Chapter 8

Estimation, testing, and prediction

Statistics is fundamentally about using data to answer inferential questions. These questions are usually hard to answer just by looking at the raw data, so our instinct is to summarise the information available in the observed sample. We might do this by computing summary statistics, such as the sample mean, sample variance, sample minimum and maximum, and so on. Each of these statistics reduces the observed sample to a single number; this process, referred to as data reduction, is the first key topic of this chapter.

We introduce three basic procedures for inference that are the topics of subsequent chapters: point estimation, interval estimation, and hypothesis testing. In all of these procedures, statistics play a dominant role. The inferential methods in this chapter are firmly rooted in classical inference; an alternative framework, Bayesian inference, is discussed in Chapter 11.

8.1 Functions of a sample

8.1.1 Statistics

The formal definition of a statistic is surprisingly loose: any function of the sample is a statistic. The function may be scalar-valued or vector-valued. As the sample is a random vector, a statistic is also a random vector.

Definition 8.1.1 (Statistic)
For a sample, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, a **statistic**, $\boldsymbol{U} = \boldsymbol{h}(\boldsymbol{Y})$, is a random vector that is a function of the sample and known constants alone. Given an observed sample, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$, we can compute the observed value of a statistic, $\boldsymbol{u} = \boldsymbol{h}(\boldsymbol{y})$.

Loose definitions crop up repeatedly in statistical inference. Our approach is to define concepts very broadly, then specify desirable characteristics that will only be found in members of a much smaller subset. Statistics are useful as devices for data reduction when their dimension is smaller than that of the sample. In particular, we will often

consider scalar statistics, that is, statistics that reduce the sample to a single random variable.

The distribution of the statistic $\boldsymbol{U}$ is often referred to as the **sampling distribution** of $\boldsymbol{U}$. Although a statistic is a function of the sample and known constants alone, the distribution of a statistic may depend on unknown parameters. The observed value of a statistic, $\boldsymbol{u}$, is just a vector of real numbers. Just as the observed sample, $\boldsymbol{y}$, is thought of as one instance of the sample, $\boldsymbol{Y}$, the value $\boldsymbol{u} = \boldsymbol{h}(\boldsymbol{y})$ is taken to be one instance of the statistic $\boldsymbol{U} = \boldsymbol{h}(\boldsymbol{Y})$.

In some cases, we can use the distributional results from previous chapters to work out the sampling distributions of statistics of interest. In others, the mathematics involved are not tractable and we need to resolve to simulation-based approaches; these are discussed in Chapter 12.

8.1.2 Pivotal functions

Consider the following situation: if $\boldsymbol{Y}$ is a random sample from an $\mathrm{N}(\mu, 1)$ distribution and $\bar{Y}$ is the sample mean, we know that $\sqrt{n}(\bar{Y} - \mu) \sim \mathrm{N}(0, 1)$ is a function of $\boldsymbol{Y}$ and μ whose distribution does not depend on μ. This is an example of a **pivotal function**. Pivotal functions play a fundamental (or, even, pivotal) role in the construction of confidence intervals. We start with a more formal definition.

Definition 8.1.2 (Pivotal function)
Consider a sample $\boldsymbol{Y}$ and a scalar parameter θ. Let $g(\boldsymbol{Y}, \theta)$ be a function of $\boldsymbol{Y}$ and θ that does not involve any unknown parameter other than θ. We say that $g(\boldsymbol{Y}, \theta)$ is a pivotal function if its distribution does not depend on θ.

Note that a pivotal function defines a random variable, say $W = g(\boldsymbol{Y}, \theta)$. By definition, the distribution of W does not depend on θ.

We illustrate the use of pivotal functions with examples. Before we start on the examples, we introduce a distribution that plays an important role in both interval estimation and hypothesis testing.

Definition 8.1.3 (t-distribution)
Suppose that Z has a standard normal distribution and V has a chi-squared distribution on k degrees of freedom, that is, $Z \sim \mathrm{N}(0, 1)$ and $V \sim \chi^2_k$. Suppose also that Z and V are independent. If we define

$$T = \frac{Z}{\sqrt{V/k}},$$

then T has a **t-distribution** on k degrees of freedom, denoted $T \sim t_k$. The density function of the t-distribution is

$$f_T(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{t^2}{k}\right)^{-(k+1)/2} \quad \text{for } -\infty < t < \infty.$$

Deriving this density is part of Exercise 8.1.

The density function of the t-distribution is a bell-shaped curve centred on the origin. It is characterised by having fatter tails (higher kurtosis) than the density of the standard normal distribution. This excess kurtosis tends to zero as the degrees of freedom tend to infinity. For values of k larger than 30, the t_k is practically indistinguishable from the standard normal (Figure 8.1).

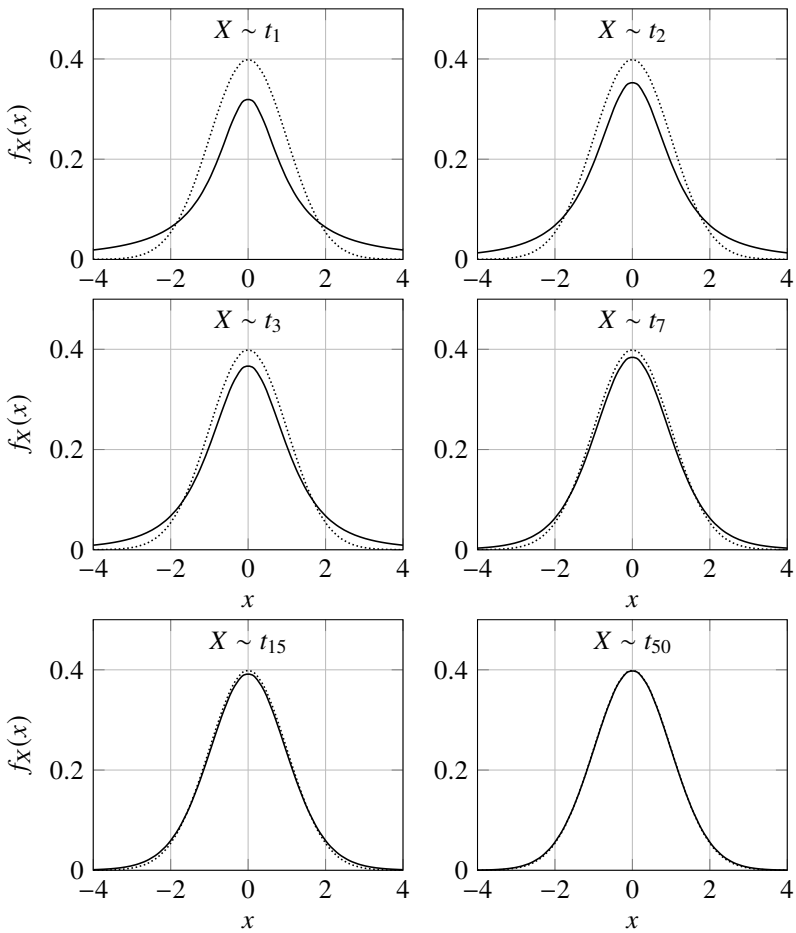

Figure 8.1 *The t_k density function for various values of k, compared to the* N(0, 1) *density function (dotted).*

The t-distribution is often referred to as Student's t. Note the capitalisation and the location of the apostrophe; in this context, "Student" is a proper noun. In fact, Student was the pseudonym of William Sealy Gosset, an employee of the Guinness brewery whose contributions to statistics were motivated by a desire to improve the quality of beer (if only we all had such noble aspirations).

Example 8.1.4 (Pivotal function for mean of normal – variance unknown)
Consider a sample $\boldsymbol{Y}$ from an $\mathrm{N}(\mu, \sigma^2)$ distribution. We know from Corollary 7.3.4 that

$$\frac{\bar{Y} - \mu}{\sqrt{\sigma^2/n}} \sim \mathrm{N}(0,1)\,.$$

However, this does not constitute a pivotal function for μ, as it involves another unknown parameter, namely σ^2. Suppose that we replace the variance, σ^2, by its estimator, the sample variance, S^2. Referring to Corollary 7.3.4 again we see that

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

Furthermore, we know from Lemma 7.3.1 that, for a normal sample, $\bar{Y}$ and S^2 (and hence their functions) are independent. Thus, by definition of the t-distribution,

$$\frac{(\bar{Y} - \mu) \Big/ \sqrt{\sigma^2/n}}{\sqrt{S^2/\sigma^2}} \sim t_{n-1}\,.$$

With some rearrangement, we conclude that

$$\frac{\bar{Y} - \mu}{\sqrt{S^2/n}} \sim t_{n-1}\,.$$

This is a pivotal function since S^2 is a function of $\boldsymbol{Y}$, and the distribution, t_{n-1}, does not depend on the value of μ.

Example 8.1.5 (Pivotal function for parameter of exponential)
Suppose that $\boldsymbol{Y}$ is a random sample from an $\mathrm{Exp}(\lambda)$ distribution. We would like to construct an interval estimator for the parameter λ. We know that $\lambda\,\mathbb{E}(Y) = 1$ so we might consider $\lambda\bar{Y}$ as a potential source of pivotal functions. In fact, we can show that

$$W = \lambda \sum_{i=1}^{n} Y_i$$

is a pivotal function. It is clear from the definition that W does not involve any unknown parameters other than λ. We use moment-generating functions to establish that the distribution of W does not depend on λ. As $Y \sim \mathrm{Exp}(\lambda)$, the moment-generating function of Y is

$$M_Y(t) = (1 - t/\lambda)^{-1}.$$

The moment-generating function of W is

$$\begin{aligned}
M_W(t) = \mathbb{E}[\exp(tW)] &= \mathbb{E}\left[\exp\left(t\lambda \sum_{i=1}^{n} Y_i\right)\right] \\
&= \{\mathbb{E}\left[\exp(t\lambda Y)\right]\}^n && \text{by independence} \\
&= \{M_Y(t\lambda)\}^n \\
&= (1-t)^{-n}. && \text{by definition of } M_Y
\end{aligned}$$

The distribution of a random variable is completely characterised by its moment-generating function. As the moment-generating function of W does not depend on λ, we conclude that the distribution of W does not depend on λ, so W is pivotal.

Exercise 8.1

1. (**Density of the t-distribution**) Find the density function of $X = \sqrt{V/k}$, where $V \sim \chi^2_k$. Now suppose that $Z \sim \mathrm{N}(0,1)$ and Z, X are independent. Using the formula for the density of a ratio of two random variables (see Exercise 4.6), show that $T = Z/X$ has the density function given in Definition 8.1.3.
2. (**Pivotal function for parameter of uniform distribution**) Suppose that $X_1, \cdots, X_n$ is a random sample from the Unif$[0,\theta]$ distribution. Find a pivotal quantity that is a function of $X_{(n)}$.

8.2 Point estimation

Consider a scalar parameter θ; here, θ is just a single unknown number. Any method for estimating the value of θ based on a sample of size n can be represented as a function $h : \mathbb{R}^n \to \mathbb{R}$. When this function is applied to the observed sample it yields a **point estimate**, $h(\boldsymbol{y})$. This estimate is just a number. In order to gain an insight into the properties of the estimation method, we consider applying the function to the sample. The resulting random variable, $h(\boldsymbol{Y})$, is referred to as a **point estimator**. Notice the rather subtle distinction here between an estimator, which is a statistic (random variable), and an estimate, which is an observed value of a statistic (just a number).

It is clear that any point estimator is a statistic. In fact, this association goes in both directions, as the following definition makes clear.

Definition 8.2.1 (Point estimator)
Any scalar statistic may be taken to be a point estimator for a parameter, θ. An observed value of this statistic is referred to as a point estimate.

Definition 8.2.1 seems rather loose; we do not mention anything about restricting our attention to point estimators that are likely to yield estimates close to the true value. In subsection 8.2.1 we introduce the concepts of **bias**, **mean squared error**, and **consistency**. These concepts allow us to formalise "likely to yield estimates close to the true value" as a desirable property of a point estimator. Some commonly used point estimators are given in the following example.

Example 8.2.2 (Some well-known point estimators)
Consider an observed sample $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ that we view as an instance of the sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$. An obvious statistic to calculate is the sample mean. The observed value of the sample mean is

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

which is clearly a function of the observed sample. Applying the same function to the sample, we get

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

The sample mean is the usual estimator for the population mean. We will consider its properties as an estimator using the results established in section 7.1.

In Table 8.1 we list some frequently used statistics, their observed sample values, and the population characteristics for which they are used as point estimators. The terminology gets a little bit cumbersome, but we will try to stick with "sample mean" and "sample variance" for statistics, and "observed sample mean" and "observed sample variance" for instances of these statistics. As always, we use $y_{(i)}$ to denote the i^{th} smallest value when $y_1, \ldots, y_n$ are placed in ascending order.

	statistic (U)	observed value (u)	estimator for
sample mean	$\bar{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$	$\bar{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$	pop. mean
sample variance	$S^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2$	$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2$	pop. variance
1st order statistic	$Y_{(1)} = \min_{1\le j\le n} Y_j$	$y_{(1)} = \min_{1\le j\le n} y_j$	pop. minimum
n^{th} order statistic	$Y_{(n)} = \max_{1\le j\le n} Y_j$	$y_{(n)} = \max_{1\le j\le n} y_j$	pop. maximum
sample median	$Y_{((n+1)/2)}$	$y_{((n+1)/2)}$	pop. median

Table 8.1 *Some frequently used point estimators, observed values, and population quantities that they are used to estimate.*

In general, we may have more than one parameter of interest. These are grouped together into a vector, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^T$. This vector can either be viewed as a single vector parameter or as a vector of scalar parameters. The statistic used as a point estimator for $\boldsymbol{\theta}$ is an $r \times 1$ random vector, and the corresponding point estimate will be an $r \times 1$ vector of real numbers (a point in $\mathbb{R}^r$). The orientation of $\boldsymbol{\theta}$ is often unimportant; row vectors and column vectors serve our purpose equally well. In these cases, we may drop the transpose and write $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$. As an example of a vector of parameters, suppose that our population is normally distributed with unknown mean, μ, and variance, σ^2, that is, $Y \sim \mathrm{N}(\mu, \sigma^2)$. We might take $\boldsymbol{\theta} = (\mu, \sigma^2)$. A point estimator for $\boldsymbol{\theta}$ would be $\hat{\boldsymbol{\theta}} = (\bar{Y}, S^2)$.

Before discussing the properties of point estimators, we reiterate the importance of the distinction between estimators and estimates. Suppose that a statistic $U = h(\boldsymbol{Y})$ is a point estimator of a parameter θ. Consider the observed sample

$$\boldsymbol{y}^{(1)} = (y_1^{(1)}, \ldots, y_n^{(1)})^T,$$

from which we evaluate the sample value of the statistic,

$$u^{(1)} = h(\boldsymbol{y}^{(1)}).$$

We could take another sample,

$$\boldsymbol{y}^{(2)} = (y_1^{(2)}, \ldots, y_n^{(2)})^T,$$

and work out the corresponding value of the statistic,

$$u^{(2)} = h(\boldsymbol{y}^{(2)}).$$

Since $u^{(2)}$ is based on a new sample, it will in general be a different value to $u^{(1)}$. We can repeat this process to build up a collection of values of the statistic, $u^{(1)}, \ldots, u^{(m)}$. Each of these values is an estimate of θ, and each one can be viewed as an instance of the estimator U. Thus, the estimator is a random variable from whose distribution the estimates are sampled; the estimator captures our uncertainty about the values of our estimates. It is the properties of an estimator that determine how useful it is. We might expect a good point estimator to have most of its mass concentrated around the value that we are trying to estimate. In general, this will ensure that the estimator has a high probability of yielding a value close to the parameter value.

8.2.1 Bias, variance, and mean squared error

For many distributions, mass is concentrated around the centre. In order to identify good estimators, we may insist that the centre of the distribution of the estimator is close to our parameter θ. Our usual measure of central tendency of a distribution is the mean. If the mean of an estimator is equal to the parameter value, we say that the estimator is **unbiased**.

Definition 8.2.3 (Bias)
Suppose that U is a statistic. The **bias** of U as a point estimator of θ is

$$\underset{\theta}{\text{Bias}}(U) = \underset{\theta}{\mathbb{E}}(U - \theta).$$

We say that U is an unbiased estimator of θ if $\text{Bias}_\theta(U) = 0$, that is, if $\mathbb{E}_\theta(U) = \theta$.

The θ subscript on $\text{Bias}_\theta(U)$ indicates that we are considering the bias of U as an estimator of θ. We also use a θ subscript on the expectation and probability operators in this context. This is because the distribution of a statistic may depend on unknown parameters; we could denote this by $U \sim F_U(.;\theta)$, where $F_U(u;\theta) = P_\theta(U \leq u)$, and so

$$\underset{\theta}{\mathbb{E}}(U) = \begin{cases} \sum_u u f_U(u;\theta) & \text{for the discrete case,} \\ \int_{-\infty}^{\infty} u f_U(u;\theta) du & \text{for the continuous case,} \end{cases}$$

where f_U is the mass or density function associated with F_U. We will drop the subscript on expectation, variance, and probability operators when the connection between the parameters and the distribution of the estimator is obvious.

An estimator centred on the parameter value does not guarantee that we have a high probability of getting an estimate close to the parameter value. The problem here is

that bias measures expected distance from θ; a statistic that attributes high probability both to values that are much larger than θ and to values that are much smaller than θ is clearly undesirable but could still be unbiased. One way of dealing with this problem is to consider the expected squared distance, that is, the **mean squared error**.

Definition 8.2.4 (Mean squared error)
Suppose that U is a statistic. The mean squared error (MSE) of U as an estimator of θ is given by

$$\underset{\theta}{\text{MSE}}(U) = \underset{\theta}{\mathbb{E}}\left[(U-\theta)^2\right].$$

As we might expect, the mean squared error of an estimator is related to its bias.

Proposition 8.2.5 (Relationship between MSE, bias, and variance)
Suppose that U is a statistic, then, for all $\theta \in \Theta$,

$$\underset{\theta}{\text{MSE}}(U) = [\underset{\theta}{\text{Bias}}(U)]^2 + \underset{\theta}{\text{Var}}(U).$$

For an unbiased estimator, the mean squared error is equal to the variance.

Proving Proposition 8.2.5 is part of Exercise 8.2. The proposition has an intuitive interpretation: an estimator with low bias and low variance is appealing. Low variance means that mass is concentrated around the centre of the distribution but, given that we also have low bias, the centre of the distribution is close to the parameter value. In summary, an estimator with low mean squared error will have a distribution whose mass is concentrated near θ. In practice, low bias and low variance are often competing demands; to reduce bias we may have to accept higher variance, whereas to have an estimator with low variance we may have to introduce some bias. This is referred to as the **bias-variance tradeoff**.

Proposition 8.2.6 (MSE of the sample mean and variance)
Suppose that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a random sample from a population distribution with mean μ, variance σ^2, and fourth central moment μ_4. The mean squared error of the sample mean and variance as estimators of the population mean and variance, respectively, are

$$\underset{\mu}{\text{MSE}}(\bar{Y}) = \frac{\sigma^2}{n},$$
$$\underset{\sigma^2}{\text{MSE}}(S^2) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\sigma^4\right).$$

These results follow directly from Propositions 7.1.2 and 7.2.5. Notice that, as both estimators are unbiased, their MSE is equal to their respective variances.

8.2.2 Consistency

If we could take a sample of infinite size, we might hope that our estimator would be perfect. In other words, we would want our estimator to yield the true parameter value

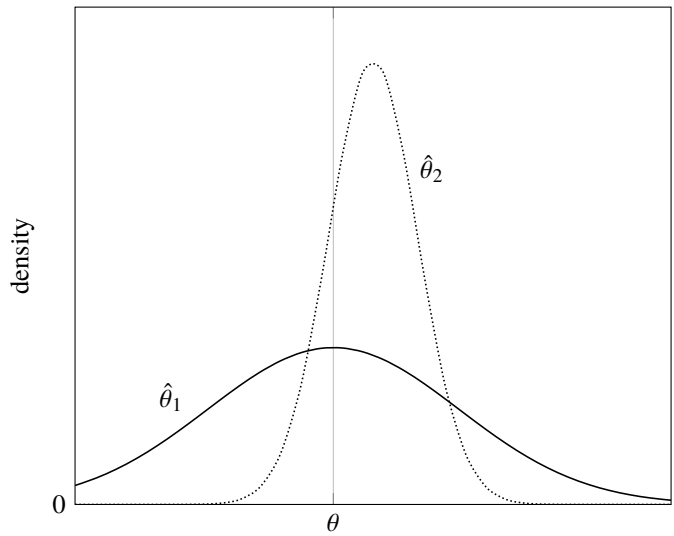

Figure 8.2 *Two estimators of the same parameter θ. $\hat{\theta}_1$ has zero bias but high variance, so it may produce estimates very far from the true value of θ. By contrast, $\hat{\theta}_2$ is biased but has much lower variance, so it would typically give more accurate estimates.*

with probability 1. This is the basic idea behind **consistency**. Consistency provides a formal expression of our hope that, as the sample size increases, the performance of our estimators improves. Some statisticians view consistency as so important that they will not even tentatively entertain an inconsistent estimator.

Consistency is an asymptotic property. In order to define consistency we first need to define a sequence of estimators. We have so far used n to denote sample size, and in defining our estimators ignored the implicit dependence on n. Our estimators are defined in a way that is valid for any n, so it is easy to make the dependence on n explicit and, thus, define a sequence of estimators. For example, if we define the sample mean for a sample of size n to be $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$, then $\bar{Y}_1 = Y_1$, $\bar{Y}_2 = \frac{1}{2}(Y_1 + Y_2)$, $\bar{Y}_3 = \frac{1}{3}(Y_1 + Y_2 + Y_3)$, and so on. For a general estimator, U_n, the subscript n denotes the size of the sample on which the estimator is based. We then define $\{U_n : n = 1, 2, \dots\}$ to be a sequence of estimators; this is often denoted as just $\{U_n\}$ for brevity. We are now in a position to define consistency.

Definition 8.2.7 (Consistency)
The sequence $\{U_n : n = 1, 2, \dots\}$ is a **consistent** sequence of estimators of the parameter θ if it converges in probability to θ, that is, for every $\delta > 0$ and $\theta \in \Theta$,

$$\lim_{n\to\infty} P_\theta(|U_n - \theta| < \delta) = 1 .$$

In most situations the role of the sample size is obvious and we would just say that U is a consistent estimator of θ. It is worth examining Definition 8.2.7 in the context of

the formal definition of convergence . This says that, given any $\varepsilon > 0$, we can find N such that, for any $n > N$,

$$1 - P_\theta(|U_n - \theta| < \delta) < \varepsilon\,.$$

Rearranging the final statement using the properties of probability, our statement of consistency becomes: given any $\varepsilon > 0$, we can find N such that, for any $n > N$,

$$P_\theta(|U_n - \theta| \geq \delta) < \varepsilon\,. \tag{8.1}$$

One possible intuitive interpretation is as follows. Suppose $\{U_n\}$ is consistent for θ. This means that, for any open interval around the parameter value, say $(\theta - \delta, \theta + \delta)$, that we choose, we can find a number N such that, whenever the sample size is larger than N, the probability of our estimator taking a value outside the interval is arbitrarily small. By arbitrarily small we mean that, if we want to reduce the probability that the estimator takes a value outside the interval, all we need to do is increase the value of N.

One of the consequences of (8.1) is an equivalent definition of a sequence of consistent estimators. The sequence $\{U_n\}$ is a consistent sequence of estimators of the parameter θ if, for every $\delta > 0$ and $\theta \in \Theta$,

$$\lim_{n\to\infty} P_\theta(|U_n - \theta| \geq \delta) = 0\,. \tag{8.2}$$

Checking for consistency directly can be tedious. Fortunately, the following proposition provides another route.

Proposition 8.2.8 (Relationship between MSE and consistency)
Suppose that $\{U_n\}$ is a sequence of statistics and θ an unknown parameter. If, for all $\theta \in \Theta$,

$$\lim_{n\to\infty} \operatorname*{MSE}_\theta(U_n) = 0\,,$$

then $\{U_n\}$ is a consistent sequence of estimators for θ.

Proof.

If $\lim_{n\to\infty} \mathrm{MSE}_\theta(U_n) = 0$ for all $\theta \in \Theta$ then, by definition,

$$\lim_{n\to\infty} \mathbb{E}_\theta[(U_n - \theta)^2] = 0 \text{ for all } \theta \in \Theta\,. \tag{8.3}$$

We now apply the Chebyshev inequality (Proposition 3.4.13) to the alternative definition of consistency (8.2) to obtain

$$P_\theta(|U_n - \theta| \geq \delta) \leq \frac{\mathbb{E}_\theta[(U_n - \theta)^2]}{\delta^2}\,. \tag{8.4}$$

From (8.3), we know that the right-hand side of (8.4) converges to zero for all $\theta \in \Theta$. The left-hand side of (8.4) is a non-negative number bounded above by something that converges to zero. Thus, $\lim_{n\to\infty} P_\theta(|U_n - \theta| \geq \delta) = 0$, the condition in (8.2) is satisfied, and $\{U_n\}$ is a consistent sequence of estimators for θ. □

The condition for consistency given in Proposition 8.2.8 is sufficient but not necessary for consistency; there are consistent estimators whose MSE does not converge to zero. Taking into account the expression for MSE in terms of bias and variance given by Proposition 8.2.5 yields the following corollary.

Corollary 8.2.9 (Relationship between consistency, bias, and variance)
Suppose that $\{U_n\}$ is a sequence of statistics and θ an unknown parameter. If, for all $\theta \in \Theta$,

$$\lim_{n\to\infty} \operatorname*{Bias}_{\theta}(U_n) = 0 \text{ and } \lim_{n\to\infty} \operatorname*{Var}_{\theta}(U_n) = 0,$$

then $\{U_n\}$ is a consistent sequence of estimators for θ.

Consider the sample mean for n observations, $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$. This is our usual estimator of the population mean. As such, we might hope that the sample mean is a consistent estimator of the population mean. For a random sample, this is easy to establish. The result is sometimes referred to as the **weak law of large numbers**.

Proposition 8.2.10 (Weak law of large numbers)
If $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a random sample with $\mathbb{E}(Y) = \mu < \infty$ and $\operatorname{Var}(Y) = \sigma^2 < \infty$, then the sample mean is a consistent estimator of the population mean, that is,

$$\bar{Y}_n \xrightarrow{p} \mu \text{ as } n \to \infty,$$

where $\bar{Y}_n = \frac{1}{n}\sum_{j=1}^{n} Y_j$ is the sample mean.

Proof.

From Proposition 8.2.6,

$$\lim_{n\to\infty} \operatorname*{MSE}_{\mu}(\bar{Y}_n) = \lim_{n\to\infty} \frac{1}{n}\sigma^2 = 0,$$

since $\sigma^2 < \infty$. By Proposition 8.2.8, $\bar{Y}_n$ is consistent for μ and, by definition, $\bar{Y}_n \xrightarrow{p} \mu$ as $n \to \infty$. □

If we make some slightly stronger assumptions, we can also show that the sample mean converges almost surely to the population mean. This result is referred to as the **strong law of large numbers** (the terminology is slightly unfortunate). Furthermore, we can deduce from Proposition 8.2.6 that, as long as the fourth central moment is finite, the sample variance is a consistent estimator of the population variance.

8.2.3 The method of moments

In general, population moments are functions of parameters that we would like to estimate. Sample moments are statistics, that is, functions of the sample. By equating

population moments with sample moments and solving the resulting set of simultaneous equations, we can generate estimators for the population parameters. These are referred to as **method-of-moments estimators**.

Consider a sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ parameterised by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^T$. Method-of-moments estimators for $\theta_1, \ldots, \theta_r$ would be generated by solving

$$\mu_i'(\hat{\theta}_1, \ldots, \hat{\theta}_r) = m_i'(Y_1, \ldots, Y_n) \quad \text{for} \quad i = 1, \ldots, r$$

to give expressions for $\hat{\theta}_1, \ldots, \hat{\theta}_r$ in terms of $Y_1, \ldots, Y_n$. Here, $\mu_i'(\hat{\theta}_1, \ldots, \hat{\theta}_r)$ is the i^{th} population moment and $m_i'(Y_1, \ldots, Y_n)$ is the i^{th} sample moment. We use this notation to emphasise that μ_i' depends on the population parameters, while m_i' is a function of the sample only.

The method of moments has the advantage of being easy to implement; for most of the probability distributions encountered in this book, the corresponding equations can be solved easily by hand. However, the resulting estimators often have undesirable properties. Beyond very simple cases, the main use of the method of moments is to provide starting values for other estimation procedures.

Example 8.2.11 (Method of moments for normal)
Suppose that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a sample from an $\mathrm{N}(\mu, \sigma^2)$ distribution, where both μ and σ^2 are unknown parameters. The method-of-moments estimators are found by solving

$$\begin{aligned} \mu_1'(\hat{\mu}, \hat{\sigma}^2) &= m_1'(Y_1, \ldots, Y_n), \\ \mu_2'(\hat{\mu}, \hat{\sigma}^2) &= m_2'(Y_1, \ldots, Y_n). \end{aligned}$$

This yields

$$\begin{aligned} \hat{\mu} &= \bar{Y}, \\ \hat{\sigma}^2 + \hat{\mu}^2 &= \frac{1}{n}\sum_{i=1}^{n} Y_i^2 \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2. \end{aligned}$$

Notice that the method-of-moments estimator of the variance is not the sample variance S^2. In fact, if $\hat{\sigma}^2$ is the method-of-moments estimator, then

$$\hat{\sigma}^2 = \frac{n-1}{n} S^2.$$

As such, $\hat{\sigma}^2$ is a *biased* estimator of the population variance.

Example 8.2.12 (Method of moments for binomial)
Consider a sample $Y_1, \cdots, Y_n$ from a $\mathrm{Bin}(r, p)$ distribution, where both r and p are unknown. We know that

$$\begin{aligned} \mathbb{E}(Y) &= rp, \\ \mathbb{E}(Y^2) &= \mathrm{Var}(Y) + \mathbb{E}(Y)^2 = rp(1-p) + r^2p^2. \end{aligned}$$

The method-of-moments estimators are found by solving

$$\hat{r}\hat{p} = \bar{Y}\,,$$
$$\hat{r}\hat{p}(1-\hat{p}) + \hat{r}^2\hat{p}^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2.$$

Solving for $\hat{r}$ by substituting $\hat{p} = \bar{Y}/\hat{r}$ yields

$$\hat{r} = \frac{\bar{Y}^2}{\bar{Y} - \frac{1}{n}\sum(Y_i - \bar{Y})^2}\,,$$
$$\hat{p} = \frac{\bar{Y}}{\hat{r}}\,.$$

Filling in the details is part of Exercise 8.2.

8.2.4 Ordinary least squares

We now consider an estimation technique that is most commonly used in the linear regression setup of section 6.3, and which you may well have already encountered. Recall that

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}\,,$$

where $\boldsymbol{Y}$ is an $n\times 1$ response vector, $\boldsymbol{X}$ is an $n\times(p+1)$ matrix of explanatory variables (including a constant term), $\boldsymbol{\beta}$ is a $(p+1)\times 1$ parameter vector, and $\boldsymbol{\varepsilon}$ is an $n\times 1$ error vector. Given an estimate $\boldsymbol{b}$ for $\boldsymbol{\beta}$, the estimated value of the i^{th} observation of the response variable is

$$\hat{Y}_i = \boldsymbol{x}_i^T\boldsymbol{b}\,,$$

where $\boldsymbol{x}_i^T$ is the i^{th} row of $\boldsymbol{X}$. In **ordinary least-squares** (OLS) estimation, we seek to minimise the **error sum of squares**,

$$S(\boldsymbol{b}) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})\,.$$

The least-squares estimator for $\boldsymbol{\beta}$ is then

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{b}} S(\boldsymbol{b})\,.$$

This has a straightforward geometric interpretation: in the simple regression case, we seek to minimise the squared vertical distance between each point (X_i, Y_i) and the regression line (Figure 8.3). In higher dimensions, this is replaced by a distance between a point in $\mathbb{R}^{p+1}$ and a p-dimensional hyperplane (Figure 8.4).

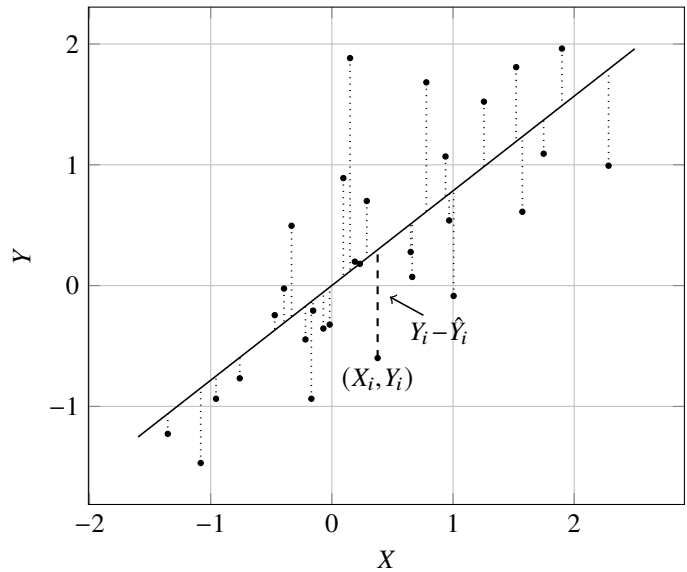

Figure 8.3 *Ordinary least squares in the simple regression case. We want to minimise the squared vertical distance between each point* (X_i, Y_i) *and the regression line.*

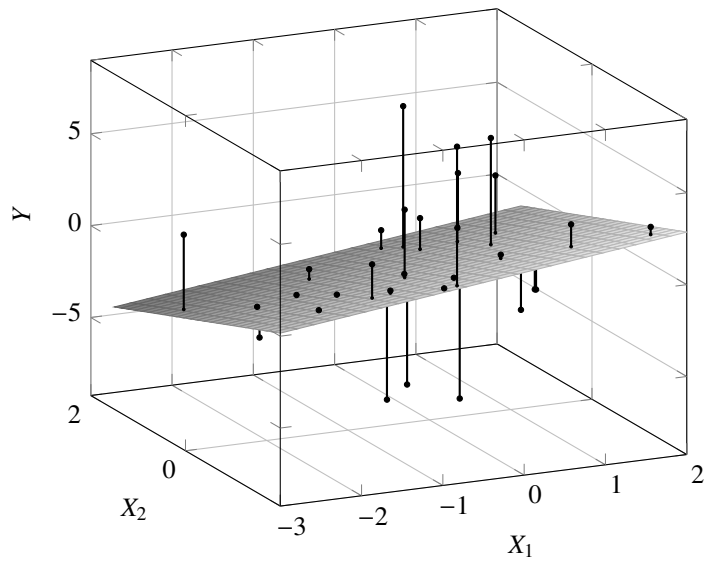

Figure 8.4 *Ordinary least squares with two explanatory variables. We now want to minimise the squared vertical distance between each point* $(X_{1,i}, X_{2,i}, Y_i)$ *and the regression plane.*

We compute $\hat{\boldsymbol{\beta}}$ by setting the derivative of $S(\boldsymbol{b})$ with respect to $\boldsymbol{b}$ equal to zero. We need a few standard results from matrix calculus; if $\boldsymbol{A}$ is an $r \times c$ constant matrix and $\boldsymbol{w}$ is a $c \times 1$ vector, we have

$$\frac{\partial \boldsymbol{A}\boldsymbol{w}}{\partial \boldsymbol{w}} = \boldsymbol{A}, \quad \frac{\partial \boldsymbol{w}^T\boldsymbol{A}}{\partial \boldsymbol{w}} = \boldsymbol{A}^T, \quad \text{and} \quad \frac{\partial \boldsymbol{w}^T\boldsymbol{A}\boldsymbol{w}}{\partial \boldsymbol{w}} = \boldsymbol{w}^T(\boldsymbol{A} + \boldsymbol{A}^T).$$

The derivative of $S(\boldsymbol{b})$ with respect to $\boldsymbol{b}$ is then

$$\begin{aligned}
\frac{\partial S(\boldsymbol{b})}{\partial \boldsymbol{b}} &= \frac{\partial}{\partial \boldsymbol{b}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}) \\
&= \frac{\partial}{\partial \boldsymbol{b}}\left(\boldsymbol{Y}^T\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{b} - \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{Y} + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{b}\right) \\
&= 0 - \boldsymbol{Y}^T\boldsymbol{X} - (\boldsymbol{X}^T\boldsymbol{Y})^T + \boldsymbol{b}^T[\boldsymbol{X}^T\boldsymbol{X} + (\boldsymbol{X}^T\boldsymbol{X})^T] \\
&= 2(-\boldsymbol{Y}^T\boldsymbol{X} + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}),
\end{aligned}$$

as $\boldsymbol{X}^T\boldsymbol{X}$ is symmetric. Setting the derivative equal to zero yields

$$\hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{Y}^T\boldsymbol{X} \;\Leftrightarrow\; \boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}^T\boldsymbol{Y} \;\Leftrightarrow\; \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y},$$

assuming that the matrix $\boldsymbol{X}^T\boldsymbol{X}$ is of full rank and can be inverted. The second derivative is

$$\frac{\partial^2 S(\boldsymbol{b})}{\partial \boldsymbol{b}^T \partial \boldsymbol{b}} = \frac{\partial}{\partial \boldsymbol{b}^T} 2(-\boldsymbol{Y}^T\boldsymbol{X} + \boldsymbol{b}^T\boldsymbol{X}^T\boldsymbol{X}) = 2\boldsymbol{X}^T\boldsymbol{X},$$

which is a positive-definite matrix; for a vector-valued function, this is the equivalent of a positive second derivative, so we conclude that $\hat{\boldsymbol{\beta}}$ minimises $S(\boldsymbol{b})$.

Notice that $\hat{\boldsymbol{\beta}}$ is a linear function of $\boldsymbol{Y}$. Later in this section, we show that $\hat{\boldsymbol{\beta}}$ is unbiased and has the lowest variance of any linear unbiased estimator of $\boldsymbol{\beta}$.

We can write

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y},$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is known as the **hat matrix**; it maps the observed values, $\boldsymbol{Y}$, to the estimated values, $\hat{\boldsymbol{Y}}$. The hat matrix is **idempotent**, that is, $\boldsymbol{H}^2 = \boldsymbol{H}$.

Proposition 8.2.13
If we assume that $\boldsymbol{X}$ is non-random, and the errors have mean $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ and covariance matrix $\mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{I}_n$, then the least-squares estimator, $\hat{\boldsymbol{\beta}}$,

i. is unbiased, $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$,
ii. has covariance matrix $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2$.

If, in addition, we assume that the errors are multivariate normal, then $\hat{\boldsymbol{\beta}}$ is also multivariate normal.

Proving this proposition is part of Exercise 8.2.

Theorem 8.2.14 (Gauss-Markov theorem)
Assume again that $\boldsymbol{X}$ *is non-random, and the errors have mean* $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ *and covariance matrix* $\mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{I}$. *If* $\hat{\boldsymbol{\beta}}$ *is the least-squares estimator of* $\boldsymbol{\beta}$, *and* $\hat{\boldsymbol{\beta}}^* \neq \hat{\boldsymbol{\beta}}$ *is any other unbiased estimator of* $\boldsymbol{\beta}$ *that is a linear function of* $\boldsymbol{Y}$, *then*

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}^*) - \mathrm{Var}(\hat{\boldsymbol{\beta}})$$

is a positive definite matrix.

Proof.

The estimator $\hat{\boldsymbol{\beta}}^*$ is linear, so we can write $\hat{\boldsymbol{\beta}}^* = \boldsymbol{CY}$ for some $(p+1) \times n$ matrix $\boldsymbol{C}$. It has expectation

$$\mathbb{E}(\hat{\boldsymbol{\beta}}^*) = \mathbb{E}(\boldsymbol{CY}) = \boldsymbol{CX\beta}$$

for all $\boldsymbol{\beta}$, and is unbiased, so $\boldsymbol{CX} = \boldsymbol{I}_{p+1}$. Its variance is

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}^*) = \mathrm{Var}(\boldsymbol{CY}) = \sigma^2 \boldsymbol{CC}^T.$$

We now define the $(p+1) \times n$ matrix $\boldsymbol{D} = \boldsymbol{C} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Notice that

$$\boldsymbol{DX} = (\boldsymbol{C} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{X} = \boldsymbol{CX} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_{p+1} - \boldsymbol{I}_{p+1} = \boldsymbol{0},$$

so we have

$$\begin{aligned}\boldsymbol{CC}^T &= [\boldsymbol{D} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T][\boldsymbol{D} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]^T \\ &= \boldsymbol{DD}^T + (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \boldsymbol{DD}^T + (\boldsymbol{X}^T\boldsymbol{X})^{-1}.\end{aligned}$$

Using the properties in Proposition 8.2.13, we can write

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}^*) - \mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2[\boldsymbol{DD}^T + (\boldsymbol{X}^T\boldsymbol{X})^{-1}] - \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \sigma^2\boldsymbol{DD}^T.$$

The matrix $\boldsymbol{D}$ cannot be zero (as that would imply $\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}}$), so $\boldsymbol{DD}^T$ must be positive definite. □

Note the significance of this theorem: if the difference between the covariance matrices is positive definite then, in the matrix sense, the variance of $\hat{\boldsymbol{\beta}}^*$ is higher than the variance of $\hat{\boldsymbol{\beta}}$. We say that the least-squares estimator is, thus, the **best linear unbiased estimator** (BLUE).

Exercise 8.2

1. (**Proving relationship between MSE, bias, and variance**) For a statistic U show that $\mathrm{MSE}_\theta(U) = [\mathrm{Bias}_\theta(U)]^2 + \mathrm{Var}_\theta(U)$ for all $\theta \in \Theta$.
2. Suppose that $Y_1, \ldots, Y_n$ is a random sample from a normally distributed population with mean μ and variance 1, that is, $Y \sim \mathrm{N}(\mu, 1)$. Consider the following statistics:
   i. $U_1 = Y_1 + (Y_2 - Y_3)^2 - 2$ as an estimator of μ,
   ii. $U_2 = \bar{Y}^2 - \frac{1}{n}$ as an estimator of μ^2,

For each estimator determine whether they are

(a) unbiased;

(b) consistent.

[You may assume that, if Y_n converges in probability to y, and g is a continuous function, then $g(Y_n)$ converges in probability to $g(y)$.]

3. Suppose that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a random sample from a Bernoulli(p) distribution, and define $S = \sum_{i=1}^{n} Y_i$. Show that

$$\hat{\theta} = \frac{S(S-1)}{n(n-1)}$$

is an unbiased estimator for $\theta = p^2$. Is this estimator consistent?

4. Suppose that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a sample from a Bin(r, p) distribution where r and p are unknown. Carefully show that the method-of-moments estimators of r and p are

$$\hat{r} = \frac{\bar{Y}^2}{\bar{Y} - \frac{1}{n}\sum_i (Y_i - \bar{Y})^2} \quad \text{and} \hat{p} = \frac{\bar{Y}}{\hat{r}}\,.$$

5. Prove Proposition 8.2.13.

8.3 Interval estimation

In the previous section we described point estimators, which provide a single value for the parameter of interest. Another important class of inferential methods are **interval estimators**. As the name suggests, an interval estimator provides a range of possible values for our unknown parameter, rather than just a single point. Interval estimates are widely used but also widely misinterpreted. Interval estimators can be viewed as a specific case of a **confidence set**. We start by clarifying some terminology.

Recall that a point estimator is a statistic, that is, a function of the sample. By definition, a point estimator is a random variable. When we replace the sample by the observed sample the result is an estimate, that is, just a number. The situation for interval estimators is completely analogous. An interval estimator is a random interval; the end-points of the interval are statistics. When we replace the sample by the observed sample we end up with an interval estimate that is simply a section of the real line.

Definition 8.3.1 (Interval estimator)
Suppose that we have a sample $\boldsymbol{Y}$ parameterised by θ. Let $U_1 = h_1(\boldsymbol{Y})$ and $U_2 = h_2(\boldsymbol{Y})$ be statistics with $U_1 \leq U_2$. The random interval $[U_1, U_2]$ is an interval estimator for θ. If the observed sample is $\boldsymbol{y}$, then the observed values of the statistics are $u_1 = h_1(\boldsymbol{y})$ and $u_2 = h_2(\boldsymbol{y})$. The interval $[u_1, u_2]$ is an interval estimate for θ.

As with the definition of a point estimator, the definition of an interval estimator is rather loose; any scalar statistics U_1 and U_2 will do for lower and upper end points,

provided $U_1 \leq U_2$. The obvious question is then what constitutes a good interval estimator. We have to balance two competing demands: we would like our interval estimate to be as narrow as possible, but we would also like the probability that it covers the true value to be high. The balance between **interval length** and **coverage probability** is discussed in section 8.3.1.

Example 8.3.2 (Interval estimator for mean of normal – variance known)
Suppose that we have a random sample, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, from an $\mathrm{N}(\mu, \sigma^2)$ distribution, so $Y_i \sim \mathrm{N}(\mu, \sigma^2)$ for $i = 1, \ldots, n$. We are interested in an interval estimator for μ assuming that σ^2 is known. We know that the sample mean is also normally distributed,

$$\bar{Y} \sim \mathrm{N}(\mu, \sigma^2/n)\,,$$

and thus,

$$\frac{\bar{Y} - \mu}{\sqrt{\sigma^2/n}} \sim \mathrm{N}(0,1)\,.$$

The quantiles of a standard normal are well known; if $Z \sim \mathrm{N}(0,1)$ then

$$\mathrm{P}(Z \leq -1.96) = 0.025 \;\text{ and }\; \mathrm{P}(Z \geq 1.96) = 0.025\,,$$

and thus,

$$\mathrm{P}(-1.96 \leq Z \leq 1.96) = 0.95\,.$$

As $(\bar{Y} - \mu)/\sqrt{\sigma^2/n}$ has a standard normal distribution, we conclude that

$$\mathrm{P}\left(-1.96 \leq \frac{\bar{Y} - \mu}{\sqrt{\sigma^2/n}} \leq 1.96\right) = 0.95\,.$$

Rearranging this expression yields

$$\mathrm{P}\left(\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95\,.$$

Note that the random variable in this expression is $\bar{Y}$, not μ. A sensible interval estimator for μ is then

$$\left[\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}\right]\,.$$

The corresponding interval estimate is

$$\left[\bar{y} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{y} + 1.96\frac{\sigma}{\sqrt{n}}\right], \tag{8.5}$$

where $\bar{y}$ is the observed sample mean.

8.3.1 Coverage probability and length

A good interval estimator should have a high probability of containing the true value of the parameter. If this were our only criterion, we would always choose the interval $(-\infty, \infty)$, as this covers the true value with probability 1. Clearly, the interval $(-\infty, \infty)$ provides no useful information about plausible values of the parameter. In choosing an interval estimator, there is a tradeoff between the probability of covering the true value and the length of the interval; we would like coverage probability to be high and length to be small. As the probability of covering the true value may depend on the parameter, we make the distinction between **coverage probability** and **confidence coefficient**.

Definition 8.3.3 (Coverage probability)
For an interval estimator $[U_1, U_2]$ of θ, the coverage probability is the probability that the interval estimator covers θ, that is, $\mathrm{P}(U_1 \leq \theta \leq U_2)$.

Definition 8.3.4 (Confidence coefficient)
For an interval estimator $[U_1, U_2]$ of θ, the confidence coefficient is the infimum over θ of the coverage probabilities, that is, $\inf_{\theta \in \Theta} \mathrm{P}(U_1 \leq \theta \leq U_2)$.

It is important to be clear that the random variables in $\mathrm{P}(U_1 \leq \theta \leq U_2)$ are U_1 and U_2, thus,

$$\mathrm{P}(U_1 \leq \theta \leq U_2) = \mathrm{P}[(U_1 \leq \theta) \cap (U_2 \geq \theta)] = 1 - \mathrm{P}(U_1 > \theta) - \mathrm{P}(U_2 < \theta),$$

since $U_1 \leq U_2$ implies that $U_1 > \theta$ and $U_2 < \theta$ are disjoint events.

In Example 8.3.2, we use the fact that $(\bar{Y} - \mu)/(\sqrt{\sigma^2/n})$ is a pivotal function of μ, as its distribution is $\mathrm{N}(0, 1)$ regardless of the value of μ. Thus, the coverage probability of the resulting interval estimator is always 0.95, and we can attribute a confidence coefficient of 0.95 to our interval estimator. In subsection 8.3.2 we formalise the procedure for obtaining confidence intervals from pivotal functions.

You will often see an interval estimator with confidence coefficient $(1-\alpha)$ referred to as a $100(1-\alpha)\%$ **confidence interval**. For example, if $\alpha = 0.05$ then the resulting interval estimator is often called the 95% confidence interval. In this context, the confidence coefficient (expressed as a percentage) is often referred to as the **confidence level**. We use the terms confidence interval and interval estimator interchangeably.

In general, the length of a confidence interval will be a random variable. One possible measure of the width of an interval is expected length.

Definition 8.3.5 (Expected length)
Consider an interval estimator $[U_1, U_2]$. The **expected length** of the interval is defined as $\mathbb{E}(U_2 - U_1)$.

A desirable feature of an interval estimator is that the coverage probability is large for all values of θ. The confidence coefficient represents the worst case scenario; by definition, for any value of θ, the coverage probability will be at least as large as the

confidence coefficient. We return to considering confidence intervals for the mean of a normal distribution to illustrate.

Example 8.3.6 (Assessing the merits of several possible interval estimators)
Suppose that we have a random sample from an $\mathrm{N}(\mu, 1)$ distribution, and we are interested in an interval estimator for μ. Let k_1 and k_2 be non-negative finite constants, that is, $0 \leq k_1 < \infty$ and $0 \leq k_2 < \infty$. Any of the following is a valid interval estimator for μ:

a. $[-k_1, k_2]$,
b. $[Y_1 - k_1, Y_1 + k_2]$,
c. $[\bar{Y} - k_1, \bar{Y} + k_2]$.

All of the intervals in this example are the same length, $k_1 + k_2$. In what follows, we work out the coverage probability and confidence coefficient associated with each one. A quick reminder of some notation: if $Z \sim \mathrm{N}(0, 1)$, we use Φ to denote the cumulative distribution function of Z, so $\Phi(z) = \mathrm{P}(Z \leq z)$.

a. This first interval does not depend on the sample. If μ is the true mean, there are two possible situations: either $\mu \in [-k_1, k_2]$ or $\mu \notin [-k_1, k_2]$. If $\mu \in [-k_1, k_2]$, then the coverage probability is 1, otherwise the coverage probability is 0. Thus, the confidence coefficient for this interval is 0.

b. We can use the fact that $Y_1 - \mu \sim \mathrm{N}(0, 1)$ to work out the coverage probability directly. We have

$$\begin{aligned}
\mathrm{P}(Y_1 - k_1 \leq \mu \leq Y_1 + k_2) &= 1 - \mathrm{P}(Y_1 - k_1 > \mu) - \mathrm{P}(Y_1 + k_2 < \mu) \\
&= 1 - \mathrm{P}(Y_1 - \mu > k_1) - \mathrm{P}(Y_1 - \mu < -k_2) \\
&= \Phi(k_1) - \Phi(-k_2) \\
&= \Phi(k_1) + \Phi(k_2) - 1 \,.
\end{aligned}$$

   This coverage probability does not depend on μ, so the confidence coefficient, which is the infimum over μ, is also $\Phi(k_1) + \Phi(k_2) - 1$.

c. Using the fact that $\sqrt{n}(\bar{Y} - \mu) \sim \mathrm{N}(0, 1)$ and similar reasoning to that given in case b., we can show that the coverage probability is

$$\mathrm{P}(Y_1 - k_1 \leq \mu \leq Y_1 + k_2) = \Phi(\sqrt{n}k_1) + \Phi(\sqrt{n}k_2) - 1 \,.$$

   As with case b., the coverage probability does not involve μ, so the confidence coefficient is equal to the coverage probability.

It is clear that the first interval, with confidence coefficient 0, is of no practical value. Consider cases b. and c. Since k_1 is positive and Φ is a non-decreasing function, we have $\sqrt{n}k_1 \geq k_1$ and thus $\Phi(\sqrt{n}k_1) \geq \Phi(k_1)$ for all $n \geq 1$ (similarly for k_2). We conclude that

$$\Phi(\sqrt{n}k_1) + \Phi(\sqrt{n}k_2) - 1 \geq \Phi(k_1) + \Phi(k_2) - 1 \,,$$

and thus, the confidence coefficient of the interval $[\bar{Y} - k_1, \bar{Y} + k_2]$ is larger than

α_1	α_2	$z_{(1-\alpha_1)}$	$z_{(1-\alpha_2)}$	$\sqrt{n}$ length
0.001	0.049	3.090	1.655	4.745
0.010	0.040	2.326	1.751	4.077
0.020	0.030	2.054	1.881	3.935
0.025	0.025	1.960	1.960	3.920

Table 8.2 *The length of* 95% *confidence intervals for the mean of a normal distribution for various lower and upper end points.*

that for $[Y_1 - k_1, Y_1 + k_2]$. If we had to choose between these intervals we would use $[\bar{Y} - k_1, \bar{Y} + k_2]$. As we will see in section 10.1, this is consistent with the principle of sufficiency ($\bar{Y}$ is sufficient for μ).

In the previous example we considered three intervals of equal length and compared their confidence coefficients. In practice, the usual approach is the reverse: we fix the desired level of confidence and try to find the smallest corresponding interval. The following example illustrates.

Example 8.3.7 (Shortest confidence interval for normal mean)
Suppose that we have a random sample from an $\mathrm{N}(\mu, 1)$ distribution, and we want an interval estimator for μ with confidence coefficient $(1 - \alpha)$. A good place to start is with the pivotal function $\sqrt{n}(\bar{Y} - \mu)$, which is standard normal. Let z_p denote the p-quantile of a standard normal, that is, $\Phi(z_p) = p$. If we choose $\alpha_1, \alpha_2 \geq 0$ such that $\alpha = \alpha_1 + \alpha_2$, we have

$$\mathrm{P}(z_{\alpha_1} \leq \sqrt{n}(\bar{Y} - \mu) \leq -z_{\alpha_2}) = 1 - \alpha_1 - \alpha_2 = 1 - \alpha\,.$$

In other words, we split the total tail probability, α, into a left tail probability, α_1, and a right tail probability, α_2. By rearrangement, and using the fact that $z_{(1-\alpha_1)} = -z_{\alpha_1}$, we can see that

$$\left[\bar{Y} - \frac{1}{\sqrt{n}} z_{(1-\alpha_2)}, \bar{Y} + \frac{1}{\sqrt{n}} z_{(1-\alpha_1)}\right] \tag{8.6}$$

is an interval estimator for μ with confidence coefficient $(1 - \alpha)$. The length of this interval is

$$\frac{1}{\sqrt{n}}(z_{(1-\alpha_1)} + z_{(1-\alpha_2)}).$$

If either α_1 or α_2 is 0, the length of the interval is infinite. Suppose that $\alpha = 0.05$, that is, we want a 95% confidence interval for μ. Table 8.2 compares the length of the confidence interval defined by equation (8.6) for various possible values of α_1 and α_2. It is clear that the shortest interval is given by taking $\alpha_1 = \alpha_2$. This illustrates a general result that coincides with our intuition; for a given confidence coefficient, the shortest confidence intervals for the population mean of a normal distribution will be symmetric about the sample mean.

8.3.2 *Constructing interval estimators using pivotal functions*

Pivotal functions provide a simple mechanism for generating interval estimators with a given confidence coefficient. Suppose that we want an interval estimator for θ with confidence coefficient $1 - \alpha$. We could use the following procedure:

1. Find a pivotal function $g(\boldsymbol{Y}, \theta)$.
2. Use the distribution of the pivotal function to find values w_1 and w_2 such that

$$\mathrm{P}(w_1 \leq g(\boldsymbol{Y}, \theta) \leq w_2) = 1 - \alpha\,. \tag{8.7}$$

   [Note that w_1 and w_2 will depend only on α.]
3. Manipulate the inequalities $g(\boldsymbol{Y}, \theta) \geq w_1$ and $g(\boldsymbol{Y}, \theta) \leq w_2$ to make θ the subject. This yields inequalities of the form

$$\theta \geq h_1(\boldsymbol{Y}, w_1, w_2) \;\text{ and }\; \theta \leq h_2(\boldsymbol{Y}, w_1, w_2)\,,$$

   for some functions h_1 and h_2.
4. We can now give

$$[h_1(\boldsymbol{Y}, w_1, w_2), h_2(\boldsymbol{Y}, w_1, w_2)]$$

   as an interval estimator for θ with confidence coefficient $(1 - \alpha)$. [In practice, the end-points of the interval are usually a function of one of w_1 or w_2 but not the other.]

If the pivotal happens to be a linear function of the unknown parameter, we can establish the optimal choice of w_1 and w_2 for a quite a wide class of distributions. We begin with a definition.

Definition 8.3.8 (Unimodality)
Let X be a random variable with mass/density function $f_X(x)$ and support D. We say that X is (strictly) **unimodal** if there exists a value x^* such that $f_X(x)$ is (strictly) increasing for $x < x^*$ and (strictly) decreasing for $x > x^*$, where $x \in D$. The value x^* is the **mode** of the distribution.

We have encountered several of these distributions so far: $\mathrm{N}(\mu, \sigma^2)$, $\mathrm{Gamma}(\alpha, \lambda)$ for $\alpha > 1$, t_k, and χ^2_k for $k > 2$ are all strictly unimodal; $\mathrm{Pois}(\lambda)$ is strictly unimodal if $\lambda \notin \mathbb{Z}^+$; and $\mathrm{Bin}(n, p)$ is strictly unimodal if $0 < p < 1$ and $(n + 1)p \notin \mathbb{Z}^+$.

Example 8.3.9 (Shortest confidence interval from a linear pivotal)
Suppose that $W = g(\boldsymbol{Y}, \theta)$ is a linear function of θ, and W is continuous and strictly unimodal, with mode w^*. The procedure for constructing a confidence interval from the pivotal function involves finding values w_1 and w_2 that satisfy (8.7). As $g(\boldsymbol{Y}, \theta)$ is linear in θ, we can write

$$g(\boldsymbol{Y}, \theta) = a(\boldsymbol{Y}) + b(\boldsymbol{Y})\theta\,,$$

where $a(\boldsymbol{Y})$ and $b(\boldsymbol{Y})$ may depend on the sample, but not on θ. Solving the inequalities

$W \geq w_1$ and $W \leq w_2$ for θ, we obtain a confidence interval for θ of length $(w_2 - w_1)/|b(\boldsymbol{Y})|$. It is clearly in our interest to minimise the difference $(w_2 - w_1)$.

Now suppose we can find values w_1, w_2 that satisfy $w_1 < w^* < w_2$ and $f_W(w_1) = f_W(w_2)$. Notice that, by the symmetry of the normal distribution, the quantiles we use in Example 8.3.2 satisfy these conditions. We can readily show that we can do no better. Let $k = w_2 - w_1$ and define the function

$$p_k(w) = \mathrm{P}(w \leq W \leq w + k),$$

the probability than an interval of length k with lower endpoint w covers W. Notice that $p_k(w_1) = \mathrm{P}(w_1 \leq W \leq w_2) = 1 - \alpha$. The derivative of this function is

$$p'_k(w) = \frac{d}{dw} \int_w^{w+k} f_W(w)dw = f_W(w + k) - f_W(w),$$

from the fundamental theorem of calculus. There is a stationary point at $w = w_1$, as $p'_k(w_1) = f_W(w_2) - f_W(w_1) = 0$. The second derivative at that point is

$$p''_k(w_1) = f'_W(w_2) - f'_W(w_1).$$

The unimodality of f_W implies that $f'_W(w_2) < 0$ and $f'_W(w_1) > 0$, so $p_k(w)$ has a maximum at $w = w_1$, which means that any interval of length k other than (w_1, w_2) will cover W with probability less than $1-\alpha$. From this we can deduce that no interval of length less than k can cover W with the required probability.

If f_W is also symmetric (as in the normal case), the resulting estimator is an equal-tail interval, in the sense that $\mathrm{P}(W < w_1) = \mathrm{P}(W > w_2) = \alpha/2$.

Example 8.3.10 (Interval estimator for mean of normal – variance unknown)
Suppose that we have a random sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ from an $\mathrm{N}(\mu, \sigma^2)$ distribution, but the variance σ^2 is unknown. We can use the pivotal function from Example 8.1.4, that is,

$$\frac{\bar{Y} - \mu}{\sqrt{S^2/n}} \sim t_{n-1}.$$

This pivotal function is linear in μ, and the t-distribution is strictly unimodal and symmetric. It follows that we should choose

$$\mathrm{P}\left(t_{n-1,\alpha/2} \leq \frac{\bar{Y} - \mu}{\sqrt{S^2/n}} \leq t_{n-1,1-\alpha/2}\right) = 1 - \alpha, \tag{8.8}$$

where $t_{n-1,p}$ denotes the p-quantile of a t_{n-1} distribution. By symmetry of the t-distribution we have

$$t_{n-1,\alpha/2} = -t_{n-1,1-\alpha/2}.$$

We now rearrange expression (8.8) to make μ the subject; this yields the interval estimator

$$\left[\bar{Y} - t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}, \bar{Y} + t_{n-1,1-\alpha/2}\frac{S}{\sqrt{n}}\right].$$

The corresponding interval estimate is

$$\left[\bar{y} - t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{y} + t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}\right], \tag{8.9}$$

where $\bar{y}$ is the observed sample mean, and s the observed sample standard deviation. As in Example 8.3.2, the confidence interval is symmetric about the sample mean.

The t-distribution has fatter tails than the standard normal so, for a 95% confidence interval, the value $t_{n-1,0.975}$ will be larger than 1.96, the corresponding quantile of $\mathrm{N}(0,1)$. In other words, all else being equal, the confidence interval (8.9) will be longer than the corresponding interval (8.5). This is a result of the additional uncertainty introduced by estimating σ^2. As the sample size increases, we know that t_{n-1} converges to $\mathrm{N}(0,1)$, so $t_{n-1,0.975} \to 1.96$ and the two estimates are indistinguishable from each other. This is unsurprising; S^2 is a consistent estimator of σ^2, so it will converge to the true value as the sample size increases.

Example 8.3.11
Suppose that $\boldsymbol{Y}$ is a random sample from an $\mathrm{Exp}(\lambda)$ distribution. In Example 8.1.5 we establish that $W = \lambda \sum_{i=1}^{n} Y_i$ is a pivotal function of λ, with moment-generating function

$$M_W(t) = (1-t)^{-n}.$$

In order to use W to construct interval estimators, we need the parametric form of the distribution of W. Recall from Exercise 7.3 that, if $V \sim \chi^2_k$, then

$$M_V(t) = (1-2t)^{-k/2}.$$

Comparing moment-generating functions, we can see that

$$2W \sim \chi^2_{2n}.$$

This pivotal function is linear in λ and, if $n > 1$, its distribution is strictly unimodal. From Example 8.3.9, constructing the optimal confidence interval involves finding values w_1 and w_2 such that

$$P\left(w_1 \le 2\lambda \sum_{i=1}^{n} Y_i \le w_2\right) = 1-\alpha \quad \text{and} \quad f_{2W}(w_1) = f_{2W}(w_2),$$

where f_{2W} is the density of a chi-squared distribution on $2n$ degrees of freedom. There is no closed-form solution to these equations, so we usually resort to constructing an equal-tail interval instead (Figure 8.5).

Exploiting the parametric form of $2W$, we have

$$P\left(\chi^2_{2n,\alpha/2} \le 2\lambda \sum_{i=1}^{n} Y_i \le \chi^2_{2n,(1-\alpha/2)}\right) = 1-\alpha,$$

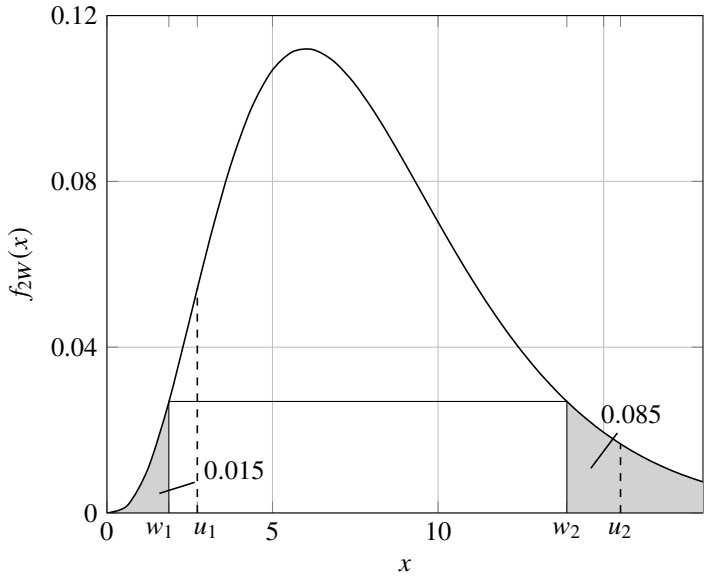

Figure 8.5 *Optimal and equal-tail confidence intervals for the rate of an Exponential, with* $n = 4$ *and* $\alpha = 0.1$. *The distribution of the pivotal function is* $2W \sim \chi^2_8$. *The values* $(u_1, u_2) = (2.73, 15.51)$ *give equal tail probabilities of 0.05, whereas* $(w_1, w_2) = (1.87, 13.89)$ *correspond to the optimal confidence interval with* $f_{2W}(w_1) = f_{2W}(w_2)$. *Notice that the two intervals are quite different when the sample size is low.*

where $\chi^2_{k,p}$ is the p-quantile of a chi-squared distribution on k degrees of freedom. Rearranging the inequalities, we conclude that

$$\left[\frac{\chi^2_{2n,\alpha/2}}{2\sum_{i=1}^n Y_i}, \frac{\chi^2_{2n,1-\alpha/2}}{2\sum_{i=1}^n Y_i} \right],$$

is an interval estimator for λ with confidence coefficient $(1 - \alpha)$.

8.3.3 Constructing interval estimators using order statistics

In the examples we have considered so far, we have given confidence intervals for the mean, variance, and other specific parameters from well-known distributions. Another useful class of population parameters are the quantiles. In section 7.4 we show that order statistics provide a natural mechanism for inference about population quantiles. It is unsurprising then that interval estimators for quantiles are also based on order statistics.

We start by establishing the coverage probability associated with an interval based on the first and last order statistics as an interval estimator for a scalar parameter θ.

Proposition 8.3.12 (Interval estimator based on first and last order statistics)
Consider a random sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ *from a distribution with scalar parameter* θ. *Let* F_Y *be the cumulative distribution function of* Y *and* $Y_{(i)}$ *the* i^{th} *order statistic. The probability that the random interval* $[Y_{(1)}, Y_{(n)})$ *covers* θ *is*

$$1 - [1 - F_Y(\theta)]^n - [F_Y(\theta)]^n.$$

Proof.

Using the properties of order statistics, we know that if $\{Y_{(1)} > \theta\}$ then $\{Y_i > \theta\}$ for all i. Similarly, if $\{Y_{(n)} \leq \theta\}$ then $\{Y_i \leq \theta\}$ for all i. Thus,

$$\begin{aligned} \mathrm{P}(Y_{(1)} > \theta) &= [1 - F_Y(\theta)]^n, \\ \mathrm{P}(Y_{(n)} \leq \theta) &= [F_Y(\theta)]^n. \end{aligned}$$

The events $\{Y_{(1)} > \theta\}$ and $\{Y_{(n)} \leq \theta\}$ are clearly mutually exclusive, so

$$\begin{aligned} \mathrm{P}((Y_{(1)} > \theta) \cup (Y_{(n)} \leq \theta)) &= \mathrm{P}(Y_{(1)} > \theta) + \mathrm{P}(Y_{(n)} \leq \theta) \\ &= [1 - F_Y(\theta)]^n + [F_Y(\theta)]^n. \end{aligned}$$

Using de Morgan's laws, the result follows,

$$\begin{aligned} \mathrm{P}((Y_{(1)} \leq \theta) \cap (Y_{(n)} > \theta)) &= 1 - \mathrm{P}((Y_{(1)} > \theta) \cup (Y_{(n)} \leq \theta)) \\ &= 1 - [1 - F_Y(\theta)]^n - [F_Y(\theta)]^n. \end{aligned}$$

□

Suppose that the parameter of interest is the β-quantile, $\theta = q_\beta$, where q_β is the point satisfying $F_Y(q_\beta) = \beta$. The coverage probability given by Proposition 8.3.12 is then a confidence coefficient. The following corollary gives details.

Corollary 8.3.13 ($[Y_{(1)}, Y_{(n)})$ as interval estimator for a quantile)
Consider a random sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$. *Let* q_β *be the* β*-quantile of* Y. *The random interval* $[Y_{(1)}, Y_{(n)})$ *is an interval estimator for* q_β *with confidence coefficient*

$$1 - (1 - \beta)^n - \beta^n.$$

In practice, we would like to be able to provide an interval estimator for q_β for any given confidence coefficient. We gain greater flexibility by considering the interval estimator associated with general order statistics $Y_{(i)}$ and $Y_{(j)}$, where $i < j$. The following proposition indicates how coverage probability is calculated.

Proposition 8.3.14 ($[Y_{(i)}, Y_{(j)})$ as an interval estimator)
Consider a random sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ *from a distribution with scalar parameter* θ. *Let* $F_{Y_{(i)}}$ *be the cumulative distribution function of the* i^{th} *order statistic. If* $i < j$, *the probability that the random interval* $[Y_{(i)}, Y_{(j)})$ *covers* θ *is*

$$F_{Y_{(i)}}(\theta) - F_{Y_{(j)}}(\theta).$$

The proof follows the same reasoning as Proposition 8.3.12; this is part of Exercise 8.3. The values from the cumulative distribution function of order statistics can be evaluated using Proposition 7.4.5, for example,

$$F_{Y_{(i)}}(\theta) = \sum_{k=i}^{n} \binom{n}{k} [F_Y(\theta)]^k [1 - F_Y(\theta)]^{n-k}.$$

This calculation involves a sum of binomial probabilities, and can be more conveniently expressed in terms of a binomial distribution function. If we define $X \sim \text{Bin}(n, F_Y(\theta))$, then

$$F_{Y_{(i)}}(\theta) = \text{P}(Y_{(i)} \leq \theta) = \text{P}(X \geq i) = 1 - F_X(i-1)\,.$$

Substituting this expression into the result from Proposition 8.3.14, we conclude that the probability that the random interval $[Y_{(i)}, Y_{(j)})$ covers θ is

$$F_X(j-1) - F_X(i-1)\,,$$

where $X \sim \text{Bin}(n, F_Y(\theta))$. Once again, this result is most usefully applied to quantiles; if $\theta = q_\beta$, then $X \sim \text{Bin}(n, \beta)$.

We illustrate these ideas by constructing interval estimators for the median.

Example 8.3.15 (Interval estimators for the median)
Consider a random sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ from any distribution. The median is the 0.5-quantile, $q_{0.5}$. By Corollary 8.3.13, the interval with end-points given by the sample minimum and maximum is an interval estimator for the median with confidence coefficient

$$1 - \left(\frac{1}{2}\right)^n - \left(\frac{1}{2}\right)^n = 1 - \left(\frac{1}{2}\right)^{n-1} = \frac{2^{n-1} - 1}{2^{n-1}}\,.$$

For a sample of size 4, this confidence coefficient is $15/16 = 0.9375$. Clearly, in a reasonable size sample, the probability that $[Y_{(1)}, Y_{(n)}]$ covers the median will be very close to 1.

Suppose that we want an interval estimator for the median with confidence coefficient $1 - \alpha$. From Proposition 8.3.14 and subsequent discussion, we want to find i and j such that

$$F_X(j-1) - F_X(i-1) = 1 - \alpha\,,$$

where $X \sim \text{Bin}(n, 0.5)$. For example, if $n = 100$ then

$$\text{P}(X \leq 39) = 0.0176\,,$$
$$\text{P}(X \leq 59) = 0.9716\,,$$

so $[Y_{(40)}, Y_{(60)})$ provides a confidence interval for the median with coverage coefficient 0.954.

8.3.4 *Confidence sets*

In the bulk of our examples we will construct confidence intervals. However, it is important to understand that a confidence interval can be viewed as a special case of a **confidence set**. Confidence sets are useful in two contexts:

i. if we are unsure that the result of a procedure is an interval,
ii. if we have a vector of parameters, in which case we may refer to our confidence set as a **confidence region**.

A confidence set with confidence coefficient $1 - \alpha$ for a vector parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is defined as the random set $C(\boldsymbol{Y}) \subseteq \boldsymbol{\Theta}$ where

$$\mathrm{P}(\boldsymbol{\theta} \in C(\boldsymbol{Y})) = 1 - \alpha\,.$$

Note that in the above expression the order in which the variables appear can cause confusion; $C(\boldsymbol{Y})$ is the random variable here (it is a function of the sample). In this instance, for an observed sample $\boldsymbol{y}$ we would have the observed confidence set $C(\boldsymbol{y})$.

Exercise 8.3

1. The confidence level for a confidence interval is sometimes described as the probability that the true parameter value lies in the interval. Why is this slightly misleading?
2. Using a computer package or statistical tables, check the entries in Table 8.2.
3. Suppose $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a random sample from an $\mathrm{N}(\mu, \sigma^2)$ distribution. Find a pivotal function for σ^2 in the case where μ is known, and the case where μ is unknown. In each case derive an interval estimator for σ^2 with confidence coefficient $1 - \alpha$ and give an expression for the expected length of the interval.
4. Suppose that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a random sample from a Bernoulli(p) distribution. Using the normal approximation to the binomial, find a function that is approximately pivotal for p. Hence, construct an interval estimator for p with approximate confidence coefficient 0.95.
5. Prove Proposition 8.3.14.
6. Suppose that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a random sample from any distribution. Find interval estimators with confidence coefficients of approximately 0.95 for the upper quartile, $q_{0.75}$, when $n = 100$.

8.4 Hypothesis testing

A hypothesis test starts with a statement about the population; this statement is referred to as the **null hypothesis**. A hypothesis test determines whether the sample provides evidence to reject the null hypothesis. The basis for making this inference is a comparison of the observed value of a **test statistic** with the sampling distribution of the statistic when we assume the null hypothesis is true. A perfect hypothesis test

would always reject a false null hypothesis and never reject a true one. In practice we have to strike a balance between the probability of rejecting a false hypothesis and the probability of rejecting a true one.

The formal setup for hypothesis testing has been extensively criticised and is somewhat out of fashion in modern statistics. However, the ideas presented here are still widely used even if the precise formality is not made explicit. News sites are full of reports on food scares, exam performance, causes of disease, and so on. These reports contain statements of the form, "there is no evidence of a risk to health", or "the change in performance is statistically significant", or "factor x is associated with disease y". These statements are rooted in ideas of hypothesis testing and draw loosely on its terminology. Even if we have doubts about the value of formal testing methodologies, it is of fundamental importance that these methodologies and their limitations are properly understood. One important role for professional statisticians is to warn against naive interpretation (and identify gross misinterpretation) of hypothesis tests.

8.4.1 Statistical hypotheses

The hypothesis tests we consider in this section are **parametric**, in the sense that the hypotheses are statements about the value of a distribution parameter; the form of the distribution itself is assumed to be known. By contrast, **nonparametric** tests do not require any assumptions about the form of the distribution. We have encountered some nonparametric techniques earlier in this book, such as the quantile-based confidence intervals in section 8.3.3.

In general, a parametric hypothesis test takes the form

$$H_* \ : \ \boldsymbol{\theta} \in \boldsymbol{\Theta}_* \ \text{ where } \ \boldsymbol{\Theta}_* \subset \boldsymbol{\Theta}. \tag{8.10}$$

H_* is just a convenient label for our hypothesis. In saying $\boldsymbol{\theta} \in \boldsymbol{\Theta}_*$, where $\boldsymbol{\Theta}_* \subset \boldsymbol{\Theta}$, we are proposing that the true parameter value, $\boldsymbol{\theta}$, lies in a specific subset, $\boldsymbol{\Theta}_*$, of the parameter space, $\boldsymbol{\Theta}$. For scalar parameter θ and known constant k, forms of hypothesis that we might consider include $\theta = k$, $\theta < k$, $\theta \leq k$, $\theta > k$, and $\theta \geq k$. Note that, consistent with (8.10), $\theta = k$ could be written $\theta \in \{k\}$, $\theta < k$ could be written $\theta \in (-\infty, k)$, $\theta \leq k$ could be written $\theta \in (-\infty, k]$, and so on. In this brief illustration we have actually described two distinct forms of hypothesis. The following definition clarifies.

Definition 8.4.1 (Simple and composite hypotheses)
A **simple hypothesis** gives an exact specification of the parameter value $\boldsymbol{\theta}$. Thus, a simple hypothesis will take the form $\boldsymbol{\theta} = \boldsymbol{k}$ for some known constant vector $\boldsymbol{k}$ of the same dimension as $\boldsymbol{\theta}$. Any hypothesis that is not simple is referred to as a **composite hypothesis**.

In the scalar parameter case it is clear that $\theta = k$ is a simple hypothesis. The other four examples do not give a specific value for θ so $\theta < k$, $\theta \leq k$, $\theta > k$, and $\theta \geq k$

are all composite hypotheses. In general, a simple hypothesis will propose that $\boldsymbol{\theta}$ is a member of a set with a single element, $\boldsymbol{\theta} \in \{\boldsymbol{k}\}$. A composite hypothesis is anything that proposes that $\boldsymbol{\theta}$ is a member of a set with more than one element, that is, $\boldsymbol{\theta} \in \boldsymbol{\Theta}_*$ where $|\boldsymbol{\Theta}_*| > 1$. The cardinality of $\boldsymbol{\Theta}_*$ may be finite, as in $\boldsymbol{\theta} \in \{\boldsymbol{k}_1, \boldsymbol{k}_2\}$ or $\boldsymbol{\theta} \in \{\boldsymbol{k}_1, \ldots, \boldsymbol{k}_n\}$; it may be countably infinite, as in $\boldsymbol{\theta} \in \{\boldsymbol{k}_1, \boldsymbol{k}_2, \ldots\}$; or it may be uncountably infinite, as in $\theta \in (k_1, k_2)$, $\theta \in (-\infty, k]$, or $\boldsymbol{\theta} \in \{\boldsymbol{x} : |\boldsymbol{x}| \leq k\}$.

In classical hypothesis testing, two hypotheses are put forward. The **null hypothesis** is the hypothesis that we would like to test; the **alternative hypothesis** informs the manner in which we construct the test of the null.

1. The null hypothesis is usually conservative, in the sense that it reflects established thinking. The null often takes the form "no change", "no difference", or "no effect". We will adopt the convention that the null hypothesis is labelled H_0.
2. The alternative reflects our suspicion about values that the parameter might take. The alternative is usually composite and often takes a form that can be interpreted as "there is a change", "there is a difference", or "there is an effect". We will use H_1 to label the alternative hypothesis.

We can now give a formal statement of a general hypothesis test. For a parameter θ, a hypothesis test takes the form

$$H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 ,$$
$$H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1 .$$

where $\boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$, $\boldsymbol{\Theta}_1 \subset \boldsymbol{\Theta}$, and $\boldsymbol{\Theta}_0 \cap \boldsymbol{\Theta}_1 = \varnothing$. Note that we insist that $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_1$ are mutually exclusive, so there is no way that H_0 and H_1 can both be true. However, there is no requirement for these sets to be exhaustive, that is, we do not require that $\boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_1 = \boldsymbol{\Theta}$.

8.4.2 Decision rules

Classical hypothesis testing involves a binary choice. The options available to us are:

a. reject the null hypothesis,
b. do not reject the null hypothesis.

Teachers of statistics are fond of pointing out that the statement *"do not reject the null hypothesis"* should not be treated as equivalent to *"accept the null"* (students are equally fond of ignoring this advice). This is a somewhat pedantic point, however it is important to realise that all a statistical procedure can do is check whether the available evidence is consistent with the null hypothesis. On the same theme, *"reject the null"* should not be interpreted as *"accept the alternative"*.

The evidence we use to make the decision whether or not to reject the null is the observed sample $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. The sample is used to construct a **test statistic**,

and the decision about H_0 is made on the basis of this statistic. For now, we will summarise the decision process in a single function, ϕ, where

$$\phi(\boldsymbol{y}) = \begin{cases} 1 & \text{when } H_0 \text{ rejected,} \\ 0 & \text{when } H_0 \text{ not rejected.} \end{cases}$$

The function ϕ is referred to as a **decision rule**. The effect of ϕ is to partition the space of possible observed sample values into two sets: $R = \{\boldsymbol{y} : \phi(\boldsymbol{y}) = 1\}$ and $R^c = \{\boldsymbol{y} : \phi(\boldsymbol{y}) = 0\}$. The set R, which corresponds to observed sample values for which H_0 will be rejected, is often referred to as the **rejection region** or **critical region**. For a given test, the decision function and critical region are equivalent, in the sense that if either one is known then the other is completely determined. If the critical region R is specified, we can write down the decision function

$$\phi(\boldsymbol{y}) = \begin{cases} 1 & \text{if } \boldsymbol{y} \in R, \\ 0 & \text{if } \boldsymbol{y} \notin R. \end{cases}$$

Similarly, for a specific decision function ϕ, the critical region is given by

$$R = \{\boldsymbol{y} \in \mathbb{R}^n : \phi(\boldsymbol{y}) = 1\}.$$

8.4.3 Types of error and the power function

In any non-trivial decision problem, there is a chance that we make the wrong decision. In hypothesis testing two wrong decisions are possible. These are usually stated as:

a. Type I error : rejecting H_0 when H_0 is true,

b. Type II error : failing to reject H_0 when H_0 is false.

One possible mechanism for assessing the performance of a test is to consider the probability that the test makes an error. The probability of a type I error is associated with the **significance level** and the **size** of the test. These terms are often used interchangeably, although the meanings are distinct, as the following definition makes clear.

Definition 8.4.2 (Significance level and size)
Consider testing the null hypothesis $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. The test has significance level α if

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} P_{\boldsymbol{\theta}}(H_0 \text{ rejected}) \leq \alpha.$$

The test has size α if

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} P_{\boldsymbol{\theta}}(H_0 \text{ rejected}) = \alpha.$$

The distinction is rather subtle and will not be important for the initial examples that we consider, as size can readily be calculated. However, in some more complex situations, size is not readily computed and we have to settle for significance level. Some notes on size and significance level follow.

1. A test of size α has significance level α. The converse is not true.
2. A test with significance level α will always be at least as conservative as a test of size α.
3. If the null hypothesis is simple (as is often the case) then the size of a test is the probability of a type I error.

Point 3 above relates to the fact that we can only specify a precise probability for type I errors if the null hypothesis is simple. Similarly, we can only specify a precise probability for type II errors if the alternative is simple. The fact that the alternative hypothesis is usually composite motivates the definition of the **power function**.

Definition 8.4.3 (Power function)
For $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ the power function is defined as

$$\beta(\boldsymbol{\theta}) = P_{\boldsymbol{\theta}}(H_0 \text{ rejected}),$$

that is, the probability that the null hypothesis is rejected if the true parameter value is $\boldsymbol{\theta}$.

The **power** of the test under a specific alternative, $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}_1$, is defined as $\beta(\boldsymbol{\theta}_1)$, that is, the value of the power function at the specific value, $\boldsymbol{\theta}_1$. The relationship with type II error is then clear, as

$$P_{\boldsymbol{\theta}_1}(\text{type II error}) = P_{\boldsymbol{\theta}_1}(H_0 \text{ not rejected}) = 1 - \beta(\boldsymbol{\theta}_1).$$

For specific parameter values under the null, the power function gives the probability of a type I error. If our test is of size α, then it is clear from the definition of size that $\beta(\boldsymbol{\theta}) \leq \alpha$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$. In fact, significance level and size can be defined in terms of the power function; if a test has significance level α, then

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \beta(\boldsymbol{\theta}) \leq \alpha,$$

and, if the test has size α, then

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \beta(\boldsymbol{\theta}) = \alpha.$$

8.4.4 Basic ideas in constructing tests

An intuitive description of the perfect hypothesis test is easy to think up; this test would never reject a true null hypothesis but would always reject a false one. Put another way, we would want to reject a true H_0 with probability 0 and reject a false H_0 with probability 1. The power function of this test would be

$$\beta(\boldsymbol{\theta}) = \begin{cases} 0 & \text{if } \boldsymbol{\theta} \in \boldsymbol{\Theta}_0, \\ 1 & \text{if } \boldsymbol{\theta} \notin \boldsymbol{\Theta}_0, \end{cases}$$

This perfect test has size 0 and power 1. It is clear that this ideal is impossible in any situation where there is uncertainty about the values of the parameters. In practice, there is a tradeoff between size and power.

The conventional approach to hypothesis testing is to control the size of the test. This is equivalent to controlling the probability of rejecting a true null hypothesis. If the consequences of rejecting a true null hypothesis are very severe, then we might fix the size of the test to be very small. If we are more relaxed about rejecting a true null then we might choose a larger value for the size. Often, size is fixed at 0.05, a situation referred to as "testing at a 5% level". Tests of size 0.1 and 0.01 are also frequently used. For a test of a given size we would like to maximise the power, that is, to maximise the probability of rejecting the null hypothesis when it is false.

8.4.5 Conclusions and p-values from tests

In principle, there are two possible conclusions of a hypothesis test: we either reject the null hypothesis or we do not reject it. In practice, some information about the weight of evidence against the null hypothesis is desirable. This may take the form of the observed sample value of the test statistic. Another useful way in which the information can be presented is as a ***p*-value**.

Definition 8.4.4 (p-value)
Consider a hypothesis test for a scalar parameter θ with test statistic $U = h(\boldsymbol{Y})$. Suppose, without loss of generality, that the null is rejected when U takes large values. For an observed sample $\boldsymbol{y}$, the corresponding p-value is

$$p(\boldsymbol{y}) = \sup_{\theta \in \Theta_0} \mathrm{P}(U \geq h(\boldsymbol{y}))\,.$$

Clearly, $0 \leq p(\boldsymbol{y}) \leq 1$.

For a simple null hypothesis, we can remove the $\sup_{\theta \in \Theta_0}$ from the definition and say that the p-value is the probability that we got the result that we did from the sample, or a more extreme result. If we have a fixed significance level α, then we can describe the critical region for a test as

$$R = \{\boldsymbol{y} : p(\boldsymbol{y}) \leq \alpha\}\,.$$

Put loosely, if there is only a small probability of observing a result at least as extreme as that observed in the sample, then we reject the null hypothesis.

Example 8.4.5 (Testing the mean of a normal when variance known)
Suppose that we have a random sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ from an $\mathrm{N}(\mu, \sigma^2)$ distribution where σ^2 is known. We would like to test

$$H_0 : \mu = \mu_0,$$
$$H_1 : \mu < \mu_0.$$

Following the basic framework we have established, we can construct a test of size α with critical region

$$R = \left\{ \boldsymbol{y} : \bar{y} < \mu_0 - z_{(1-\alpha)} \frac{\sigma}{\sqrt{n}} \right\}.$$

Notice that, since $z_\alpha = -z_{(1-\alpha)}$, we can write this critical region as

$$R = \left\{ \boldsymbol{y} : \bar{y} < \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} \right\}.$$

The power function for the test is given by

$$\begin{aligned} \beta(\mu) &= P_\mu(\boldsymbol{Y} \in R) \\ &= \mathrm{P}\left(\frac{\bar{Y} - \mu_0}{\sqrt{\sigma^2/n}} < z_\alpha \right) \\ &= \mathrm{P}\left(\frac{\bar{Y} - \mu}{\sqrt{\sigma^2/n}} < z_\alpha - \frac{\mu - \mu_0}{\sqrt{\sigma^2/n}} \right) \\ &= \Phi\left(z_\alpha - \frac{\mu - \mu_0}{\sqrt{\sigma^2/n}} \right), \end{aligned}$$

where Φ is the cumulative distribution function of a standard normal. As we might expect, the power is a strictly decreasing function of μ that takes the value α when $\mu = \mu_0$ (Figure 8.6).

The p-value associated with the test is

$$p(\boldsymbol{y}) = \Phi\left(\frac{\bar{y} - \mu_0}{\sqrt{\sigma^2/n}} \right).$$

Again, as we might expect, small p-values arise when $\bar{y}$ is much smaller than μ_0 (Figure 8.7).

Example 8.4.6 (Testing the mean of a normal when variance unknown)
Suppose that we have a random sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ from an $\mathrm{N}(\mu, \sigma^2)$ distribution where both μ and σ^2 are unknown. We would like to test

$$\begin{aligned} H_0 &: \mu = \mu_0, \\ H_1 &: \mu > \mu_0. \end{aligned}$$

In order to construct a test we will exploit the pivotal function derived in Example 8.1.4; under H_0,

$$\frac{\bar{Y} - \mu_0}{\sqrt{S^2/n}} \sim t_{n-1}.$$

The critical region for a test of size α is given by

$$R = \left\{ \boldsymbol{y} : \bar{y} > \mu_0 + t_{n-1,1-\alpha} \frac{s}{\sqrt{n}} \right\}.$$

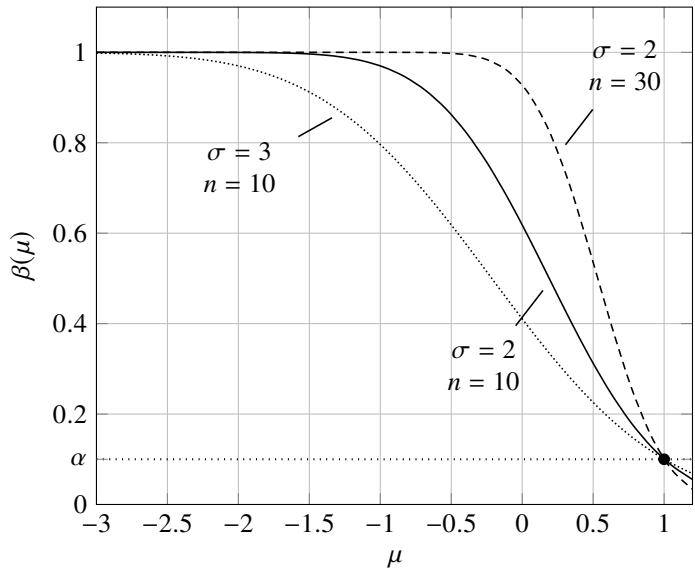

Figure 8.6 *Example of this power function with* $\mu_0 = 1$ *and* $\alpha = 0.1$*, for various values of n and* σ*. Notice that the curve is flatter (lower power) for higher* σ *and steeper (higher power) for larger n.*

Exercise 8.4

1. Someone tells you that the p-value is the probability that the null hypothesis is true. How would you explain why this view is mistaken?
2. Consider the pivotal functions of section 8.1.2. How can these be used in hypothesis testing?
3. In the hypothesis test from Example 8.4.5, for a given value of μ (and assuming μ_0 is fixed and unknown), how could you increase $\beta(\mu)$, the power of the test?

8.5 Prediction

So far, we have focused on inference for unknown parameters. In many problems in statistics our real interest lies in generating sensible estimates for potential sample members that we have not observed; these may include missing values and future values. We will start by supposing that we have a random sample, $\boldsymbol{X} = (X_1, \ldots, X_n)^T$, and we are interested in generating an estimator of a random variable Y based on this sample. In other words, we want to find a statistic $h(\boldsymbol{X})$ that provides us with a reasonable estimator for Y. Clearly, there must be some association between $\boldsymbol{X}$ and Y for this problem to be interesting. An obvious question arises: how do we judge what constitutes a good estimator? We can borrow an idea from point estimation and use mean squared error,

$$\underset{Y}{\mathrm{MSE}}(h(\boldsymbol{X})) = \mathbb{E}\left[(Y - h(\boldsymbol{X}))^2\right],$$

where the expectation is taken with respect to the joint distribution of $\boldsymbol{X}$ and Y.

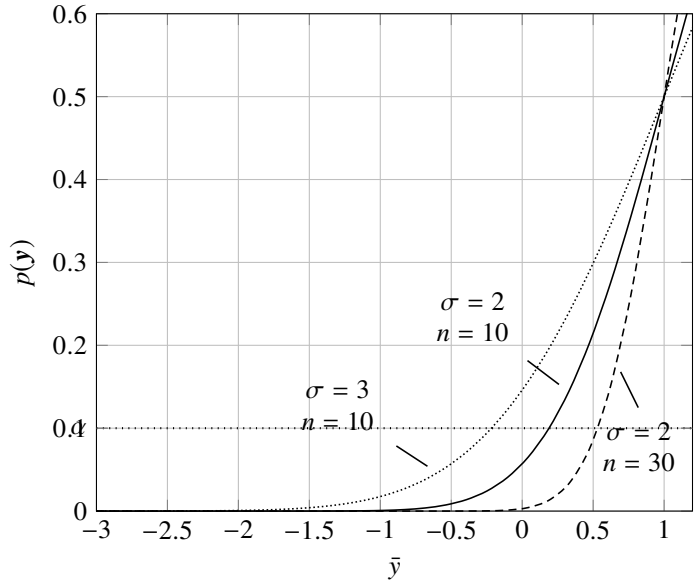

Figure 8.7 *The p-value of this test with $\mu_0 = 1$ and $\alpha = 0.1$. The critical region is where $p(\boldsymbol{y}) < \alpha$, so we require lower (more extreme) values of $\bar{y}$ to reject H_0 when σ is higher or n is smaller.*

An intuitively reasonable choice of estimator is the conditional expectation $\mathbb{E}(Y|\boldsymbol{X})$. In fact, this turns out to be optimal in the mean squared error context. Before we prove optimality, we establish an important property of conditional expectation that has a geometric interpretation.

Lemma 8.5.1 (Conditional expectation as projection)
For any reasonably well-behaved function $g : \mathbb{R}^n \to \mathbb{R}$, the random variable $g(\boldsymbol{X})$ is uncorrelated with $Y - \mathbb{E}(Y|\boldsymbol{X})$.

Proof.

By iterated expectations, we know that $\mathbb{E}[Y - \mathbb{E}(Y|\boldsymbol{X})] = 0$. Thus, the covariance between $g(\boldsymbol{X})$ and $Y - \mathbb{E}(Y|\boldsymbol{X})$ can be written as the expectation of their product,

$$\begin{aligned}
\operatorname{Cov}[g(\boldsymbol{X}), Y - \mathbb{E}(Y|\boldsymbol{X})] &= \mathbb{E}[g(\boldsymbol{X})(Y - \mathbb{E}(Y|\boldsymbol{X}))] \\
&= \mathbb{E}\left\{\mathbb{E}[g(\boldsymbol{X})(Y - \mathbb{E}(Y|\boldsymbol{X}))|\boldsymbol{X}]\right\} \\
&= \mathbb{E}\left\{g(\boldsymbol{X})\,\mathbb{E}[(Y - \mathbb{E}(Y|\boldsymbol{X}))|\boldsymbol{X}]\right\} \\
&= \mathbb{E}\left\{g(\boldsymbol{X})[\mathbb{E}(Y|\boldsymbol{X}) - \mathbb{E}(Y|\boldsymbol{X})]\right\} \\
&= 0.
\end{aligned}$$

□

We are now in a position to prove that the conditional expectation provides the minimum-mean-square estimator (MMSE).

Proposition 8.5.2 (Minimum-mean-square estimator)
The conditional expectation, $\mathbb{E}(Y|\boldsymbol{X})$, is the function of $\boldsymbol{X}$ that has minimum mean squared error as an estimator of Y.

Proof.

Let $h(\boldsymbol{X})$ be any reasonably well-behaved function of $\boldsymbol{X}$. Our method of proof is to show that the mean squared error of $h(\boldsymbol{X})$ is smallest when $h(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X})$. We have

$$\begin{aligned}\underset{Y}{\mathrm{MSE}}(h(\boldsymbol{X})) &= \mathbb{E}\left[(Y - h(\boldsymbol{X}))^2\right] \\ &= \mathbb{E}\left[\{(Y - \mathbb{E}(Y|\boldsymbol{X})) + (\mathbb{E}(Y|\boldsymbol{X}) - h(\boldsymbol{X}))\}^2\right] \\ &= \mathbb{E}\left[(Y - \mathbb{E}(Y|\boldsymbol{X}))^2\right] + \mathbb{E}\left[(Y - \mathbb{E}(Y|\boldsymbol{X}))(\mathbb{E}(Y|\boldsymbol{X}) - h(\boldsymbol{X}))\right] \\ &\quad + \mathbb{E}\left[(\mathbb{E}(Y|\boldsymbol{X}) - h(\boldsymbol{X}))^2\right]. \end{aligned} \tag{8.11}$$

The term $\mathbb{E}(Y|\boldsymbol{X}) - h(\boldsymbol{X})$ is a function of $\boldsymbol{X}$ alone, so (by Lemma 8.5.1) it is uncorrelated with $Y - \mathbb{E}(Y|\boldsymbol{X})$. We conclude that the middle term in equation (8.11) is zero, and thus

$$\underset{Y}{\mathrm{MSE}}(h(\boldsymbol{X})) = \mathbb{E}\left[(Y - \mathbb{E}(Y|\boldsymbol{X}))^2\right] + \mathbb{E}\left[(\mathbb{E}(Y|\boldsymbol{X}) - h(\boldsymbol{X}))^2\right].$$

As the expectation of a non-negative random variable,

$$\mathbb{E}\left[(\mathbb{E}(Y|\boldsymbol{X}) - h(\boldsymbol{X}))^2\right] \geq 0\,.$$

We conclude that $\mathrm{MSE}_Y(h(\boldsymbol{X}))$ is minimised by taking $h(\boldsymbol{X}) = \mathbb{E}(Y|\boldsymbol{X})$. □

Conditional expectation has a number of desirable properties. In particular, it is unbiased and has mean squared error that is less than or equal to the variance of Y. These properties are summarised by the following proposition; the proof is part of Exercise 8.5.

Proposition 8.5.3 (Properties of MMSE)
The MMSE of Y, $\mathbb{E}(Y|\boldsymbol{X})$,

i. is unbiased, $\mathbb{E}[Y - \mathbb{E}(Y|\boldsymbol{X})] = 0$.
ii. has mean squared error $\mathrm{MSE}_Y[\mathbb{E}(Y|\boldsymbol{X})] = \mathrm{Var}(Y) - \mathrm{Var}(\mathbb{E}(Y|\boldsymbol{X}))$.

In many instances, the conditional expectation $\mathbb{E}(Y|\boldsymbol{X})$ is hard to construct. However, if we restrict our attention to estimators that are linear functions of the sample, the estimator that minimises mean squared error can be expressed in terms of the joint moments of $\boldsymbol{X}$ and Y. This estimator is referred to as the **minimum-mean-square linear estimator** (MMSLE). We will use the following notation:

$$\begin{aligned}\boldsymbol{\mu}_X &= \mathbb{E}(\boldsymbol{X}), \quad \mu_Y = \mathbb{E}(Y), \\ \boldsymbol{\Sigma}_X &= \mathrm{Var}(\boldsymbol{X}), \quad \Sigma_Y = \mathrm{Var}(Y), \quad \boldsymbol{\Sigma}_{YX} = \mathrm{Cov}(Y, \boldsymbol{X})\,.\end{aligned}$$

Note that $\mathrm{Cov}(\boldsymbol{X}, Y) = \boldsymbol{\Sigma}_{XY} = \boldsymbol{\Sigma}_{YX}^T$. We start by considering the zero-mean case.

Lemma 8.5.4 (MMSLE zero-mean case)
Suppose that $\boldsymbol{\mu}_X = 0$ *and* $\mu_Y = 0$. *If*

$$\tilde{Y} = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{X}\,,$$

then $\tilde{Y}$ *is the minimum-mean-square linear estimator of* Y *based on* $\boldsymbol{X}$. *The mean squared error of this estimator is*

$$\underset{Y}{\mathrm{MSE}}(\tilde{Y}) = \Sigma_Y - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{XY}\,.$$

Proof.

Consider the estimator $\boldsymbol{a}^T\boldsymbol{X}$, which is the linear function of $\boldsymbol{X}$ generated by the vector of coefficients $\boldsymbol{a} = (a_1, \ldots, a_n)^T$. The mean squared error of this estimator is

$$\begin{aligned}\underset{Y}{\mathrm{MSE}}(\boldsymbol{a}^T\boldsymbol{X}) &= \mathbb{E}\left[(Y - \boldsymbol{a}^T\boldsymbol{X})(Y - \boldsymbol{a}^T\boldsymbol{X})^T\right] \\ &= \mathbb{E}\left[YY^T - \boldsymbol{a}^T\boldsymbol{X}Y^T - Y\boldsymbol{X}^T\boldsymbol{a} + \boldsymbol{a}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{a}\right] \\ &= \Sigma_Y - \boldsymbol{a}^T\boldsymbol{\Sigma}_{XY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{a} + \boldsymbol{a}^T\boldsymbol{\Sigma}_X\boldsymbol{a}^T\,. \end{aligned} \tag{8.12}$$

In order to find the MMSLE, we need the value of $\boldsymbol{a}$ that minimises $\mathrm{MSE}_Y(\boldsymbol{a}^T\boldsymbol{X})$. To calculate this value, we differentiate (8.12) with respect to $\boldsymbol{a}$, set the derivative equal to zero, and solve for $\boldsymbol{a}$. Using the matrix calculus results from subsection 8.2.4, we have

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{a}} \underset{Y}{\mathrm{MSE}}(\boldsymbol{a}^T\boldsymbol{X}) &= -\boldsymbol{\Sigma}_{XY} - \boldsymbol{\Sigma}_{XY} + \boldsymbol{a}^T(\boldsymbol{\Sigma}_X + \boldsymbol{\Sigma}_X^T) \\ -2\boldsymbol{\Sigma}_{YX} + 2\boldsymbol{a}^T\boldsymbol{\Sigma}_X\,.\end{aligned}$$

Setting this derivative equal to zero and solving yields

$$\boldsymbol{a} = \boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{XY}\,, \tag{8.13}$$

and, thus,

$$\tilde{Y} = \boldsymbol{a}^T\boldsymbol{X} = \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{X}\,.$$

The expression for the mean squared error is an immediate consequence of substituting from equation (8.13) in equation (8.12). □

The general case is now a straightforward consequence of Lemma 8.5.4. The proof is part of Exercise 8.5.

Proposition 8.5.5 (MMSLE general case)
The minimum-mean-square linear estimator of Y *based on a sample* $\boldsymbol{X}$ *is given by*

$$\tilde{Y} = \mu_Y + \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_X^{-1}(\boldsymbol{X} - \boldsymbol{\mu}_X)\,.$$

The mean squared error of this estimator is

$$\underset{Y}{\mathrm{MSE}}(\tilde{Y}) = \Sigma_Y - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_X^{-1}\boldsymbol{\Sigma}_{XY}\,.$$

Furthermore, if $\boldsymbol{X}$ and Y are jointly normally distributed, the MMSLE and its MSE are identical to the mean and variance of Y conditional on $\boldsymbol{X} = \boldsymbol{x}$ (Proposition 5.7.2). In other words, in the normal case, the MMSLE is also the MMSE.

Proposition 8.5.5 has a strong connection with linear regression models. For the model $Y = \boldsymbol{X}\boldsymbol{\beta} + \sigma\varepsilon$, the least-squares estimator of Y is

$$\hat{Y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\underbrace{\boldsymbol{X}^T\boldsymbol{X}}_{\boldsymbol{S}_X})^{-1}\underbrace{\boldsymbol{X}^T Y}_{\boldsymbol{S}_{XY}},$$

where $\boldsymbol{S}_X$ and $\boldsymbol{S}_{XY}$ are the sample analogues of $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_{XY}$, respectively.

Exercise 8.5

1. Prove Proposition 8.5.3 and hence show that

$$\underset{Y}{\mathrm{MSE}}(\mathbb{E}(Y|\boldsymbol{X})) = \mathbb{E}(\mathrm{Var}(Y|\boldsymbol{X})).$$

2. Prove Proposition 8.5.5.

8.6 Further exercises

1. Suppose that $Y_1, \ldots, Y_n$ is a random sample from $\mathrm{N}(\mu, 1)$. Define X_n to be a discrete random variable with support $\{0, n\}$ and mass function

$$f_X(x) = \begin{cases} 1 - \frac{1}{n} & \text{for } x = 0\,, \\ \frac{1}{n} & \text{for } x = n\,. \end{cases}$$

   Show that the intuitively unappealing estimator

$$U_n = \bar{Y} + X_n$$

   is consistent for μ. What is the limit of the mean squared error of this estimator as $n \to \infty$?

2. Suppose that $Y_1, \ldots, Y_n$ is a sample from a population with mean μ and variance σ^2. The sample is not random, $\mathrm{Cov}(Y_i, Y_j) = \rho\sigma^2$ for $i \neq j$. Let $U = \sum_{i=1}^n a_i Y_i$. Give the condition on the constants $a_1, \ldots, a_n$ for U to be an unbiased estimator of μ. Under this condition, calculate $\mathrm{MSE}_\mu(U)$.

3. Find the bias and mean squared error for the method-of-moments estimator of the variance,

$$\hat{\sigma}^2 = \frac{n-1}{n}S^2,$$

   based on a random sample from a normal distribution. On the basis of MSE, would you prefer to use $\hat{\sigma}^2$ or S^2 to estimate σ^2? Now consider all estimators of the form

$$\hat{\sigma}_k^2 = \frac{1}{k}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\,.$$

   Which value of k minimises the MSE of $\hat{\sigma}_k^2$ as an estimator of σ^2?

4. You are watching a marathon race; from your vantage point, you observe k competitors run past, with numbers $X_1, X_2, \ldots, X_k$. Suppose that the runners are randomly assigned numbers $1, 2, \ldots, n$, where n is the total number of participants. Find the method-of-moments estimator of n. Is this a sensible estimator?

5. Consider a random sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ from an $\mathrm{N}(\mu, \sigma^2)$ distribution where μ is known and σ^2 is unknown. Two possible interval estimators for σ^2 are

$$\left[\frac{1}{a_1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \frac{1}{a_2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right], \text{ and}$$
$$\left[\frac{1}{b_1}\sum_{i=1}^{n}(Y_i - \mu)^2, \frac{1}{b_2}\sum_{i=1}^{n}(Y_i - \mu)^2\right],$$

   where a_1, a_2, b_1, and b_2 are constants. For $n = 10$, find the values of a_1, a_2, b_1, and b_2 that give intervals with confidence coefficient 0.9. Compare the expected lengths of the intervals and comment.

6. If $V \sim \chi^2_n$, $W \sim \chi^2_m$, and V and W are independent, then the ratio

$$\frac{V/n}{W/m}$$

   has an **F-distribution** on (n, m) degrees of freedom (denoted $F_{n,m}$). Suppose that $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ are random samples from $\mathrm{N}(\mu_X, \sigma^2_X)$ and $\mathrm{N}(\mu_Y, \sigma^2_Y)$, respectively, where μ_X and μ_Y are known. Find a pivotal function for the ratio σ^2_X/σ^2_Y and hence give an expression for an interval estimator for σ^2_X/σ^2_Y with confidence coefficient $1 - \alpha$.

7. (**F-statistic for one-way ANOVA**) For the one-way ANOVA model described in section 6.3.3 and Exercise 6.3, we typically want to test the null hypothesis $H_0 : \beta_1 = \ldots = \beta_{k-1} = 0$ (all group means are equal) against the alternative H_1 : "at least one $\beta_j \neq 0$". Use the distributional results from Exercise 7.5 to show that, under this null hypothesis, the ratio of the between-group sum of squares over the within-group sum of squares (scaled by an appropriate constant) follows an F-distribution. What are its degrees of freedom? How would you perform the test?

8. Consider a random sample from an $\mathrm{N}(\mu, \sigma^2)$ distribution with σ^2 known. Show that the random set

$$\left(-\infty, \bar{Y} - 0.126\frac{\sigma}{\sqrt{n}}\right] \cup \left[\bar{Y} + 0.126\frac{\sigma}{\sqrt{n}}, \infty\right)$$

   is a confidence set for the mean with confidence coefficient 0.95. Would you use this in practice?

9. Let $X_1, \cdots, X_n$ be a random sample from the $\mathrm{Unif}[0, \theta]$ distribution. How would you construct a confidence interval for θ?

8.7 Chapter summary

On completion of this chapter you should be able to:

- describe how statistics play a fundamental role in data reduction,
- write down pivotal functions for standard cases,
- distinguish between estimators and estimates,
- explain the association between bias, variance, and mean squared error,
- prove that, under certain conditions, the sample mean and sample variance are consistent estimators of the population mean and population variance,
- explain the distinction between an interval estimate and an interval estimator,
- interpret interval estimates,
- define the terms coverage probability and coverage coefficient,
- calculate the expected length of an interval estimator,
- use pivotal functions to derive confidence intervals,
- compute the confidence coefficient for an interval estimator based on order statistics.
- explain the distinction between simple and composite hypotheses, and the distinction between size and significance level of a test,
- define type I and type II errors, and describe the association between these errors and the notions of size and power,
- discuss the use of p-values in hypothesis testing, and
- construct hypothesis tests using pivotal functions.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

Chapter 9

Likelihood-based inference

In parametric statistics, we sometimes denote the joint density of the sample by $f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)$, in order to make explicit its dependence on the unknown parameter θ. We can treat this density as a function of θ, known as the likelihood. Despite the fact that density and likelihood have exactly the same functional form, it is often useful to work with likelihood, particularly when we are interested in what happens across a range of parameter values.

In this chapter, we take a more technical approach to parameter estimation and hypothesis testing, based on the likelihood and its functions, such as the score and information. We introduce maximum-likelihood estimation, a key inferential technique, and explore the properties of the resulting estimators. In addition, we discuss various techniques for likelihood maximisation, such as the Newton-Raphson method and the EM algorithm.

We also look at some types of likelihood-based hypothesis tests – the likelihood-ratio, score, and Wald tests – which are widely applicable, as they do not require us to work out the exact distribution of the test statistics.

9.1 Likelihood function and log-likelihood function

We start by considering a sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ from a population with a known parametric form (distribution) with a single, unknown, scalar parameter θ.

Definition 9.1.1 (Likelihood function)
Consider a sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ with density $f_{\boldsymbol{Y}}(\boldsymbol{y};\theta)$. The **likelihood (function)** has the same functional form as the joint probability mass/density, but is viewed as a function of θ rather than $\boldsymbol{y}$. We use the notation $L_{\boldsymbol{Y}}$ to denote likelihood,

$$L_{\boldsymbol{Y}}(\theta;\boldsymbol{y}) = f_{\boldsymbol{Y}}(\boldsymbol{y};\theta).$$

We may sometimes refer to the likelihood as $L_{\boldsymbol{Y}}(\theta)$ when the dependence on $\boldsymbol{y}$ is not important, or as $L(\theta;\boldsymbol{y})$ when this can be done without ambiguity. It is important to

understand that the likelihood is not a mass/density as it does not sum/integrate to 1 over θ. However, the likelihood value $L_{\boldsymbol{Y}}(\theta; \boldsymbol{y})$ may be used as a measure of the plausibility of the parameter value θ for a given set of observations $\boldsymbol{y}$.

We will use likelihood in a number of ways. In addition to the obvious applications (maximum-likelihood estimation in section 9.3 and likelihood-ratio tests in section 9.4), likelihood arises during the discussion of unbiased estimators in section 10.2. In the remainder of this section we introduce a number of definitions, and prove results that will be used in several different contexts. The first of these definitions is the **log-likelihood function**. As the name suggests, this is just the (natural) logarithm of the likelihood function.

Definition 9.1.2 (Log-likelihood function)
If $L_{\boldsymbol{Y}}$ is a likelihood function, we define the log-likelihood function $\ell_{\boldsymbol{Y}}$ as

$$\ell_{\boldsymbol{Y}}(\theta; \boldsymbol{y}) \equiv \log L_{\boldsymbol{Y}}(\theta; \boldsymbol{y}),$$

wherever $L_{\boldsymbol{Y}}(\theta; \boldsymbol{y}) \neq 0$.

Taking logs is a monotone transformation; the location of the maxima and minima of a function are not changed by a log transformation. However, the functional form of the log-likelihood is often more convenient to work with than that of the likelihood.

Simplification for random samples

If $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a random sample, we can express the likelihood and log-likelihood in terms of the shared marginal mass/density, $f_Y(y; \theta)$. We denote the likelihood and log-likelihood associated with a single observation as $L_Y(\theta; y)$ and $\ell_Y(\theta; y)$. Note the importance of the distinction between the vector $\boldsymbol{Y}$, denoting the whole sample, and the scalar Y, representing a single observation. For a random sample, we can exploit independence to write the joint density as a product of the marginal masses/densities. Since mass/density and likelihood have the same functional form, we have

$$L_{\boldsymbol{Y}}(\theta; \boldsymbol{y}) = \prod_{i=1}^{n} L_Y(\theta; y_i).$$

Taking logs yields

$$\ell_{\boldsymbol{Y}}(\theta; \boldsymbol{y}) = \sum_{i=1}^{n} \ell_Y(\theta; y_i). \tag{9.1}$$

In most of the cases we consider, the sample will be random and the likelihood can be written as a product of individual likelihood functions.

Vector parameter case

Now consider the case where there is more than one unknown parameter. The likelihood can be viewed as a function of a single vector parameter or as a function of several scalar parameters; in practice, these two viewpoints are equivalent. Suppose that we have a sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ and parameter(s) $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$. The

likelihood and log-likelihood can be written as

$$L_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{y}) = L_{\boldsymbol{Y}}(\theta_1, \ldots, \theta_r; \boldsymbol{y}),$$
$$\ell_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{y}) = \ell_{\boldsymbol{Y}}(\theta_1, \ldots, \theta_r; \boldsymbol{y}).$$

If $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a random sample then

$$L_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} L_Y(\theta_1, \ldots, \theta_r; y_i),$$
$$\ell_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \ell_Y(\theta_1, \ldots, \theta_r; y_i).$$

Some examples follow.

Example 9.1.3 (Exponential likelihood)
Suppose $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a random sample from an Exp(λ) distribution. The density function of each individual sample member is

$$f_Y(y_i; \lambda) = \lambda \exp(-\lambda y_i) \text{ for } y_i \geq 0.$$

By independence, the joint density of the sample is

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}; \lambda) &= \prod_{i=1}^{n} \lambda \exp(-\lambda y_i) \\ &= \lambda^n \exp\left(-\lambda \sum_{i=1}^{n} y_i\right) \\ &= \lambda^n \exp(-\lambda n \bar{y}) \text{ for } y_1 \geq 0, \ldots, y_n \geq 0. \end{aligned}$$

The likelihood has exactly the same functional form as the density, so

$$L_{\boldsymbol{Y}}(\lambda; \boldsymbol{y}) = \lambda^n \exp(-\lambda n \bar{y}) \text{ for } \lambda > 0.$$

Taking logs yields the log-likelihood,

$$\ell_{\boldsymbol{Y}}(\lambda; \boldsymbol{y}) = n \log(\lambda) - n\lambda \bar{y} \text{ for } \lambda > 0.$$

Example 9.1.4 (Normal likelihood)
Suppose that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a random sample from an N(μ, σ^2) distribution. We define the parameter vector $\boldsymbol{\theta} = (\mu, \sigma^2)$. The likelihood function is then

$$L_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{y}) = L_{\boldsymbol{Y}}(\mu, \sigma^2; \boldsymbol{y}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2\right].$$

Taking logs yields the log-likelihood,

$$\ell_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{y}) = \ell_{\boldsymbol{Y}}(\mu, \sigma^2; \boldsymbol{y}) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2.$$

Exercise 9.1

1. Find the likelihood and log-likelihood function of the parameters for a random sample of size n from each of the following distributions:
   (a) Bernoulli: $f_Y(y) = p^y(1-p)^y$ for $y = 0, 1$ and $0 < p < 1$.
   (b) Pareto: $f_Y(y) = a/(1+y)^{a+1}$ for $0 < y < \infty$ and $a > 0$.
   (c) Weibull: $f_Y(y) = c\tau y^{\tau-1}\exp(-cy^\tau)$ for $0 < y < \infty$, $\tau > 0$ and $c > 0$.
2. Let $X_1, \ldots, X_n$ be a random sample from a Unif$[0,\theta]$ distribution, with density function
$$f_X(x) = \frac{1}{\theta}\mathbf{1}_{[0,\theta]}(x).$$
   Give an expression for the joint density that is a function of θ and $y_{(n)}$ (the sample maximum) only. Sketch the likelihood function.

9.2 Score and information

In section 9.3 we will be maximising the likelihood function, so it is natural that we are interested in derivatives of the likelihood. In fact, we define the **score** to be the first derivative of the log-likelihood. Working with derivatives of the log-likelihood has a number of advantages that will become apparent later on.

Definition 9.2.1 (Score function)
The score function associated with the log-likelihood $\ell_{\boldsymbol{Y}}(\theta; \boldsymbol{y})$ is defined as

$$s_{\mathbf{Y}}(\theta; \boldsymbol{y}) = \frac{\partial}{\partial\theta}\ell_{\boldsymbol{Y}}(\theta; \boldsymbol{y}).$$

By the properties of logs and the chain rule, we have

$$s_{\mathbf{Y}}(\theta; \boldsymbol{y}) = \frac{\frac{\partial}{\partial\theta}L_{\boldsymbol{Y}}(\theta; \boldsymbol{y})}{L_{\boldsymbol{Y}}(\theta; \boldsymbol{y})}. \tag{9.2}$$

The score is defined to make dependence on $\boldsymbol{y}$ explicit. Of course we could apply this function to the random variable $\boldsymbol{Y}$. It turns out that the expected value of $s_{\mathbf{Y}}(\theta; \boldsymbol{Y})$ is zero. In proving this result, and in subsequent proofs, we assume $\boldsymbol{Y}$ is continuous. Similar proofs hold for the discrete case.

Lemma 9.2.2 (A property of the score function)
Assuming sufficient regularity to allow us to differentiate under integral signs, $\mathbb{E}[s_{\mathbf{Y}}(\theta; \boldsymbol{Y})] = 0$.

Proof.

To avoid cluttering the notation, we will drop the $\boldsymbol{Y}$ subscript for this proof, so we

write $L(\theta; \boldsymbol{y}) = L_{\boldsymbol{Y}}(\theta; \boldsymbol{y})$, $f(\boldsymbol{y}; \theta) = f_{\boldsymbol{Y}}(\boldsymbol{y}; \theta)$ and $s(\theta; \boldsymbol{y}) = s_{\boldsymbol{Y}}(\theta; \boldsymbol{y})$. We have

$$\begin{aligned}
\mathbb{E}[s(\theta; \boldsymbol{Y})] &= \int_{\mathbb{R}^n} s(\theta; \boldsymbol{y}) f(\boldsymbol{y}; \theta) d\boldsymbol{y} \\
&= \int_{\mathbb{R}^n} \frac{\frac{\partial}{\partial \theta} L(\theta; \boldsymbol{y})}{L(\theta; \boldsymbol{y})} f(\boldsymbol{y}; \theta) d\boldsymbol{y} && \text{from (9.2)} \\
&= \int_{\mathbb{R}^n} \frac{\partial}{\partial \theta} f(\boldsymbol{y}; \theta) d\boldsymbol{y} && \text{since } f(\boldsymbol{y}; \theta) = L(\theta; \boldsymbol{y}) \\
&= \frac{d}{d\theta} \int_{\mathbb{R}^n} f(\boldsymbol{y}; \theta) d\boldsymbol{y} \, .
\end{aligned}$$

We know that $\int_{\mathbb{R}^n} f(\boldsymbol{y}; \theta) d\boldsymbol{y} = 1$, so the derivative of $\int_{\mathbb{R}^n} f(\boldsymbol{y}; \theta) d\boldsymbol{y}$ with respect to θ is zero. Thus, $\mathbb{E}[s(\theta; \boldsymbol{Y})] = 0$. □

For a given sample $\boldsymbol{y}$, we can plot the likelihood $L_{\boldsymbol{Y}}(\theta; \boldsymbol{y})$ (or equivalently the log-likelihood $\ell_{\boldsymbol{Y}}(\theta; \boldsymbol{y})$) against θ. Consider two extreme cases:

i. The likelihood has a sharp peak: this suggests that a small subset of values of θ are far more plausible than other values.
ii. The likelihood is flat: a large set of values of θ are equally plausible.

In the first case the likelihood provides us with a lot of information; it allows us to narrow down plausible values of θ to a small subset. In the second case, looking at the likelihood does not help us make inferences about θ. The first case is characterised by a sharp peak in the likelihood. This arises when the likelihood changes rapidly with θ, that is, the absolute values of the first derivatives are large. These ideas can be formalised by defining the concept of **Fisher information**. In fact, Fisher information is found by looking at the square (rather than absolute value) of the derivative of the log-likelihood, and taking expectations across sample values.

Definition 9.2.3 (Fisher information)
Suppose that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a sample from a population whose distribution is parameterised by θ, and that $\ell_{\boldsymbol{Y}}(\theta; \boldsymbol{y})$ is the log-likelihood function. The Fisher information about parameter θ in the sample $\boldsymbol{Y}$ is defined as

$$\mathcal{I}_{\boldsymbol{Y}}(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \ell_{\boldsymbol{Y}}(\theta; \boldsymbol{Y})\right)^2\right].$$

Many texts drop the "Fisher" and refer to this quantity simply as the information.

The quantity defined by 9.2.3 is sometimes referred to as **total Fisher information** (or just total information). This is to indicate that we are looking at information in the entire sample. We will encounter Fisher information associated with individual sample members shortly.

Fisher information is a quantity that you might see expressed in a number of different ways. From the definition of the score, we have

$$\mathcal{I}_{\boldsymbol{Y}}(\theta) = \mathbb{E}\left[s_{\mathbf{Y}}(\theta;\boldsymbol{Y})^2\right] .$$

Perhaps more usefully, we can exploit the fact that $\mathbb{E}[s_{\mathbf{Y}}(\theta;\boldsymbol{Y})] = 0$, and thus

$$\mathrm{Var}[s_{\mathbf{Y}}(\theta;\boldsymbol{Y})] = \mathbb{E}[s_{\mathbf{Y}}(\theta;\boldsymbol{Y})^2],$$

to write

$$\mathcal{I}_{\boldsymbol{Y}}(\theta) = \mathrm{Var}\left[s_{\mathbf{Y}}(\theta;\boldsymbol{Y})\right] .$$

Another form of Fisher information, that is often computationally convenient, is given by the following lemma.

Lemma 9.2.4 (Alternative form for Fisher information)
Consider $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, a sample from a population whose distribution is parameterised by θ, and suppose that $\ell_{\boldsymbol{Y}}(\theta;\boldsymbol{y})$ is the log-likelihood function. An alternative representation of the Fisher information about parameter θ in the sample $\boldsymbol{Y}$ is

$$\mathcal{I}_{\boldsymbol{Y}}(\theta) = -\,\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ell_{\boldsymbol{Y}}(\theta;\boldsymbol{Y})\right].$$

Proof.

This result is obtained by differentiating $\mathbb{E}[s(\theta;\boldsymbol{Y})]$ with respect to θ. Once again, in the interests of uncluttered notation we will drop $\boldsymbol{Y}$ subscripts on the density and the score. We have

$$\begin{aligned}
&\frac{d}{d\theta}\,\mathbb{E}[s(\theta;\boldsymbol{Y})] \\
&= \frac{d}{d\theta}\int_{\mathbb{R}^n} s(\theta;\boldsymbol{y}) f(\boldsymbol{y};\theta) d\boldsymbol{y} \\
&= \int_{\mathbb{R}^n} \frac{d}{d\theta}\left(s(\theta;\boldsymbol{y}) f(\boldsymbol{y};\theta)\right) d\boldsymbol{y} \\
&= \int_{\mathbb{R}^n} \left[\left(\frac{\partial}{\partial\theta}s(\theta;\boldsymbol{y})\right) f(\boldsymbol{y};\theta) + s(\theta;\boldsymbol{y})\left(\frac{\partial}{\partial\theta}f(\boldsymbol{y};\theta)\right)\right] d\boldsymbol{y} && \text{product rule} \\
&= \int_{\mathbb{R}^n} \left[\frac{\partial}{\partial\theta}s(\theta;\boldsymbol{y}) + s(\theta;\boldsymbol{y})^2\right] f(\boldsymbol{y};\theta) d\boldsymbol{y} && \text{by 9.2} \\
&= \mathbb{E}\left[\frac{\partial}{\partial\theta}s(\theta;\boldsymbol{Y}) + s(\theta;\boldsymbol{Y})^2\right].
\end{aligned}$$

By Lemma 9.2.2, $\mathbb{E}[s(\theta;\boldsymbol{Y})] = 0$, so $\frac{d}{d\theta}\,\mathbb{E}[s(\theta;\boldsymbol{Y})] = 0$. We thus have

$$\mathbb{E}\left[s(\theta;\boldsymbol{Y})^2\right] = -\,\mathbb{E}\left[\frac{\partial}{\partial\theta}s(\theta;\boldsymbol{Y})\right].$$

Recalling that $s(\theta;\boldsymbol{y})$ is defined as $\frac{\partial}{\partial\theta}\ell_{\boldsymbol{Y}}(\theta;\boldsymbol{y})$ gives the final result. □

We now have two key expressions for Fisher information, and several possible variations using the score; it is important to understand that these all refer to the same quantity. From the original definition (9.2.3) we have

$$\mathcal{I}_{\boldsymbol{Y}}(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\ell_{\boldsymbol{Y}}(\theta;\boldsymbol{Y})\right)^2\right] = \mathbb{E}\left[s_{\mathbf{Y}}(\theta;\boldsymbol{Y})^2\right] = \mathrm{Var}\left[s_{\mathbf{Y}}(\theta;\boldsymbol{Y})\right],$$

and from the alternative form given by Lemma 9.2.4,

$$\mathcal{I}_{\boldsymbol{Y}}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ell_{\boldsymbol{Y}}(\theta;\boldsymbol{Y})\right] = -\mathbb{E}\left[\frac{\partial}{\partial\theta}s_{\mathbf{Y}}(\theta;\boldsymbol{Y})\right].$$

The quantity

$$\hat{\mathcal{I}}_{\boldsymbol{Y}}(\theta;\boldsymbol{Y}) = -\frac{\partial^2}{\partial\theta^2}\ell_{\boldsymbol{Y}}(\theta;\boldsymbol{Y})$$

is known as the **observed information**. It is the sample analogue of the Fisher information, in the sense that it is a function of the random sample $\boldsymbol{Y}$. Notice that

$$\mathbb{E}\left[\hat{\mathcal{I}}_{\boldsymbol{Y}}(\theta;\boldsymbol{Y})\right] = \mathcal{I}_{\boldsymbol{Y}}(\theta).$$

Simplification for random samples

In the previous subsection we saw that if $\boldsymbol{Y} = (Y_1,\ldots,Y_n)^T$ is a random sample, we can express the likelihood as a product of individual likelihoods, and the log-likelihood as a sum of individual log-likelihoods. Similar results hold for the score and information. We denote the score and information associated with a single observation as $s_Y(\theta;y)$ and $\mathcal{I}_Y(\theta)$ respectively. Differentiating (9.1) yields

$$s_{\mathbf{Y}}(\theta;\boldsymbol{y}) = \sum_{i=1}^{n} s_Y(\theta;y_i).$$

We can also readily show that

$$\mathcal{I}_{\boldsymbol{Y}}(\theta) = n\mathcal{I}_Y(\theta). \tag{9.3}$$

Proving this relationship is part of Exercise 9.2. Equation (9.3) indicates that, for a random sample, total Fisher information is just the Fisher information for a single observation multiplied by the sample size; in other words, information is additive, and every observation contains the same amount of information about the unknown parameter.

Vector parameter case

Suppose that we have a vector of parameters $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_r)$. In this case the score will be an $r\times 1$ vector, and the Fisher information will be an $r\times r$ matrix. We introduce the following definitions:

Definition 9.2.5 (Score vector and information matrix)
Consider the sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ with log-likelihood $\ell_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{Y})$. The **score vector** is defined as

$$s_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{y}) = \nabla_{\boldsymbol{\theta}} \ell_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{y}) = \left(\frac{\partial}{\partial \theta_1} \ell_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{y}), \ldots, \frac{\partial}{\partial \theta_r} \ell_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{y}) \right)^T ,$$

and the information matrix is given by

$$\mathcal{I}_{\boldsymbol{Y}}(\boldsymbol{\theta}) = \mathbb{E}[s_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{Y}) s_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{Y})^T] .$$

The symbol ∇ denotes the **del** operator, which is a generalisation of the derivative to higher dimensions.

Consider the function $g : \mathbb{R}^k \to \mathbb{R}$, so if $\mathbf{x} = (x_1, \ldots, x_k)^T$ is a k-dimensional vector, $g(\mathbf{x})$ is a scalar. We use the del operator to represent the partial derivatives of $g(\mathbf{x})$ with respect to each element of $\mathbf{x}$, that is,

$$\nabla_{\mathbf{x}} g(\mathbf{x}) = \left(\frac{\partial}{\partial x_1} g(\mathbf{x}), \ldots, \frac{\partial}{\partial x_k} g(\mathbf{x}) \right)^T .$$

Notice that $\nabla_{\mathbf{x}} g(\mathbf{x})$ has the same dimension as $\mathbf{x}$.

The $(i, j)^{\text{th}}$ element of the information matrix is

$$[\mathcal{I}_{\boldsymbol{Y}}(\boldsymbol{\theta})]_{i,j} = \mathbb{E}\left[\frac{\partial}{\partial \theta_i} \ell_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{Y}) \frac{\partial}{\partial \theta_j} \ell_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{Y}) \right] .$$

Proving that this relationship holds is part of Exercise 9.3.

We can derive properties for the score vector and information matrix that are very similar to those given in the scalar case; the following proposition summarises them.

Proposition 9.2.6
Consider the sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ *parameterised by* $\boldsymbol{\theta}$*, with log-likelihood* $\ell_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{Y})$*. The following relationships hold:*

i. $\mathbb{E}[s_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{Y})] = \mathbf{0}$,
ii. $\mathcal{I}_{\boldsymbol{Y}}(\boldsymbol{\theta}) = \mathrm{Var}[s_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{Y})]$,
iii. $\mathcal{I}_{\boldsymbol{Y}}(\boldsymbol{\theta}) = -\mathbb{E}\left[\nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}}^T \ell_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{Y}) \right] = -\mathbb{E}[\nabla_{\boldsymbol{\theta}} s_{\boldsymbol{Y}}(\boldsymbol{\theta}; \boldsymbol{Y})]$.

The third of these statements says that the $(i, j)^{\text{th}}$ element of the information matrix can be written as

$$[\mathcal{I}_{\boldsymbol{Y}}(\boldsymbol{\theta})]_{i,j} = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f_{\boldsymbol{Y}}(\boldsymbol{Y}; \boldsymbol{\theta}) \right] .$$

Proving Proposition 9.2.6 is part of Exercise 9.2.

Example 9.2.7 (Score and information for exponential)
Suppose $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a random sample from an Exp(λ) distribution. By differentiating the expression for the log-likelihood found in Example 9.1.3, we obtain the score,

$$s_{\mathbf{Y}}(\lambda; \boldsymbol{y}) = \frac{n}{\lambda} - \sum_{i=1}^{n} y_i \, .$$

This is a case in which the information is much easier to derive using the alternative form given by Lemma 9.2.4,

$$\frac{\partial}{\partial \lambda} s_{\mathbf{Y}}(\lambda; \boldsymbol{y}) = -\frac{n}{\lambda^2} \, ,$$

so

$$\mathcal{I}_{\boldsymbol{Y}}(\lambda) = \mathbb{E}\left[-\frac{\partial}{\partial \lambda} s_{\mathbf{Y}}(\lambda; \boldsymbol{Y}) \right] = \frac{n}{\lambda^2} \, .$$

Example 9.2.8 (Score and information for mean of normal with known variance)
Consider $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, a random sample from an N(μ, σ^2) distribution where σ^2 is known. The score is then just the scalar function given by differentiating the log-likelihood from Example 9.1.4 with respect to μ,

$$s_{\mathbf{Y}}(\mu; \boldsymbol{y}) = \frac{\partial}{\partial \mu} \ell(\mu; \boldsymbol{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \mu) \, .$$

Using Lemma 9.2.4, the information is

$$\mathcal{I}_{\boldsymbol{Y}}(\mu) = \mathbb{E}\left[-\frac{\partial}{\partial \mu} s_{\mathbf{Y}}(\mu; \boldsymbol{y}) \right] = \mathbb{E}\left[\frac{1}{\sigma^2} \sum_{i=1}^{n} 1 \right] = \frac{n}{\sigma^2} \, .$$

This expression is intuitively appealing as it says that the information about μ available from the sample increases with the sample size and decreases as the variance grows.

Exercise 9.2

1. Given a random sample $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, give the likelihood, log-likelihood, score, and information for the parameter in each of the following cases:
   (a) $Y \sim \text{Pois}(\lambda)$,
   (b) $Y \sim \text{N}(0, \sigma^2)$,
   (c) $Y \sim \text{Geometric}(p)$,
   (d) $Y \sim \text{N}(\mu, \sigma^2)$, with both μ and σ^2 unknown.
2. Suppose that $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ is a random sample parameterised by θ. Show that

   $$\mathcal{I}_{\boldsymbol{Y}}(\theta) = n\mathcal{I}_Y(\theta)$$

   where $\mathcal{I}_{\boldsymbol{Y}}(\theta)$ is the total information and $\mathcal{I}_Y(\theta)$ is the information associated with a single observation.

3. Given the setup in Example 9.2.7, derive the information using Definition 9.2.3. Given the setup in Example 9.2.8, derive the information using the expression given in 9.2.4. In each case, show that you arrive at the same result starting from the information associated with a single observation.
4. Prove Proposition 9.2.6.

9.3 Maximum-likelihood estimation

Recall from section 9.1 that the likelihood is used as a measure of the plausibility of parameter values for a given sample. It is therefore reasonable to take as a point estimator the value that maximises the likelihood function. Such estimators are referred to (unsurprisingly) as **maximum-likelihood estimators**. We start by considering the scalar parameter case.

Definition 9.3.1 (Maximum-likelihood estimate)
Consider a sample $\boldsymbol{Y}$ parameterised by θ, and let $L_{\boldsymbol{Y}}$ be the likelihood function. Given an observed sample $\boldsymbol{y}$, the **maximum-likelihood estimate** (MLE) of θ is the value, $\hat{\theta}$, that maximises $L_{\boldsymbol{Y}}(\theta; \boldsymbol{y})$ as a function of θ.

Some notes on maximum-likelihood estimates and maximum-likelihood estimators follow.

1. The notation is rather unfortunate; $\hat{\theta}$ defined above is the maximum-likelihood estimate, which is just a number. However, by convention, $\hat{\theta}$ is also used to denote the corresponding statistic. We attempt to clarify below.
   - If $\hat{\theta} = h(\boldsymbol{y})$, then $\hat{\theta}$ is the maximum-likelihood estimate of θ. This is the point estimate (a number) that we calculate in practice given an observed sample $\boldsymbol{y}$.
   - If $\hat{\theta} = h(\boldsymbol{Y})$, then $\hat{\theta}$ is the maximum-likelihood estimator of θ. This is a function of the sample. It is the theoretical quantity that we consider when determining the properties of maximum likelihood.

   It should be clear from the context when $\hat{\theta}$ is an estimate and when it is an estimator.
2. Let Θ be the parameter space. The maximum-likelihood estimate of θ given an observed sample $\boldsymbol{y}$ is the value $\hat{\theta}$ such that
$$L_{\boldsymbol{Y}}(\hat{\theta}; \boldsymbol{y}) = \sup_{\theta \in \Theta} L_{\boldsymbol{Y}}(\theta; \boldsymbol{y}) .$$
   Notice that, if Θ is an open set, this definition admits the possibility that $\hat{\theta}$ is not a member of Θ. We will deal with regular cases in which this issue does not arise.
3. The definition of the maximum-likelihood estimate ensures that no other possible parameter value has a larger likelihood. If $\hat{\theta}$ is the maximum-likelihood estimate based on an observed sample $\boldsymbol{y}$, then
$$L_{\boldsymbol{Y}}(\theta; \boldsymbol{y}) \leq L_{\boldsymbol{Y}}(\hat{\theta}; \boldsymbol{y}) ,$$
   for all $\theta \in \Theta$.