and the estimated standard error is

$$\widehat{\mathsf{se}}_{\text{boot}} = \sqrt{\frac{\sum_{b=1}^{B}(\widehat{\tau}_b^* - \widehat{\tau})^2}{B}}. \quad \blacksquare$$

The bootstrap is much easier than the delta method. On the other hand, the delta method has the advantage that it gives a closed form expression for the standard error.

9.12 Checking Assumptions

If we assume the data come from a parametric model, then it is a good idea to check that assumption. One possibility is to check the assumptions informally by inspecting plots of the data. For example, if a histogram of the data looks very bimodal, then the assumption of Normality might be questionable. A formal way to test a parametric model is to use a **goodness-of-fit test.** See Section 10.8.

9.13 Appendix

9.13.1 Proofs

PROOF OF THEOREM 9.13. Since $\widehat{\theta}_n$ maximizes $M_n(\theta)$, we have $M_n(\widehat{\theta}_n) \geq M_n(\theta_\star)$. Hence,

$$\begin{aligned} M(\theta_\star) - M(\widehat{\theta}_n) &= M_n(\theta_\star) - M(\widehat{\theta}_n) + M(\theta_\star) - M_n(\theta_\star) \\ &\leq M_n(\widehat{\theta}_n) - M(\widehat{\theta}_n) + M(\theta_\star) - M_n(\theta_\star) \\ &\leq \sup_\theta |M_n(\theta) - M(\theta)| + M(\theta_\star) - M_n(\theta_\star) \\ &\xrightarrow{\text{P}} 0 \end{aligned}$$

where the last line follows from (9.7). It follows that, for any $\delta > 0$,

$$\mathbb{P}\left(M(\widehat{\theta}_n) < M(\theta_\star) - \delta\right) \to 0.$$

Pick any $\epsilon > 0$. By (9.8), there exists $\delta > 0$ such that $|\theta - \theta_\star| \geq \epsilon$ implies that $M(\theta) < M(\theta_\star) - \delta$. Hence,

$$\mathbb{P}(|\widehat{\theta}_n - \theta_\star| > \epsilon) \leq \mathbb{P}\left(M(\widehat{\theta}_n) < M(\theta_\star) - \delta\right) \to 0. \quad \blacksquare$$

Next we want to prove Theorem 9.18. First we need a lemma.

9.31 Lemma. *The score function satisfies*

$$\mathbb{E}_\theta\left[s(X;\theta)\right] = 0.$$

PROOF. Note that $1 = \int f(x;\theta)dx$. Differentiate both sides of this equation to conclude that

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\int f(x;\theta)dx = \int \frac{\partial}{\partial\theta} f(x;\theta)dx \\ &= \int \frac{\frac{\partial f(x;\theta)}{\partial\theta}}{f(x;\theta)} f(x;\theta)dx = \int \frac{\partial \log f(x;\theta)}{\partial\theta} f(x;\theta)dx \\ &= \int s(x;\theta)f(x;\theta)dx = \mathbb{E}_\theta s(X;\theta). \quad \blacksquare \end{aligned}$$

PROOF OF THEOREM 9.18. Let $\ell(\theta) = \log \mathcal{L}(\theta)$. Then,

$$0 = \ell'(\widehat{\theta}) \approx \ell'(\theta) + (\widehat{\theta} - \theta)\ell''(\theta).$$

Rearrange the above equation to get $\widehat{\theta} - \theta = -\ell'(\theta)/\ell''(\theta)$ or, in other words,

$$\sqrt{n}(\widehat{\theta} - \theta) = \frac{\frac{1}{\sqrt{n}}\ell'(\theta)}{-\frac{1}{n}\ell''(\theta)} \equiv \frac{\text{TOP}}{\text{BOTTOM}}.$$

Let $Y_i = \partial \log f(X_i;\theta)/\partial\theta$. Recall that $\mathbb{E}(Y_i) = 0$ from the previous lemma and also $\mathbb{V}(Y_i) = I(\theta)$. Hence,

$$\text{TOP} = n^{-1/2}\sum_i Y_i = \sqrt{n}\overline{Y} = \sqrt{n}(\overline{Y} - 0) \rightsquigarrow W \sim N(0, I(\theta))$$

by the central limit theorem. Let $A_i = -\partial^2 \log f(X_i;\theta)/\partial\theta^2$. Then $\mathbb{E}(A_i) = I(\theta)$ and

$$\text{BOTTOM} = \overline{A} \xrightarrow{\text{P}} I(\theta)$$

by the law of large numbers. Apply Theorem 5.5 part (e), to conclude that

$$\sqrt{n}(\widehat{\theta} - \theta) \rightsquigarrow \frac{W}{I(\theta)} \stackrel{d}{=} N\left(0, \frac{1}{I(\theta)}\right).$$

Assuming that $I(\theta)$ is a continuous function of θ, it follows that $I(\widehat{\theta}_n) \xrightarrow{\text{P}} I(\theta)$. Now

$$\begin{aligned} \frac{\widehat{\theta}_n - \theta}{\widehat{\text{se}}} &= \sqrt{n}I^{1/2}(\widehat{\theta}_n)(\widehat{\theta}_n - \theta) \\ &= \left\{\sqrt{n}I^{1/2}(\theta)(\widehat{\theta}_n - \theta)\right\}\sqrt{\frac{I(\widehat{\theta}_n)}{I(\theta)}}. \end{aligned}$$

The first term tends in distribution to N(0,1). The second term tends in probability to 1. The result follows from Theorem 5.5 part (e). ■

Outline of Proof of Theorem 9.24. Write

$$\widehat{\tau}_n = g(\widehat{\theta}_n) \approx g(\theta) + (\widehat{\theta}_n - \theta)g'(\theta) = \tau + (\widehat{\theta}_n - \theta)g'(\theta).$$

Thus,

$$\sqrt{n}(\widehat{\tau}_n - \tau) \approx \sqrt{n}(\widehat{\theta}_n - \theta)g'(\theta),$$

and hence

$$\frac{\sqrt{nI(\theta)}(\widehat{\tau}_n - \tau)}{g'(\theta)} \approx \sqrt{nI(\theta)}(\widehat{\theta}_n - \theta).$$

Theorem 9.18 tells us that the right-hand side tends in distribution to a N(0,1). Hence,

$$\frac{\sqrt{nI(\theta)}(\widehat{\tau}_n - \tau)}{g'(\theta)} \rightsquigarrow N(0,1)$$

or, in other words,

$$\widehat{\tau}_n \approx N\left(\tau, \mathsf{se}^2(\widehat{\tau}_n)\right),$$

where

$$\mathsf{se}^2(\widehat{\tau}_n) = \frac{(g'(\theta))^2}{nI(\theta)}.$$

The result remains true if we substitute $\widehat{\theta}_n$ for θ by Theorem 5.5 part (e). ■

9.13.2 Sufficiency

A **statistic** is a function $T(X^n)$ of the data. A sufficient statistic is a statistic that contains all the information in the data. To make this more formal, we need some definitions.

9.32 Definition. *Write $x^n \leftrightarrow y^n$ if $f(x^n;\theta) = c\, f(y^n;\theta)$ for some constant c that might depend on x^n and y^n but not θ. A statistic $T(x^n)$ is* **sufficient** *if $T(x^n) \leftrightarrow T(y^n)$ implies that $x^n \leftrightarrow y^n$.*

Notice that if $x^n \leftrightarrow y^n$, then the likelihood function based on x^n has the same shape as the likelihood function based on y^n. Roughly speaking, a statistic is sufficient if we can calculate the likelihood function knowing only $T(X^n)$.

9.33 Example. Let $X_1, \ldots, X_n \sim$ Bernoulli(p). Then $\mathcal{L}(p) = p^S(1-p)^{n-S}$ where $S = \sum_i X_i$, so S is sufficient. ■

9.34 Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma)$ and let $T = (\overline{X}, S)$. Then

$$f(X^n; \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{nS^2}{2\sigma^2}\right\} \exp\left\{-\frac{n(\overline{X}-\mu)^2}{2\sigma^2}\right\}$$

where S^2 is the sample variance. The last expression depends on the data only through T and therefore, $T = (\overline{X}, S)$ is a sufficient statistic. Note that $U = (17\,\overline{X}, S)$ is also a sufficient statistic. If I tell you the value of U then you can easily figure out T and then compute the likelihood. Sufficient statistics are far from unique. Consider the following statistics for the $N(\mu, \sigma^2)$ model:

$$\begin{aligned} T_1(X^n) &= (X_1, \ldots, X_n) \\ T_2(X^n) &= (\overline{X}, S) \\ T_3(X^n) &= \overline{X} \\ T_4(X^n) &= (\overline{X}, S, X_3). \end{aligned}$$

The first statistic is just the whole data set. This is sufficient. The second is also sufficient as we proved above. The third is not sufficient: you can't compute $\mathcal{L}(\mu, \sigma)$ if I only tell you $\overline{X}$. The fourth statistic T_4 is sufficient. The statistics T_1 and T_4 are sufficient but they contain redundant information. Intuitively, there is a sense in which T_2 is a "more concise" sufficient statistic than either T_1 or T_4. We can express this formally by noting that T_2 is a function of T_1 and similarly, T_2 is a function of T_4. For example, $T_2 = g(T_4)$ where $g(a_1, a_2, a_3) = (a_1, a_2)$. ■

9.35 Definition. *A statistic T is* **minimal sufficient** *if (i) it is sufficient; and (ii) it is a function of every other sufficient statistic.*

9.36 Theorem. *T is minimal sufficient if the following is true:*

$$T(x^n) = T(y^n) \text{ if and only if } x^n \leftrightarrow y^n.$$

A statistic induces a partition on the set of outcomes. We can think of sufficiency in terms of these partitions.

9.37 Example. Let $X_1, X_2 \sim Bernoulli(\theta)$. Let $V = X_1$, $T = \sum_i X_i$ and $U = (T, X_1)$. Here is the set of outcomes and the statistics:

X_1	X_2	V	T	U
0	0	0	0	(0,0)
0	1	0	1	(1,0)
1	0	1	1	(1,1)
1	1	1	2	(2,1)

The partitions induced by these statistics are:

$$\begin{aligned} V &\longrightarrow \{(0,0),(0,1)\},\ \{(1,0),(1,1)\} \\ T &\longrightarrow \{(0,0)\},\ \{(0,1),(1,0)\},\ \{(1,1)\} \\ U &\longrightarrow \{(0,0)\},\ \{(0,1)\},\ \{(1,0)\},\ \{(1,1)\}. \end{aligned}$$

Then V is not sufficient but T and U are sufficient. T is minimal sufficient; U is not minimal since if $x^n = (1,0)$ and $y^n = (0,1)$, then $x^n \leftrightarrow y^n$ yet $U(x^n) \neq U(y^n)$. The statistic $W = 17\,T$ generates the same partition as T. It is also minimal sufficient. ■

9.38 Example. For a $N(\mu, \sigma^2)$ model, $T = (\overline{X}, S)$ is a minimal sufficient statistic. For the Bernoulli model, $T = \sum_i X_i$ is a minimal sufficient statistic. For the Poisson model, $T = \sum_i X_i$ is a minimal sufficient statistic. Check that $T = (\sum_i X_i, X_1)$ is sufficient but not minimal sufficient. Check that $T = X_1$ is not sufficient. ■

I did not give the usual definition of sufficiency. The usual definition is this: T is sufficient if the distribution of X^n given $T(X^n) = t$ does not depend on θ. In other words, T is sufficient if $f(x_1, \ldots, x_n | t; \theta) = h(x_1, \ldots, x_n, t)$ where h is some function that does not depend on θ.

9.39 Example. Two coin flips. Let $X = (X_1, X_2) \sim \text{Bernoulli}(p)$. Then $T = X_1 + X_2$ is sufficient. To see this, we need the distribution of (X_1, X_2) given $T = t$. Since T can take 3 possible values, there are 3 conditional distributions to check. They are: (i) the distribution of (X_1, X_2) given $T = 0$:

$$P(X_1 = 0, X_2 = 0 | t = 0) = 1, P(X_1 = 0, X_2 = 1 | t = 0) = 0,$$

$$P(X_1 = 1, X_2 = 0 | t = 0) = 0, P(X_1 = 1, X_2 = 1 | t = 0) = 0;$$

(ii) the distribution of (X_1, X_2) given $T = 1$:

$$P(X_1 = 0, X_2 = 0 | t = 1) = 0, P(X_1 = 0, X_2 = 1 | t = 1) = \frac{1}{2},$$

$$P(X_1 = 1, X_2 = 0 | t = 1) = \frac{1}{2}, P(X_1 = 1, X_2 = 1 | t = 1) = 0; \text{ and}$$

(iii) the distribution of (X_1, X_2) given $T = 2$:

$$P(X_1 = 0, X_2 = 0 | t = 2) = 0, P(X_1 = 0, X_2 = 1 | t = 2) = 0,$$

$$P(X_1 = 1, X_2 = 0 | t = 2) = 0, P(X_1 = 1, X_2 = 1 | t = 2) = 1.$$

None of these depend on the parameter p. Thus, the distribution of $X_1, X_2 | T$ does not depend on θ, so T is sufficient. ■

9.40 Theorem (Factorization Theorem). *T is sufficient if and only if there are functions $g(t,\theta)$ and $h(x)$ such that $f(x^n;\theta) = g(t(x^n),\theta)h(x^n)$.*

9.41 Example. Return to the two coin flips. Let $t = x_1 + x_2$. Then

$$\begin{aligned} f(x_1,x_2;\theta) &= f(x_1;\theta)f(x_2;\theta) \\ &= \theta^{x_1}(1-\theta)^{1-x_1}\theta^{x_2}(1-\theta)^{1-x_2} \\ &= g(t,\theta)h(x_1,x_2) \end{aligned}$$

where $g(t,\theta) = \theta^t(1-\theta)^{2-t}$ and $h(x_1,x_2) = 1$. Therefore, $T = X_1 + X_2$ is sufficient. ■

Now we discuss an implication of sufficiency in point estimation. Let $\widehat{\theta}$ be an estimator of θ. The Rao-Blackwell theorem says that an estimator should only depend on the sufficient statistic, otherwise it can be improved. Let $R(\theta,\widehat{\theta}) = \mathbb{E}_\theta(\theta - \widehat{\theta})^2$ denote the MSE of the estimator.

9.42 Theorem (Rao-Blackwell). *Let $\widehat{\theta}$ be an estimator and let T be a sufficient statistic. Define a new estimator by*

$$\widetilde{\theta} = \mathbb{E}(\widehat{\theta}|T).$$

Then, for every θ, $R(\theta,\widetilde{\theta}) \leq R(\theta,\widehat{\theta})$.

9.43 Example. Consider flipping a coin twice. Let $\widehat{\theta} = X_1$. This is a well-defined (and unbiased) estimator. But it is not a function of the sufficient statistic $T = X_1 + X_2$. However, note that $\widetilde{\theta} = \mathbb{E}(X_1|T) = (X_1 + X_2)/2$. By the Rao-Blackwell Theorem, $\widetilde{\theta}$ has MSE at least as small as $\widehat{\theta} = X_1$. The same applies with n coin flips. Again define $\widehat{\theta} = X_1$ and $T = \sum_i X_i$. Then $\widetilde{\theta} = \mathbb{E}(X_1|T) = n^{-1}\sum_i X_i$ has improved MSE. ■

9.13.3 Exponential Families

Most of the parametric models we have studied so far are special cases of a general class of models called exponential families. We say that $\{f(x;\theta) : \theta \in \Theta\}$ is a **one-parameter exponential family** if there are functions $\eta(\theta)$, $B(\theta)$, $T(x)$ and $h(x)$ such that

$$f(x;\theta) = h(x)e^{\eta(\theta)T(x) - B(\theta)}.$$

It is easy to see that $T(X)$ is sufficient. We call T the **natural sufficient statistic**.

9.44 Example. Let $X \sim \text{Poisson}(\theta)$. Then

$$f(x;\theta) = \frac{\theta^x e^{-\theta}}{x!} = \frac{1}{x!} e^{x \log \theta - \theta}$$

and hence, this is an exponential family with $\eta(\theta) = \log\theta$, $B(\theta) = \theta$, $T(x) = x$, $h(x) = 1/x!$. ■

9.45 Example. Let $X \sim \text{Binomial}(n, \theta)$. Then

$$f(x;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} = \binom{n}{x}\exp\left\{x\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right\}.$$

In this case,

$$\eta(\theta) = \log\left(\frac{\theta}{1-\theta}\right), B(\theta) = -n\log(\theta)$$

and

$$T(x) = x, h(x) = \binom{n}{x}.$$

■

We can rewrite an exponential family as

$$f(x;\eta) = h(x)e^{\eta T(x) - A(\eta)}$$

where $\eta = \eta(\theta)$ is called the **natural parameter** and

$$A(\eta) = \log \int h(x)e^{\eta T(x)}dx.$$

For example a Poisson can be written as $f(x;\eta) = e^{\eta x - e^\eta}/x!$ where the natural parameter is $\eta = \log\theta$.

Let $X_1, \ldots, X_n$ be IID from an exponential family. Then $f(x^n;\theta)$ is an exponential family:

$$f(x^n;\theta) = h_n(x^n)h_n(x^n)e^{\eta(\theta)T_n(x^n) - B_n(\theta)}$$

where $h_n(x^n) = \prod_i h(x_i)$, $T_n(x^n) = \sum_i T(x_i)$ and $B_n(\theta) = nB(\theta)$. This implies that $\sum_i T(X_i)$ is sufficient.

9.46 Example. Let $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$. Then

$$f(x^n;\theta) = \frac{1}{\theta^n} I(x_{(n)} \leq \theta)$$

where I is 1 if the term inside the brackets is true and 0 otherwise, and $x_{(n)} = max\{x_1, \ldots, x_n\}$. Thus $T(X^n) = max\{X_1, \ldots, X_n\}$ is sufficient. But since $T(X^n) \neq \sum_i T(X_i)$, this cannot be an exponential family. ■

9.47 Theorem. *Let X have density in an exponential family. Then,*

$$\mathbb{E}(T(X)) = A'(\eta),\ \mathbb{V}(T(X)) = A''(\eta).$$

If $\theta = (\theta_1, \ldots, \theta_k)$ is a vector, then we say that $f(x;\theta)$ has exponential family form if

$$f(x;\theta) = h(x) \exp\left\{ \sum_{j=1}^{k} \eta_j(\theta) T_j(x) - B(\theta) \right\}.$$

Again, $T = (T_1, \ldots, T_k)$ is sufficient. An IID sample of size n also has exponential form with sufficient statistic $(\sum_i T_1(X_i), \ldots, \sum_i T_k(X_i))$.

9.48 Example. Consider the normal family with $\theta = (\mu, \sigma)$. Now,

$$f(x;\theta) = \exp\left\{ \frac{\mu}{\sigma^2} x - \frac{x^2}{2\sigma^2} - \frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2) \right) \right\}.$$

This is exponential with

$$\eta_1(\theta) = \frac{\mu}{\sigma^2},\ T_1(x) = x$$

$$\eta_2(\theta) = -\frac{1}{2\sigma^2},\ T_2(x) = x^2$$

$$B(\theta) = \frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2) \right),\ h(x) = 1.$$

Hence, with n IID samples, $(\sum_i X_i, \sum_i X_i^2)$ is sufficient. ∎

As before we can write an exponential family as

$$f(x;\eta) = h(x) \exp\left\{ T^T(x)\eta - A(\eta) \right\},$$

where $A(\eta) = \log \int h(x) e^{T^T(x)\eta} dx$. It can be shown that

$$\mathbb{E}(T(X)) = \dot{A}(\eta)\ \ \mathbb{V}(T(X)) = \ddot{A}(\eta),$$

where the first expression is the vector of partial derivatives and the second is the matrix of second derivatives.

9.13.4 Computing Maximum Likelihood Estimates

In some cases we can find the MLE $\widehat{\theta}$ analytically. More often, we need to find the MLE by numerical methods. We will briefly discuss two commonly

used methods: (i) Newton-Raphson, and (ii) the EM algorithm. Both are iterative methods that produce a sequence of values $\theta^0, \theta^1, \ldots$ that, under ideal conditions, converge to the MLE $\widehat{\theta}$. In each case, it is helpful to use a good starting value θ^0. Often, the method of moments estimator is a good starting value.

NEWTON-RAPHSON. To motivate Newton-Raphson, let's expand the derivative of the log-likelihood around θ^j:

$$0 = \ell'(\widehat{\theta}) \approx \ell'(\theta^j) + (\widehat{\theta} - \theta^j)\ell''(\theta^j).$$

Solving for $\widehat{\theta}$ gives

$$\widehat{\theta} \approx \theta^j - \frac{\ell'(\theta^j)}{\ell''(\theta^j)}.$$

This suggests the following iterative scheme:

$$\widehat{\theta}^{j+1} = \theta^j - \frac{\ell'(\theta^j)}{\ell''(\theta^j)}.$$

In the multiparameter case, the mle $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ is a vector and the method becomes

$$\widehat{\theta}^{j+1} = \theta^j - H^{-1}\ell'(\theta^j)$$

where $\ell'(\theta^j)$ is the vector of first derivatives and H is the matrix of second derivatives of the log-likelihood.

THE EM ALGORITHM. The letters EM stand for Expectation-Maximization. The idea is to iterate between taking an expectation then maximizing. Suppose we have data Y whose density $f(y;\theta)$ leads to a log-likelihood that is hard to maximize. But suppose we can find another random variable Z such that $f(y;\theta) = \int f(y,z;\theta)\,dz$ and such that the likelihood based on $f(y,z;\theta)$ is easy to maximize. In other words, the model of interest is the marginal of a model with a simpler likelihood. In this case, we call Y the observed data and Z the hidden (or latent or missing) data. If we could just "fill in" the missing data, we would have an easy problem. Conceptually, the EM algorithm works by filling in the missing data, maximizing the log-likelihood, and iterating.

9.49 Example (Mixture of Normals)**.** Sometimes it is reasonable to assume that the distribution of the data is a mixture of two normals. Think of heights of people being a mixture of men and women's heights. Let $\phi(y;\mu,\sigma)$ denote a normal density with mean μ and standard deviation σ. The density of a mixture of two Normals is

$$f(y;\theta) = (1-p)\phi(y;\mu_0,\sigma_0) + p\phi(y;\mu_1,\sigma_1).$$

The idea is that an observation is drawn from the first normal with probability p and the second with probability $1-p$. However, we don't know which Normal it was drawn from. The parameters are $\theta = (\mu_0, \sigma_0, \mu_1, \sigma_1, p)$. The likelihood function is

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \left[(1-p)\phi(y_i; \mu_0, \sigma_0) + p\phi(y_i; \mu_1, \sigma_1)\right].$$

Maximizing this function over the five parameters is hard. Imaging that we were given extra information telling us which of the two normals every observation came from. These "complete" data are of the form $(Y_1, Z_1), \ldots, (Y_n, Z_n)$, where $Z_i = 0$ represents the first normal and $Z_i = 1$ represents the second. Note that $\mathbb{P}(Z_i = 1) = p$. We shall soon see that the likelihood for the complete data $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ is much simpler than the likelihood for the observed data $Y_1, \ldots, Y_n$. ■

Now we describe the EM algorithm.

The EM Algorithm

(0) Pick a starting value θ^0. Now for $j = 1, 2, \ldots$, repeat steps 1 and 2 below:

(1) (The E-step): Calculate

$$J(\theta|\theta^j) = \mathbb{E}_{\theta^j}\left(\log \frac{f(Y^n, Z^n; \theta)}{f(Y^n, Z^n; \theta^j)} \;\middle|\; Y^n = y^n\right).$$

The expectation is over the missing data Z^n treating θ^i and the observed data Y^n as fixed.

(2) Find θ^{j+1} to maximize $J(\theta|\theta^j)$.

We now show that the EM algorithm always increases the likelihood, that is, $\mathcal{L}(\theta^{j+1}) \geq \mathcal{L}(\theta^j)$. Note that

$$\begin{aligned} J(\theta^{j+1}|\theta^j) &= \mathbb{E}_{\theta^j}\left(\log \frac{f(Y^n, Z^n; \theta^{j+1})}{f(Y^n, Z^n; \theta^j)} \;\middle|\; Y^n = y^n\right) \\ &= \log \frac{f(y^n; \theta^{j+1})}{f(y^n; \theta^j)} + \mathbb{E}_{\theta^j}\left(\log \frac{f(Z^n|Y^n; \theta^{j+1})}{f(Z^n|Y^n; \theta^j)} \;\middle|\; Y^n = y^n\right) \end{aligned}$$

and hence

$$\frac{\mathcal{L}(\theta^{j+1})}{\mathcal{L}(\theta^j)} = \log \frac{f(y^n; \theta^{j+1})}{f(y^n; \theta^j)}$$

$$
\begin{aligned}
&= J(\theta^{j+1}|\theta^j) - \mathbb{E}_{\theta^j}\left(\log \frac{f(Z^n|Y^n;\theta^{j+1})}{f(Z^n|Y^n;\theta^j)} \;\middle|\; Y^n = y^n\right) \\
&= J(\theta^{j+1}|\theta^j) + K(f_j, f_{j+1})
\end{aligned}
$$

where $f_j = f(y^n;\theta^j)$ and $f_{j+1} = f(y^n;\theta^{j+1})$ and $K(f,g) = \int f(x)\log(f(x)/g(x))\,dx$ is the Kullback-Leibler distance. Now, θ^{j+1} was chosen to maximize $J(\theta|\theta^j)$. Hence, $J(\theta^{j+1}|\theta^j) \geq J(\theta^j|\theta^j) = 0$. Also, by the properties of Kullback-Leibler divergence, $K(f_j, f_{j+1}) \geq 0$. Hence, $\mathcal{L}(\theta^{j+1}) \geq \mathcal{L}(\theta^j)$ as claimed.

9.50 Example (Continuation of Example 9.49)**.** Consider again the mixture of two normals but, for simplicity assume that $p = 1/2$, $\sigma_1 = \sigma_2 = 1$. The density is

$$f(y;\mu_1,\mu_2) = \frac{1}{2}\phi(y;\mu_0,1) + \frac{1}{2}\phi(y;\mu_1,1).$$

Directly maximizing the likelihood is hard. Introduce latent variables $Z_1,\ldots,Z_n$ where $Z_i = 0$ if Y_i is from $\phi(y;\mu_0,1)$, and $Z_i = 1$ if Y_i is from $\phi(y;\mu_1,1)$, $\mathbb{P}(Z_i = 1) = P(Z_i = 0) = 1/2$, $f(y_i|Z_i = 0) = \phi(y;\mu_0,1)$ and $f(y_i|Z_i = 1) = \phi(y;\mu_1,1)$. So $f(y) = \sum_{z=0}^{1} f(y,z)$ where we have dropped the parameters from the density to avoid notational overload. We can write

$$f(z,y) = f(z)f(y|z) = \frac{1}{2}\phi(y;\mu_0,1)^{1-z}\phi(y;\mu_1,1)^z.$$

Hence, the complete likelihood is

$$\prod_{i=1}^{n} \phi(y_i;\mu_0,1)^{1-z_i}\phi(y_i;\mu_1,1)^{z_i}.$$

The complete log-likelihood is then

$$\widetilde{\ell} = -\frac{1}{2}\sum_{i=1}^{n}(1-z_i)(y_i-\mu_0) - \frac{1}{2}\sum_{i=1}^{n} z_i(y_i-\mu_1).$$

And so

$$J(\theta|\theta^j) = -\frac{1}{2}\sum_{i=1}^{n}(1-\mathbb{E}(Z_i|y^n,\theta^j))(y_i-\mu_0) - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}(Z_i|y^n,\theta^j))(y_i-\mu_1).$$

Since Z_i is binary, $\mathbb{E}(Z_i|y^n,\theta^j) = \mathbb{P}(Z_i = 1|y^n,\theta^j)$ and, by Bayes' theorem,

$$
\begin{aligned}
\mathbb{P}(Z_i = 1|y^n,\theta^i) &= \frac{f(y^n|Z_i = 1;\theta^j)\mathbb{P}(Z_i = 1)}{f(y^n|Z_i = 1;\theta^j)\mathbb{P}(Z_i = 1) + f(y^n|Z_i = 0;\theta^j)\mathbb{P}(Z_i = 0)} \\
&= \frac{\phi(y_i;\mu_1^j,1)\frac{1}{2}}{\phi(y_i;\mu_1^j,1)\frac{1}{2} + \phi(y_i;\mu_0^j,1)\frac{1}{2}} \\
&= \frac{\phi(y_i;\mu_1^j,1)}{\phi(y_i;\mu_1^j,1) + \phi(y_i;\mu_0^j,1)} \\
&= \tau(i).
\end{aligned}
$$

Take the derivative of $J(\theta|\theta^j)$ with respect to μ_1 and μ_2, set them equal to 0 to get

$$\widehat{\mu}_1^{j+1} = \frac{\sum_{i=1}^n \tau_i y_i}{\sum_{i=1}^n \tau_i}$$

and

$$\widehat{\mu}_0^{j+1} = \frac{\sum_{i=1}^n (1-\tau_i) y_i}{\sum_{i=1}^n (1-\tau_i)}.$$

We then recompute τ_i using $\widehat{\mu}_1^{j+1}$ and $\widehat{\mu}_0^{j+1}$ and iterate. ■

9.14 Exercises

1. Let $X_1, \ldots, X_n \sim \text{Gamma}(\alpha, \beta)$. Find the method of moments estimator for α and β.

2. Let $X_1, \ldots, X_n \sim \text{Uniform}(a, b)$ where a and b are unknown parameters and $a < b$.

   (a) Find the method of moments estimators for a and b.

   (b) Find the MLE $\widehat{a}$ and $\widehat{b}$.

   (c) Let $\tau = \int x\, dF(x)$. Find the MLE of τ.

   (d) Let $\widehat{\tau}$ be the MLE of τ. Let $\widetilde{\tau}$ be the nonparametric plug-in estimator of $\tau = \int x\, dF(x)$. Suppose that $a = 1$, $b = 3$, and $n = 10$. Find the MSE of $\widehat{\tau}$ by simulation. Find the MSE of $\widetilde{\tau}$ analytically. Compare.

3. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Let τ be the .95 percentile, i.e. $\mathbb{P}(X < \tau) = .95$.

   (a) Find the MLE of τ.

   (b) Find an expression for an approximate $1-\alpha$ confidence interval for τ.

   (c) Suppose the data are:

```
3.23 -2.50  1.88 -0.68  4.43  0.17
1.03 -0.07 -0.01  0.76  1.76  3.18
0.33 -0.31  0.30 -0.61  1.52  5.43
1.54  2.28  0.42  2.33 -1.03  4.00
0.39
```

   Find the MLE $\widehat{\tau}$. Find the standard error using the delta method. Find the standard error using the parametric bootstrap.

4. Let $X_1, \ldots, X_n \sim \text{Uniform}(0, \theta)$. Show that the MLE is consistent. Hint: Let $Y = \max\{X_1, ..., X_n\}$. For any c, $\mathbb{P}(Y < c) = \mathbb{P}(X_1 < c, X_2 < c, ..., X_n < c) = \mathbb{P}(X_1 < c)\mathbb{P}(X_2 < c)...\mathbb{P}(X_n < c)$.

5. Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. Find the method of moments estimator, the maximum likelihood estimator and the Fisher information $I(\lambda)$.

6. Let $X_1, ..., X_n \sim N(\theta, 1)$. Define

$$Y_i = \begin{cases} 1 & \text{if } X_i > 0 \\ 0 & \text{if } X_i \leq 0. \end{cases}$$

   Let $\psi = \mathbb{P}(Y_1 = 1)$.

   (a) Find the maximum likelihood estimator $\widehat{\psi}$ of ψ.

   (b) Find an approximate 95 percent confidence interval for ψ.

   (c) Define $\widetilde{\psi} = (1/n)\sum_i Y_i$. Show that $\widetilde{\psi}$ is a consistent estimator of ψ.

   (d) Compute the asymptotic relative efficiency of $\widetilde{\psi}$ to $\widehat{\psi}$. Hint: Use the delta method to get the standard error of the MLE. Then compute the standard error (i.e. the standard deviation) of $\widetilde{\psi}$.

   (e) Suppose that the data are not really normal. Show that $\widehat{\psi}$ is not consistent. What, if anything, does $\widehat{\psi}$ converge to?

7. (Comparing two treatments.) n_1 people are given treatment 1 and n_2 people are given treatment 2. Let X_1 be the number of people on treatment 1 who respond favorably to the treatment and let X_2 be the number of people on treatment 2 who respond favorably. Assume that $X_1 \sim \text{Binomial}(n_1, p_1)$ $X_2 \sim \text{Binomial}(n_2, p_2)$. Let $\psi = p_1 - p_2$.

   (a) Find the MLE $\widehat{\psi}$ for ψ.

   (b) Find the Fisher information matrix $I(p_1, p_2)$.

   (c) Use the multiparameter delta method to find the asymptotic standard error of $\widehat{\psi}$.

   (d) Suppose that $n_1 = n_2 = 200$, $X_1 = 160$ and $X_2 = 148$. Find $\widehat{\psi}$. Find an approximate 90 percent confidence interval for ψ using (i) the delta method and (ii) the parametric bootstrap.

8. Find the Fisher information matrix for Example 9.29.

9. Let $X_1, ..., X_n \sim \text{Normal}(\mu, 1)$. Let $\theta = e^\mu$ and let $\widehat{\theta} = e^{\overline{X}}$ be the MLE. Create a data set (using $\mu = 5$) consisting of n=100 observations.

(a) Use the delta method to get $\widehat{\mathsf{se}}$ and a 95 percent confidence interval for θ. Use the parametric bootstrap to get $\widehat{\mathsf{se}}$ and 95 percent confidence interval for θ. Use the nonparametric bootstrap to get $\widehat{\mathsf{se}}$ and 95 percent confidence interval for θ. Compare your answers.

(b) Plot a histogram of the bootstrap replications for the parametric and nonparametric bootstraps. These are estimates of the distribution of $\widehat{\theta}$. The delta method also gives an approximation to this distribution namely, Normal$(\widehat{\theta}, \mathsf{se}^2)$. Compare these to the true sampling distribution of $\widehat{\theta}$ (which you can get by simulation). Which approximation — parametric bootstrap, bootstrap, or delta method — is closer to the true distribution?

10. Let $X_1, ..., X_n \sim \text{Uniform}(0, \theta)$. The MLE is $\widehat{\theta} = X_{(n)} = \max\{X_1, ..., X_n\}$. Generate a dataset of size 50 with $\theta = 1$.

(a) Find the distribution of $\widehat{\theta}$ analytically. Compare the true distribution of $\widehat{\theta}$ to the histograms from the parametric and nonparametric bootstraps.

(b) This is a case where the nonparametric bootstrap does very poorly. Show that for the parametric bootstrap $\mathbb{P}(\widehat{\theta}^* = \widehat{\theta}) = 0$, but for the nonparametric bootstrap $\mathbb{P}(\widehat{\theta}^* = \widehat{\theta}) \approx .632$. Hint: show that, $\mathbb{P}(\widehat{\theta}^* = \widehat{\theta}) = 1 - (1 - (1/n))^n$ then take the limit as n gets large. What is the implication of this?

10
Hypothesis Testing and p-values

Suppose we want to know if exposure to asbestos is associated with lung disease. We take some rats and randomly divide them into two groups. We expose one group to asbestos and leave the second group unexposed. Then we compare the disease rate in the two groups. Consider the following two hypotheses:

The Null Hypothesis: The disease rate is the same in the two groups.

The Alternative Hypothesis: The disease rate is not the same in the two groups.

If the exposed group has a much higher rate of disease than the unexposed group then we will reject the null hypothesis and conclude that the evidence favors the alternative hypothesis. This is an example of hypothesis testing.

More formally, suppose that we partition the parameter space Θ into two disjoint sets Θ_0 and Θ_1 and that we wish to test

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1. \tag{10.1}$$

We call H_0 the **null hypothesis** and H_1 the **alternative hypothesis**.

Let X be a random variable and let $\mathcal{X}$ be the range of X. We test a hypothesis by finding an appropriate subset of outcomes $R \subset \mathcal{X}$ called the **rejection**

	Retain Null	Reject Null
H_0 true	$\surd$	type I error
H_1 true	type II error	$\surd$

TABLE 10.1. Summary of outcomes of hypothesis testing.

region. If $X \in R$ we reject the null hypothesis, otherwise, we do not reject the null hypothesis:

$$\begin{aligned} X \in R &\implies \text{reject } H_0 \\ X \notin R &\implies \text{retain (do not reject) } H_0 \end{aligned}$$

Usually, the rejection region R is of the form

$$R = \Big\{ x : \ T(x) > c \Big\} \tag{10.2}$$

where T is a **test statistic** and c is a **critical value**. The problem in hypothesis testing is to find an appropriate test statistic T and an appropriate critical value c.

Warning! There is a tendency to use hypothesis testing methods even when they are not appropriate. Often, estimation and confidence intervals are better tools. Use hypothesis testing only when you want to test a well-defined hypothesis.

Hypothesis testing is like a legal trial. We assume someone is innocent unless the evidence strongly suggests that he is guilty. Similarly, we retain H_0 unless there is strong evidence to reject H_0. There are two types of errors we can make. Rejecting H_0 when H_0 is true is called a **type I error**. Retaining H_0 when H_1 is true is called a **type II error**. The possible outcomes for hypothesis testing are summarized in Tab. 10.1.

10.1 Definition. *The* **power function** *of a test with rejection region R is defined by*

$$\beta(\theta) = \mathbb{P}_\theta(X \in R). \tag{10.3}$$

The **size** *of a test is defined to be*

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta). \tag{10.4}$$

A test is said to have **level** α *if its size is less than or equal to α.*

A hypothesis of the form $\theta = \theta_0$ is called a **simple hypothesis**. A hypothesis of the form $\theta > \theta_0$ or $\theta < \theta_0$ is called a **composite hypothesis**. A test of the form

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0$$

is called a **two-sided test**. A test of the form

$$H_0 : \theta \leq \theta_0 \quad \text{versus} \quad H_1 : \theta > \theta_0$$

or

$$H_0 : \theta \geq \theta_0 \quad \text{versus} \quad H_1 : \theta < \theta_0$$

is called a **one-sided test**. The most common tests are two-sided.

10.2 Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma)$ where σ is known. We want to test $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$. Hence, $\Theta_0 = (-\infty, 0]$ and $\Theta_1 = (0, \infty)$. Consider the test:

$$\text{reject } H_0 \text{ if } T > c$$

where $T = \overline{X}$. The rejection region is

$$R = \Big\{(x_1, \ldots, x_n) : \ T(x_1, \ldots, x_n) > c\Big\}.$$

Let Z denote a standard Normal random variable. The power function is

$$\begin{aligned}
\beta(\mu) &= \mathbb{P}_\mu\left(\overline{X} > c\right) \\
&= \mathbb{P}_\mu\left(\frac{\sqrt{n}(\overline{X} - \mu)}{\sigma} > \frac{\sqrt{n}(c - \mu)}{\sigma}\right) \\
&= \mathbb{P}\left(Z > \frac{\sqrt{n}(c - \mu)}{\sigma}\right) \\
&= 1 - \Phi\left(\frac{\sqrt{n}(c - \mu)}{\sigma}\right).
\end{aligned}$$

This function is increasing in μ. See Figure 10.1. Hence

$$\text{size} = \sup_{\mu \leq 0} \beta(\mu) = \beta(0) = 1 - \Phi\left(\frac{\sqrt{n}c}{\sigma}\right).$$

For a size α test, we set this equal to α and solve for c to get

$$c = \frac{\sigma\, \Phi^{-1}(1 - \alpha)}{\sqrt{n}}.$$

We reject when $\overline{X} > \sigma\, \Phi^{-1}(1 - \alpha)/\sqrt{n}$. Equivalently, we reject when

$$\frac{\sqrt{n}\,(\overline{X} - 0)}{\sigma} > z_\alpha.$$

where $z_\alpha = \Phi^{-1}(1 - \alpha)$. ■

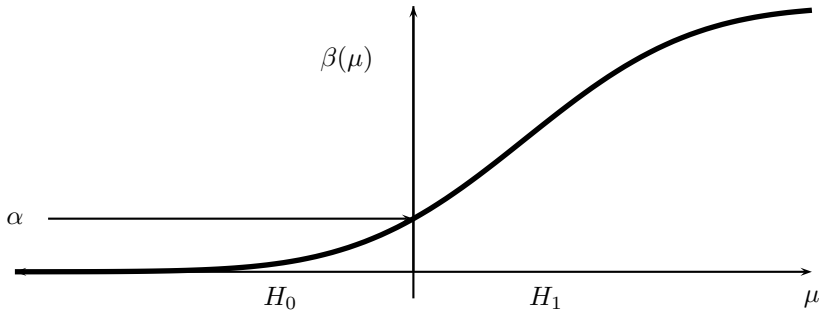

FIGURE 10.1. The power function for Example 10.2. The size of the test is the largest probability of rejecting H_0 when H_0 is true. This occurs at $\mu = 0$ hence the size is $\beta(0)$. We choose the critical value c so that $\beta(0) = \alpha$.

It would be desirable to find the test with highest power under H_1, among all size α tests. Such a test, if it exists, is called **most powerful**. Finding most powerful tests is hard and, in many cases, most powerful tests don't even exist. Instead of going into detail about when most powerful tests exist, we'll just consider four widely used tests: the Wald test,[1] the χ^2 test, the permutation test, and the likelihood ratio test.

10.1 The Wald Test

Let θ be a scalar parameter, let $\widehat{\theta}$ be an estimate of θ and let $\widehat{\mathsf{se}}$ be the estimated standard error of $\widehat{\theta}$.

[1]The test is named after Abraham Wald (1902–1950), who was a very influential mathematical statistician. Wald died in a plane crash in India in 1950.

10.3 Definition. The Wald Test

Consider testing

$$H_0 : \theta = \theta_0 \quad \text{versus} \quad H_1 : \theta \neq \theta_0.$$

Assume that $\widehat{\theta}$ *is asymptotically Normal:*

$$\frac{(\widehat{\theta} - \theta_0)}{\widehat{\mathsf{se}}} \rightsquigarrow N(0,1).$$

The size α **Wald test** *is: reject* H_0 *when* $|W| > z_{\alpha/2}$ *where*

$$W = \frac{\widehat{\theta} - \theta_0}{\widehat{\mathsf{se}}}. \tag{10.5}$$

10.4 Theorem. *Asymptotically, the Wald test has size* α*, that is,*

$$\mathbb{P}_{\theta_0}\left(|W| > z_{\alpha/2}\right) \to \alpha$$

as $n \to \infty$*.*

PROOF. Under $\theta = \theta_0$, $(\widehat{\theta} - \theta_0)/\widehat{\mathsf{se}} \rightsquigarrow N(0,1)$. Hence, the probability of rejecting when the null $\theta = \theta_0$ is true is

$$\begin{aligned} \mathbb{P}_{\theta_0}\left(|W| > z_{\alpha/2}\right) &= \mathbb{P}_{\theta_0}\left(\frac{|\widehat{\theta} - \theta_0|}{\widehat{\mathsf{se}}} > z_{\alpha/2}\right) \\ &\to \mathbb{P}\left(|Z| > z_{\alpha/2}\right) \\ &= \alpha \end{aligned}$$

where $Z \sim N(0,1)$. ■

10.5 Remark. An alternative version of the Wald test statistic is $W = (\widehat{\theta} - \theta_0)/\mathsf{se}_0$ where se_0 is the standard error computed at $\theta = \theta_0$. Both versions of the test are valid.

Let us consider the power of the Wald test when the null hypothesis is false.

10.6 Theorem. *Suppose the true value of* θ *is* $\theta_\star \neq \theta_0$*. The power* $\beta(\theta_\star)$ *— the probability of correctly rejecting the null hypothesis — is given (approximately) by*

$$1 - \Phi\left(\frac{\theta_0 - \theta_\star}{\widehat{\mathsf{se}}} + z_{\alpha/2}\right) + \Phi\left(\frac{\theta_0 - \theta_\star}{\widehat{\mathsf{se}}} - z_{\alpha/2}\right). \tag{10.6}$$

Recall that $\widehat{\mathsf{se}}$ tends to 0 as the sample size increases. Inspecting (10.6) closely we note that: (i) the power is large if $\theta_\star$ is far from θ_0, and (ii) the power is large if the sample size is large.

10.7 Example (Comparing Two Prediction Algorithms)**.** We test a prediction algorithm on a test set of size m and we test a second prediction algorithm on a second test set of size n. Let X be the number of incorrect predictions for algorithm 1 and let Y be the number of incorrect predictions for algorithm 2. Then $X \sim \text{Binomial}(m, p_1)$ and $Y \sim \text{Binomial}(n, p_2)$. To test the null hypothesis that $p_1 = p_2$ write

$$H_0 : \delta = 0 \quad \text{versus} \quad H_1 : \delta \neq 0$$

where $\delta = p_1 - p_2$. The MLE is $\widehat{\delta} = \widehat{p}_1 - \widehat{p}_2$ with estimated standard error

$$\widehat{\mathsf{se}} = \sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}}.$$

The size α Wald test is to reject H_0 when $|W| > z_{\alpha/2}$ where

$$W = \frac{\widehat{\delta} - 0}{\widehat{\mathsf{se}}} = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{m} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n}}}.$$

The power of this test will be largest when p_1 is far from p_2 and when the sample sizes are large.

What if we used the same test set to test both algorithms? The two samples are no longer independent. Instead we use the following strategy. Let $X_i = 1$ if algorithm 1 is correct on test case i and $X_i = 0$ otherwise. Let $Y_i = 1$ if algorithm 2 is correct on test case i, and $Y_i = 0$ otherwise. Define $D_i = X_i - Y_i$. A typical dataset will look something like this:

Test Case	X_i	Y_i	$D_i = X_i - Y_i$
1	1	0	1
2	1	1	0
3	1	1	0
4	0	1	-1
5	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	0	1	-1

Let

$$\delta = \mathbb{E}(D_i) = \mathbb{E}(X_i) - \mathbb{E}(Y_i) = \mathbb{P}(X_i = 1) - \mathbb{P}(Y_i = 1).$$

The nonparametric plug-in estimate of δ is $\widehat{\delta} = \overline{D} = n^{-1}\sum_{i=1}^n D_i$ and $\widehat{\mathsf{se}}(\widehat{\delta}) = S/\sqrt{n}$, where $S^2 = n^{-1}\sum_{i=1}^n (D_i - \overline{D})^2$. To test $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$

we use $W = \widehat{\delta}/\widehat{\mathsf{se}}$ and reject H_0 if $|W| > z_{\alpha/2}$. This is called a **paired comparison.** ■

10.8 Example (Comparing Two Means)**.** Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two independent samples from populations with means μ_1 and μ_2, respectively. Let's test the null hypothesis that $\mu_1 = \mu_2$. Write this as $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ where $\delta = \mu_1 - \mu_2$. Recall that the nonparametric plug-in estimate of δ is $\widehat{\delta} = \overline{X} - \overline{Y}$ with estimated standard error

$$\widehat{\mathsf{se}} = \sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}$$

where s_1^2 and s_2^2 are the sample variances. The size α Wald test rejects H_0 when $|W| > z_{\alpha/2}$ where

$$W = \frac{\widehat{\delta} - 0}{\widehat{\mathsf{se}}} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}}. \quad ■$$

10.9 Example (Comparing Two Medians)**.** Consider the previous example again but let us test whether the medians of the two distributions are the same. Thus, $H_0 : \delta = 0$ versus $H_1 : \delta \neq 0$ where $\delta = \nu_1 - \nu_2$ where ν_1 and ν_2 are the medians. The nonparametric plug-in estimate of δ is $\widehat{\delta} = \widehat{\nu}_1 - \widehat{\nu}_2$ where $\widehat{\nu}_1$ and $\widehat{\nu}_2$ are the sample medians. The estimated standard error $\widehat{\mathsf{se}}$ of $\widehat{\delta}$ can be obtained from the bootstrap. The Wald test statistic is $W = \widehat{\delta}/\widehat{\mathsf{se}}$. ■

There is a relationship between the Wald test and the $1 - \alpha$ asymptotic confidence interval $\widehat{\theta} \pm \widehat{\mathsf{se}}\, z_{\alpha/2}$.

10.10 Theorem. *The size α Wald test rejects $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$ if and only if $\theta_0 \notin C$ where*

$$C = (\widehat{\theta} - \widehat{\mathsf{se}}\, z_{\alpha/2},\ \widehat{\theta} + \widehat{\mathsf{se}}\, z_{\alpha/2}).$$

Thus, testing the hypothesis is equivalent to checking whether the null value is in the confidence interval.

Warning! When we reject H_0 we often say that the result is **statistically significant.** A result might be statistically significant and yet the size of the effect might be small. In such a case we have a result that is statistically significant but not scientifically or practically significant. The difference between statistical significance and scientific significance is easy to understand in light of Theorem 10.10. Any confidence interval that excludes θ_0 corresponds to rejecting H_0. But the values in the interval could be close to θ_0 (not scientifically significant) or far from θ_0 (scientifically significant). See Figure 10.2.

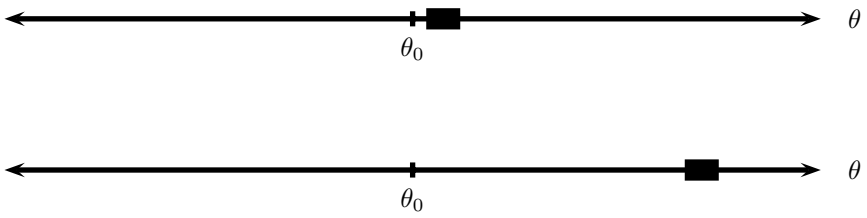

FIGURE 10.2. Scientific significance versus statistical significance. A level α test rejects $H_0 : \theta = \theta_0$ if and only if the $1 - \alpha$ confidence interval does not include θ_0. Here are two different confidence intervals. Both exclude θ_0 so in both cases the test would reject H_0. But in the first case, the estimated value of θ is close to θ_0 so the finding is probably of little scientific or practical value. In the second case, the estimated value of θ is far from θ_0 so the finding is of scientific value. This shows two things. First, statistical significance does not imply that a finding is of scientific importance. Second, confidence intervals are often more informative than tests.

10.2 p-values

Reporting "reject H_0" or "retain H_0" is not very informative. Instead, we could ask, for every α, whether the test rejects at that level. Generally, if the test rejects at level α it will also reject at level $\alpha' > \alpha$. Hence, there is a smallest α at which the test rejects and we call this number the p-value. See Figure 10.3.

10.11 Definition. *Suppose that for every $\alpha \in (0, 1)$ we have a size α test with rejection region R_α. Then,*

$$\text{p-value} = \inf\Big\{\alpha : \ T(X^n) \in R_\alpha\Big\}.$$

That is, the p-value is the smallest level at which we can reject H_0.

Informally, the p-value is a measure of the evidence against H_0: the smaller the p-value, the stronger the evidence against H_0. Typically, researchers use the following evidence scale:

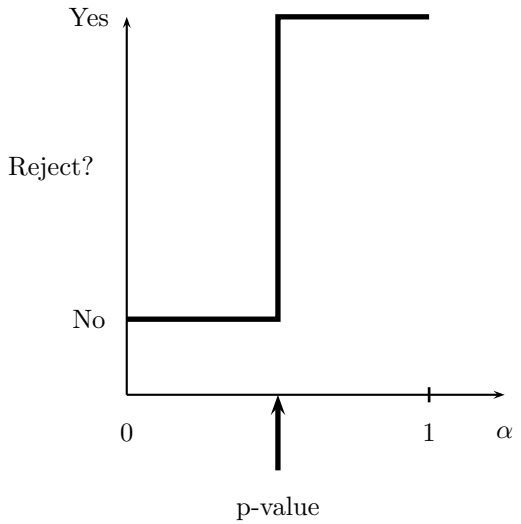

FIGURE 10.3. p-values explained. For each α we can ask: does our test reject H_0 at level α? The p-value is the smallest α at which we do reject H_0. If the evidence against H_0 is strong, the p-value will be small.

p-value	evidence
$< .01$	very strong evidence against H_0
$.01 - .05$	strong evidence against H_0
$.05 - .10$	weak evidence against H_0
$> .1$	little or no evidence against H_0

Warning! A large p-value is not strong evidence in favor of H_0. A large p-value can occur for two reasons: (i) H_0 is true or (ii) H_0 is false but the test has low power.

Warning! Do not confuse the p-value with $\mathbb{P}(H_0|\text{Data})$. [2] **The p-value is not the probability that the null hypothesis is true.**

The following result explains how to compute the p-value.

[2] We discuss quantities like $\mathbb{P}(H_0|\text{Data})$ in the chapter on Bayesian inference.

10.12 Theorem. *Suppose that the size α test is of the form*

$$\text{reject } H_0 \text{ if and only if } T(X^n) \geq c_\alpha.$$

Then,

$$\text{p-value} = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(X^n) \geq T(x^n))$$

where x^n is the observed value of X^n. If $\Theta_0 = \{\theta_0\}$ then

$$\text{p-value} = \mathbb{P}_{\theta_0}(T(X^n) \geq T(x^n)).$$

We can express Theorem 10.12 as follows:

> The p-value is the probability (under H_0) of observing a value of the test statistic the same as or more extreme than what was actually observed.

10.13 Theorem. *Let $w = (\widehat{\theta} - \theta_0)/\widehat{\mathsf{se}}$ denote the observed value of the Wald statistic W. The p-value is given by*

$$\text{p} - \text{value} = \mathbb{P}_{\theta_0}(|W| > |w|) \approx \mathbb{P}(|Z| > |w|) = 2\Phi(-|w|) \tag{10.7}$$

where $Z \sim N(0,1)$.

To understand this last theorem, look at Figure 10.4.
Here is an important property of p-values.

10.14 Theorem. *If the test statistic has a continuous distribution, then under $H_0 : \theta = \theta_0$, the p-value has a Uniform (0,1) distribution. Therefore, if we reject H_0 when the p-value is less than α, the probability of a type I error is α.*

In other words, if H_0 is true, the p-value is like a random draw from a $\text{Unif}(0,1)$ distribution. If H_1 is true, the distribution of the p-value will tend to concentrate closer to 0.

10.15 Example. Recall the cholesterol data from Example 7.15. To test if the means are different we compute

$$W = \frac{\widehat{\delta} - 0}{\widehat{\mathsf{se}}} = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{s_1^2}{m} + \frac{s_2^2}{n}}} = \frac{216.2 - 195.3}{\sqrt{5^2 + 2.4^2}} = 3.78.$$

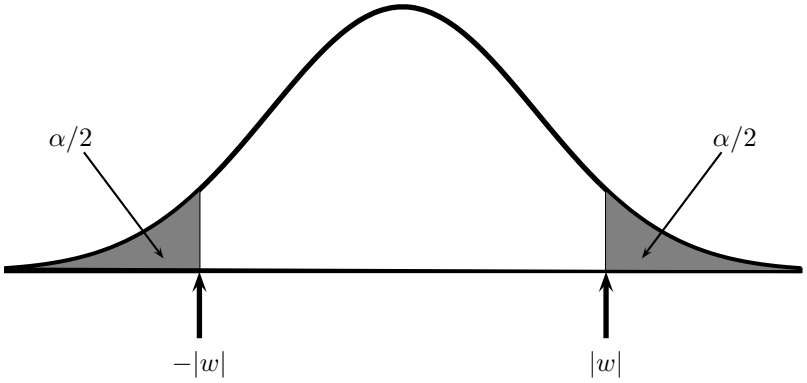

FIGURE 10.4. The p-value is the smallest α at which you would reject H_0. To find the p-value for the Wald test, we find α such that $|w|$ and $-|w|$ are just at the boundary of the rejection region. Here, w is the observed value of the Wald statistic: $w = (\widehat{\theta} - \theta_0)/\widehat{\mathsf{se}}$. This implies that the p-value is the tail area $\mathbb{P}(|Z| > |w|)$ where $Z \sim N(0,1)$.

To compute the p-value, let $Z \sim N(0,1)$ denote a standard Normal random variable. Then,

$$\text{p-value} = \mathbb{P}(|Z| > 3.78) = 2\mathbb{P}(Z < -3.78) = .0002$$

which is very strong evidence against the null hypothesis. To test if the medians are different, let $\widehat{\nu}_1$ and $\widehat{\nu}_2$ denote the sample medians. Then,

$$W = \frac{\widehat{\nu}_1 - \widehat{\nu}_2}{\widehat{\mathsf{se}}} = \frac{212.5 - 194}{7.7} = 2.4$$

where the standard error 7.7 was found using the bootstrap. The p-value is

$$\text{p-value} = \mathbb{P}(|Z| > 2.4) = 2\mathbb{P}(Z < -2.4) = .02$$

which is strong evidence against the null hypothesis. ■

10.3 The χ^2 Distribution

Before proceeding we need to discuss the χ^2 distribution. Let $Z_1, \ldots, Z_k$ be independent, standard Normals. Let $V = \sum_{i=1}^k Z_i^2$. Then we say that V has a χ^2 distribution with k degrees of freedom, written $V \sim \chi^2_k$. The probability density of V is

$$f(v) = \frac{v^{(k/2)-1} e^{-v/2}}{2^{k/2}\Gamma(k/2)}$$

for $v > 0$. It can be shown that $\mathbb{E}(V) = k$ and $\mathbb{V}(V) = 2k$. We define the upper α quantile $\chi^2_{k,\alpha} = F^{-1}(1-\alpha)$ where F is the CDF. That is, $\mathbb{P}(\chi^2_k > \chi^2_{k,\alpha}) = \alpha$.