the population variance σ^2.

NEW RESULT: That S^2 is an unbiased estimator of σ^2 holds in general; that is, as long as $Y_1, Y_2, ..., Y_n$ is an iid sample with mean μ and variance σ^2,

$$E(S^2) = \sigma^2;$$

that is, S^2 is an unbiased estimator of σ^2, *regardless of the population distribution*, as long as $\sigma^2 < \infty$. The proof of this result is given on pp 398-399 in WMS.

8.5 Error bounds and the Empirical Rule

TERMINOLOGY: We are often interested in understanding how close our estimator $\widehat{\theta}$ is to a population parameter θ. Of course, in real life, θ is unknown, so we can never know for sure. However, we can make probabilistic statements regarding the closeness of $\widehat{\theta}$ and θ. We call $\epsilon = |\widehat{\theta} - \theta|$ the **error in estimation**.

THE EMPIRICAL RULE: Suppose the estimator $\widehat{\theta}$ has an approximate normal **sampling distribution** with mean θ and variance $\sigma^2_{\widehat{\theta}}$. It follows then that

- about 68 percent of the values of $\widehat{\theta}$ will fall between $\theta \pm \sigma_{\widehat{\theta}}$
- about 95 percent of the values of $\widehat{\theta}$ will fall between $\theta \pm 2\sigma_{\widehat{\theta}}$
- about 99.7 percent (or nearly all) of the values of $\widehat{\theta}$ will fall between $\theta \pm 3\sigma_{\widehat{\theta}}$.

These facts follow directly from the normal distribution. For example, with $Z \sim \mathcal{N}(0,1)$, we compute

$$\begin{aligned} P\left(\theta - \sigma_{\widehat{\theta}} < \widehat{\theta} < \theta + \sigma_{\widehat{\theta}}\right) &= P\left(\frac{\theta - \sigma_{\widehat{\theta}} - \theta}{\sigma_{\widehat{\theta}}} < \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} < \frac{\theta + \sigma_{\widehat{\theta}} - \theta}{\sigma_{\widehat{\theta}}}\right) \\ &\approx P(-1 < Z < 1) \\ &= F_Z(1) - F_Z(-1) \\ &= 0.8413 - 0.1587 = 0.6826, \end{aligned}$$

where $F_Z(\cdot)$ denotes the cdf of the standard normal distribution.

REMARK: Most estimators $\widehat{\theta}$, with probability "in the vicinity of" 0.95, will fall within two standard deviations (standard errors) of its mean. Thus, if $\widehat{\theta}$ is an unbiased estimator of θ, or is approximately unbiased, then $b = 2\sigma_{\widehat{\theta}}$ serves as a good approximate **upper bound** for the error in estimation; that is, $\epsilon = |\widehat{\theta} - \theta| \leq 2\sigma_{\widehat{\theta}}$ with "high" probability.

Example 8.6. In an agricultural experiment, we observe an iid sample of n yields, say, $Y_1, Y_2, ..., Y_n$, measured in kg/area per plot. We can estimate the (population) mean yield

μ with $\overline{Y}$, the sample mean; from the Central Limit Theorem, we know that

$$\overline{Y} \sim \mathcal{AN}\left(\mu, \frac{\sigma^2}{n}\right),$$

for large n. Thus, $b = 2\sigma/\sqrt{n}$ serves as an approximate 95 percent bound on the error in estimation $\epsilon = |\overline{Y} - \mu|$. $\square$

Example 8.7. In a public-health study involving intravenous drug users, subjects are tested for HIV. Denote the HIV statuses by $Y_1, Y_2, ..., Y_n$ and assume these statuses are iid Bernoulli(p) random variables (e.g., 1, if positive; 0, otherwise). The sample proportion of HIV infecteds, then, is given by

$$\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Recall that for n large,

$$\widehat{p} \sim \mathcal{AN}\left[p, \frac{p(1-p)}{n}\right].$$

Thus, $b = 2\sqrt{p(1-p)/n}$ serves as an approximate 95 percent bound on the error in estimation $\epsilon = |\widehat{p} - p|$. $\square$

REMARK: To use the Empirical Rule, we need the sampling distribution of $\widehat{\theta}$ to be normally distributed, or, at least, approximately normally distributed. Otherwise, the Empirical Rule may provide incorrect results. If we have an estimator $\widehat{\theta}$ that does not follow a normal distribution, we could use Chebyshev's Inequality to put a bound on the error in estimation ϵ. Recall that **Chebyshev's Inequality** says

$$P(|\widehat{\theta} - \theta| < k\sigma_{\widehat{\theta}}) \geq 1 - \frac{1}{k^2},$$

for any value $k > 0$. For example, if $k = 2$, then $b = 2\sigma_{\widehat{\theta}}$ is an **at least** 75 percent bound on the error in estimation $\epsilon = |\widehat{\theta} - \theta|$.

8.6 Confidence intervals and pivotal quantities

REMARK: A **point estimator** $\widehat{\theta}$ provides a "one-shot guess" of the value of an unknown parameter θ. On the other hand, an interval estimator, or **confidence interval**, provides a range of values that is likely to contain θ.

Table 8.1: *Manufacturing part length data. These observations are modeled as $n = 10$ realizations from a $\mathcal{N}(\mu, \sigma^2)$ distribution.*

12.2	12.0	12.2	11.9	12.4	12.6	12.1	12.2	12.9	12.4

Example 8.8. The length of a critical part, measured in mm, in a manufacturing process varies according to a $\mathcal{N}(\mu, \sigma^2)$ distribution (this is a model assumption). Engineers plan to observe an iid sample of $n = 10$ parts and record $Y_1, Y_2, ..., Y_{10}$. The observed data from the experiment are given in Table 8.1.

POINT ESTIMATES: The sample mean computed with the observed data is $\overline{y} = 12.3$ and sample variance is $s^2 = 0.09$ (verify!). The sample mean $\overline{y} = 12.3$ is a **point estimate** for the population mean μ. Similarly, the sample variance $s^2 = 0.09$ is a **point estimate** for the population variance σ^2. However, neither of these estimates has a measure of variability associated with it; that is, both estimates are just single "one-number" values.

TERMINOLOGY: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a population distribution (probability model) described by $f_Y(y; \theta)$. *Informally, a confidence interval is an interval of plausible values for a parameter θ.* More specifically, if θ is our parameter of interest, then we call $(\widehat{\theta}_L, \widehat{\theta}_U)$ a $100(1-\alpha)$ **percent confidence interval** for θ if

$$P(\widehat{\theta}_L \leq \theta \leq \widehat{\theta}_U) = 1 - \alpha,$$

where $0 < \alpha < 1$. We call $1 - \alpha$ the **confidence level**. In practice, we would like the confidence level $1 - \alpha$ to be large (e.g., 0.90, 0.95, 0.99, etc.).

IMPORTANT: Before we observe $Y_1, Y_2, ..., Y_n$, the interval $(\widehat{\theta}_L, \widehat{\theta}_U)$ is a **random** interval. This is true because $\widehat{\theta}_L$ and $\widehat{\theta}_U$ are random quantities as they will be functions of $Y_1, Y_2, ..., Y_n$. On the other hand, θ is a **fixed** parameter; it's value does not change. After we see the data $Y_1 = y_1, Y_2 = y_2, ..., Y_n = y_n$, like those data in Table 8.1, the numerical interval $(\widehat{\theta}_L, \widehat{\theta}_U)$ based on the realizations $y_1, y_2, ..., y_n$ is no longer random.

Example 8.9. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\mu, \sigma_0^2)$ sample, where the mean μ is unknown and variance σ_0^2 is **known**. In this example, we focus on the population mean μ. From past results, we know that $\overline{Y} \sim \mathcal{N}(\mu, \sigma_0^2/n)$. Thus,

$$Z = \frac{\overline{Y} - \mu}{\sigma_0/\sqrt{n}} \sim \mathcal{N}(0,1);$$

i.e., Z has a standard normal distribution. We know there exists a value $z_{\alpha/2}$ such that

$$\begin{aligned}
1-\alpha &= P(-z_{\alpha/2} < Z < z_{\alpha/2}) \\
&= P\left(-z_{\alpha/2} < \frac{\overline{Y}-\mu}{\sigma_0/\sqrt{n}} < z_{\alpha/2}\right) \\
&= P\left(-z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} < \overline{Y} - \mu < z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) \\
&= P\left(z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} > \mu - \overline{Y} > -z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) \\
&= P\left(\overline{Y} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}} > \mu > \overline{Y} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}\right) \\
&= P\Bigg(\underbrace{\overline{Y} - z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}}_{\widehat{\theta}_L} < \mu < \underbrace{\overline{Y} + z_{\alpha/2}\frac{\sigma_0}{\sqrt{n}}}_{\widehat{\theta}_U}\Bigg).
\end{aligned}$$

These calculations show that

$$\overline{Y} \pm z_{\alpha/2}\left(\frac{\sigma_0}{\sqrt{n}}\right)$$

is a $100(1-\alpha)$ percent confidence interval for the population mean μ. The probability that the **random** interval $(\widehat{\theta}_L, \widehat{\theta}_U)$ includes the mean μ is $1-\alpha$. $\square$

Example 8.8 (revisited). In Example 8.8, suppose that the population variance for the distribution of part lengths is $\sigma_0^2 = 0.1$, so that $\sigma_0 \approx 0.32$ (we did not make this assumption before) and that we would like to construct a 95 percent confidence interval for μ, the mean length. From the data in Table 8.1, we have $n = 10$, $\overline{y} = 12.3$, $\alpha = 0.05$, and $z_{0.025} = 1.96$ (z-table). A 95 percent confidence interval for μ is

$$12.3 \pm 1.96\left(\frac{0.32}{\sqrt{10}}\right) \Longrightarrow (12.1, 12.5).$$

INTERPRETATION: We are 95 percent confident that the population mean length μ is between 12.1 and 12.5 mm. $\square$

NOTE: The interval $(12.1, 12.5)$ is no longer random! Thus, it is not theoretically appropriate to say that "the mean length μ is between 12.1 and 12.5 with probability 0.95." A confidence interval, after it has been computed with actual data (like above), no longer possesses any randomness. We only attach probabilities to events involving random quantities.

INTERPRETATION: Instead of attaching the concept of probability to the interpretation of a confidence interval, here is how one must think about them. *In repeated sampling, approximately* $100(1-\alpha)$ *percent of the confidence intervals will contain the true parameter* θ*. Our calculated interval is just one of these.*

TERMINOLOGY: We call the quantity Q a **pivotal quantity**, or a **pivot**, if its sampling distribution does not depend on any unknown parameters. Note that Q can depend on unknown parameters, but its sampling distribution can not. *Pivots help us derive confidence intervals.* Illustrative examples now follow.

Example 8.10. In Example 8.9, the quantity

$$Z = \frac{\overline{Y} - \mu}{\sigma_0/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Since the standard normal distribution does not depend on any unknown parameters, Z is a pivot. We used this fact to derive a $100(1-\alpha)$ confidence interval for the population mean μ, when $\sigma^2 = \sigma_0^2$ was known. $\square$

Example 8.11. The time (in seconds) for a certain chemical reaction to take place is assumed to follow a $\mathcal{U}(0,\theta)$ distribution. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of such times and that we would like to derive a $100(1-\alpha)$ percent confidence interval for θ, the maximum possible time. Intuitively, the largest order statistic $Y_{(n)}$ should be "close" to θ, so let's use $Y_{(n)}$ as an estimator. From Example 8.2, the pdf of $Y_{(n)}$ is given by

$$f_{Y_{(n)}}(y) = \begin{cases} n\theta^{-n}y^{n-1}, & 0 < y < \theta \\ 0, & \text{otherwise.} \end{cases}$$

As we will now show,

$$Q = \frac{Y_{(n)}}{\theta}$$

is a pivot. We can show this using a transformation argument. With $q = y_{(n)}/\theta$, the inverse transformation is given by $y_{(n)} = q\theta$ and the Jacobian is $dy_{(n)}/dq = \theta$. Thus, the pdf of Q, for values of $0 < q < 1$ (why?), is given by

$$\begin{aligned} f_Q(q) &= f_{Y_{(n)}}(q\theta) \times |\theta| \\ &= n\theta^{-n}(q\theta)^{n-1} \times \theta \\ &= nq^{n-1}. \end{aligned}$$

You should recognize that $Q \sim \text{beta}(n, 1)$. Since Q has a distribution free of unknown parameters, Q is a pivot, as claimed.

USING THE PIVOT: Define b as the value that satisfies $P(Q > b) = 1 - \alpha$. That is, b solves

$$1 - \alpha = P(Q > b) = \int_b^1 nq^{n-1}dq = 1 - b^n,$$

so that $b = \alpha^{1/n}$. Recognizing that $P(Q > b) = P(b < Q < 1)$, it follows that

$$\begin{aligned} 1 - \alpha = P(\alpha^{1/n} < Q < 1) &= P\left(\alpha^{1/n} < \frac{Y_{(n)}}{\theta} < 1\right) \\ &= P\left(\alpha^{-1/n} > \frac{\theta}{Y_{(n)}} > 1\right) \\ &= P\left(Y_{(n)} < \theta < \alpha^{-1/n}Y_{(n)}\right). \end{aligned}$$

This argument shows that

$$(Y_{(n)}, \alpha^{-1/n}Y_{(n)})$$

is a $100(1 - \alpha)$ percent confidence interval for the unknown parameter θ. □

Example 8.11 (revisited). Table 8.2 contains $n = 36$ chemical reaction times, modeled as iid $\mathcal{U}(0, \theta)$ realizations. The largest order statistic is $y_{(36)} = 9.962$. With $\alpha = 0.05$, a 95 percent confidence interval for θ is

$$(9.962,\ (0.05)^{-1/36} \times 9.962) \Longrightarrow (9.962, 10.826).$$

Thus, we are 95 percent confident that the maximum reaction time θ is between 9.962 and 10.826 seconds. □

Table 8.2: *Chemical reaction data. These observations are modeled as $n = 36$ realizations from $\mathcal{U}(0, \theta)$ distribution.*

0.478	0.787	1.102	0.851	8.522	5.272	4.113	7.921	3.457
3.457	9.159	6.344	6.481	4.448	5.756	0.076	3.462	**9.962**
2.938	3.281	5.481	1.232	5.175	5.864	8.176	2.031	1.633
4.803	8.249	8.991	7.358	2.777	5.905	7.762	8.563	7.619

Example 8.12. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from an exponential distribution with mean θ and that we would like to estimate θ with a $100(1-\alpha)$ percent confidence interval. Recall that

$$T = \sum_{i=1}^{n} Y_i \sim \text{gamma}(n, \theta)$$

and that

$$Q = \frac{2T}{\theta} \sim \chi^2(2n).$$

Thus, since Q has a distribution free of unknown parameters, Q is a pivot. Because $Q \sim \chi^2(2n)$, we can trap Q between two quantiles from the $\chi^2(2n)$ distribution with probability $1-\alpha$. In particular, let $\chi^2_{2n,1-\alpha/2}$ and $\chi^2_{2n,\alpha/2}$ denote the lower and upper $\alpha/2$ quantiles of a $\chi^2(2n)$ distribution; that is, $\chi^2_{2n,1-\alpha/2}$ solves

$$P(Q < \chi^2_{2n,1-\alpha/2}) = \alpha/2$$

and $\chi^2_{2n,\alpha/2}$ solves

$$P(Q > \chi^2_{2n,\alpha/2}) = \alpha/2.$$

Recall that the χ^2 distribution is tabled in Table 6 (WMS); the quantiles $\chi^2_{2n,1-\alpha/2}$ and $\chi^2_{2n,\alpha/2}$ can be found in this table (or by using R). We have that

$$\begin{aligned} 1-\alpha = P(\chi^2_{2n,1-\alpha/2} < Q < \chi^2_{2n,\alpha/2}) &= P\left(\chi^2_{2n,1-\alpha/2} < \frac{2T}{\theta} < \chi^2_{2n,\alpha/2}\right) \\ &= P\left(\frac{1}{\chi^2_{2n,1-\alpha/2}} > \frac{\theta}{2T} > \frac{1}{\chi^2_{2n,\alpha/2}}\right) \\ &= P\left(\frac{2T}{\chi^2_{2n,\alpha/2}} < \theta < \frac{2T}{\chi^2_{2n,1-\alpha/2}}\right). \end{aligned}$$

Table 8.3: *Observed explosion data. These observations are modeled as $n = 8$ realizations from an exponential distribution with mean θ.*

3.690	14.091	1.989	0.047	8.114	4.996	20.734	6.975

This argument shows that

$$\left(\frac{2T}{\chi^2_{2n,\alpha/2}}, \frac{2T}{\chi^2_{2n,1-\alpha/2}}\right)$$

is a $100(1-\alpha)$ percent confidence interval for θ. $\square$

Example 8.12 (revisited). Explosive devices used in mining operations produce nearly circular craters when detonated. The radii of these craters, measured in feet, follow an exponential distribution with mean θ. An iid sample of $n = 8$ explosions is observed and the radii observed in the explosions are catalogued in Table 8.3. With these data, we would like to write a 90 percent confidence interval for θ. The sum of the radii is $t = \sum_{i=1}^{8} y_i = 60.636$. With $n = 8$ and $\alpha = 0.10$, we find (from WMS, Table 6),

$$\begin{aligned} \chi^2_{16,0.95} &= 7.96164 \\ \chi^2_{16,0.05} &= 26.2962. \end{aligned}$$

A 90 percent confidence interval for θ based on these data is

$$\left(\frac{2 \times 60.636}{26.2962}, \frac{2 \times 60.636}{7.96164}\right) \Longrightarrow (4.612, 15.232).$$

Thus, we are 90 percent confident that the mean crater radius θ is between 4.612 and 15.232 feet. $\square$

8.7 Large-sample confidence intervals

TERMINOLOGY: The terms "large-sample" and/or "asymptotic" are used to describe confidence intervals that are constructed from asymptotic theory. Of course, the main asymptotic result we have seen so far is the **Central Limit Theorem**. This theorem provides the basis for the large-sample intervals studied in this subsection.

GOALS: In particular, we will present **large-sample confidence intervals** for

1. one population mean μ
2. one population proportion p
3. the difference of two population means $\mu_1 - \mu_2$
4. the difference of two population proportions $p_1 - p_2$.

Because these are "large-sample" confidence intervals, this means that the intervals are approximate, so their true confidence levels are "close" to $1-\alpha$ for large sample sizes.

LARGE-SAMPLE APPROACH: In each of the situations listed above, we will use a point estimator, say, $\widehat{\theta}$, which satisfies

$$Z = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} \sim \mathcal{AN}(0,1),$$

for large sample sizes. In this situation, we say that Z is an **asymptotic pivot** because its large-sample distribution is free of all unknown parameters. Because Z follows an approximate standard normal distribution, we can find a value $z_{\alpha/2}$ that satisfies

$$P\left(-z_{\alpha/2} < \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}} < z_{\alpha/2}\right) \approx 1-\alpha,$$

which, after straightforward algebra (verify!), can be restated as

$$P\left(\widehat{\theta} - z_{\alpha/2}\sigma_{\widehat{\theta}} < \theta < \widehat{\theta} + z_{\alpha/2}\sigma_{\widehat{\theta}}\right) \approx 1-\alpha.$$

This shows that

$$\widehat{\theta} \pm z_{\alpha/2}\sigma_{\widehat{\theta}}$$

is an **approximate** $100(1-\alpha)$ percent confidence interval for the parameter θ.

PROBLEM: As we will see shortly, the standard error $\sigma_{\widehat{\theta}}$ will often depend on unknown parameters (either θ itself or other unknown parameters). This is a problem, because we are trying to compute a confidence interval for θ, and the standard error $\sigma_{\widehat{\theta}}$ depends on population parameters which are not known.

SOLUTION: If we can substitute a "good" estimator for $\sigma_{\widehat{\theta}}$, say, $\widehat{\sigma}_{\widehat{\theta}}$, then the interval

$$\widehat{\theta} \pm z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}}$$

should remain valid in large samples. The theoretical justification as to why this approach is, in fact, reasonable will be seen in the next chapter.

"GOOD" ESTIMATOR: In the preceding paragraph, the term "good" is used to describe an estimator $\widehat{\sigma}_{\widehat{\theta}}$ that "approaches" the true standard error $\sigma_{\widehat{\theta}}$ (in some sense) as the sample size(s) become(s) large. Thus, we have two approximations at play:

- the Central Limit Theorem that approximates the true sampling distribution of
$$Z = \frac{\widehat{\theta} - \theta}{\sigma_{\widehat{\theta}}}$$
- the approximation arising from using $\widehat{\sigma}_{\widehat{\theta}}$ as an estimate of $\sigma_{\widehat{\theta}}$.

TERMINOLOGY: I like to call $\widehat{\sigma}_{\widehat{\theta}}$ the **estimated standard error**. It is simply a point estimate of the true standard error $\sigma_{\widehat{\theta}}$.

APPROXIMATE CONFIDENCE INTERVALS: We will use

$$\widehat{\theta} \pm z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}}$$

as an **approximate** $100(1-\alpha)$ percent confidence interval for θ. We now present this interval in the context of our four scenarios described earlier. *Each of the following intervals is valid for large sample sizes.* These intervals may not be valid for small sample sizes.

8.7.1 One population mean

SITUATION: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample with mean μ and variance σ^2 and that interest lies in estimating the population mean μ. In this situation, the Central Limit Theorem says that

$$Z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{AN}(0,1)$$

for large n. Here,

$$\begin{aligned} \theta &= \mu \\ \widehat{\theta} &= \overline{Y} \\ \sigma_{\widehat{\theta}} &= \frac{\sigma}{\sqrt{n}} \\ \widehat{\sigma}_{\widehat{\theta}} &= \frac{S}{\sqrt{n}}, \end{aligned}$$

where S denotes the sample standard deviation. Thus,

$$\overline{Y} \pm z_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$$

is an approximate $100(1-\alpha)$ percent confidence interval for the population mean μ.

Example 8.13. The administrators for a hospital would like to estimate the mean number of days required for in-patient treatment of patients between the ages of 25 and 34 years. A random sample of $n = 500$ hospital patients between these ages produced the following sample statistics:

$$\begin{aligned} \overline{y} &= 5.4 \text{ days} \\ s &= 3.1 \text{ days.} \end{aligned}$$

Construct a 90 percent confidence interval for μ, the (population) mean length of stay for this cohort of patients.

SOLUTION. Here, $n = 500$ and $z_{0.10/2} = z_{0.05} = 1.65$. Thus, a 90 percent confidence interval for μ is

$$5.4 \pm 1.65\left(\frac{3.1}{\sqrt{500}}\right) \implies (5.2, 5.6) \text{ days.}$$

We are 90 percent confident that the true mean length of stay, for patients aged 25-34, is between 5.2 and 5.6 days. □

8.7.2 One population proportion

SITUATION: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid Bernoulli(p) sample, where $0 < p < 1$, and that interest lies in estimating the population proportion p. In this situation, the

Central Limit Theorem says that

$$Z = \frac{\widehat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathcal{AN}(0,1)$$

for large n, where $\widehat{p}$ denotes the sample proportion. Here,

$$\begin{aligned} \theta &= p \\ \widehat{\theta} &= \widehat{p} \\ \sigma_{\widehat{\theta}} &= \sqrt{\frac{p(1-p)}{n}} \\ \widehat{\sigma}_{\widehat{\theta}} &= \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}. \end{aligned}$$

Thus,

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

is an approximate $100(1-\alpha)$ percent confidence interval for p.

Example 8.14. The Women's Interagency HIV Study (WIHS) is a large observational study funded by the National Institutes of Health to investigate the effects of HIV infection in women. The WIHS study reports that a total of 1288 HIV-infected women were recruited to examine the prevalence of childhood abuse. Of the 1288 HIV positive women, a total of 399 reported that, in fact, they had been a victim of childhood abuse. Find a 95 percent confidence interval for p, the true proportion of HIV infected women who are victims of childhood abuse.

SOLUTION. Here, $n = 1288$, $z_{0.05/2} = z_{0.025} = 1.96$, and the sample proportion of HIV childhood abuse victims is

$$\widehat{p} = \frac{399}{1288} \approx 0.31.$$

Thus, a 95 percent confidence interval for p is

$$0.31 \pm 1.96\sqrt{\frac{0.31(1-0.31)}{1288}} \implies (0.28, 0.34).$$

We are 95 percent confident that the true proportion of HIV infected women who are victims of childhood abuse is between 0.28 and 0.34. □

8.7.3 Difference of two population means

SITUATION: Suppose that we have two **independent** samples; i.e.,

Sample 1: $Y_{11}, Y_{12}, ..., Y_{1n_1}$ are iid with mean μ_1 and variance σ_1^2

Sample 2: $Y_{21}, Y_{22}, ..., Y_{2n_2}$ are iid with mean μ_2 and variance σ_2^2

and that interest lies in estimating the population mean difference $\mu_1 - \mu_2$. In this situation, the Central Limit Theorem says that

$$Z = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{AN}(0,1)$$

for large n_1 and n_2. Here,

$$\begin{aligned} \theta &= \mu_1 - \mu_2 \\ \widehat{\theta} &= \overline{Y}_{1+} - \overline{Y}_{2+} \\ \sigma_{\widehat{\theta}} &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ \widehat{\sigma}_{\widehat{\theta}} &= \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \end{aligned}$$

where S_1^2 and S_2^2 are the respective sample variances. Thus,

$$(\overline{Y}_{1+} - \overline{Y}_{2+}) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

is an approximate $100(1-\alpha)$ percent confidence interval for the mean difference $\mu_1 - \mu_2$.

Example 8.15. A botanist is interested in comparing the growth response of dwarf pea stems to different levels of the hormone indoleacetic acid (IAA). Using 16-day-old pea plants, the botanist obtains 5-millimeter sections and floats these sections on solutions with different hormone concentrations to observe the effect of the hormone on the growth of the pea stem. Let Y_1 and Y_2 denote, respectively, the independent growths that can be attributed to the hormone during the first 26 hours after sectioning for $\frac{1}{2}10^{-4}$ and 10^{-4} levels of concentration of IAA (measured in mm). Summary statistics from the study are given in Table 8.4.

Table 8.4: *Botany data. Summary statistics for pea stem growth by hormone treatment.*

Treatment	Sample size	Sample mean	Sample standard deviation
$\frac{1}{2}10^{-4}$ mm IAA	$n_1 = 53$	$\overline{y}_{1+} = 1.03$	$s_1 = 0.49$
10^{-4} mm IAA	$n_2 = 51$	$\overline{y}_{2+} = 1.66$	$s_2 = 0.59$

The researcher would like to construct a 99 percent confidence interval for $\mu_1 - \mu_2$, the mean difference in growths for the two IAA levels. This confidence interval is

$$(1.03 - 1.66) \pm 2.58\sqrt{\frac{(0.49)^2}{53} + \frac{(0.59)^2}{51}} \implies (-0.90, -0.36).$$

That is, we are 99 percent confident that the mean difference $\mu_1 - \mu_2$ is between -0.90 and -0.36. Note that, because this interval does not conclude 0, this analysis suggests that the two (population) means are, in fact, truly different. □

8.7.4 Difference of two population proportions

SITUATION: Suppose that we have two **independent** samples; i.e.,

$$\begin{aligned} &\text{Sample 1:} \quad Y_{11}, Y_{12}, ..., Y_{1n_1} \text{ are iid Bernoulli}(p_1) \\ &\text{Sample 2:} \quad Y_{21}, Y_{22}, ..., Y_{2n_2} \text{ are iid Bernoulli}(p_2) \end{aligned}$$

and that interest lies in estimating the population proportion difference $p_1 - p_2$. In this situation, the Central Limit Theorem says that

$$Z = \frac{(\widehat{p}_1 - \widehat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim \mathcal{AN}(0, 1)$$

for large n_1 and n_2, where $\widehat{p}_1$ and $\widehat{p}_2$ are the sample proportions. Here,

$$\begin{aligned} \theta &= p_1 - p_2 \\ \widehat{\theta} &= \widehat{p}_1 - \widehat{p}_2 \\ \sigma_{\widehat{\theta}} &= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \\ \widehat{\sigma}_{\widehat{\theta}} &= \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}. \end{aligned}$$

Thus,

$$(\widehat{p}_1 - \widehat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$$

is an approximate $100(1-\alpha)$ percent confidence interval for the population proportion difference $p_1 - p_2$.

Example 8.16. An experimental type of chicken feed, Ration 1, contains a large amount of an ingredient that enables farmers to raise heavier chickens. However, this feed may be too strong and the mortality rate may be higher than that with the usual feed. One researcher wished to compare the mortality rate of chickens fed Ration 1 with the mortality rate of chickens fed the current best-selling feed, Ration 2. Denote by p_1 and p_2 the population mortality rates (proportions) for Ration 1 and Ration 2, respectively. She would like to get a 95 percent confidence interval for $p_1 - p_2$. Two hundred chickens were randomly assigned to each ration; of those fed Ration 1, 24 died within one week; of those fed Ration 2, 16 died within one week.

- Sample 1: 200 chickens fed Ration 1 $\Longrightarrow \widehat{p}_1 = 24/200 = 0.12$
- Sample 2: 200 chickens fed Ration 2 $\Longrightarrow \widehat{p}_2 = 16/200 = 0.08$.

An approximate 95 percent confidence interval for the true difference $p_1 - p_2$ is

$$(0.12 - 0.08) \pm 1.96\sqrt{\frac{0.12(1-0.12)}{200} + \frac{0.08(1-0.08)}{200}} \Longrightarrow (-0.02, 0.10).$$

Thus, we are 95 percent confident that the true difference in mortality rates is between -0.02 and 0.10. Note that this interval does include 0, so we do not have strong (statistical) evidence that the mortality rates (p_1 and p_2) are truly different. □

8.8 Sample size determinations

MOTIVATION: In many research investigations, it is of interest to determine how many observations are needed to write a $100(1-\alpha)$ percent confidence interval with a given precision. For example, we might want to construct a 95 percent confidence interval for a

population mean in a way so that the confidence interval length is no more than 5 units (e.g., days, inches, dollars, etc.). Sample-size determinations ubiquitously surface in agricultural experiments, clinical trials, engineering investigations, epidemiological studies, etc., and, in most real problems, there is no "free lunch." Collecting more data costs money! Thus, one must be cognizant not only of the statistical issues associated with sample-size determination, but also of the practical issues like cost, time spent in data collection, personnel training, etc.

8.8.1 One population mean

SIMPLE SETTING: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma_0^2)$ population, where σ_0^2 is known. In this situation, an exact $100(1-\alpha)$ percent confidence interval for μ is given by

$$\overline{Y} \pm \underbrace{z_{\alpha/2}\left(\frac{\sigma_0}{\sqrt{n}}\right)}_{=B, \text{ say}},$$

where B denotes the bound on the error in estimation; this bound is called the **margin of error**.

SAMPLE SIZE FORMULA: In the setting described above, it is possible to determine the sample size n necessary once we specify these two pieces of information:

- the confidence level, $100(1-\alpha)$
- the margin of error, B.

This is true because

$$B = z_{\alpha/2}\left(\frac{\sigma_0}{\sqrt{n}}\right) \iff n = \left(\frac{z_{\alpha/2}\sigma_0}{B}\right)^2.$$

Example 8.17. In a biomedical experiment, we would like to estimate the mean remaining life of healthy rats that are given a high dose of a toxic substance. This may be done in an early phase clinical trial by researchers trying to find a maximum tolerable dose for

humans. Suppose that we would like to write a 99 percent confidence interval for μ with a margin of error equal to $B = 2$ days. From past studies, remaining rat lifetimes are well-approximated by a normal distribution with standard deviation $\sigma_0 = 8$ days. How many rats should we use for the experiment?

SOLUTION. Here, $z_{0.01/2} = z_{0.005} = 2.58$, $B = 2$, and $\sigma_0 = 8$. Thus,

$$n = \left(\frac{z_{\alpha/2}\sigma_0}{B}\right)^2 = \left(\frac{2.58 \times 8}{2}\right)^2 \approx 106.5.$$

Thus, we would need $n = 107$ rats to achieve these goals. □

8.8.2 One population proportion

SITUATION: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid Bernoulli(p) sample, where $0 < p < 1$, and that interest lies in writing a confidence interval for p with a prescribed length. In this situation, we know that

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

is an approximate $100(1-\alpha)$ percent confidence interval for p.

SAMPLE SIZE: To determine the sample size for estimating p with a $100(1-\alpha)$ percent confidence interval, we need to specify the **margin of error** that we desire; i.e.,

$$B = z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}.$$

We would like to solve this equation for n. However, note that B depends on $\widehat{p}$, which, in turn, depends on n! This is a small problem, but we can overcome it by replacing $\widehat{p}$ with p^*, a **guess** for the value of p. Doing this, the last expression becomes

$$B = z_{\alpha/2}\sqrt{\frac{p^*(1-p^*)}{n}},$$

and solving this equation for n, we get

$$n = \left(\frac{z_{\alpha/2}}{B}\right)^2 p^*(1-p^*).$$

This is the desired sample size to find a $100(1-\alpha)$ percent confidence interval for p with a prescribed margin of error equal to B.

Example 8.18. In a Phase II clinical trial, it is posited that the proportion of patients responding to a certain drug is $p^* = 0.4$. To engage in a larger Phase III trial, the researchers would like to know how many patients they should recruit into the study. Their resulting 95 percent confidence interval for p, the true population proportion of patients responding to the drug, should have a margin of error no greater than $B = 0.03$. What sample size do they need for the Phase III trial?

SOLUTION. Here, we have $B = 0.03$, $p^* = 0.4$, and $z_{0.05/2} = z_{0.025} = 1.96$. The desired sample size is

$$n = \left(\frac{z_{\alpha/2}}{B}\right)^2 p^*(1-p^*) = \left(\frac{1.96}{0.03}\right)^2 (0.4)(1-0.4) \approx 1024.43.$$

Thus, their Phase III trial should recruit around 1025 patients. □

CONSERVATIVE APPROACH: If there is no sensible guess for p available, use $p^* = 0.5$. In this situation, the resulting value for n will be as large as possible. Put another way, using $p^* = 0.5$ gives the most **conservative** solution (i.e., the largest sample size, n). This is true because

$$n = n(p^*) = \left(\frac{z_{\alpha/2}}{B}\right)^2 p^*(1-p^*),$$

when viewed as a function of p^*, is maximized when $p^* = 0.5$.

8.9 Small-sample confidence intervals for normal means

RECALL: We have already discussed how one can use large-sample arguments to justify the use of **large-sample confidence intervals** like

$$\overline{Y} \pm z_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$$

for estimating a single population mean, μ, and

$$(\overline{Y}_{1+} - \overline{Y}_{2+}) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

for estimating the difference of two population means, $\mu_1 - \mu_2$.

CURIOSITY: What happens if the sample size n (or the sample sizes n_1 and n_2 in the two-sample case) is/are not large? How appropriate are these intervals? Unfortunately, neither of these confidence intervals is preferred when dealing with small sample sizes. Thus, we need to treat small-sample problems differently. In doing so, we will assume (at least initially) that we are dealing with normally distributed data.

8.9.1 One population mean

SETTING: Suppose that $Y_1, Y_2, ...Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population. If $\sigma^2 = \sigma_0^2$ is **known**, we have already seen that

$$\overline{Y} \pm z_{\alpha/2}\left(\frac{\sigma_0}{\sqrt{n}}\right)$$

is an exact $100(1-\alpha)$ percent confidence interval for μ.

PROBLEM: In most real problems, rarely will anyone tell us the value of σ^2. That is, it is almost always the case that the population variance σ^2 is an unknown parameter. One might think to try using S as a point estimator for σ and substituting it into the confidence interval formula above. This is certainly not illogical, but, the sample standard deviation S is not an unbiased estimator for σ! Thus, if the sample size is small, there is no guarantee that sample standard deviation S will be "close" to the population standard deviation σ (it likely will be "close" if the sample size n is large). Furthermore, when the sample size n is small, the bias and variability associated with S (as an estimator of σ) could be large. To obviate this difficulty, we recall the following result.

RECALL: Suppose that $Y_1, Y_2, ...Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ population. From past results, we know that

$$T = \frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

Note that since the sampling distribution of T is free of all unknown parameters, T a **pivotal quantity**. So, just like before, we can use this fact to derive an exact $100(1-\alpha)$ percent confidence interval for μ.

DERIVATION: Let $t_{n-1,\alpha/2}$ denote the upper $\alpha/2$ quantile of the $t(n-1)$ distribution. Then, because $T \sim t(n-1)$, we can write

$$\begin{aligned}
1-\alpha &= P(-t_{n-1,\alpha/2} < T < t_{n-1,\alpha/2}) \\
&= P\left(-t_{n-1,\alpha/2} < \frac{\overline{Y}-\mu}{S/\sqrt{n}} < t_{n-1,\alpha/2}\right) \\
&= P\left(-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} < \overline{Y}-\mu < t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) \\
&= P\left(t_{n-1,\alpha/2}\frac{S}{\sqrt{n}} > \mu-\overline{Y} > -t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}\right) \\
&= P\Bigg(\underbrace{\overline{Y}-t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}}_{\widehat{\theta}_L} < \mu < \underbrace{\overline{Y}+t_{n-1,\alpha/2}\frac{S}{\sqrt{n}}}_{\widehat{\theta}_U}\Bigg).
\end{aligned}$$

This argument shows that

$$\overline{Y} \pm t_{n-1,\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$$

is an exact $100(1-\alpha)$ percent confidence interval for the population mean μ. This interval is "exact" only if the underlying probability distribution is normal. □

Example 8.19. In an agricultural experiment, a random sample of $n = 10$ plots produces the yields below (measured in kg per plot). From past studies, it has been observed that plot yields vary according to a normal distribution. The goal is to write a 95 percent confidence interval for μ, the population mean yield. Here are the sample yields:

```
23.2   20.1   18.8   19.3   24.6   27.1   33.7   24.7   32.4   17.3
```

From these data, we compute $\overline{y} = 24.1$ and $s = 5.6$. Also, with $n = 10$, the degrees of freedom is $n - 1 = 9$, and $t_{n-1,\alpha/2} = t_{9,0.025} = 2.262$ (WMS Table 5). The 95 percent confidence interval is

$$24.1 \pm 2.262\left(\frac{5.6}{\sqrt{10}}\right) \implies (20.1, 28.1).$$

Thus, based on these data, we are 95 percent confident that the population mean yield μ is between 20.1 and 28.1 kg/plot. □

8.9.2 Difference of two population means

TWO-SAMPLE SETTING: Suppose that we have two **independent** samples:

$$\begin{aligned} \text{Sample 1}: &\quad Y_{11}, Y_{12}, ..., Y_{1n_1} \sim \text{iid } \mathcal{N}(\mu_1, \sigma_1^2) \\ \text{Sample 2}: &\quad Y_{21}, Y_{22}, ..., Y_{2n_2} \sim \text{iid } \mathcal{N}(\mu_2, \sigma_2^2) \end{aligned}$$

and that we would like to construct a $100(1-\alpha)$ percent confidence interval for the difference of population means $\mu_1 - \mu_2$. As before, we define the statistics

$$\begin{aligned} \overline{Y}_{1+} = \frac{1}{n_1}\sum_{j=1}^{n_1} Y_{1j} &= \text{sample mean for sample 1} \\ \overline{Y}_{2+} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_{2j} &= \text{sample mean for sample 2} \\ S_1^2 = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(Y_{1j} - \overline{Y}_{1+})^2 &= \text{sample variance for sample 1} \\ S_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(Y_{2j} - \overline{Y}_{2+})^2 &= \text{sample variance for sample 2.} \end{aligned}$$

We know that

$$\overline{Y}_{1+} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \text{and} \quad \overline{Y}_{2+} \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right).$$

Furthermore, since $\overline{Y}_{1+}$ and $\overline{Y}_{2+}$ are both normally distributed, the difference $\overline{Y}_{1+} - \overline{Y}_{2+}$ is too since it is just a linear combination of $\overline{Y}_{1+}$ and $\overline{Y}_{2+}$. By straightforward calculation, it follows that

$$\overline{Y}_{1+} - \overline{Y}_{2+} \sim \mathcal{N}\left(\mu_1 - \mu_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

Standardizing, we get

$$Z = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0,1).$$

Also recall that $(n_1-1)S_1^2/\sigma_1^2 \sim \chi^2(n_1-1)$ and that $(n_2-1)S_2^2/\sigma_2^2 \sim \chi^2(n_2-1)$. It follows that

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} + \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_1+n_2-2).$$

REMARK: The population variances are **nuisance parameters** in the sense that they are not the parameters of interest here. Still, they have to be estimated. We want to write a confidence interval for $\mu_1 - \mu_2$, but exactly how this interval is constructed depends on the true values of σ_1^2 and σ_2^2. In particular, we consider two cases:

- $\sigma_1^2 = \sigma_2^2 = \sigma^2$; that is, the two population variances are **equal**
- $\sigma_1^2 \neq \sigma_2^2$; that is, the two population variances are **not equal**.

EQUAL-VARIANCE ASSUMPTION: When $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we have

$$Z = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim \mathcal{N}(0,1)$$

and

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} + \frac{(n_2-1)S_2^2}{\sigma_2^2} = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2} \sim \chi^2(n_1+n_2-2).$$

Thus,

$$\frac{\frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{\sigma^2}/(n_1+n_2-2)}} = \frac{\text{“}\mathcal{N}(0,1)\text{”}}{\text{“}\chi^2(n_1+n_2-2)\text{”}/(n_1+n_2-2)} \sim t(n_1+n_2-2).$$

The last distribution results because the numerator and denominator are independent (why?). But, algebraically, the last expression reduces to

$$T = \frac{(\overline{Y}_{1+} - \overline{Y}_{2+}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1+n_2-2),$$

where

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

is the **pooled sample variance estimator** of the common population variance σ^2.

PIVOTAL QUANTITY: Since T has a sampling distribution that is free of all unknown parameters, it is a pivotal quantity. We can use this fact to construct a $100(1-\alpha)$ percent

confidence interval for mean difference $\mu_1 - \mu_2$. In particular, because $T \sim t(n_1 + n_2 - 2)$, we can find the value $t_{n_1+n_2-2,\alpha/2}$ that satisfies

$$P(-t_{n_1+n_2-2,\alpha/2} < T < t_{n_1+n_2-2,\alpha/2}) = 1 - \alpha.$$

Substituting T into the last expression and performing the usual algebraic manipulations (verify!), we can conclude that

$$(\overline{Y}_{1+} - \overline{Y}_{2+}) \pm t_{n_1+n_2-2,\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

is an exact $100(1-\alpha)$ percent confidence interval for the mean difference $\mu_1 - \mu_2$. $\square$

Example 8.20. In the vicinity of a nuclear power plant, marine biologists at the EPA would like to determine whether there is a difference between the mean weight in two species of a certain fish. To do this, they will construct a 90 percent confidence interval for the mean difference $\mu_1 - \mu_2$. Two independent random samples were taken, and here are the recorded weights (in ounces):

- Species 1: `29.9, 11.4, 25.3, 16.5, 21.1`
- Species 2: `26.6, 23.7, 28.5, 14.2, 17.9, 24.3`

Out of necessity, the scientists assume that each sample arises from a normal distribution with $\sigma_1^2 = \sigma_2^2 = \sigma^2$ (i.e., they assume a common population variance). Here, we have $n_1 = 5$, $n_2 = 6$, $n_1 + n_2 - 2 = 9$, and $t_{9,0.05} = 1.833$. Straightforward computations show that $\overline{y}_{1+} = 20.84$, $s_1^2 = 52.50$, $\overline{y}_{2+} = 22.53$, $s_2^2 = 29.51$, and that

$$s_p^2 = \frac{4(52.50) + 5(29.51)}{9} = 39.73.$$

Thus, the 90 percent confidence interval for $\mu_1 - \mu_2$, based on these data, is given by

$$(20.84 - 22.53) \pm 1.833\sqrt{39.73}\sqrt{\frac{1}{5} + \frac{1}{6}} \implies (-8.69, 5.31).$$

We are 90 percent confident that the mean difference $\mu_1 - \mu_2$ is between -8.69 and 5.31 ounces. Since this interval includes 0, this analysis does not suggest that the mean species weights, μ_1 and μ_2, are truly different. $\square$

UNEQUAL-VARIANCE ASSUMPTION: When $\sigma_1^2 \neq \sigma_2^2$, the problem of constructing a $100(1-\alpha)$ percent confidence interval for $\mu_1 - \mu_2$ becomes markedly more difficult. The reason why this is true stems from the fact that there is no "obvious" pivotal quantity to construct (go back to the equal-variance case and see how this assumption simplified the derivation). However, in this situation, we can still write an **approximate** confidence interval for $\mu_1 - \mu_2$; this interval is given by

$$(\overline{Y}_{1+} - \overline{Y}_{2+}) \pm t_{\nu,\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}},$$

where the degree of freedom parameter ν is approximated by

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}}.$$

This formula for ν is called **Satterwaite's formula**. The derivation of this interval is left to another day.

8.9.3 Robustness of the t procedures

REMARK: In the derivation of the one and two-sample confidence intervals for normal means (based on the t distribution), we have explicitly assumed that the underlying population distribution(s) was/were normal. Under the normality assumption,

$$\overline{Y} \pm t_{n-1,\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$$

is an **exact** $100(1-\alpha)$ percent confidence interval for the population mean μ. Under the normal, independent sample, and constant variance assumptions,

$$(\overline{Y}_{1+} - \overline{Y}_{2+}) \pm t_{n_1+n_2-2,\alpha/2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

is an **exact** $100(1-\alpha)$ percent confidence interval for the mean difference $\mu_1 - \mu_2$. Of course, the natural question arises:

*"What if the data are **not** normally distributed?"*

ROBUSTNESS: A statistical inference procedure (like constructing a confidence interval) is said to be **robust** if the quality of the procedure is not affected by a departure from the assumptions made.

IMPORTANT: The t confidence interval procedures are based on the population distribution being normal. However, these procedures are fairly robust to departures from normality; i.e., even if the population distribution(s) is/are nonnormal, we can still use the t procedures and get approximate results. The following guidelines are common:

- $n < 15$: Use t procedures only if the population distribution appears normal and there are no outliers.
- $15 \leq n \leq 40$: Be careful about using t procedures if there is strong skewness and/or outliers present.
- $n > 40$: t procedures should be fine regardless of the population distribution shape.

REMARK: These are just guidelines and should not be taken as "truth." Of course, if we know the distribution of $Y_1, Y_2, ..., Y_n$ (e.g., Poisson, exponential, etc.), then we might be able to derive an exact $100(1-\alpha)$ percent confidence interval for the mean directly by finding a suitable pivotal quantity. In such cases, it may be better to avoid the t procedures altogether.

8.10 Confidence intervals for variances

MOTIVATION: In many experimental settings, the researcher is concerned not with the mean of the underlying population, but with the population variance σ^2 instead. For example, in a laboratory setting, chemists might wish to estimate the variability associated with a measurement system (e.g., scale, caliper, etc.) or to estimate the unit-to-unit variation of vitamin tablets. In large-scale field trials, agronomists are often likely to compare variability levels for different cultivars or genetically-altered varieties. In clinical trials, the FDA is often concerned whether or not there is significant variation among various clinic sites. We examine the one and two-sample problems here.

8.10.1 One population variance

RECALL: Suppose $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution. In this case, we know

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Because Q has a distribution that is free of all unknown parameters, Q is a pivot. We will use this pivot to derive an exact $100(1-\alpha)$ percent confidence interval for σ^2.

DERIVATION: Let $\chi^2_{n-1,\alpha/2}$ denote the upper $\alpha/2$ quantile and let $\chi^2_{n-1,1-\alpha/2}$ denote the lower $\alpha/2$ quantile of the $\chi^2(n-1)$ distribution; i.e., $\chi^2_{n-1,\alpha/2}$ and $\chi^2_{n-1,1-\alpha/2}$ satisfy

$$P[\chi^2(n-1) > \chi^2_{n-1,\alpha/2}] = \alpha/2 \quad \text{and} \quad P[\chi^2(n-1) < \chi^2_{n-1,1-\alpha/2}] = \alpha/2,$$

respectively. Then, because $Q \sim \chi^2(n-1)$,

$$\begin{aligned} 1-\alpha &= P\left[\chi^2_{n-1,1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1,\alpha/2}\right] \\ &= P\left[\frac{1}{\chi^2_{n-1,1-\alpha/2}} > \frac{\sigma^2}{(n-1)S^2} > \frac{1}{\chi^2_{n-1,\alpha/2}}\right] \\ &= P\left[\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}} > \sigma^2 > \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}\right]. \end{aligned}$$

This argument shows that

$$\left[\frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}\right]$$

is an exact $100(1-\alpha)$ percent confidence interval for the population variance σ^2. □

NOTE: Taking the square root of both endpoints in the $100(1-\alpha)$ percent confidence interval for σ^2 gives a $100(1-\alpha)$ percent confidence interval for σ.

Example 8.21. Entomologists studying the bee species *Euglossa mandibularis* Friese measure the wing-stroke frequency for $n = 4$ bees for a fixed time. The data are

235 225 190 188

Assuming that these data are an iid sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, find a 90 percent confidence interval for σ^2.

SOLUTION. Here, $n = 4$ and $\alpha = 0.10$, so we need $\chi^2_{3,0.95} = 0.351846$ and $\chi^2_{3,0.05} = 7.81473$ (Table 6, WMS). I used R to compute $s^2 = 577.6667$. The 90 percent confidence interval is thus

$$\left[\frac{3(577.6667)}{7.81473}, \frac{3(577.6667)}{0.351846}\right] \Longrightarrow (221.76, 4925.45).$$

That is, we are 90 percent confident that the true population variance σ^2 is between 221.76 and 4925.45; i.e., that the true population standard deviation σ is between 14.89 and 70.18. Of course, both of these intervals are quite wide, but remember that $n = 4$, so we shouldn't expect notably precise intervals. □

8.10.2 Ratio of two variances

TWO-SAMPLE SETTING: Suppose that we have two **independent** samples:

$$\begin{aligned} \text{Sample 1}: &\quad Y_{11}, Y_{12}, ..., Y_{1n_1} \sim \text{iid } \mathcal{N}(\mu_1, \sigma_1^2) \\ \text{Sample 2}: &\quad Y_{21}, Y_{22}, ..., Y_{2n_2} \sim \text{iid } \mathcal{N}(\mu_2, \sigma_2^2) \end{aligned}$$

and that we would like to construct a $100(1-\alpha)$ percent confidence interval for $\theta = \sigma_2^2/\sigma_1^2$, the **ratio** of the population variances. Under these model assumptions, we know that $(n_1-1)S_1^2/\sigma_1^2 \sim \chi^2(n_1-1)$, that $(n_2-1)S_2^2/\sigma_2^2 \sim \chi^2(n_2-1)$, and that these two quantities are independent. It follows that

$$F = \frac{\frac{(n_1-1)S_1^2}{\sigma_1^2}/(n_1-1)}{\frac{(n_2-1)S_2^2}{\sigma_2^2}/(n_2-1)} \sim \frac{\text{“}\chi^2(n_1-1)\text{”}/(n_1-1)}{\text{“}\chi^2(n_2-1)\text{”}/(n_2-1)} \sim F(n_1-1, n_2-1).$$

Because F has a distribution that is free of all unknown parameters, F is a pivot, and we can use it to derive $100(1-\alpha)$ percent confidence interval for $\theta = \sigma_2^2/\sigma_1^2$. Let $F_{n_1-1,n_2-1,\alpha/2}$ denote the upper $\alpha/2$ quantile and let $F_{n_1-1,n_2-1,1-\alpha/2}$ denote the lower $\alpha/2$ quantile of the $F(n_1-1, n_2-1)$ distribution. Because $F \sim F(n_1-1, n_2-1)$, we can write

$$\begin{aligned} 1-\alpha &= P\left(F_{n_1-1,n_2-1,1-\alpha/2} < F < F_{n_1-1,n_2-1,\alpha/2}\right) \\ &= P\left(F_{n_1-1,n_2-1,1-\alpha/2} < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < F_{n_1-1,n_2-1,\alpha/2}\right) \\ &= P\left(\frac{S_2^2}{S_1^2} \times F_{n_1-1,n_2-1,1-\alpha/2} < \frac{\sigma_2^2}{\sigma_1^2} < \frac{S_2^2}{S_1^2} \times F_{n_1-1,n_2-1,\alpha/2}\right). \end{aligned}$$

This argument shows that

$$\left(\frac{S_2^2}{S_1^2} \times F_{n_1-1,n_2-1,1-\alpha/2}, \frac{S_2^2}{S_1^2} \times F_{n_1-1,n_2-1,\alpha/2}\right)$$

is an exact $100(1-\alpha)$ percent confidence interval for the ratio $\theta = \sigma_2^2/\sigma_1^2$. $\square$

Example 8.22. Snout beetles cause millions of dollars worth of damage each year to cotton crops. Two different chemical treatments are used to control this beetle population using 13 randomly selected plots. Below are the percentages of cotton plants with beetle damage (after treatment) for the plots:

- Treatment 1: `22.3, 19.5, 18.6, 24.3, 19.9, 20.4`
- Treatment 2: `9.8, 12.3, 16.2, 14.1, 15.3, 10.8, 18.3`

Under normality, and assuming that these two samples are independent, find a 95 percent confidence interval for $\theta = \sigma_2^2/\sigma_1^2$, the ratio of the two treatment variances.

SOLUTION. Here, $n_1 = 6$, $n_2 = 7$, and $\alpha = 0.05$, so that $F_{5,6,0.025} = 5.99$ (WMS, Table 7). To find $F_{5,6,0.975}$, we can use the fact that

$$F_{5,6,0.975} = \frac{1}{F_{6,5,0.025}} = \frac{1}{6.98} \approx 0.14$$

(WMS, Table 7). Again, I used R to compute $s_1^2 = 4.40$ and $s_2^2 = 9.27$. Thus, a 95 percent confidence interval for $\theta = \sigma_2^2/\sigma_1^2$ is given by

$$\left(\frac{9.27}{4.40} \times 0.14, \frac{9.27}{4.40} \times 5.99\right) \implies (0.29, 12.62).$$

We are 95 percent confident that the ratio of variances $\theta = \sigma_2^2/\sigma_1^2$ is between 0.29 and 12.62. Since this interval includes 1, we can not conclude that the two treatment variances are significantly different. $\square$

NOTE: Unlike the t confidence intervals for means, the confidence interval procedures for one and two population variances are **not robust** to departures from normality. Thus, one who uses these confidence intervals is placing strong faith in the underlying normality assumption.

9 Properties of Point Estimators and Methods of Estimation

Complementary reading: Chapter 9 (WMS).

9.1 Introduction

RECALL: In many problems, we are able to observe an iid sample $Y_1, Y_2, ..., Y_n$ from a population distribution $f_Y(y;\theta)$, where θ is regarded as an **unknown parameter** that is to be estimated with the observed data. From the last chapter, we know that a "good" estimator $\widehat{\theta} = T(Y_1, Y_2, ..., Y_n)$ has the following properties:

- $\widehat{\theta}$ is **unbiased**; i.e., $E(\widehat{\theta}) = \theta$, for all θ
- $\widehat{\theta}$ has **small variance**.

In our quest to find a good estimator for θ, we might have several "candidate estimators" to consider. For example, suppose that $\widehat{\theta}_1 = T_1(Y_1, Y_2, ..., Y_n)$ and $\widehat{\theta}_2 = T_2(Y_1, Y_2, ..., Y_n)$ are two estimators for θ. Which estimator is better? Is there a "best" estimator available? If so, how do we find it? This chapter largely addresses this issue.

TERMINOLOGY: Suppose that $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are **unbiased** estimators for θ. We call

$$\text{eff}(\widehat{\theta}_1, \widehat{\theta}_2) = \frac{V(\widehat{\theta}_2)}{V(\widehat{\theta}_1)}$$

the **relative efficiency** of $\widehat{\theta}_2$ to $\widehat{\theta}_1$. This is simply a ratio of the variances. If

$$\begin{aligned} V(\widehat{\theta}_1) = V(\widehat{\theta}_2) &\iff \text{eff}(\widehat{\theta}_1, \widehat{\theta}_2) = 1 \\ V(\widehat{\theta}_1) > V(\widehat{\theta}_2) &\iff \text{eff}(\widehat{\theta}_1, \widehat{\theta}_2) < 1 \\ V(\widehat{\theta}_1) < V(\widehat{\theta}_2) &\iff \text{eff}(\widehat{\theta}_1, \widehat{\theta}_2) > 1. \end{aligned}$$

NOTE: It only makes sense to use this measure when both $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are **unbiased**.

Example 9.1. Suppose that Y_1, Y_2, Y_3 is an iid sample of $n = 3$ Poisson observations with mean θ. Consider the two candidate estimators:

$$\begin{aligned}\widehat{\theta}_1 &= \overline{Y} \\ \widehat{\theta}_2 &= \frac{1}{6}(Y_1 + 2Y_2 + 3Y_3).\end{aligned}$$

It is easy to see that both $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are unbiased estimators of θ (verify!). In deciding which estimator is better, we thus should compare $V(\widehat{\theta}_1)$ and $V(\widehat{\theta}_2)$. Straightforward calculations show that $V(\widehat{\theta}_1) = V(\overline{Y}) = \theta/3$ and

$$V(\widehat{\theta}_2) = \frac{1}{36}(\theta + 4\theta + 9\theta) = \frac{7\theta}{18}.$$

Thus,

$$\text{eff}(\widehat{\theta}_1, \widehat{\theta}_2) = \frac{V(\widehat{\theta}_2)}{V(\widehat{\theta}_1)} = \frac{7\theta/18}{\theta/3} = \frac{7}{6} \approx 1.17.$$

Since this value is larger than 1, $\widehat{\theta}_1$ is a better estimator than $\widehat{\theta}_2$. In other words, the estimator $\widehat{\theta}_2$ is only $100(6/7) \approx 86$ percent as efficient as $\widehat{\theta}_1$. $\square$

NOTE: There is not always a clear-cut winner when comparing two (or more) estimators. One estimator may perform better for certain values of θ, but be worse for other values of θ. Of course, it would be nice to have an estimator perform uniformly better than all competitors. This begs the question: Can we find the **best** estimator for the parameter θ? How should we define "best?"

CONVENTION: We will define the "best" estimator as one that is **unbiased** and has the **smallest possible variance** among all unbiased estimators.

9.2 Sufficiency

INTRODUCTION: No concept in the theory of point estimation is more important than that of sufficiency. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y; \theta)$ and that the goal is to find the best estimator for θ based on $Y_1, Y_2, ..., Y_n$. We will soon see that best estimators, if they exist, are always functions of **sufficient statistics**. For now, we will assume that θ is a scalar (we'll relax this assumption later).

TERMINOLOGY: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from the population distribution $f_Y(y;\theta)$. We call $U = g(Y_1, Y_2, ..., Y_n)$ a **sufficient statistic** for θ if the conditional distribution of $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$, given U, does not depend on θ.

ESTABLISHING SUFFICIENCY DIRECTLY: To show that U is sufficient, it suffices to show that the ratio

$$f_{\boldsymbol{Y}|U}(\boldsymbol{y}|u) = \frac{f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)}{f_U(u;\theta)}$$

does not depend on θ. Recall that since $Y_1, Y_2, ..., Y_n$ is an iid sample, the joint distribution of $\boldsymbol{Y}$ is the product of the marginal density (mass) functions; i.e.,

$$f_{\boldsymbol{Y}}(\boldsymbol{y};\theta) = \prod_{i=1}^{n} f_Y(y_i;\theta).$$

Example 9.2. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of Poisson observations with mean θ. Show that $U = \sum_{i=1}^{n} Y_i$ is a sufficient statistic for θ.

SOLUTION. A moment-generating function argument shows that $U \sim \text{Poisson}(n\theta)$; thus, the pdf of U is given by

$$f_U(u;\theta) = \begin{cases} \frac{(n\theta)^u e^{-n\theta}}{u!}, & u = 0, 1, 2, ..., \\ 0, & \text{otherwise.} \end{cases}$$

The joint distribution of the data $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$ is the product of the marginal Poisson mass functions; i.e.,

$$f_{\boldsymbol{Y}}(\boldsymbol{y};\theta) = \prod_{i=1}^{n} f_Y(y_i;\theta) = \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} = \frac{\theta^{\sum_{i=1}^{n} y_i} e^{-n\theta}}{\prod_{i=1}^{n} y_i!},$$

Therefore, the conditional distribution of $\boldsymbol{Y}$, given U, is equal to

$$\begin{aligned} f_{\boldsymbol{Y}|U}(\boldsymbol{y}|u) &= \frac{f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)}{f_U(u;\theta)} \\ &= \frac{\frac{\theta^u e^{-n\theta}}{\prod_{i=1}^{n} y_i!}}{(n\theta)^u e^{-n\theta}/u!} \\ &= \frac{u!}{n^u \prod_{i=1}^{n} y_i!}. \end{aligned}$$

Since $f_{\boldsymbol{Y}|U}(\boldsymbol{y}|u)$ does not depend on the unknown parameter θ, it follows (from the definition of sufficiency) that $U = \sum_{i=1}^{n} Y_i$ is a sufficient statistic for θ. $\square$

HEURISTIC INTERPRETATION: In a profound sense, sufficient statistics summarize all the information about the unknown parameter θ. That is, we can reduce our sample $Y_1, Y_2, ..., Y_n$ to a sufficient statistic U and not lose any information about θ. To illustrate, in Example 9.2, suppose that we have two experimenters:

- Experimenter 1 keeps $Y_1, Y_2, ..., Y_n$; i.e., s/he keeps all the data
- Experimenter 2 records $Y_1, Y_2, ..., Y_n$, but only keeps $U = \sum_{i=1}^{n} Y_i$; i.e., s/he keeps the sum, but forgets the original values of $Y_1, Y_2, ..., Y_n$.

RESULT: If both experimenters wanted to estimate θ, Experimenter 2 has just as much information with U as Experimenter 1 does with the entire sample of data!

9.2.1 The likelihood function

BACKGROUND: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y; \theta)$. After we observe the data $y_1, y_2, ..., y_n$; i.e., the realizations of $Y_1, Y_2, ..., Y_n$, we can think of the function

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \theta) = \prod_{i=1}^{n} f_Y(y_i; \theta)$$

in two different ways:

(1) as the multivariate probability density/mass function of $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$, for a fixed (but unknown) value of θ, or

(2) as a function of θ, given the observed data $\boldsymbol{y} = (y_1, y_2, ..., y_n)$.

In (1), we write

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \theta) = \prod_{i=1}^{n} f_Y(y_i; \theta).$$

In (2), we write

$$L(\theta|\boldsymbol{y}) = L(\theta|y_1, y_2, ..., y_n) = \prod_{i=1}^{n} f_Y(y_i; \theta).$$

Table 9.5: *Number of stoplights until the first stop is required. These observations are modeled as $n = 10$ realizations from geometric distribution with parameter θ.*

4	3	1	3	6	5	4	2	7	1

REALIZATION: The two functions $f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)$ and $L(\theta|\boldsymbol{y})$ are the same function! The only difference is in the **interpretation** of it. In (1), we fix the parameter θ and think of $f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)$ as a multivariate function of $\boldsymbol{y}$. In (2), we fix the data $\boldsymbol{y}$ and think of $L(\theta|\boldsymbol{y})$ as a function of the parameter θ.

TERMINOLOGY: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y;\theta)$ and that $y_1, y_2, ..., y_n$ are the n observed values. The **likelihood function** for θ is given by

$$L(\theta|\boldsymbol{y}) \equiv L(\theta|y_1, y_2, ..., y_n) = \prod_{i=1}^{n} f_Y(y_i;\theta).$$

Example 9.3. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of geometric random variables with parameter $0 < \theta < 1$; i.e., Y_i counts the number of Bernoulli trials until the 1st success is observed. Recall that the geometric(θ) pmf is given by

$$f_Y(y;\theta) = \begin{cases} \theta(1-\theta)^{y-1}, & y = 1, 2, ..., \\ 0, & \text{otherwise.} \end{cases}$$

The likelihood function for θ, given the data $\boldsymbol{y} = (y_1, y_2, ..., y_n)$, is

$$L(\theta|\boldsymbol{y}) = \prod_{i=1}^{n} f_Y(y_i;\theta) = \prod_{i=1}^{n} \theta(1-\theta)^{y_i-1} = \theta^n(1-\theta)^{\sum_{i=1}^{n} y_i - n}.$$

Using the data from Table 9.5, we have $n = 10$ and $\sum_{i=1}^{10} y_i = 36$. Thus, the likelihood function $L(\theta|\boldsymbol{y})$, for $0 < \theta < 1$, is given by

$$\begin{aligned} L(\theta|\boldsymbol{y}) = L(\theta|y_1, y_2, ..., y_{10}) &= \theta^{10}(1-\theta)^{36-10} \\ &= \theta^{10}(1-\theta)^{26}. \end{aligned}$$

This likelihood function is plotted in Figure 9.7. In a sense, the likelihood function describes which values of θ are more consistent with the observed data $\boldsymbol{y}$. Which values of θ are more consistent with the data in Example 9.3?

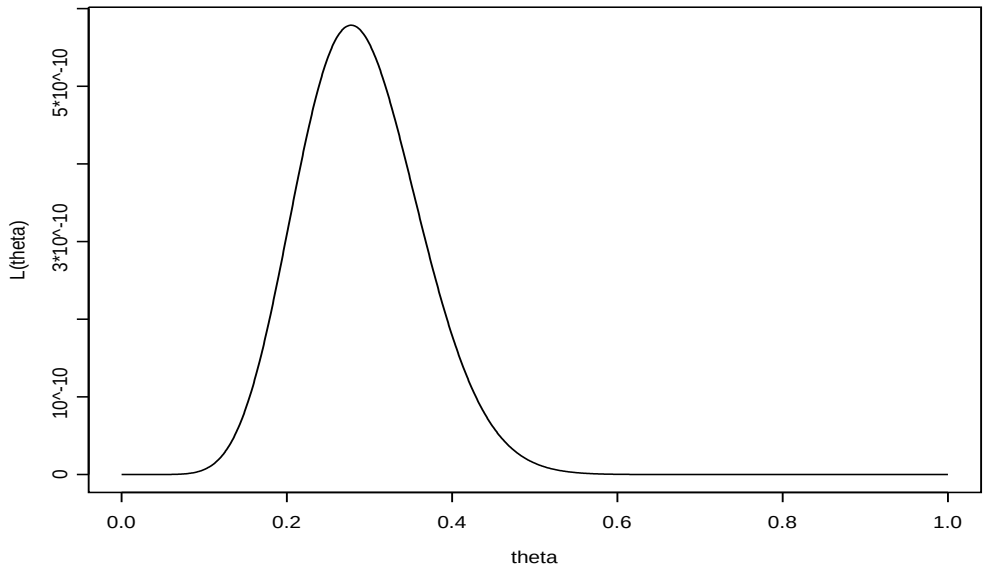

Figure 9.7: *Likelihood function* $L(\theta|\boldsymbol{y})$ *in Example* 9.3.

9.2.2 Factorization Theorem

RECALL: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y;\theta)$. We have already learned how to directly show that a statistic U is sufficient for θ; namely, we can show that the conditional distribution of the data $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$, given U, does not depend on θ. It turns out that there is an easier way to show that a statistic U is sufficient for θ.

FACTORIZATION THEOREM: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y;\theta)$ and that U is a statistic. If the likelihood function for θ, $L(\theta|\boldsymbol{y})$, can be expressed as the product of two nonnegative functions $g(u,\theta)$ and $h(y_1, y_2, ..., y_n)$, where

- $g(u,\theta)$ is only a function of u and θ, and
- $h(y_1, y_2, ..., y_n)$ is only a function of $y_1, y_2, ..., y_n$,

then U is a **sufficient statistic** for θ.

REMARK: The Factorization Theorem makes getting sufficient statistics easy! All we have to do is be able to write the likelihood function

$$L(\theta|\boldsymbol{y}) = g(u,\theta) \times h(y_1, y_2, ..., y_n)$$

for nonnegative functions g and h. Now that we have the Factorization Theorem, there will rarely be a need to work directly with the conditional distribution $f_{\boldsymbol{Y}|U}(\boldsymbol{y}|u)$; i.e., to establish sufficiency using the definition.

Example 9.4. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of Poisson observations with mean θ. Our goal is to show that $U = \sum_{i=1}^n Y_i$ is a sufficient statistic for θ using the Factorization Theorem. You'll recall that in Example 9.2, we showed that $U = \sum_{i=1}^n Y_i$ is sufficient by appealing to the definition of sufficiency directly. The likelihood function for θ is given by

$$\begin{aligned} L(\theta|\boldsymbol{y}) = \prod_{i=1}^n f_Y(y_i;\theta) &= \prod_{i=1}^n \frac{\theta^{y_i} e^{-\theta}}{y_i!} \\ &= \frac{\theta^{\sum_{i=1}^n y_i} e^{-n\theta}}{\prod_{i=1}^n y_i!} \\ &= \underbrace{\theta^{\sum_{i=1}^n y_i} e^{-n\theta}}_{g(u,\theta)} \times \underbrace{\left(\prod_{i=1}^n y_i!\right)^{-1}}_{h(y_1,y_2,...,y_n)}. \end{aligned}$$

Both $g(u,\theta)$ and $h(y_1, y_2, ..., y_n)$ are nonnegative functions. Thus, by the Factorization Theorem, $U = \sum_{i=1}^n Y_i$ is a sufficient statistic for θ. $\square$

Example 9.5. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of $\mathcal{N}(0, \sigma^2)$ observations. The likelihood function for σ^2 is given by

$$\begin{aligned} L(\sigma^2|\boldsymbol{y}) = \prod_{i=1}^n f_Y(y_i;\sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-y_i^2/2\sigma^2} \\ &= \underbrace{\left(\frac{1}{\sqrt{2\pi}}\right)^n}_{h(y_1,y_2,...,y_n)} \times \underbrace{\left(\sigma^{-n} e^{-\sum_{i=1}^n y_i^2/2\sigma^2}\right)}_{g(u,\sigma^2)}. \end{aligned}$$

Both $g(u,\sigma^2)$ and $h(y_1, y_2, ..., y_n)$ are nonnegative functions. Thus, by the Factorization Theorem, $U = \sum_{i=1}^n Y_i^2$ is a sufficient statistic for σ^2. $\square$

Example 9.6. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a beta$(1, \theta)$ distribution. The likelihood function for θ is given by

$$\begin{aligned} L(\theta|\boldsymbol{y}) = \prod_{i=1}^{n} f_Y(y_i;\theta) &= \prod_{i=1}^{n} \theta(1-y_i)^{\theta-1} \\ &= \theta^n \prod_{i=1}^{n} (1-y_i)^{\theta-1} \\ &= \underbrace{\theta^n \left[\prod_{i=1}^{n}(1-y_i)\right]^{\theta}}_{g(u,\theta)} \times \underbrace{\left[\prod_{i=1}^{n}(1-y_i)\right]^{-1}}_{h(y_1,y_2,...,y_n)}. \end{aligned}$$

Both $g(u, \theta)$ and $h(y_1, y_2, ..., y_n)$ are nonnegative functions. Thus, by the Factorization Theorem, $U = \prod_{i=1}^{n}(1 - Y_i)$ is a sufficient statistic for θ. $\square$

SOME NOTES ON SUFFICIENCY:

(1) The sample itself $\boldsymbol{Y} = (Y_1, Y_2, ..., Y_n)$ is always sufficient for θ, of course, but this provides no data reduction!

(2) The order statistics $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ are sufficient for θ.

(3) If g is a one-to-one function over the set of all possible values of θ and if U is a sufficient statistic, then $g(U)$ is also sufficient.

Example 9.7. In Example 9.4, we showed that $U = \sum_{i=1}^{n} Y_i$ is a sufficient statistic for θ, the mean of Poisson distribution. Thus,

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

is also a sufficient statistic for θ since $g(u) = u/n$ is a one-to-one function. In Example 9.6, we showed that $U = \prod_{i=1}^{n}(1-Y_i)$ is a sufficient statistic for θ in the beta$(1, \theta)$ family. Thus,

$$\log\left[\prod_{i=1}^{n}(1-Y_i)\right] = \sum_{i=1}^{n}\log(1-Y_i)$$

is also a sufficient statistic for θ since $g(u) = \log u$ is a one-to-one function. $\square$

MULTIDIMENSIONAL EXTENSION: As you might expect, we can generalize the Factorization Theorem to the case wherein θ is vector-valued. To emphasize this, we will write $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)$, a p-dimensional parameter. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y; \boldsymbol{\theta})$. The likelihood function for $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)$ is given by

$$L(\boldsymbol{\theta}|\boldsymbol{y}) \equiv L(\boldsymbol{\theta}|y_1, y_2, ..., y_n) = \prod_{i=1}^{n} f_Y(y_i; \boldsymbol{\theta}).$$

If one can express $L(\boldsymbol{\theta}|\boldsymbol{y})$ as

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = g(u_1, u_2, ..., u_p; \boldsymbol{\theta}) \times h(y_1, y_2, ..., y_n),$$

where g is a nonnegative function of $u_1, u_2, ..., u_p$ and $\boldsymbol{\theta}$ alone, and h is a nonnegative function of the data only, then we call $\boldsymbol{U} = (U_1, U_2, ..., U_p)$ a sufficient statistic for θ. In other words, $U_1, U_2, ..., U_p$ are p **jointly sufficient statistics** for $\theta_1, \theta_2, ..., \theta_p$.

Example 9.8. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of gamma(α, β) observations. We would like to find a $p = 2$ dimensional sufficient statistic for $\boldsymbol{\theta} = (\alpha, \beta)$. The likelihood function for $\boldsymbol{\theta}$ is given by

$$\begin{aligned} L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} f_Y(y_i; \alpha, \beta) &= \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^\alpha} y_i^{\alpha-1} e^{-y_i/\beta} \\ &= \left[\frac{1}{\Gamma(\alpha)\beta^\alpha}\right]^n \left(\prod_{i=1}^{n} y_i\right)^{\alpha-1} e^{-\sum_{i=1}^{n} y_i/\beta} \\ &= \underbrace{\left(\prod_{i=1}^{n} y_i\right)^{-1}}_{h(y_1, y_2, ..., y_n)} \times \underbrace{\left[\frac{1}{\Gamma(\alpha)\beta^\alpha}\right]^n \left(\prod_{i=1}^{n} y_i\right)^{\alpha} e^{-\sum_{i=1}^{n} y_i/\beta}}_{g(u_1, u_2; \alpha, \beta)}. \end{aligned}$$

Both $g(u_1, u_2; \alpha, \beta)$ and $h(y_1, y_2, ..., y_n)$ are nonnegative functions. Thus, by the Factorization Theorem, $\boldsymbol{U} = (\prod_{i=1}^{n} Y_i, \sum_{i=1}^{n} Y_i)$ is a sufficient statistic for $\boldsymbol{\theta} = (\alpha, \beta)$. $\square$

Example 9.9. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\mu, \sigma^2)$ sample. We can use the multidimensional Factorization Theorem to show $\boldsymbol{U} = (\sum_{i=1}^{n} Y_i, \sum_{i=1}^{n} Y_i^2)$ is a sufficient statistic for $\boldsymbol{\theta} = (\mu, \sigma^2)$. Because $\boldsymbol{U}^* = (\overline{Y}, S^2)$ is a one-to-one function of $\boldsymbol{U}$, it follows that $\boldsymbol{U}^*$ is also sufficient for $\boldsymbol{\theta} = (\mu, \sigma^2)$. $\square$

9.3 The Rao-Blackwell Theorem

PREVIEW: One of the main goals of this chapter is to find the **best** possible estimator for θ based on an iid sample $Y_1, Y_2, ..., Y_n$ from $f_Y(y;\theta)$. The Rao-Blackwell Theorem will help us see how to find a best estimator, provided that it exists.

RAO-BLACKWELL: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y;\theta)$, and let $\widehat{\theta}$ be an **unbiased estimator** of θ; i.e.,

$$E(\widehat{\theta}) = \theta.$$

In addition, suppose that U is a sufficient statistic for θ, and define

$$\widehat{\theta}^* = E(\widehat{\theta}|U),$$

the conditional expectation of $\widehat{\theta}$ given U (which we know, most importantly, is a function of U). Then, for all θ, $E(\widehat{\theta}^*) = \theta$ and $V(\widehat{\theta}^*) \leq V(\widehat{\theta})$.

Proof. That $E(\widehat{\theta}^*) = \theta$ (i.e., that $\widehat{\theta}^*$ is unbiased) follows from the iterated law for expectation (see Section 5.11 WMS):

$$E(\widehat{\theta}^*) = E\big[E(\widehat{\theta}|U)\big] = E(\widehat{\theta}) = \theta.$$

That $V(\widehat{\theta}^*) \leq V(\widehat{\theta})$ follows from Adam's Rule (i.e., the iterated law for variances; see Section 5.11 WMS):

$$V(\widehat{\theta}) = E\big[V(\widehat{\theta}|U)\big] + V\big[E(\widehat{\theta}|U)\big] = E\big[V(\widehat{\theta}|U)\big] + V(\widehat{\theta}^*).$$

Since $V(\widehat{\theta}|U) \geq 0$, this implies that $E\big[V(\widehat{\theta}|U)\big] \geq 0$ as well. Thus, $V(\widehat{\theta}) \geq V(\widehat{\theta}^*)$, and the result follows. □

INTERPRETATION: What does the Rao-Blackwell Theorem tell us? To use the result, some students think that they have to find $\widehat{\theta}$, an unbiased estimator for θ, obtain the conditional distribution of $\widehat{\theta}$ given U, and then compute the mean of this conditional distribution. This is not the case at all! *The Rao-Blackwell Theorem simply convinces us that in our search for the best possible estimator for θ, we can restrict our search to those*

estimators that are functions of sufficient statistics. That is, best estimators, provided they exist, will always be functions of sufficient statistics.

TERMINOLOGY: The **minimum-variance unbiased estimator** (MVUE) for θ is the best estimator for θ. The two conditions for an estimator $\widehat{\theta}$ to be MVUE are that

- the estimator $\widehat{\theta}$ is **unbiased**; i.e., $E(\widehat{\theta}) = \theta$,
- among all unbiased estimators of θ, $\widehat{\theta}$ has the **smallest** possible variance.

REMARK: If an MVUE exists (in some problems it may not), it is **unique**. The proof of this claim is slightly beyond the scope of this course. In practice, how do we find the MVUE for θ, or the MVUE for $\tau(\theta)$, a function of θ?

STRATEGY FOR FINDING MVUE's: The Rao-Blackwell Theorem says that best estimators are always functions of the sufficient statistic U. Thus, **first find a sufficient statistic** U (this is the starting point).

- Then, find a function of U that is unbiased for the parameter θ. This function of U is the MVUE for θ.
- If we need to find the MVUE for a function of θ, say, $\tau(\theta)$, then find a function of U that unbiased for $\tau(\theta)$; this function will then the MVUE for $\tau(\theta)$.

MATHEMATICAL ASIDE: You should know that this strategy works often (it will work for the examples we consider in this course). However, there are certain situations where this approach fails. The reason that it can fail is that the sufficient statistic U may not be **complete**. The concept of completeness is slightly beyond the scope of this course too, but, nonetheless, it is very important when finding MVUE's. This is not an issue we will discuss again, but you should be aware that in higher-level discussions (say, in a graduate-level theory course), this would be an issue. For us, we will only consider examples where completeness is guaranteed. Thus, we can adopt the strategy above for finding best estimators (i.e., MVUEs).

Example 9.10. Suppose that $Y_1, Y_2, ..., Y_n$ are iid Poisson observations with mean θ. We have already shown (in Examples 9.2 and 9.4) that $U = \sum_{i=1}^n Y_i$ is a sufficient statistic for θ. Thus, Rao-Blackwell says that the MVUE for θ is a function of U. Consider

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i,$$

the sample mean. Clearly, $\overline{Y}$ is a function of the sufficient statistic U. Furthermore, we know that $E(\overline{Y}) = \theta$. Since $\overline{Y}$ is unbiased and is a function of the sufficient statistic, it must be the MVUE for θ. $\square$

Example 9.11. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of $\mathcal{N}(0, \sigma^2)$ observations. From Example 9.5, we know that $U = \sum_{i=1}^n Y_i^2$ is a sufficient statistic for σ^2. Thus, Rao-Blackwell says that the MVUE for σ^2 is a function of U. Let's first compute $E(U)$:

$$E(U) = E\left(\sum_{i=1}^n Y_i^2\right) = \sum_{i=1}^n E(Y_i^2).$$

Now, for each i,

$$E(Y_i^2) = V(Y_i) + [E(Y_i)]^2 = \sigma^2 + 0^2 = \sigma^2.$$

Thus,

$$E(U) = \sum_{i=1}^n E(Y_i^2) = \sum_{i=1}^n \sigma^2 = n\sigma^2$$

which implies that

$$E\left(\frac{U}{n}\right) = E\left(\frac{1}{n}\sum_{i=1}^n Y_i^2\right) = \sigma^2.$$

Since $\frac{1}{n}\sum_{i=1}^n Y_i^2$ is a function of the sufficient statistic U, and is unbiased, it must be the MVUE for σ^2. $\square$

Example 9.12. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of exponential observations with mean θ and that the goal is to find the MVUE for $\tau(\theta) = \theta^2$, the population variance. We start by finding U, a sufficient statistic. The likelihood function for θ is given by

$$\begin{aligned} L(\theta|\boldsymbol{y}) = \prod_{i=1}^n f_Y(y_i;\theta) &= \prod_{i=1}^n \left(\frac{1}{\theta}\right) e^{-y_i/\theta} \\ &= \underbrace{\frac{1}{\theta^n} e^{-\sum_{i=1}^n y_i/\theta}}_{g(u,\theta)} \times h(\boldsymbol{y}), \end{aligned}$$

where $h(\boldsymbol{y}) = 1$. Thus, by the Factorization Theorem, $U = \sum_{i=1}^{n} Y_i$ is a sufficient statistic for θ. Now, to estimate $\tau(\theta) = \theta^2$, consider the "candidate estimator" $\overline{Y}^2$ (clearly, $\overline{Y}^2$ is a function of U). It follows that

$$E(\overline{Y}^2) = V(\overline{Y}) + [E(\overline{Y})]^2 = \frac{\theta^2}{n} + \theta^2 = \left(\frac{n+1}{n}\right)\theta^2.$$

Thus,

$$E\left(\frac{n\overline{Y}^2}{n+1}\right) = \left(\frac{n}{n+1}\right)E(\overline{Y}^2) = \left(\frac{n}{n+1}\right)\left(\frac{n+1}{n}\right)\theta^2 = \theta^2.$$

Since $n\overline{Y}^2/(n+1)$ is unbiased for $\tau(\theta) = \theta^2$ and is a function of the sufficient statistic U, it must be the MVUE for $\tau(\theta) = \theta^2$. □

Example 9.13. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of $\mathcal{N}(\mu, \sigma^2)$ observations. Try to prove each of these results:

- If σ^2 is known, then $\overline{Y}$ is MVUE for μ.
- If μ is known, then
$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)^2$$
is MVUE for σ^2.
- If both μ and σ^2 are unknown, then $(\overline{Y}, S^2)$ is MVUE for $\boldsymbol{\theta} = (\mu, \sigma^2)$. □

SUMMARY: Sufficient statistics are very good statistics to deal with because they contain all the information in the sample. Best (point) estimators are always functions of sufficient statistics. Not surprisingly, the best confidence intervals and hypothesis tests (STAT 513) almost always depend on sufficient statistics too. Statistical procedures which are not based on sufficient statistics usually are not the best available procedures.

PREVIEW: We now turn our attention to studying two additional techniques which provide point estimators:

- method of moments
- method of maximum likelihood.

9.4 Method of moments estimators

METHOD OF MOMENTS: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from $f_Y(y; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a p-dimensional parameter. The **method of moments** (MOM) approach to point estimation says to equate population moments to sample moments and solve the resulting system for all unknown parameters. To be specific, define the kth **population moment** to be

$$\mu_k' = E(Y^k),$$

and the kth **sample moment** to be

$$m_k' = \frac{1}{n}\sum_{i=1}^{n} Y_i^k.$$

Let p denote the number of parameters to be estimated; i.e., p equals the dimension of $\boldsymbol{\theta}$. The method of moments (MOM) procedure uses the following system of p equations and p unknowns:

$$\begin{aligned} \mu_1' &= m_1' \\ \mu_2' &= m_2' \\ &\vdots \\ \mu_p' &= m_p'. \end{aligned}$$

Estimators are obtained by solving the system for $\theta_1, \theta_2, ..., \theta_p$ (the population moments $\mu_1', \mu_2', ..., \mu_p'$ will almost always be functions of $\boldsymbol{\theta}$). The resulting estimators are called **method of moments estimators**. If θ is a scalar (i.e., $p = 1$), then we only need one equation. If $p = 2$, we will need 2 equations, and so on.

Example 9.14. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of $\mathcal{U}(0, \theta)$ observations. Find the MOM estimator for θ.

SOLUTION. The first population moment is $\mu_1' = \mu = E(Y) = \theta/2$, the population mean. The first sample moment is

$$m_1' = \frac{1}{n}\sum_{i=1}^{n} Y_i = \overline{Y},$$

the sample mean. To find the MOM estimator of θ, we simply set

$$\mu_1' = \frac{\theta}{2} \stackrel{\text{set}}{=} \overline{Y} = m_1'$$

and solve for θ. The MOM estimator for θ is $\widehat{\theta} = 2\overline{Y}$. $\square$

Example 9.15. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of gamma(α, β) observations. Here, there are $p = 2$ unknown parameters. The first two population moments are

$$\begin{aligned} \mu_1' &= E(Y) = \alpha\beta \\ \mu_2' &= E(Y^2) = V(Y) + [E(Y)]^2 = \alpha\beta^2 + (\alpha\beta)^2. \end{aligned}$$

Our 2×2 system becomes

$$\begin{aligned} \alpha\beta &\stackrel{\text{set}}{=} \overline{Y} \\ \alpha\beta^2 + (\alpha\beta)^2 &\stackrel{\text{set}}{=} m_2', \end{aligned}$$

where $m_2' = \frac{1}{n}\sum_{i=1}^n Y_i^2$. Substituting the first equation into the second, we get

$$\alpha\beta^2 = m_2' - \overline{Y}^2.$$

Solving for β in the first equation, we get $\beta = \overline{Y}/\alpha$; substituting this into the last equation, we get

$$\widehat{\alpha} = \frac{\overline{Y}^2}{m_2' - \overline{Y}^2}.$$

Substituting $\widehat{\alpha}$ into the original system (the first equation), we get

$$\widehat{\beta} = \frac{m_2' - \overline{Y}^2}{\overline{Y}}.$$

These are the MOM estimators of α and β, respectively. From Example 9.8 (notes), we can see that $\widehat{\alpha}$ and $\widehat{\beta}$ are not functions of the sufficient statistic $\boldsymbol{U} = (\prod_{i=1}^n Y_i, \sum_{i=1}^n Y_i)$; i.e., if you knew the value of $\boldsymbol{U}$, you could not compute $\widehat{\alpha}$ and $\widehat{\beta}$. From Rao-Blackwell, we know that the MOM estimators are not the best available estimators of α and β. $\square$

REMARK: The method of moments approach is one of the oldest methods of finding estimators. It is a "quick and dirty" approach (we are simply equating sample and population moments); however, it is sometimes a good place to start. Method of moments estimators are usually not functions of sufficient statistics, as we have just seen.

9.5 Maximum likelihood estimation

INTRODUCTION: The method of maximum likelihood is, by far, the most popular technique for estimating parameters in practice. The method is intuitive; namely, we estimate θ with $\widehat{\theta}$, the value that **maximizes** the likelihood function $L(\theta|\boldsymbol{y})$. Loosely speaking, $L(\theta|\boldsymbol{y})$ can be thought of as "the probability of the data," (in the discrete case, this makes sense; in the continuous case, this interpretation is somewhat awkward), so, we are choosing the value of θ that is "most likely" to have produced the data $y_1, y_2, ..., y_n$.

MAXIMUM LIKELIHOOD: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from the population distribution $f_Y(y;\theta)$. The **maximum likelihood estimator** (**MLE**) for θ, denoted $\widehat{\theta}$, is the value of θ that maximizes the likelihood function $L(\theta|\boldsymbol{y})$; that is,

$$\widehat{\theta} = \arg\max_{\theta} L(\theta|\boldsymbol{y}).$$

Example 9.16. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\theta, 1)$ sample. Find the MLE of θ.

SOLUTION. The likelihood function of θ is given by

$$\begin{aligned} L(\theta|\boldsymbol{y}) = \prod_{i=1}^{n} f_Y(y_i;\theta) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_i-\theta)^2} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^2}. \end{aligned}$$

Taking derivatives with respect to θ, we get

$$\frac{\partial}{\partial\theta} L(\theta|\boldsymbol{y}) = \underbrace{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(y_i-\theta)^2}}_{\text{this is always positive}} \times \sum_{i=1}^{n}(y_i - \theta).$$

The only value of θ that makes this derivative equal to 0 is $\overline{y}$; this is true since

$$\sum_{i=1}^{n}(y_i - \theta) = 0 \Longleftrightarrow \theta = \overline{y}.$$

Furthermore, it is possible to show that

$$\left.\frac{\partial^2}{\partial\theta^2} L(\theta|\boldsymbol{y})\right|_{\theta=\overline{y}} < 0,$$

(verify!) showing us that, in fact, $\overline{y}$ maximizes $L(\theta|\boldsymbol{y})$. We have shown that $\widehat{\theta} = \overline{Y}$ is the maximum likelihood estimator (MLE) of θ. $\square$

MAXIMIZING TRICK: For all $x > 0$, the function $r(x) = \ln x$ is an increasing function. This follows since $r'(x) = 1/x > 0$ for $x > 0$. How is this helpful? When maximizing a likelihood function $L(\theta|\boldsymbol{y})$, we will often be able to use differentiable calculus (i.e., find the first derivative, set it equal to zero, solve for θ, and verify the solution is a maximizer by verifying appropriate second order conditions). However, it will often be "friendlier" to work with $\ln L(\theta|\boldsymbol{y})$ instead of $L(\theta|\boldsymbol{y})$. Since the log function is increasing, $L(\theta|\boldsymbol{y})$ and $\ln L(\theta|\boldsymbol{y})$ are maximized at the same value of θ; that is,

$$\widehat{\theta} = \arg\max_{\theta} L(\theta|\boldsymbol{y}) = \arg\max_{\theta} \ln L(\theta|\boldsymbol{y}).$$

So, without loss, we can work with $\ln L(\theta|\boldsymbol{y})$ instead if it simplifies the calculus.

Example 9.17. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of Poisson observations with mean θ. Find the MLE of θ.

SOLUTION. The likelihood function of θ is given by

$$\begin{aligned} L(\theta|\boldsymbol{y}) = \prod_{i=1}^{n} f_Y(y_i;\theta) &= \prod_{i=1}^{n} \frac{\theta^{y_i} e^{-\theta}}{y_i!} \\ &= \frac{\theta^{\sum_{i=1}^n y_i} e^{-n\theta}}{\prod_{i=1}^n y_i!}. \end{aligned}$$

This function is difficult to maximize analytically. It is much easier to work with the log-likelihood function; i.e.,

$$\ln L(\theta|\boldsymbol{y}) = \sum_{i=1}^{n} y_i \ln\theta - n\theta - \ln\left(\prod_{i=1}^{n} y_i!\right).$$

Its derivative is equal to

$$\frac{\partial}{\partial\theta} \ln L(\theta|\boldsymbol{y}) = \frac{\sum_{i=1}^n y_i}{\theta} - n \stackrel{\text{set}}{=} 0.$$

Setting this derivative equal to 0 and solving for θ, we get

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y_i = \overline{y}.$$

REMINDER: Whenever we derive an MLE, we should always check the appropriate second-order conditions to verify that our solution is, indeed, a **maximum**, and not a minimum. It suffices to calculate the second derivative of $\ln L(\theta|\boldsymbol{y})$ and show that

$$\left.\frac{\partial^2}{\partial\theta^2} \ln L(\theta|\boldsymbol{y})\right|_{\theta=\widehat{\theta}} < 0.$$

In this example, it is easy to show that

$$\frac{\partial^2}{\partial\theta^2}\ln L(\theta|\boldsymbol{y})\bigg|_{\theta=\overline{y}} = -\frac{\sum_{i=1}^n y_i}{\overline{y}^2} = -\frac{n}{\overline{y}} < 0.$$

Thus, we know that $\overline{y}$ is, indeed, a maximizer (as opposed to being a minimizer). We have shown that $\widehat{\theta} = \overline{Y}$ is the maximum likelihood estimator (MLE) of θ. $\square$

Example 9.18. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a gamma distribution with parameters $\alpha = 2$ and $\beta = \theta$; i.e., the pdf of Y is given by

$$f_Y(y;\theta) = \begin{cases} \frac{1}{\theta^2}ye^{-y/\theta}, & y > 0 \\ 0, & \text{otherwise.} \end{cases}$$

Find the MLE of θ.

SOLUTION. The likelihood function for θ is given by

$$\begin{aligned} L(\theta|\boldsymbol{y}) = \prod_{i=1}^n f_Y(y_i;\theta) &= \prod_{i=1}^n \frac{1}{\theta^2}y_i e^{-y_i/\theta} \\ &= \left(\frac{1}{\theta^2}\right)^n \left(\prod_{i=1}^n y_i\right) e^{-\sum_{i=1}^n y_i/\theta}. \end{aligned}$$

This function is very difficult to maximize analytically. It is much easier to work with the log-likelihood function; i.e.,

$$\ln L(\theta|\boldsymbol{y}) = -2n\ln\theta + \ln\left(\prod_{i=1}^n y_i\right) - \frac{\sum_{i=1}^n y_i}{\theta}.$$

Its derivative is equal to

$$\frac{\partial}{\partial\theta}\ln L(\theta|\boldsymbol{y}) = \frac{-2n}{\theta} + \frac{\sum_{i=1}^n y_i}{\theta^2} \stackrel{\text{set}}{=} 0 \Longrightarrow -2n\theta + \sum_{i=1}^n y_i = 0.$$

Solving this equation gives

$$\widehat{\theta} = \frac{1}{2n}\sum_{i=1}^n y_i = \overline{y}/2.$$

Because

$$\frac{\partial^2}{\partial\theta^2}\ln L(\theta|\boldsymbol{y})\bigg|_{\theta=\widehat{\theta}} < 0,$$

(verify!) it follows that $\widehat{\theta} = \overline{Y}/2$ is the maximum likelihood estimator (MLE) of θ. $\square$

Example 9.19. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a $\mathcal{U}(0, \theta)$ distribution. The likelihood function for θ is given by

$$L(\theta|\boldsymbol{y}) = \prod_{i=1}^{n} f_Y(y_i; \theta) = \prod_{i=1}^{n} \frac{1}{\theta} = \frac{1}{\theta^n},$$

for $0 < y_i < \theta$, and 0, otherwise. In this example, we can not differentiate the likelihood (or the log-likelihood) because the derivative will never be zero. We have to obtain the MLE in another way. Note that $L(\theta|\boldsymbol{y})$ is a **decreasing** function of θ, since

$$\frac{\partial}{\partial\theta} L(\theta|\boldsymbol{y}) = -n/\theta^{n+1} < 0,$$

for $\theta > 0$. Furthermore, we know that if any y_i value exceeds θ, the likelihood function is equal to zero, since the value of $f_Y(y_i; \theta)$ for that particular y_i would be zero. So, we have a likelihood function that is decreasing, but is only nonzero as long as $\theta > y_{(n)}$, the largest order statistic. Thus, the likelihood function must attain its maximum value when $\theta = y_{(n)}$. This argument shows that $\widehat{\theta} = Y_{(n)}$ is the MLE of θ. $\square$

LINK WITH SUFFICIENCY: Are maximum likelihood estimators good estimators? It turns out that they are always functions of sufficient statistics. Suppose that U is a sufficient statistic for θ. We know by the Factorization Theorem that the likelihood function for θ can be written as

$$L(\theta|\boldsymbol{y}) = \prod_{i=1}^{n} f_Y(y_i; \theta) = g(u, \theta) \times h(\boldsymbol{y}),$$

for nonnegative functions g and h. Thus, when we maximize $L(\theta|\boldsymbol{y})$, or its logarithm, we see that the MLE will always depend on U through the g function.

PUNCHLINE: In our quest to find the MVUE for a parameter θ, we could simply (1) derive the MLE for θ and (2) try to find a function of the MLE that is unbiased. Since the MLE will always be a function of the sufficient statistic U, this unbiased function will be the MVUE for θ.

MULTIDIMENSIONAL SITUATION: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from the population distribution $f_Y(y; \boldsymbol{\theta})$, where the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)$. Conceptually, finding the MLE of $\boldsymbol{\theta}$ is the same as when θ is a scalar parameter; namely, we

still maximize the log-likelihood function $\ln L(\boldsymbol{\theta}|\boldsymbol{y})$. This can be done by solving

$$\begin{aligned}
\frac{\partial}{\partial\theta_1}\ln L(\boldsymbol{\theta}|\boldsymbol{y}) &= 0 \\
\frac{\partial}{\partial\theta_2}\ln L(\boldsymbol{\theta}|\boldsymbol{y}) &= 0 \\
&\vdots \\
\frac{\partial}{\partial\theta_p}\ln L(\boldsymbol{\theta}|\boldsymbol{y}) &= 0
\end{aligned}$$

jointly for $\theta_1, \theta_2, ..., \theta_p$. The solution to this system, say $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2, ..., \widehat{\theta}_p)$, is the maximum likelihood estimator of $\boldsymbol{\theta}$, provided that appropriate second-order conditions hold.

Example 9.20. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid $\mathcal{N}(\mu, \sigma^2)$ sample, where both parameters are unknown. Find the MLE of $\boldsymbol{\theta} = (\mu, \sigma^2)$.

SOLUTION. The likelihood function of $\boldsymbol{\theta} = (\mu, \sigma^2)$ is given by

$$\begin{aligned}
L(\boldsymbol{\theta}|\boldsymbol{y}) = L(\mu, \sigma^2|\boldsymbol{y}) = \prod_{i=1}^{n} f_Y(y_i; \mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu)^2} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}.
\end{aligned}$$

The log-likelihood function of $\boldsymbol{\theta} = (\mu, \sigma^2)$ is

$$\ln L(\mu, \sigma^2|\boldsymbol{y}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2,$$

and the two partial derivatives of $\ln L(\mu, \sigma^2|\boldsymbol{y})$ are

$$\begin{aligned}
\frac{\partial}{\partial\mu}\ln L(\mu, \sigma^2|\boldsymbol{y}) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\mu) \\
\frac{\partial}{\partial\sigma^2}\ln L(\mu, \sigma^2|\boldsymbol{y}) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\mu)^2.
\end{aligned}$$

Setting the first equation equal to zero and solving for μ we get $\widehat{\mu} = \overline{y}$. Plugging $\widehat{\mu} = \overline{y}$ into the second equation, we are then left to solve

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(y_i-\overline{y})^2 = 0$$

for σ^2; this solution is $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2$. One can argue that

$$\begin{aligned}
\widehat{\mu} &= \overline{y} \\
\widehat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(y_i-\overline{y})^2 \equiv s_b^2
\end{aligned}$$

Table 9.6: *Maximum* 24*-hour precipitation recorded for* 36 *inland hurricanes* (1900-1969).

Year	Location	Precip.	Year	Location	Precip.
1969	Tye River, VA	31.00	1932	Ceasars Head, SC	4.75
1968	Hickley, NY	2.82	1932	Rockhouse, NC	6.85
1965	Haywood Gap, NC	3.98	1929	Rockhouse, NC	6.25
1960	Cairo, NY	4.02	1928	Roanoke, VA	3.42
1959	Big Meadows, VA	9.50	1928	Ceasars Head, SC	11.80
1957	Russels Point, OH	4.50	1923	Mohonk Lake, NY	0.80
1955	Slide, Mt., NY	11.40	1923	Wappingers Falls, NY	3.69
1954	Big Meadows, VA	10.71	1920	Landrum, SC	3.10
1954	Eagles Mere, PA	6.31	1916	Altapass, NC	22.22
1952	Bloserville, PA	4.95	1916	Highlands, NC	7.43
1949	North Ford, NC	5.64	1915	Lookout Mt., TN	5.00
1945	Crossnore, NC	5.51	1915	Highlands, NC	4.58
1942	Big Meadows, VA	13.40	1912	Norcross, GA	4.46
1940	Rodhiss Dam, NC	9.72	1906	Horse Cove, NC	8.00
1939	Ceasars Head, SC	6.47	1902	Sewanee, TN	3.73
1938	Hubbardston, MA	10.16	1901	Linville, NC	3.50
1934	Balcony Falls, VA	4.21	1900	Marrobone, KY	6.20
1933	Peekamoose, NY	11.60	1900	St. Johnsbury, VT	0.67

is indeed a maximizer (although I'll omit the second order details). This argument shows that $\widehat{\boldsymbol{\theta}} = (\overline{Y}, S_b^2)$ is the MLE of $\boldsymbol{\theta} = (\mu, \sigma^2)$. $\square$

REMARK: In some problems, the likelihood function (or log-likelihood function) can not be maximized analytically because its derivative(s) does/do not exist in closed form. In such situations (which are common in real life), maximum likelihood estimators must be computed numerically.

Example 9.21. The U.S. Weather Bureau confirms that during 1900-1969, a total of 36 hurricanes moved as far inland as the Appalachian Mountains. The data in Table 9.6 are the 24-hour precipitation levels (in inches) recorded for those 36 storms during the time they were over the mountains. Suppose that we decide to model these data as iid gamma(α, β) realizations. The likelihood function for $\boldsymbol{\theta} = (\alpha, \beta)$ is given by

$$L(\alpha, \beta|\boldsymbol{y}) = \prod_{i=1}^{36} \frac{1}{\Gamma(\alpha)\beta^\alpha} y_i^{\alpha-1} e^{-y_i/\beta} = \left[\frac{1}{\Gamma(\alpha)\beta^\alpha}\right]^{36} \left(\prod_{i=1}^{36} y_i\right)^{\alpha-1} e^{-\sum_{i=1}^{36} y_i/\beta}.$$

The log-likelihood function is given by

$$\ln L(\alpha, \beta|\boldsymbol{y}) = -36 \ln \Gamma(\alpha) - 36\alpha \ln \beta + (\alpha - 1) \sum_{i=1}^{36} \ln y_i - \frac{\sum_{i=1}^{36} y_i}{\beta}.$$

This log-likelihood can not be maximized analytically; the gamma function $\Gamma(\cdot)$ messes things up. However, we can maximize $\ln L(\alpha, \beta|\boldsymbol{y})$ numerically using R.

```
############################################################
## Name: Joshua M. Tebbs
## Date: 7 Apr 2007
## Purpose: Fit gamma model to hurricane data
############################################################

# Enter data
y<-c(31,2.82,3.98,4.02,9.5,4.5,11.4,10.71,6.31,4.95,5.64,5.51,13.4,9.72,
6.47,10.16,4.21,11.6,4.75,6.85,6.25,3.42,11.8,0.8,3.69,3.1,22.22,7.43,5,
4.58,4.46,8,3.73,3.5,6.2,0.67)

## Second sample (uncentred) moment; needed for MOM
m2<-(1/36)*sum(y**2)

# MOM estimates (see Example 9.15 notes)
alpha.mom<-(mean(y))**2/(m2-(mean(y))**2)
beta.mom<-(m2-(mean(y))**2)/mean(y)

# Sufficient statistics
t1<-sum(log(y))
t2<-sum(y)

# Negative loglikelihood function (to be minimised)
# x1 = alpha
# x2 = beta
loglike<-function(x){
    x1<-x[1]
    x2<-x[2]
    36*log(gamma(x1))+36*x1*log(x2)-t1*(x1-1)+t2/x2
    }

# Use "optim" function to maximise the loglikelihood function
mle<-optim(par=c(alpha.mom,beta.mom),fn=loglike)

# look at the qq-plot to assess the fit of the gamma model
plot(qgamma(ppoints(y),mle$par[1],1/mle$par[2]),sort(y),pch=16,
```

```
    xlab="gamma percentiles",ylab="observed values")
```

Here is the output from running the program:

```
> alpha.mom
[1] 1.635001
> beta.mom
[1] 4.457183
> mle

$par
[1] 2.186535 3.332531

$value
[1] 102.3594
```

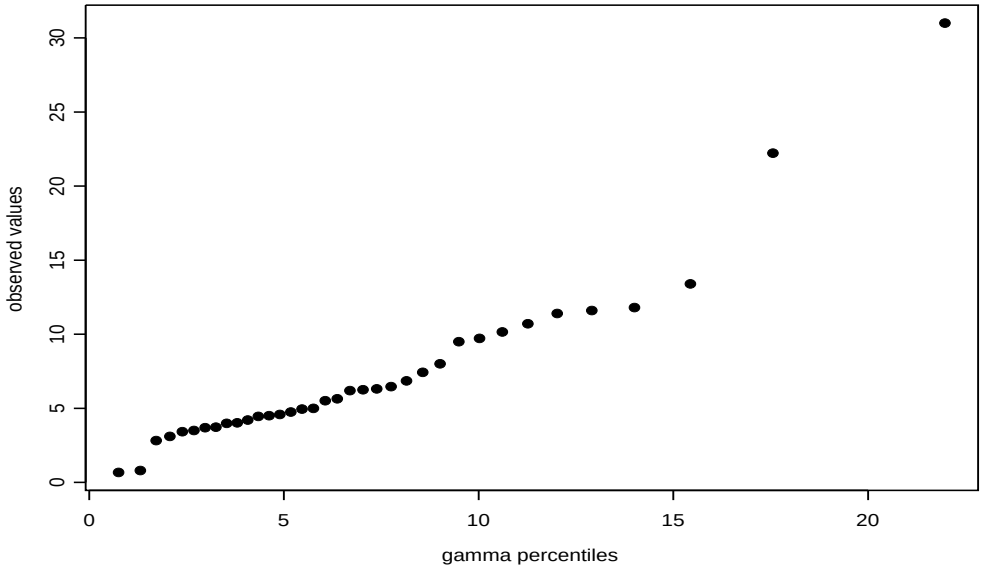

Figure 9.8: *Gamma qq-plot for the hurricane data in Example* 9.21.

ANALYSIS: First, note the difference in the MOM and the maximum likelihood estimates for these data. Which estimates would you rather report? Also, the two-parameter gamma distribution is not a bad model for these data; note that the qq-plot is somewhat linear (although there are two obvious outliers on each side). □

INVARIANCE: Suppose that $\widehat{\theta}$ is the MLE of θ, and let g be any real function, possibly vector-valued. Then $g(\widehat{\theta})$ is the MLE of $g(\theta)$.

Example 9.22. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of Poisson observations with mean $\theta > 0$. In Example 9.17, we showed that the MLE of θ is $\widehat{\theta} = \overline{Y}$. The invariance property of maximum likelihood estimators says, for example, that

- $\overline{Y}^2$ is the MLE for θ^2
- $\sin \overline{Y}$ is the MLE for $\sin \theta$
- $e^{-\overline{Y}}$ is the MLE for $e^{-\theta}$.

9.6 Asymptotic properties of point estimators

IMPORTANCE: In many problems, exact (i.e., finite-sample) distributional results are not available. In the absence of exact calculations, or when finite sample results are intractable, one may be able to obtain approximate results by using **large-sample theory**. Statistical methods based on large-sample theory are pervasive in research and practice. To emphasize a point estimator's dependence on the sample size n, we often write $\widehat{\theta} = \widehat{\theta}_n$. This is common notation when discussing asymptotic results.

9.6.1 Consistency and the Weak Law of Large Numbers

TERMINOLOGY: An estimator $\widehat{\theta}_n$ is said to be a **consistent** estimator of θ if, for all $\epsilon > 0$,

$$\lim_{n\to\infty} P(|\widehat{\theta}_n - \theta| > \epsilon) = 0;$$

that is, the sequence of real numbers $P(|\widehat{\theta}_n - \theta| > \epsilon) \to 0$, as $n \to \infty$. Consistency is a desirable large-sample property. If $\widehat{\theta}_n$ is consistent, then the probability that the estimator $\widehat{\theta}_n$ differs from the true θ becomes small as the sample size n increases. On the other hand, if you have an estimator that is not consistent, then no matter how many data you collect, the estimator $\widehat{\theta}_n$ may never "converge" to θ.

TERMINOLOGY: If an estimator $\widehat{\theta}_n$ is consistent, we say that $\widehat{\theta}_n$ **converges in probability** to θ and write $\widehat{\theta}_n \xrightarrow{p} \theta$.

Example 9.23. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a shifted-exponential distribution

$$f_Y(y;\theta) = \begin{cases} e^{-(y-\theta)}, & y > \theta \\ 0, & \text{otherwise.} \end{cases}$$

Show that the first order statistic $\widehat{\theta}_n = Y_{(1)}$ is a consistent estimator of θ.

SOLUTION. As you might suspect, we first have to find the pdf of $Y_{(1)}$. Recall from Chapter 6 (WMS) that

$$f_{Y_{(1)}}(y;\theta) = nf_Y(y;\theta)[1 - F_Y(y;\theta)]^{n-1}.$$

It is easy to show (verify!) that the cdf of Y is

$$F_Y(y;\theta) = \begin{cases} 0, & y \leq \theta \\ 1 - e^{-(y-\theta)}, & y > \theta. \end{cases}$$

Thus, the pdf of $Y_{(1)}$, for $y > \theta$, is

$$f_{Y_{(1)}}(y;\theta) = ne^{-(y-\theta)}\left\{1 - \left[1 - e^{-(y-\theta)}\right]\right\}^{n-1} = ne^{-n(y-\theta)}.$$

Using the definition of consistency, for $\epsilon > 0$, we have that

$$\begin{aligned} P(|Y_{(1)} - \theta| > \epsilon) &= \underbrace{P(Y_{(1)} < \theta - \epsilon)}_{=0} + P(Y_{(1)} > \theta + \epsilon) \\ &= \int_{\theta+\epsilon}^{\infty} ne^{-n(y-\theta)}dy \\ &= n\left[-\frac{1}{n}e^{-n(y-\theta)}\Big|_{\theta+\epsilon}^{\infty}\right] \\ &= e^{-n(y-\theta)}\Big|_{\infty}^{\theta+\epsilon} = e^{-n(\theta+\epsilon-\theta)} - 0 = e^{-n\epsilon} \to 0, \end{aligned}$$

as $n \to \infty$. Thus, $\widehat{\theta}_n = Y_{(1)}$ is a consistent estimator for θ. $\square$

RESULT: Suppose that $\widehat{\theta}_n$ is an estimator of θ. If both $B(\widehat{\theta}_n) \to 0$ and $V(\widehat{\theta}_n) \to 0$, as $n \to \infty$, then $\widehat{\theta}_n$ is a **consistent** estimator for θ. In many problems, it will be

much easier to show that $B(\widehat{\theta}_n) \to 0$ and $V(\widehat{\theta}_n) \to 0$, as $n \to \infty$, rather than showing $P(|\widehat{\theta}_n - \theta| > \epsilon) \to 0$; i.e., appealing directly to the definition of consistency.

THE WEAK LAW OF LARGE NUMBERS: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a population with mean μ and variance $\sigma^2 < \infty$. Then, the sample mean $\overline{Y}_n$ is a consistent estimator for μ; that is, $\overline{Y}_n \xrightarrow{p} \mu$, as $n \to \infty$.

Proof. Clearly, $B(\overline{Y}_n) = 0$, since $\overline{Y}_n$ is an unbiased estimator of μ. Also, $V(\overline{Y}_n) = \sigma^2/n \to 0$, as $n \to \infty$. □

RESULT: Suppose that $\widehat{\theta}_n \xrightarrow{p} \theta$ and $\widehat{\theta}'_n \xrightarrow{p} \theta'$. Then,

(a) $\widehat{\theta}_n + \widehat{\theta}'_n \xrightarrow{p} \theta + \theta'$

(b) $\widehat{\theta}_n \widehat{\theta}'_n \xrightarrow{p} \theta\theta'$

(c) $\widehat{\theta}_n / \widehat{\theta}'_n \xrightarrow{p} \theta/\theta'$, for $\theta' \neq 0$

(d) $g(\widehat{\theta}_n) \xrightarrow{p} g(\theta)$, for any continuous function g.

NOTE: We will omit the proofs of the above facts. Statements (a), (b), and (c) can be shown by appealing to the limits of sequences of real numbers. Proving statement (d) is somewhat more involved.

Example 9.24. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of gamma$(2, \theta)$ observations and that we want to find a consistent estimator for the scale parameter $\theta > 0$. From the Weak Law of Large Numbers (WLLN), we know that $\overline{Y}_n$ is a consistent estimator for $\mu = 2\theta$; i.e., $\overline{Y}_n \xrightarrow{p} 2\theta$. Since $g(s) = s/2$ is a continuous function, as $n \to \infty$,

$$\overline{Y}_n/2 = g(\overline{Y}_n) \xrightarrow{p} g(2\theta) = \theta.$$

That is, $\overline{Y}_n/2$ is consistent for θ. Furthermore,

$$\frac{\overline{Y}_n^2}{2} = 2\left(\frac{\overline{Y}_n}{2}\right)^2 \xrightarrow{p} 2\theta^2,$$

since $h(t) = 2t^2$ is a continuous function. Thus, $\overline{Y}_n^2/2$ is a consistent estimator of the population variance $\sigma^2 = 2\theta^2$. □

9.6.2 Slutsky's Theorem

SLUTSKY'S THEOREM: Suppose that U_n is a sequence of random variables that **converges in distribution** to a standard normal distribution; i.e., $U_n \xrightarrow{d} \mathcal{N}(0,1)$, as $n \to \infty$. In addition, suppose that $W_n \xrightarrow{p} 1$, as $n \to \infty$. Then, U_n/W_n converges to a standard normal distribution as well; that is, $U_n/W_n \xrightarrow{d} \mathcal{N}(0,1)$, as $n \to \infty$.

RECALL: When we say that "U_n converges in distribution to a $\mathcal{N}(0,1)$ distribution," we mean that the distribution function of U_n, $F_{U_n}(t)$, viewed as a sequence of real functions indexed by n, converges pointwise to the cdf of the $\mathcal{N}(0,1)$ distribution, for all t; i.e.,

$$F_{U_n}(t) \to \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy,$$

as $n \to \infty$, for all $-\infty < t < \infty$. Slutsky's Theorem says that, in the limit, U_n and U_n/W_n will have the same distribution.

Example 9.25. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from a population with mean μ and variance σ^2. Let S^2 denote the usual sample variance. By the CLT, we know that

$$U_n = \sqrt{n}\left(\frac{\overline{Y} - \mu}{\sigma}\right) \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$. From Example 9.3 (WMS) we know that $S^2 \xrightarrow{p} \sigma^2$ and $S^2/\sigma^2 \xrightarrow{p} 1$, as $n \to \infty$. Since $g(t) = \sqrt{t}$ is a continuous function, for $t > 0$,

$$W_n = g\left(\frac{S^2}{\sigma^2}\right) = \sqrt{\frac{S^2}{\sigma^2}} = \frac{S}{\sigma} \xrightarrow{p} g(1) = 1.$$

Finally, by Slutsky's Theorem,

$$\sqrt{n}\left(\frac{\overline{Y} - \mu}{S}\right) = \frac{\sqrt{n}\left(\frac{\overline{Y}-\mu}{\sigma}\right)}{S/\sigma} \xrightarrow{d} \mathcal{N}(0,1),$$

as $n \to \infty$. This result provides the theoretical justification as to why

$$\overline{Y} \pm z_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$$

serves as an approximate $100(1-\alpha)$ percent confidence interval for the population mean μ when the sample size is large.

REMARK: Slutsky's Theorem can also be used to explain why

$$\widehat{p} \pm z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

serves as an approximate $100(1-\alpha)$ percent confidence interval for a population proportion p when the sample size is large.

9.6.3 Large-sample properties of maximum likelihood estimators

REMARK: Another advantage of maximum likelihood estimators is that, under suitable "regularity conditions," they have very desirable large-sample properties. Succinctly put, maximum likelihood estimators are consistent and asymptotically normal.

IMPORTANT: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from the population distribution $f_Y(y;\theta)$ and that $\widehat{\theta}$ is the MLE for θ. It can be shown (under certain regularity conditions which we will omit) that

- $\widehat{\theta} \xrightarrow{p} \theta$, as $n \to \infty$; i.e., $\widehat{\theta}$ is a **consistent** estimator of θ
- $\widehat{\theta} \sim \mathcal{AN}(\theta, \sigma^2_{\widehat{\theta}})$, where

$$\sigma^2_{\widehat{\theta}} = \left\{ nE\left[-\frac{\partial^2}{\partial\theta^2} \ln f_Y(Y;\theta) \right] \right\}^{-1},$$

  for large n. That is, $\widehat{\theta}$ is **approximately normal** when the sample size is large. The quantity

$$\left\{ nE\left[-\frac{\partial^2}{\partial\theta^2} \ln f_Y(Y;\theta) \right] \right\}^{-1}$$

  is called the **Cramer-Rao Lower Bound**. This quantity has great theoretical importance in upper-level discussions on MLE theory.

LARGE-SAMPLE CONFIDENCE INTERVALS: To construct a large-sample confidence interval for θ, we need to be able to find a good large-sample estimator of $\sigma^2_{\widehat{\theta}}$. Define

$$\widehat{\sigma}^2_{\widehat{\theta}} = \left\{ nE\left[-\frac{\partial^2}{\partial\theta^2} \ln f_Y(Y;\theta) \right] \right\}^{-1} \Bigg|_{\theta=\widehat{\theta}}.$$

Since $\widehat{\theta} \xrightarrow{p} \theta$, it follows (by continuity) that

$$E\left[-\frac{\partial^2}{\partial\theta^2}\ln f_Y(Y;\theta)\right]\Bigg|_{\theta=\widehat{\theta}} \xrightarrow{p} E\left[-\frac{\partial^2}{\partial\theta^2}\ln f_Y(Y;\theta)\right]$$

so that $\sigma_{\widehat{\theta}}/\widehat{\sigma}_{\widehat{\theta}} \xrightarrow{p} 1$. Slutsky's Theorem allows us to conclude

$$Q_n = \frac{\widehat{\theta}-\theta}{\widehat{\sigma}_{\widehat{\theta}}} = \frac{\widehat{\theta}-\theta}{\sigma_{\widehat{\theta}}}\left(\frac{\sigma_{\widehat{\theta}}}{\widehat{\sigma}_{\widehat{\theta}}}\right) \xrightarrow{d} \mathcal{N}(0,1);$$

i.e., Q_n is asymptotically pivotal, so that

$$P\left(-z_{\alpha/2} < \frac{\widehat{\theta}-\theta}{\widehat{\sigma}_{\widehat{\theta}}} < z_{\alpha/2}\right) \approx 1-\alpha.$$

It follows that

$$\widehat{\theta} \pm z_{\alpha/2}\widehat{\sigma}_{\widehat{\theta}}$$

is an approximate $100(1-\alpha)$ percent confidence interval for θ.

Example 9.26. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of Poisson observations with mean $\theta > 0$. In Example 9.17, we showed that the MLE of θ is $\widehat{\theta} = \overline{Y}$. The natural logarithm of the Poisson(θ) mass function, for $y = 0, 1, 2, ...$, is

$$\ln f(y;\theta) = y\ln\theta - \theta - \ln y!.$$

The first and second derivatives of $\ln f(y;\theta)$ are, respectively,

$$\begin{aligned}\frac{\partial}{\partial\theta}\ln f(y;\theta) &= \frac{y}{\theta} - 1 \\ \frac{\partial^2}{\partial\theta^2}\ln f(y;\theta) &= -\frac{y}{\theta^2},\end{aligned}$$

so that

$$E\left[-\frac{\partial^2}{\partial\theta^2}\ln f(Y;\theta)\right] = E\left(\frac{Y}{\theta^2}\right) = \frac{1}{\theta}.$$

The Cramer-Rao Lower Bound is given by

$$\sigma^2_{\widehat{\theta}} = \left\{nE\left[-\frac{\partial^2}{\partial\theta^2}\ln f_Y(Y;\theta)\right]\right\}^{-1} = \frac{\theta}{n}.$$

From the asymptotic properties of maximum likelihood estimators, we know that

$$\overline{Y} \sim \mathcal{AN}\left(\theta, \frac{\theta}{n}\right).$$

To find an approximate confidence interval for θ, note that $\overline{Y} \xrightarrow{p} \theta$ and that the estimated large-sample variance of $\overline{Y}$ is

$$\widehat{\sigma}^2_{\widehat{\theta}} = \frac{\overline{Y}}{n}.$$

Thus, an approximate $100(1-\alpha)$ percent confidence interval for θ is given by

$$\overline{Y} \pm z_{\alpha/2}\sqrt{\frac{\overline{Y}}{n}}.$$

9.6.4 Delta Method

DELTA METHOD: Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample from the population distribution $f_Y(y;\theta)$. In addition, suppose that $\widehat{\theta}$ is the MLE of θ and let g be a real differentiable function. It can be shown (under certain regularity conditions) that, for large n,

$$g(\widehat{\theta}) \sim \mathcal{AN}\left\{g(\theta), [g'(\theta)]^2\sigma^2_{\widehat{\theta}}\right\},$$

where $g'(\theta) = \partial g(\theta)/\partial\theta$ and

$$\sigma^2_{\widehat{\theta}} = \left\{nE\left[-\frac{\partial^2}{\partial\theta^2}\ln f_Y(Y;\theta)\right]\right\}^{-1}.$$

The Delta Method is a useful asymptotic result. It enables us to state large-sample distributions of functions of maximum likelihood estimators.

LARGE-SAMPLE CONFIDENCE INTERVALS: The Delta Method makes getting large-sample confidence intervals for $g(\theta)$ easy. We know that, for n large,

$$\frac{g(\widehat{\theta}) - g(\theta)}{g'(\theta)\sigma_{\widehat{\theta}}} \sim \mathcal{AN}(0,1)$$

and that $g'(\theta)\sigma_{\widehat{\theta}}$ can be consistently estimated by $g'(\widehat{\theta})\widehat{\sigma}_{\widehat{\theta}}$. These two facts, along with Slutsky's Theorem, allow us to conclude that

$$g(\widehat{\theta}) \pm z_{\alpha/2}[g'(\widehat{\theta})\widehat{\sigma}_{\widehat{\theta}}]$$

is an approximate $100(1-\alpha)$ percent confidence interval for $g(\theta)$.

Example 9.27. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid Bernoulli(p) sample of observations, where $0 < p < 1$. A quantity often used in categorical data analysis is the function

$$g(p) = \ln\left(\frac{p}{1-p}\right),$$

which is the **log-odds**. The goal of this example is to derive an approximate $100(1-\alpha)$ percent confidence interval for $g(p)$.

SOLUTION. We first derive the MLE of p. The likelihood function for p is

$$L(p|\boldsymbol{y}) = \prod_{i=1}^{n} f_Y(y_i; p) = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i} = p^{\sum_{i=1}^n y_i}(1-p)^{n-\sum_{i=1}^n y_i},$$

and the log-likelihood function of p is

$$\ln L(p|\boldsymbol{y}) = \sum_{i=1}^{n} y_i \ln p \ + \left(n - \sum_{i=1}^{n} y_i\right) \ln(1-p).$$

The partial derivative of $\ln L(p|\boldsymbol{y})$ is given by

$$\frac{\partial}{\partial p} \ln L(p|\boldsymbol{y}) = \frac{\sum_{i=1}^n y_i}{p} - \frac{n - \sum_{i=1}^n y_i}{1-p}.$$

Setting this derivative equal to zero, and solving for p gives $\widehat{p} = \overline{y}$, the sample proportion. The second-order conditions hold (verify!) so that $\widehat{p} = \overline{Y}$ is the MLE of p. By invariance, the MLE of the log-odds $g(p)$ is given by

$$g(\widehat{p}) = \ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right).$$

The derivative of g with respect to p is

$$g'(p) = \frac{\partial}{\partial p}\left[\ln\left(\frac{p}{1-p}\right)\right] = \frac{1}{p(1-p)}.$$

It can be shown (verify!) that

$$\sigma^2_{\widehat{p}} = \frac{p(1-p)}{n};$$

thus, the large-sample variance of $g(\widehat{p})$ is

$$[g'(p)]^2 \sigma^2_{\widehat{p}} = \left[\frac{1}{p(1-p)}\right]^2 \times \frac{p(1-p)}{n} = \frac{1}{np(1-p)},$$

which is estimated by $1/n\widehat{p}(1-\widehat{p})$. Thus,

$$\ln\left(\frac{\widehat{p}}{1-\widehat{p}}\right) \pm z_{\alpha/2}\left[\frac{1}{\sqrt{n\widehat{p}(1-\widehat{p})}}\right]$$

is an approximate $100(1-\alpha)$ percent confidence interval for $g(p) = \ln[p/(1-p)]$. $\square$