4 Properties of Point Estimators and Methods of Estimation

Complementary reading: Chapter 9 (WMS).

4.1 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 4.1. 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 $\widehat{\theta}_1$. □

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.

4.2 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 4.2. 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. □

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 4.2, 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!

4.2.1 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 4.5: *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 4.3. 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?

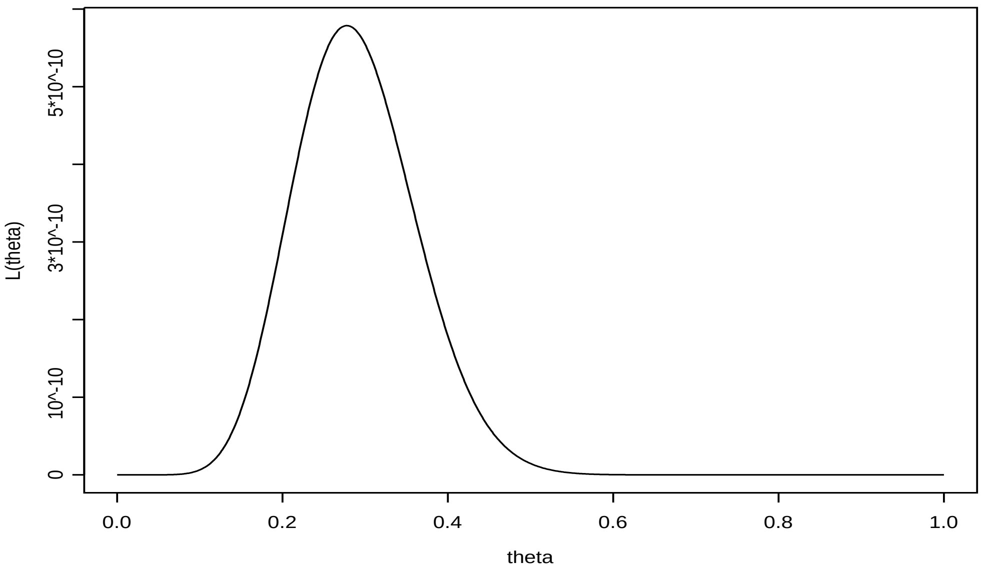

Figure 4.7: *Likelihood function* $L(\theta|\boldsymbol{y})$ *in Example* 4.3.

4.2.2 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 4.4. 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. □

Example 4.5. 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. □

Example 4.6. 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. □

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 4.7. In Example 4.4, 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. □

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 4.8. 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 $\boldsymbol{\theta} = (\alpha, \beta)$. □

Example 4.9. 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 $\boldsymbol{\theta} = (\mu, \sigma^2)$. □

4.3 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 4.10. 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. □

Example 4.11. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of $\mathcal{N}(0, \sigma^2)$ observations. From Example 4.5, we know that $U = \sum_{i=1}^{n} Y_i 2$ is a sufficient statistic for $\sigma 2$. 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. □

Example 4.12. 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 4.13. 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.

4.4 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 4.14. Supposethat $Y_1, Y_2, ..., Y_n$ isaniidsampleof $\mathcal{U}(0, \theta)$ 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

$$\mu_1' = \frac{\theta}{2} \overset{\text{set}}{=} \overline{Y} = m_1'$$

and solve for θ. The MOM estimator for θ is $\widehat{\theta} = 2\overline{Y}$. □

Example 4.15. 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

$$\begin{aligned} \alpha\beta &\overset{\text{set}}{=} \overline{Y} \\ \alpha\beta^2 + (\alpha\beta)^2 &\overset{\text{set}}{=} m_2', \end{aligned}$$

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 4.8 (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. □

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.

4.5 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 4.16. Supposethat $Y_1, Y_2, ..., Y_n$ isaniid $\mathcal{N}(\theta, 1)$ sample.FindtheMLEof 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. □

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 4.17. 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

$$\frac{\partial}{\partial\theta} \ln L(\theta|\boldsymbol{y}) = \frac{\sum_{i=1}^{n} y_i}{\theta} - n \overset{\text{set}}{=} 0.$$

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. □

Example 4.18. 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

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

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

$$\frac{\partial}{\partial\theta}\ln L(\theta|\boldsymbol{y}) = \frac{-2n}{\theta} + \frac{\sum_{i=1}^n y_i}{\theta^2} \overset{\text{set}}{=} 0 \Longrightarrow -2n\theta + \sum_{i=1}^n y_i = 0.$$

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. □

Example 4.19. 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. □

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 4.20. 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 4.6: *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 $\boldsymbol{\theta} = (\mu, \sigma^2)$. □

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 4.21. 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.

```
############################################################
## R codes to 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
```

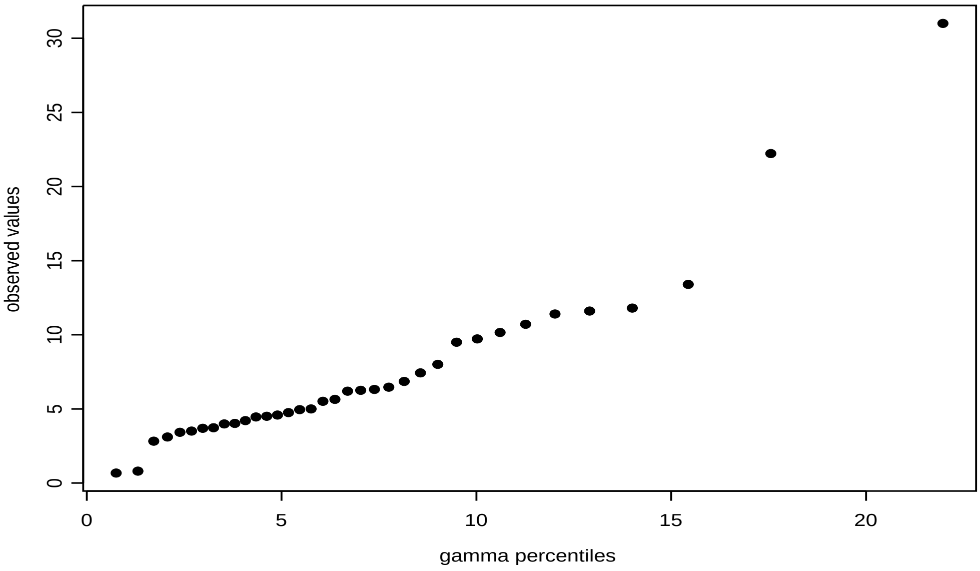

Figure 4.8: *Gamma qq-plot for the hurricane data in Example* 4.21.

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 4.22. 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 $e^{-\theta}$.

4.6 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.

4.6.1 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 4.23. 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

$$\begin{aligned} P(|Y_{(1)} - \theta| > \epsilon) &= \underbrace{P(Y_{(1)} < \theta - \epsilon)}_{=0} + P(Y_{(1)} > \theta + \epsilon) \\ &= \int_{\theta+\epsilon}^{\infty} ne^{-n(y-\theta)} dy \\ &= n\left[-\frac{1}{n}e^{-n(y-\theta)}\Big|_{\theta+\epsilon}^{\infty}\right] \\ &= e^{-n(y-\theta)}\Big|_{\infty}^{\theta+\epsilon} = e^{-n(\theta+\epsilon-\theta)} - 0 = e^{-n\epsilon} \longrightarrow 0, \end{aligned}$$

as $n \to \infty$. Thus, $\widehat{\theta}_n = Y_{(1)}$ is a consistent estimator for θ. □

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 $B(\widehat{\theta}_n) \rightarrow 0$ and $V(\widehat{\theta}_n) \rightarrow 0$, as $n \rightarrow \infty$, rather than showing $P(|\widehat{\theta}_n - \theta| > \epsilon) \rightarrow 0$; 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 $n \rightarrow \infty$.

Proof. Clearly, $B(\overline{Y}_n) = 0$, since $\overline{Y}_n$ is an unbiased estimator of μ. Also, $V(\overline{Y}_n) = \sigma^2/n \rightarrow 0$, as $n \rightarrow \infty$. □

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 4.24. 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 $n \rightarrow \infty$,

$$\overline{Y}_n/2 = g(\overline{Y}_n) \xrightarrow{p} g(2\theta) = \theta.$$

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 $\sigma^2 = 2\theta^2$. □

4.6.2 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 4.25. 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.

4.6.3 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 4.26. Suppose that $Y_1, Y_2, ..., Y_n$ is an iid sample of Poisson observations with mean $\theta > 0$. In Example 4.17, we showed that the MLE of θ is $\theta\widehat{\ } = \overline{Y}$. 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

$$\overline{Y} \pm z_{\alpha/2}\sqrt{\frac{\overline{Y}}{n}}.$$

4.6.4 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 4.27. 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 $g(p) = \ln[p/(1-p)]$. □