Chapter 8

Convergence Concepts Continued

8.1 Multivariate Convergence Concepts

When we covered convergence concepts before (Chapter 6 of these notes), we only did the scalar case because of the semester system. Logically, this chapter goes with Chapter 6, but if we had done it then, students transferring into this section this semester would be lost because the material isn't in Lindgren. Then we only covered convergence in probability and in distribution of *scalar* random variables. Now we want to cover the same ground but this time for random *vectors*. It will also be a good review.

8.1.1 Convergence in Probability to a Constant

Recall that *convergence in probability to a constant* has a definition (Definition 6.1.2 in Chapter 6 of these notes), but we never used the definition. Instead we obtained all of our convergence in probability results, either directly or indirectly from the law of large numbers (LLN).

Now we want to discuss convergence of random vectors, which we can also call *multivariate convergence in probability to a constant.* It turns out, that the multivariate concept is a trivial generalization of the univariate concept.

Definition 8.1.1 (Convergence in Probability to a Constant).
A sequence of random vectors

$$\mathbf{X}_n = (X_{n1}, \ldots, X_{nm}), \qquad n = 1, 2, \ldots$$

converges in probability to a constant vector

$$\mathbf{a} = (a_1, \ldots, a_m)$$

written

$$\mathbf{X}_n \xrightarrow{P} \mathbf{a}, \qquad \text{as } n \to \infty$$

if the corresponding components converge in probability, that is, if

$$X_{ni} \xrightarrow{P} a_i, \qquad \text{as } n \to \infty$$

for each i.

The reader should be warned that this isn't the usual definition, but it is equivalent to the usual definition (we have defined the usual concept but not in the usual way).

8.1.2 The Law of Large Numbers

The componentwise nature of convergence in probability to a constant makes the multivariate law of large numbers a trivial extension of the univariate law (Theorem 6.3 in Chapter 6 of these notes).

Theorem 8.1 (Multivariate Law of Large Numbers). *If* $\mathbf{X}_1$, $\mathbf{X}_2$, ... *is a sequence of independent, identically distributed random vectors having mean vector* $\boldsymbol{\mu}$, *and*

$$\overline{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i$$

is the sample mean for sample size n, *then*

$$\overline{\mathbf{X}}_n \xrightarrow{P} \boldsymbol{\mu}, \qquad \text{as } n \to \infty. \tag{8.1}$$

The only requirement is that the mean $\boldsymbol{\mu}$ exist. No other property of the distribution of the $\mathbf{X}_i$ matters.

We will use the abbreviation LLN for either theorem. The multivariate theorem is only interesting for giving us a notational shorthand that allows us to write the law of large numbers for all the components at once. It has no mathematical content over and above the univariate LLN's for each component. Convergence in probability (to a constant) of random vectors says no more than the statement that each component converges. In the case of the LLN, each statement about a component is just the univariate LLN.

8.1.3 Convergence in Distribution

Convergence in distribution is different. Example 8.1.1 below will show that, unlike convergence in probability to a constant, convergence in distribution for random vectors is not just convergence in distribution of each component.

Univariate convergence in distribution has a definition (Theorem 6.1.1 of these notes), but the definition was not used except in Problem 7-7 in Chapter 7 of these notes, which is an odd counterexample to the usual behavior of statistical estimators.

Instead we obtained all of our convergence in distribution results, either directly or indirectly, from the central limit theorem (CLT), which is Theorem 6.2 of Chapter 6 of these notes. Multivariate convergence in distribution has a definition is much the same, we will obtain almost all such results directly or indirectly from the (multivariate) CLT. Hence here we define multivariate convergence in distribution in terms of univariate convergence in distribution.

Definition 8.1.2 (Convergence in Distribution).
A sequence of random vectors $\mathbf{X}_1$*,* $\mathbf{X}_2$*, …* converges in distribution *to a random vector* $\mathbf{X}$*, written*

$$\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}, \qquad \text{as } n \to \infty.$$

if

$$\mathbf{t}'\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{t}'\mathbf{X}, \qquad \text{for all constant vectors } \mathbf{t}. \tag{8.2}$$

Again, the reader should be warned that this isn't the usual definition, but it is equivalent to the usual definition (we have defined the usual concept but not in the usual way, the equivalence of our definition and the usual definition is called the Cramér-Wold Theorem).

This shows us in what sense the notion of multivariate convergence in distribution is determined by the univariate notion. The multivariate convergence in distribution $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$ happens if and only if the univariate convergence in distribution $\mathbf{t}'\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{t}'\mathbf{X}$ happens for every constant vector $\mathbf{t}$. The following example shows that convergence in distribution of each component of a random vector is not enough to imply convergence of the vector itself.

Example 8.1.1.
Let $\mathbf{X}_n = (U_n, V_n)$ be defined as follows. Define U_n to be standard normal for all n. Then trivially $U_n \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$. Define $V_n = (-1)^n U_n$ for all n. Then V_n is also standard normal for all n, and hence trivially $V_n \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$. Thus both components of $\mathbf{X}_n$ converge in distribution. But if $\mathbf{t}' = \begin{pmatrix} 1 & 1 \end{pmatrix}$, then

$$\mathbf{t}'\mathbf{X}_n = U_n + V_n = \begin{cases} 2U_n, & n \text{ even} \\ 0, & n \text{ odd} \end{cases}$$

This clearly does not converge in distribution, since the even terms all have one distribution, $\mathcal{N}(0,4)$ (and hence trivially converge to that distribution), and the odd terms all have another, the distribution concentrated at zero (and hence trivially converge to that distribution).

Thus, unlike convergence in probability to a constant, multivariate convergence in distribution entails more than univariate convergence of each component. Another way to say the same thing is that *marginal* convergence in distribution does not imply *joint* convergence in distribution. Of course, the converse does hold: joint convergence in distribution does imply marginal convergence in distribution (just take a vector $\mathbf{t}$ in the definition having all but one component equal to zero).

8.1.4 The Central Limit Theorem

We can now derive the multivariate central limit theorem from the univariate theorem (Theorem 6.2 of Chapter 6 of these notes).

Theorem 8.2 (Multivariate Central Limit Theorem). *If $\mathbf{X}_1$, $\mathbf{X}_2$, ... is a sequence of independent, identically distributed random vectors having mean vector $\boldsymbol{\mu}$ and variance matrix $\mathbf{M}$ and*

$$\overline{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i$$

is the sample mean for sample size n, then

$$\sqrt{n} \left(\overline{\mathbf{X}}_n - \boldsymbol{\mu} \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M}). \tag{8.3}$$

The only requirement is that second moments (the elements of $\mathbf{M}$) exist (this implies first moments also exist by Theorem 2.44 of Chapter 2 of these notes). No other property of the distribution of the $\mathbf{X}_i$ matters.

We often write the univariate CLT as

$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n} \right) \tag{8.4}$$

and the multivariate CLT as

$$\overline{\mathbf{X}}_n \approx \mathcal{N}\left(\boldsymbol{\mu}, \frac{\mathbf{M}}{n} \right) \tag{8.5}$$

These are simpler to interpret (though less precise and harder to use theoretically). We often say that the right hand side of one of these equations is the *asymptotic distribution* of the left hand side.

Derivation of the multivariate CLT from the univariate. For each constant vector $\mathbf{t}$, the scalar random variables $\mathbf{t}'(\mathbf{X}_n - \boldsymbol{\mu})$ have mean 0 and variance $\mathbf{t}'\mathbf{M}\mathbf{t}$ and hence obey the univariate CLT

$$\sqrt{n}\mathbf{t}' \left(\overline{\mathbf{X}}_n - \boldsymbol{\mu} \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{t}'\mathbf{M}\mathbf{t}).$$

The right hand side is the distribution of $\mathbf{t}'\mathbf{Z}$ where $\mathbf{Z} \sim \mathcal{N}(0, \mathbf{M})$, hence (8.3) follows by our definition of multivariate convergence in distribution. □

Example 8.1.2.
(This continues Example 5.1.1.) Let X_1, X_2, ... be a sequence of i. i. d. random variables, and define random vectors

$$\mathbf{Z}_i = \begin{pmatrix} X_i \\ X_i^2 \end{pmatrix}$$

Then $\mathbf{Z}_1$, $\mathbf{Z}_2$, ... is a sequence of i. i. d. random vectors having mean vector $\boldsymbol{\mu}$ given by (5.6) and variance matrix $\mathbf{M}$ given by (5.7), and the CLT applies

$$\sqrt{n} \left(\overline{\mathbf{Z}}_n - \boldsymbol{\mu} \right) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M}).$$

8.1.5 Slutsky and Related Theorems

As with univariate convergence in distribution, we are forced to state a number of theorems about multivariate convergence in distribution without proof. The proofs are just too hard for this course.

Theorem 8.3. *If* $\mathbf{a}$ *is a constant, then* $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{a}$ *if and only if* $\mathbf{X}_n \xrightarrow{P} \mathbf{a}$.

Thus, as was true in the univariate case, convergence in probability to a constant and convergence in distribution to a constant are equivalent concepts. We could dispense with one, but tradition and usage do not allow it. We must be able to recognize both in order to read the literature.

Theorem 8.4 (Slutsky). *If* $g(\mathbf{x}, \mathbf{y})$ *is a function jointly continuous at every point of the form* $(\mathbf{x}, \mathbf{a})$ *for some fixed* $\mathbf{a}$, *and if* $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$ *and* $\mathbf{Y}_n \xrightarrow{P} \mathbf{a}$, *then*

$$g(\mathbf{X}_n, \mathbf{Y}_n) \xrightarrow{\mathcal{D}} g(\mathbf{X}, \mathbf{a}).$$

The function g here can be either scalar or vector valued. The continuity hypothesis means that $g(\mathbf{x}_n, \mathbf{y}_n) \to g(\mathbf{x}, \mathbf{a})$ for all nonrandom sequences $\mathbf{x}_n \to \mathbf{x}$ and $\mathbf{y}_n \to \mathbf{a}$.

Sometimes Slutsky's theorem is used in a rather trivial way with the sequence "converging in probability" being nonrandom. This uses the following lemma.

Lemma 8.5. *If* $\mathbf{a}_n \to \mathbf{a}$ *considered as a nonrandom sequence, then* $\mathbf{a}_n \xrightarrow{P} \mathbf{a}$ *considered as a sequence of constant random vectors.*

This is an obvious consequence of the definition of convergence in probability (Definition 6.1.2 in Chapter 6 of these notes).

Example 8.1.3.
The so-called sample variance S_n^2 defined on p. 204 in Lindgren or in (7.17) in Chapter 7 of these notes is asymptotically equivalent to the variance of the empirical distribution V_n also defined on p. 204 in Lindgren or in (7.4) in Chapter 7 of these notes. The two estimators are related by

$$S_n^2 = \frac{n}{n-1} V_n.$$

We know the asymptotics of V_n because it is a sample moment. By Theorem 7.15 of Chapter 7 of these notes

$$V_n \xrightarrow{P} \sigma^2 \tag{8.6}$$

and by Theorem 7.16 of Chapter 7 of these notes

$$\sqrt{n}(V_n - \sigma^2) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mu_4 - \mu_2^2) \tag{8.7}$$

(strictly speaking, we don't actually know this last fact yet, because we haven't proved Theorem 7.16 yet, but we will).

One application of Slutsky's theorem shows that

$$S_n^2 \xrightarrow{P} \sigma^2 \tag{8.8}$$

because

$$\frac{n}{n-1} V_n \xrightarrow{P} 1 \cdot \sigma^2$$

because of (8.6) and

$$\frac{n}{n-1} \to 1$$

and Lemma 8.5.

To get the next level of asymptotics we write

$$\begin{aligned} \sqrt{n}(S_n^2 - \sigma^2) &= \sqrt{n} \left(\frac{n}{n-1} V_n - \sigma^2 \right) \\ &= \frac{n}{n-1} \left[\sqrt{n} \left(V_n - \sigma^2 \right) + \frac{\sqrt{n}}{n-1} \sigma^2 \right] \end{aligned} \tag{8.9}$$

Then two applications of Slutsky's theorem give us what we want. Let W be a random variable having the distribution of the right hand side of (8.7) so that equation can be rewritten

$$\sqrt{n}(V_n - \sigma^2) \xrightarrow{\mathcal{D}} W.$$

Then one application of Slutsky's theorem (and the corollary following it in Chapter 6 of these notes) shows that the term in square brackets in (8.9) also converges to W

$$\sqrt{n} \left(V_n - \sigma^2 \right) + \frac{\sqrt{n}}{n-1} \sigma^2 \xrightarrow{\mathcal{D}} W + 0$$

because of

$$\frac{\sqrt{n}}{n-1} \sigma^2 \to 0$$

and Lemma 8.5. Then another application of Slutsky's theorem shows what we want

$$\frac{n}{n-1} \times \text{term in square brackets} \xrightarrow{\mathcal{D}} 1 \cdot W.$$

The special cases of Slutsky's theorem which we only have only one sequence of random variables converging in distribution or in probability are known as "continuous mapping theorems."

Theorem 8.6 (Continuous Mapping, Convergence in Distribution). *If g is an everywhere continuous function and $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$, then $g(\mathbf{X}_n) \xrightarrow{\mathcal{D}} g(\mathbf{X})$.*

The function g here can be either scalar or vector valued. The only requirement is that it be continuous, that is, $g(\mathbf{x}_n) \to g(\mathbf{x})$ for any point $\mathbf{x}$ and any sequence $\mathbf{x}_n \to \mathbf{x}$.

Theorem 8.7 (Continuous Mapping, Convergence in Probability). *If g is a function continuous at the point $\mathbf{a}$ and $\mathbf{X}_n \xrightarrow{P} \mathbf{a}$, then $g(\mathbf{X}_n) \xrightarrow{P} g(\mathbf{a})$.*

The function g here can be either scalar or vector valued. The only requirement is that it be continuous at $\mathbf{a}$, that is, $g(\mathbf{x}_n) \to g(\mathbf{a})$ for any sequence $\mathbf{x}_n \to \mathbf{a}$.

Example 8.1.4.
Suppose (8.8) holds for some sequence of random variables S_n^2 and $\sigma > 0$, then the continuous mapping theorem for convergence in probability immediately gives many other convergence in probability results, for example,

$$
\begin{aligned}
S_n &\xrightarrow{P} \sigma \qquad (8.10) \\
\frac{1}{S_n} &\xrightarrow{P} \frac{1}{\sigma} \\
\log(S_n) &\xrightarrow{P} \log(\sigma)
\end{aligned}
$$

All of these applications are fairly obvious. These conclusions seem so natural that it is hard to remember that we need the continuous mapping theorem to tell us that they hold.

We will use the continuous mapping theorem for convergence in probability many times in our study of statistics. In contrast, our uses of the continuous mapping theorem for convergence in distribution will all be rather trivial. We will only use it to see that we can divide both sides of a convergence in distribution statement by the same constant, or add the same constant to both sides, and so forth.

Example 8.1.5.
This was assigned for homework (Problem 6-2 of Chapter 6 of these notes) last semester. We will see many other examples later, but all will be similar to this one, which is by far the most important in statistics. Suppose X_1, X_2, $\ldots$ are i. i. d. random variables with mean μ and variance σ^2, suppose that S_n is any sequence of random variables satisfying (8.10), and suppose $\sigma > 0$, then

$$
\frac{\overline{X}_n - \mu}{S_n / \sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1). \qquad (8.11)
$$

The most obvious choice of a sequence S_n satisfying (8.10) is the sample standard deviation. That's what Examples 8.1.3 and 8.1.4 showed. But the exact way S_n is defined isn't important for this example. In fact there are many sequences of random variables having this property. The only thing that is important is that such sequences exist.

How do we show (8.11) using the CLT and Slutsky's theorem? First the CLT says

$$
\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{\mathcal{D}} Y
$$

where $Y \sim \mathcal{N}(0, \sigma^2)$. Define the function

$$g(u, v) = \frac{u}{v}.$$

This is continuous everywhere except where $v = 0$, where it is undefined. Now define

$$U_n = \sqrt{n}\left(\overline{X}_n - \mu\right)$$

and apply Slutsky's theorem to $g(U_n, S_n)$. The first argument converges in distribution and the second argument converges to a constant, so Slutsky's theorem does hold and says

$$\sqrt{n}\frac{\overline{X}_n - \mu}{S_n} = g(U_n, S_n) \xrightarrow{\mathcal{D}} g(Y, \sigma) = \frac{Y}{\sigma}$$

and the right hand side does have a standard normal distribution, as asserted, by the rule giving the variance of a linear transformation (5.15b).

8.2 The Delta Method

8.2.1 The Univariate Delta Method

Suppose T_n is any sequence of random variables converging in probability to a constant θ. Many such examples arise in statistics. Particular cases are $\overline{X}_n \xrightarrow{P} \mu$ (which is the LLN) and $S_n \xrightarrow{P} \sigma$ (which we showed in Examples 8.1.3 and 8.1.4). It is conventional in statistics to use T_n as a default notation for all such sequences and θ as a default notation for all the constants.

The continuous mapping theorem for convergence in probability tells us that $g(T_n) \xrightarrow{P} g(\theta)$ for any function g that is continuous at the point θ. Many different functions g arise in applications. The continuity requirement is not very restrictive. Almost any function will do.

What this says is that $g(T_n)$ gets closer and closer to $g(\theta)$ as n gets large. The obvious next question is "How close?" We want a statement analogous to the CLT that tells us the distribution of the "error" $g(T_n) - g(\theta)$. The continuous mapping theorem for convergence in distribution doesn't do this. It would tell us, for example, that

$$g\Big(\sqrt{n}\left(\overline{X}_n - \mu\right)\Big) \xrightarrow{\mathcal{D}} g(Y) \tag{8.12a}$$

where $Y \sim \mathcal{N}(0, \sigma^2)$, but that's not what we want. We want a convergence in distribution result for

$$\sqrt{n}\big(g(T_n) - g(\theta)\big). \tag{8.12b}$$

If g is a linear function, then (8.12b) and the left hand side of (8.12a) are equal. If g is not linear, then they aren't, and the continuous mapping theorem is of no use. The delta method does do what we want: a convergence in distribution result for (8.12b).

Theorem 8.8 (Univariate Delta Method). *Suppose*

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}} Y \tag{8.13}$$

and g is any function differentiable at θ, then

$$\sqrt{n}\bigl(g(T_n) - g(\theta)\bigr) \xrightarrow{\mathcal{D}} g'(\theta)Y. \tag{8.14}$$

By far the most important applications of the delta method have Y normally distributed with mean zero, say $Y \sim \mathcal{N}(0, \sigma_Y^2)$. In that case, we can put (8.14) in "sloppy form" with "double squiggle" notation like (8.4) or (8.5). It becomes

$$g(T_n) \approx \mathcal{N}\left(g(\theta), \frac{g'(\theta)^2 \sigma_Y^2}{n}\right)$$

and we say that the right hand side is the *asymptotic distribution* of $g(T_n)$.

It is called the "delta method" because of the important role played by the derivative. The "delta" is supposed to remind you of

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$$

the triangles being capital Greek letter deltas, and the fraction on the right being pronounced "delta y over delta x." The earlier term for this concept, used throughout the nineteenth century and still used by some people, was "propagation of errors."

It is important to understand that the delta method does not produce a convergence in distribution result out of thin air. It turns one convergence in distribution statement (8.13) into another (8.14). In order to use the delta method we must already have one convergence in distribution result. Usually that comes either from the CLT or from a previous application of the delta method.

Example 8.2.1.
Suppose X_1, X_2, $\ldots$ are i. i. d. Exp(λ). Then

$$\begin{aligned} E(X_i) &= \frac{1}{\lambda} \\ \operatorname{var}(X_i) &= \frac{1}{\lambda^2} \end{aligned}$$

the LLN says

$$\overline{X}_n \xrightarrow{P} \frac{1}{\lambda}$$

and the CLT says

$$\sqrt{n}\left(\overline{X}_n - \frac{1}{\lambda}\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{\lambda^2}\right) \tag{8.15}$$

That's all well and good, but it seems more natural to look at the sequence of random variables

$$W_n = \frac{1}{\overline{X}_n} \tag{8.16}$$

because then the continuous mapping theorem for convergence in probability gives

$$W_n \xrightarrow{P} \lambda.$$

So what is the asymptotic distribution of W_n?

We want to apply the delta method. To do that we already need one convergence in distribution result. What we have is the CLT (8.15). This tells we want to use the delta method with $T_n = \overline{X}_n$ and $\theta = 1/\lambda$. Then, since we want $g(T_n) = W_n$, we must have

$$g(t) = \frac{1}{t}$$

and hence

$$g'(t) = -\frac{1}{t^2}$$

So

$$g(\theta) = \lambda$$

and

$$g'(\theta) = -\lambda^2.$$

And the delta method tells us that W_n is asymptotically normally distributed with mean λ and variance

$$\frac{g'(\theta)^2 \sigma^2}{n} = \frac{\bigl(-\lambda^2\bigr)^2}{n\lambda^2} = \frac{\lambda^2}{n}$$

The argument is a bit involved, but in the end we arrive at the fairly simple statement

$$W_n \approx \mathcal{N}\left(\lambda, \frac{\lambda^2}{n}\right).$$

Proof of the Univariate Delta Method. By definition, the derivative is

$$g'(\theta) = \lim_{t \to \theta} \frac{g(t) - g(\theta)}{t - \theta}$$

To be useful in our proof we need to rewrite this slightly. For $t \neq \theta$ define the function

$$w(t) = \frac{g(t) - g(\theta)}{t - \theta} - g'(\theta) \tag{8.17}$$

then the definition of differentiation says that $w(t) \to 0$ as $t \to \theta$, which is the same thing as saying that w is continuous at the point θ if we define $w(\theta) = 0$. (The reason for phrasing things in terms of continuity rather than limits is because Slutsky's theorem uses continuity.) Then (8.17) can be rewritten as

$$g(t) - g(\theta) = g'(\theta)(t - \theta) + w(t)(t - \theta). \tag{8.18}$$

Now plug in T_n for t and multiply by $\sqrt{n}$ giving

$$\sqrt{n}\big(g(T_n) - g(\theta)\big) = g'(\theta)\sqrt{n}(T_n - \theta) + w(T_n)\sqrt{n}(T_n - \theta).$$

By the continuous mapping theorem, the first term on the right hand side converges in distribution to $g'(\theta)Y$. By Slutsky's theorem, the second term on the right hand side converges in distribution to $w(\theta)Y = 0$. Hence by another application of Slutsky's theorem, the right hand side converges to $g'(\theta)Y$, which is the assertion of the delta method. □

8.2.2 The Multivariate Delta Method

The multivariate delta method is a straightforward extension of the univariate delta method, obvious if you know about derivatives of general vector-valued functions. You already know this material because it was used in the change of variable theorem for random vectors (Section 1.6.2 in Chapter 1 of these notes). You may need to go back and review that.

In brief, that section introduced the derivative of a vector-valued function of a vector variable $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$, which can also be thought of a a vector of scalar-valued functions

$$\mathbf{g}(\mathbf{x}) = \begin{pmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_m(\mathbf{x}) \end{pmatrix}$$

The derivative of the function $\mathbf{g}$ at the point $\mathbf{x}$ (assuming it exists) is the matrix of partial derivatives. It is written $\nabla\mathbf{g}(\mathbf{x})$ and pronounced "del g of x." Throughout this section we will also write it as the single letter $\mathbf{G}$. So

$$\mathbf{G} = \nabla\mathbf{g}(\mathbf{x})$$

is the matrix with elements

$$g_{ij} = \frac{\partial g_i(\mathbf{x})}{\partial x_j}$$

Note that if $\mathbf{g}$ maps n-dimensional vectors to m-dimensional vectors, then it is an $m \times n$ matrix (rather than the other way around). A concrete example that may help you visualize the idea is Example 1.6.1 in Chapter 1 of these notes.

Theorem 8.9 (Multivariate Delta Method). *Suppose*

$$\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathbf{Y} \tag{8.19}$$

and $\mathbf{g}$ *is any function differentiable*[1] *at* $\boldsymbol{\theta}$*, then*

$$\sqrt{n}\big(\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta})\big) \xrightarrow{\mathcal{D}} \nabla\mathbf{g}(\boldsymbol{\theta})\mathbf{Y}. \tag{8.21}$$

[1] The notion of multivariate differentiability of is actually a bit complicated. We presented only a simplified version of the facts, which is not completely correct. Here are the facts. Most readers will only want to know the first item below, maybe the second. The third is the pedantically correct mathematical definition of multivariate differentiability, which is of theoretical interest only. It won't help you do any problems. You are free to ignore it.

By far the most important applications of the delta method have $\mathbf{Y}$ normally distributed with mean zero, say $\mathbf{Y} \sim \mathcal{N}(0, \mathbf{M}_{\mathbf{Y}})$. In that case, we can put (8.21) in "sloppy form" with "double squiggle" notation like (8.4) or (8.5). It becomes

$$\mathbf{g}(\mathbf{T}_n) \approx \mathcal{N}\left(\mathbf{g}(\boldsymbol{\theta}), \frac{\mathbf{G}\mathbf{M}_{\mathbf{Y}}\mathbf{G}'}{n}\right),$$

where, as we said we would, we are now defining $\mathbf{G} = \nabla \mathbf{g}(\boldsymbol{\theta})$ to simplify notation. We say that the right hand side is the *asymptotic distribution* of $\mathbf{g}(\mathbf{T}_n)$.

Example 8.2.2.
Suppose

$$\mathbf{Z}_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix}, \qquad i = 1, 2, \ldots$$

are an i. i. d. sequence of random vectors with mean vector $\boldsymbol{\mu}$ and variance matrix $\mathbf{M}$. Suppose we are interested in the parameter

$$\omega = \log\left(\frac{\mu_1}{\mu_2}\right) = \log(\mu_1) - \log(\mu_2)$$

The continuous mapping theorem applied to the LLN gives

$$W_n = \log(\overline{X}_n) - \log(\overline{Y}_n) \xrightarrow{P} \omega$$

and we want to use the delta method to find the asymptotic distribution of the difference $W_n - \omega$. The "g" involved is

$$g(x, y) = \log(x) - \log(y)$$

which has partial derivatives

$$\begin{aligned} \frac{\partial g(x,y)}{\partial x} &= \frac{1}{x} \\ \frac{\partial g(x,y)}{\partial y} &= -\frac{1}{y} \end{aligned}$$

1. If a function is differentiable, then the derivative is the matrix of partial derivatives.
2. If the partial derivatives exist and are continuous, then the function is differentiable.
3. A function can be differentiable without the partial derivatives being continuous. The exact condition required is the multivariate analog of (8.18) in the proof of the univariate delta method

$$\mathbf{g}(\mathbf{t}) - \mathbf{g}(\boldsymbol{\theta}) = \mathbf{G}(\mathbf{t} - \boldsymbol{\theta}) + \|\mathbf{t} - \boldsymbol{\theta}\| \mathbf{w}(\mathbf{t}) \tag{8.20}$$

where the double vertical bars indicate the norm of a vector

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

A function $\mathbf{g}$ is differentiable at $\boldsymbol{\theta}$ if there exists a matrix $\mathbf{G}$ and a function $\mathbf{w}$ that is continuous at $\boldsymbol{\theta}$ with $\mathbf{w}(\boldsymbol{\theta}) = 0$ such that (8.20) holds, in which case $\mathbf{G}$ is the derivative matrix $\nabla \mathbf{g}(\boldsymbol{\theta})$.

The nice thing about this definition, pedantic though it may be, is that it makes the proof of the multivariate delta method just like the proof of the univariate proof. Start from (8.20) and proceed just like the univariate proof, changing notation as necessary.

Thus the derivative matrix is

$$\nabla g(x, y) = \begin{pmatrix} \frac{1}{x} & -\frac{1}{y} \end{pmatrix}$$

Evaluating at $\boldsymbol{\mu}$, we get

$$\mathbf{G} = \begin{pmatrix} \frac{1}{\mu_1} & -\frac{1}{\mu_2} \end{pmatrix}$$

As always, the asymptotic distribution produced by the delta method has mean $g(\boldsymbol{\mu}) = \omega$ and variance $\mathbf{GMG}'/n$. We just have to work out the latter

$$\begin{pmatrix} \frac{1}{\mu_1} & -\frac{1}{\mu_2} \end{pmatrix} \begin{pmatrix} m_{11} & m_{12} \\ m_{12} & m_{22} \end{pmatrix} \begin{pmatrix} \frac{1}{\mu_1} \\ -\frac{1}{\mu_2} \end{pmatrix} = \frac{m_{11}}{\mu_1^2} - \frac{2m_{12}}{\mu_1 \mu_2} + \frac{m_{22}}{\mu_2^2}$$

If you prefer to phrase everything in terms of the usual notation for the moments of the components X and Y, this becomes

$$\sigma_W^2 = \frac{\sigma_X^2}{\mu_X^2} - \frac{2\rho_{X,Y}\sigma_X\sigma_Y}{\mu_X\mu_Y} + \frac{\sigma_Y^2}{\mu_Y^2}$$

Thus the result of applying the delta method is

$$\sqrt{n}(W_n - \omega) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_W^2),$$

where the asymptotic variance σ_W^2 is defined above.

8.2.3 Asymptotics for Sample Moments

This section supplies the proof of Theorem 7.16 which we stated in Chapter 7 but could not prove because it requires the multivariate delta method.

Proof of Theorem 7.16. For ordinary moments, this is a homework problem (Problem 7-17 in Lindgren).

For $M_{k,n}$ we proceed as in the proof of Theorem 7.15, using (7.26b), which implies

$$\sqrt{n}(M_{k,n} - \mu_k) = \sqrt{n}(M'_{k,n} - \mu_k) + \sum_{j=1}^{k} \binom{k}{j} (-1)^j \sqrt{n}(\overline{X}_n - \mu)^j M'_{k-j,n}$$

the first term arising from the $j = 0$ term in (7.26b). Now the CLT says

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathcal{D}} Z$$

where $Z \sim \mathcal{N}(0, \mu_2)$, because $\mu_2 = \sigma^2$. Then Slutsky's theorem implies

$$\sqrt{n}(\overline{X}_n - \mu)^j \xrightarrow{\mathcal{D}} 0$$

for any $j > 1$ (Problem 7-15). Thus all the terms for $j > 1$ make no contribution to the asymptotics, and we only need to figure out the asymptotics of the sum of the first two ($j = 0$ and $j = 1$) terms

$$\sqrt{n}(M'_{k,n} - \mu_k) - k\sqrt{n}(\overline{X}_n - \mu)M'_{k-1,n}.$$

By Slutsky's theorem and (7.29) this converges to

$$W - k\mu_{k-1}Z \tag{8.22}$$

where W and Z are defined by the multivariate CLT

$$\sqrt{n}\begin{pmatrix} \overline{X}_n - \mu \\ M'_{k,n} - \mu_k \end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix} Z \\ W \end{pmatrix} \sim \mathcal{N}(0, \mathbf{M})$$

where

$$\mathbf{M} = \operatorname{var}\begin{pmatrix} X_i - \mu \\ (X_i - \mu)^k \end{pmatrix} = \begin{pmatrix} \mu_2 & \mu_{k+1} \\ \mu_{k+1} & \mu_{2k} - \mu_k^2 \end{pmatrix}$$

Now calculating the variance of (8.22) using the usual formulas for the variance of a sum gives the asserted asymptotic variance in (7.31). □

8.2.4 Asymptotics of Independent Sequences

In several places throughout the course we will need the following result. In particular, we will use it in the section immediately following this one.

Theorem 8.10. *Suppose*

$$\begin{aligned} X_n &\xrightarrow{\mathcal{D}} X \\ Y_n &\xrightarrow{\mathcal{D}} Y \end{aligned} \tag{8.23}$$

and all of the X_i are independent of all of the Y_i, and suppose

$$\begin{aligned} k_n &\to \infty \\ m_n &\to \infty \end{aligned}$$

Then

$$\begin{pmatrix} X_{k_n} \\ Y_{m_n} \end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix} X \\ Y \end{pmatrix}$$

where the right hand side denotes the random vector having independent components having the same marginal distributions as the variables in (8.23)*.*

As with many of the theorems in this section, we omit the proof.[2] The theorem seems very obvious. In fact, the marginal laws must be as stated in the theorem by the continuous mapping theorem (the map that takes a vector to one of its components being continuous). So the only nontrivial assertion of the theorem is that the joint distribution of the limiting random variable has

[2] It It can be proved fairly easily from the relationship between characteristic functions and convergence in distribution, slightly misstated as Theorem 26 of Chapter 4 in Lindgren and the characteristic function uniqueness theorem, Theorem 25 of Chapter 4 in Lindgren, or more precisely from the multivariate versions of these theorems, but since we gave no proof of those theorems and didn't even state their multivariate versions, there seems no point in proofs using them.

independent components. That seems obvious. What else could happen? The only point of stating the theorem is to point out that this actually needs a proof, which is given in texts on advanced probability theory.

The conclusion of the theorem is sometimes stated a bit less precisely as

$$\begin{pmatrix} X_k \\ Y_m \end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix} X \\ Y \end{pmatrix} \qquad \text{as } k \to \infty \text{ and } m \to \infty$$

(you have to imagine the sequences k_n and m_n if you want a more precise statement), the point being that whenever k and m are both large the distribution of the left hand side is close to the distribution of the right hand side (so the latter can be used as an approximation of the former).

Corollary 8.11. *Under the hypotheses of the theorem*

$$X_k + Y_m \xrightarrow{\mathcal{D}} X + Y, \qquad \textit{as } k \to \infty \textit{ and } m \to \infty,$$

where X and Y are independent random variables having the same marginal distributions as the variables in (8.23)*.*

This follows directly from the theorem by the continuous mapping theorem for multivariate convergence in distribution (addition of components of a vector being a continuous operation).

8.2.5 Asymptotics of Sample Quantiles

In this section we give a proof of Theorem 7.27, which we were also unable to give in Chapter 7 because it too requires the multivariate delta method. We give a proof not because it represents a useful technique. The proof is a rather specialized trick that works only for this particular theorem. The reason we give the proof is to show how asymptotic normality arises even when there are no obvious averages anywhere in sight. After all, sample quantiles have nothing to do with any averages. Still, asymptotic normality arises anyway. This is typical. Most statistics that arise in practice are asymptotically normal whether or not there is any obvious connection with the CLT. There are exceptions (Problem 7-7), but they arise rarely in practice.

Before we begin the proof, we take a closer look at the relationship between the general case and the $\mathcal{U}(0, 1)$ case. It turns out that the latter can be derived from the former using the so-called quantile transformation.

Lemma 8.12. *Suppose X is a continuous random variable having an invertible c. d. f. F, then $F(X)$ has the $\mathcal{U}(0, 1)$ distribution. Conversely if $U \sim \mathcal{U}(0, 1)$, then $F^{-1}(U)$ has the same distribution as X.*

The first assertion is Theorem 9 of Chapter 3 in Lindgren. The second assertion is a special case of Problem 3-35 in Lindgren. The transformation $X = F^{-1}(U)$ is called the *quantile transformation* because it maps p to the p-th quantile x_p, and the transformation $U = F(X)$ is called the *inverse quantile* *transformation.* These transformations are a bit odd at first sight because they use F two different ways, both as a c. d. f. and as a change-of-variable function. From calculus, we know that these two transformations have derivatives that are inverses of each other, that is, if $u = F(x)$, so $x = F^{-1}(u)$, then

$$f(x) = F'(x) = \frac{d}{dx} F(x)$$

and

$$\frac{d}{du} F^{-1}(u) = \frac{1}{f(x)}. \tag{8.24}$$

Because we want to use the quantile transformation, we need to add an additional condition to the theorem, that the variables have an invertible c. d. f., which will be the case when $f(x) > 0$ for all x by Lemma 7.4 (the theorem is true without the additional condition, the proof is just a bit messier).

Proof of Theorem 7.27 assuming an invertible c. d. f. First we show how to derive the general case from the $\mathcal{U}(0,1)$ case

$$\sqrt{n}\bigl(U_{(k_n)} - p\bigr) \xrightarrow{\mathcal{D}} \mathcal{N}\bigl(0, p(1-p)\bigr), \tag{8.25}$$

where the $U_{(k)}$ are the order statistics of a sample from the $\mathcal{U}(0,1)$ distribution. Apply the quantile transformation so that $F^{-1}\bigl(U_{(k_n)}\bigr)$ has the same distribution as $X_{(k_n)}$ and apply the delta method with the derivative of the transformation given by (8.24). The result is assertion of the theorem (7.38). Thus it is enough to prove (8.25).

Now we use some facts about the relationships between various "brand name" distributions. The distribution of $U_{(k)}$ is $\text{Beta}(k, n-k+1)$. By Theorem 4.2, this distribution is the same as the distribution of $V_k/(V_k + W_k)$, where V_k and W_k are independent and

$$\begin{aligned} V_k &\sim \text{Gam}(k, \lambda) \\ W_k &\sim \text{Gam}(n-k+1, \lambda) \end{aligned}$$

where λ can have any value, for simplicity chose $\lambda = 1$. Then we use the normal approximation for the gamma distribution (Appendix C of these notes) which arises from the addition rule for the gamma distribution and the CLT

$$\begin{aligned} V_k &\approx \mathcal{N}(k, k) \\ W_k &\approx \mathcal{N}(n-k+1, n-k+1) \end{aligned}$$

So

$$\sqrt{n}\left(\frac{V_{k_n}}{n} - \frac{k_n}{n}\right) \approx \mathcal{N}\left(0, \frac{k_n}{n}\right)$$

and because of the assumption (7.37) and Slutsky's theorem we can replace k_n/n on both sides by p giving

$$\sqrt{n}\left(\frac{V_{k_n}}{n} - p\right) \approx \mathcal{N}\left(0, p\right),$$

and, similarly,

$$\sqrt{n}\left(\frac{W_{k_n}}{n} - (1-p)\right) \approx \mathcal{N}\left(0, 1-p\right).$$

Because of the assumed independence of V_k and W_k we can use Theorem 8.10 to get a multivariate CLT for the joint distribution of these two random vectors

$$\sqrt{n}\begin{pmatrix} \frac{V_{k_n}}{n} - p \\ \frac{W_{k_n}}{n} - (1-p) \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M})$$

where

$$\mathbf{M} = \begin{pmatrix} p & 0 \\ 0 & 1-p \end{pmatrix}$$

Note that we can write this as

$$\sqrt{n}(\mathbf{T}_n - \boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M})$$

where

$$\mathbf{T}_n = \frac{1}{n}\begin{pmatrix} V_{k_n} \\ W_{k_n} \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\theta} = \begin{pmatrix} p \\ 1-p \end{pmatrix}$$

So we apply the multivariate delta method to this convergence in distribution result. Note that $g(V_{k_n}/n, W_{k_n}/n)$ has the same distribution as $U_{(k)}$. Hence we want to use the transformation

$$g(v, w) = \frac{v}{v+w},$$

which has partial derivatives

$$\begin{aligned} \frac{\partial g(v,w)}{\partial v} &= \frac{w}{(v+w)^2} \\ \frac{\partial g(v,w)}{\partial w} &= -\frac{v}{(v+w)^2} \end{aligned}$$

and derivative matrix

$$\mathbf{G} = \nabla g(\boldsymbol{\theta}) = \begin{pmatrix} 1-p & -p \end{pmatrix}$$

Thus, finally, we see that $U_{(k)}$ is asymptotically normal with mean

$$g(\boldsymbol{\theta}) = p$$

and variance

$$\mathbf{G}\mathbf{M}\mathbf{G}' = \begin{pmatrix} 1-p & -p \end{pmatrix}\begin{pmatrix} p & 0 \\ 0 & 1-p \end{pmatrix}\begin{pmatrix} 1-p \\ -p \end{pmatrix} = p(1-p)$$

and we are done. □

Problems

8-1. Suppose X is a random scalar with ordinary moments $\alpha_k = E(X^k)$.

(a) What are the mean vector and variance matrix of the random vector

$$\mathbf{Z} = \begin{pmatrix} X \\ X^2 \\ X^3 \end{pmatrix}$$

(b) Suppose $\mathbf{Z}_1$, $\mathbf{Z}_2$, $\ldots$ is an i. i. d. sequence of random vectors having the same distribution as $\mathbf{Z}$. What are the mean vector and variance matrix of $\overline{\mathbf{Z}}_n$?

8-2. Suppose Y is a random scalar having mean μ and variance σ^2 and $\mathbf{Z}$ is a random vector with i. i. d. components Z_i having mean zero and variance τ^2, and suppose also that Y is independent of $\mathbf{Z}$. Define $\mathbf{X} = Y + \mathbf{Z}$ (that is, $\mathbf{X}$ has components $X_i = Y + Z_i$).

(a) What are the mean vector and variance matrix of $\mathbf{X}$?

(b) Suppose $\mathbf{X}_1$, $\mathbf{X}_2$, $\ldots$ is an i. i. d. sequence of random vectors having the same distribution as $\mathbf{X}$. What is the asymptotic distribution of $\overline{\mathbf{X}}_n$?

8-3. Suppose

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}} Y.$$

Use the delta method to find convergence in distribution results for

(a) $\log(T_n)$

(b) $\sqrt{T_n}$

(c) $\dfrac{\exp(T_n)}{1 + \exp(T_n)}$

Note: In (a) and (b) we need to assume $\theta > 0$.

8-4. Suppose X_1, X_2, X_3, $\ldots$ is an i. i. d. sequence of $\text{Ber}(p)$ random variables.

(a) State the LLN for $\overline{X}_n$, expressing all constants in terms of the parameter p (that is, don't use μ and σ, express them as functions of p).

(b) State the CLT for $\overline{X}_n$, expressing all constants in terms of the parameter p.

(c) To what does $\overline{X}_n(1 - \overline{X}_n)$ converge in probability? What theorem allows you to conclude this?

(d) Use the delta method to determine the asymptotic distribution of the random variable $\overline{X}_n(1 - \overline{X}_n)$. (Note: there is something funny about the case $p = 1/2$. Theorem 8.8 applies but its conclusion doesn't satisfactorily describe the "asymptotic distribution".)

8-5. Suppose $X_n \xrightarrow{\mathcal{D}} X$ where $X \sim \mathcal{N}(0, 1)$. To what does X_n^2 converge in distribution? What is the *name* of the limiting distribution (it is some "brand name" distribution). What theorem allows you to conclude this?

8-6. Suppose X_1, X_2, X_3, $\ldots$ is an i. i. d. sequence of random variables with mean μ and variance σ^2, and $\overline{X}_n$ is the sample mean. To what does

$$\sqrt{n} \left(\overline{X}_n - \mu \right)^2$$

converge in probability? (**Hint:** Use the CLT, the continuous mapping theorem for convergence in distribution, Slutsky's theorem, and Lemma 8.5.)

8-7. Suppose X_1, X_2, X_3, $\ldots$ is an i. i. d. sequence of random variables with mean μ and variance σ^2, and $\overline{X}_n$ is the sample mean. Define

$$Y_i = a + bX_i,$$

where a and b are constants, and

$$\overline{Y}_n = a + b\overline{X}_n.$$

Derive the asymptotic distribution of $\overline{Y}_n$ in two different ways.

(a) Use the delta method with $g(u) = a + bu$.

(b) Use the CLT applied to the sequence Y_1, Y_2, $\ldots$.

8-8. Suppose

$$\mathbf{Z}_i = \begin{pmatrix} X_i \\ Y_i \end{pmatrix}, \qquad i = 1, 2, \ldots$$

are an i. i. d. sequence of random vectors with mean vector $\boldsymbol{\mu}$ and variance matrix $\mathbf{M}$. What is the asymptotic distribution of

$$W_n = \frac{\overline{X}_n}{\overline{Y}_n}$$

assuming $\mu_2 \neq 0$.

8-9. Suppose

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}} Y$$

Show that

$$T_n \xrightarrow{P} \theta.$$

(**Hint:** Lemma 8.5 and the continuous mapping theorem.)

8-10. In Example 8.1.3 we showed that S_n^2 and V_n have the same asymptotic distribution, which is given by Corollary 7.17 in Chapter 7 of these notes. Find the asymptotic distribution of S_n.

8-11. (This problem has nothing to do with convergence concepts. It is a lemma for the following problem.)

Consider two arbitrary events A and B, and, as usual, let I_A and I_B denote their indicator functions. Show that

$$\operatorname{cov}(I_A, I_B) = P(A \cap B) - P(A)P(B).$$

Hint: $I_{A \cap B} = I_A I_B$.

8-12. (This is the bivariate analog of Problem 6-4 of Chapter 6 of these notes.)

Suppose X_1, X_2, $\ldots$ are i. i. d. with common probability measure P, and define

$$\begin{aligned} Y_n &= I_A(X_n) \\ Z_n &= I_B(X_n) \end{aligned}$$

for some events A and B. Find the asymptotic distribution of the vector $(\overline{Y}_n, \overline{Z}_n)$.

Chapter 9

Frequentist Statistical Inference

9.1 Introduction

9.1.1 Inference

Statistics is Probability done backwards.

Probability theory allows us to do calculations given a probability model. If we assume a random variable X has a particular distribution, then we can calculate, at least in principle, $P(|X| \geq c)$ or $E(X)$ or $\operatorname{var}(X)$. Roughly speaking, given the distribution of X we can say some things about X. Statistics tries to solve the inverse problem: given an observation of X, say some things about the distribution of X. This is called *statistical inference*.

Needless to say, what statistics can say about the distribution of random data is quite limited. If we ask too much, the problem is impossible. In a typical situation, any value of the observed data is possible under any of the distributions being considered as possible models. Thus an observation of the data does not completely rule out any model. However, the observed data will be more probable under some models and less probable under others. So we ought to be able to say the data favor some distributions more than others or that some distributions seem very unlikely (although not impossible).

9.1.2 The Sample and the Population

Usually the data for a statistical problem can be considered a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ representing a sample from a population. The statistical problem is to infer the population distribution (which determines the probability distribution of $\mathbf{X}$). The simple-minded approach to this problem, which seems natural to those who have not been exposed to formal statistical thinking, is to just treat the sample as if it were the population. This is clearly a mistake, obvious as soon as it is described in these words. It violates the fundamental issue that statistics must deal with.

> *The sample is not the population.*

In introductory courses I try to repeat this slogan so often that it is drilled into students by sheer force of repetition. Despite its simplicity, it is very important to remember. I often see it violated by scientists who are trained in formal statistics and think they are correctly applying their statistical training. It's easy to do. Just a little bit of confusion obscures the issues enough to make one rely on intuition, which is always wrong.

> *Intuition is always wrong about statistics.*
>
> *Statistics courses don't develop good intuition about statistics. They teach you how to calculate instead of using intuition.*
>
> *Expert intuition is no better than anyone else's. Experts are just better at calculating and knowing what to calculate.*

Intuition treats the sample as the population. Whenever you ignore "the sample is not the population," you will say stupid things and do stupid things.

9.1.3 Frequentist versus Bayesian Inference

The word "frequentist" in the chapter title refers to a great philosophical divide in statistical inference. All statistical inference is divided into two parts. One part is generally called "Bayesian" because of the prominent role played by Bayes' rule. It will be covered in a later chapter. The other part has no generally accepted name. We could call it non-Bayesian, but that sounds negative and is also inaccurate because it also sometimes uses Bayes' rule. We could call it "statistical inference based on sampling distributions," which would be accurate but too long for everyday use.

Devotees of Bayesian inference generally call the other camp "frequentist." This is intended to be pejorative, saddling the enemy with the philosophical baggage of the frequentist philosophy of probability, which says that it only makes sense to talk about probabilities in an infinite sequence of identical random experiments. The idea is that it only makes sense to apply "frequentist" statistics to actual infinite sequences of identical random experiments. Since that doesn't describe any real data, one should never use "frequentist" statistics.

These days, however, no one is really a "frequentist" about probability. Probably no one, except a few philosophers, ever was. Everyone is a formalist, holding the view that anything that satisfies the probability axioms is probability (if it waddles like a duck and quacks like a duck, then it is a duck). This takes all the sting out of the "frequentist" label. No one minds the label, because everyone knows it isn't accurate. As we will see, the only thing required for so-called "frequentist" inference is probability models for data. It doesn't matter what you think probability really is.

9.2 Models, Parameters, and Statistics

A *statistical model* is a family of probability distributions. This differs from a *probability model*, which is only a single distribution.

9.2.1 Parametric Models

Often we consider probability models having an adjustable constant in the formula for the density.[1] Generically, we refer to such a constant as a *parameter* of the distribution. This notion was introduced back in Section 3.1 of these notes, and all of the "brand name distributions" described in Chapter 4 of these notes and Chapter 6 of Lindgren and summarized in Appendix B of these notes are examples of parametric models.

Usually, though not always, we use Greek letters for parameters to distinguish them from random variables (large Roman letters) and possible values of random variables (small Roman letters). Among the brand-name distributions (Appendix B of these notes), the only parameters we do not use Greek letters for are the success probability p occurring in the Bernoulli, binomial, geometric, and negative binomial distributions and the analogous vector parameter $\mathbf{p}$ of the multinomial distribution, the parameters s and t of the beta distribution, and the parameters a and b of the uniform distribution on the interval (a, b). All the rest are Greek letters.

When we say let X be a random variable having density f_θ, this means that for each fixed value of the parameter θ the function f_θ is a probability density, which means it satisfies (3.1a) and (3.1b) of Chapter 3 of these notes. Of course, we didn't use θ for the parameter of any brand-name distribution. The idea is that θ can stand for any parameter (for μ of the Poisson, for λ of the exponential, and so forth).

Each different value of the parameter θ gives a different probability distribution. As θ ranges over its possible values, which we call the *parameter space*, often denoted Θ when the parameter is denoted θ, we get a *parametric family* of densities

$$\{ f_\theta : \theta \in \Theta \}$$

Even the notation for parametric families is controversial. How can that be? Mere notation generate controversy? You can see it in the conflict between the notation $f_\theta(x)$ used in these notes and the notation $f(x \mid \theta)$ used in Lindgren.

Lindgren uses the same notation that one uses for conditional probability densities. The reason he uses that notation is because he belongs to the Bayesian

[1] In these notes and in the lecture we use the term *density* to refer to either of what Lindgren calls the *probability function* (p. f.) of a discrete distribution of the *probability density function* (p. d. f.) of a continuous distribution. We have two reasons for what seems at first sight a somewhat eccentric notion (failing to draw a terminological distinction between these two rather different concepts). First, these two concepts are special cases of a more general concept, also called *density*, explained in more advanced probability courses. Second, and even more important, these two concepts are used the same way in statistics, and it is a great convenience to say "density" rather than "p. f. or p. d. f." over and over.

camp, and as a matter of philosophical principle is committed to the notion that all parameters are random variables. Bayesians consider $f(x \mid \theta)$ the conditional distribution of the random variable X given the random variable θ. We can't use the "big X" and "little x" distinction for Greek letters because we often use the corresponding capital letters for something else. In particular, as explained above, we use Θ for the parameter space, not for the parameter considered as a random variable. But regardless of whether the "big X" and "little x" convention can be applied, the important point is that Bayesians do consider $f(x \mid \theta)$ a conditional density. The rest of the statistical profession, call them "non-Bayesians" is at least willing to think of parameters as not being random. They typically use the notation $f_\theta(x)$ to show that θ is an adjustable constant and is not being treated like a random variable.

It is also important to realize that in a statistical model, probabilities and expectations depend on the actual value of the parameter. Thus it is ambiguous to write $P(A)$ or $E(X)$. Sometimes we need to explicitly denote the dependence on the parameter by writing $P_\theta(A)$ and $E_\theta(X)$, just as we write $f_\theta(x)$ for densities.

Location-Scale Families

Location-scale families were introduced in Section 4.1 of Chapter 4 of these notes. The only brand name location-scale families are $\mathcal{U}(a, b)$, $\mathcal{N}(\mu, \sigma^2)$, and $\text{Cauchy}(\mu, \sigma)$. But, as shown in Example 4.1.3 in Chapter 4 of these notes, there are many more "non-brand-name" location scale families. In fact, every probability distribution of a real-valued random variable generates a location-scale family.

The only reason for bring up location-scale families here is to make the point that a statistical model (family of probability distributions) can have many different parameterizations. Example 4.1.1 of Chapter 4 of these works out the relation between the usual parameters a and b of the $\mathcal{U}(a, b)$ distribution and the mean μ and variance σ, which can also be used to parameterize this family of distributions. The relation between the two parameterizations is given by unnumbered displayed equations in that example, which we repeat here

$$\begin{aligned} \mu &= \frac{a+b}{2} \\ \sigma &= \sqrt{\frac{(b-a)^2}{12}} \end{aligned}$$

This is an invertible change of parameters, the inverse transformation being

$$\begin{aligned} a &= \mu - \sigma\sqrt{3} \\ b &= \mu + \sigma\sqrt{3} \end{aligned}$$

This illustrates a very important principle.

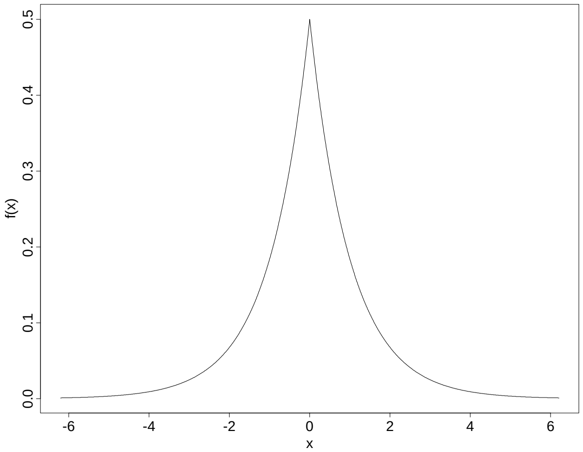
A single statistical model can have many different parameterizations.

We often change parameters, using the parameterization that seems simplest in a particular problem.

Example 9.2.1 (Laplace Distributions).
The density

$$f(x) = \frac{1}{2} e^{-|x|}, \qquad -\infty < x < +\infty \tag{9.1}$$

is called a *Laplace* or *double exponential* density. It is two Exp(1) densities back to back. The density is graphed below.



The Laplace location-scale family thus has densities

$$f_{\mu,\sigma}(x) = \frac{1}{2\sigma} e^{-|x-\mu|/\sigma}, \qquad -\infty < x < +\infty \tag{9.2}$$

Note (σ^2 is *not* the variance, see Problem 9-1).

Models and Submodels

Any set of probability distributions is a statistical model. A statistical model need not include *all* distributions of a certain type. It might have only a subset of them. We then say we have a *submodel* of the larger family. In parametric families, we specify submodels by specifying their parameter spaces.

Example 9.2.2 (All Normal Distributions).
The family of $\mathcal{N}(\mu, \sigma^2)$ distributions for $-\infty < \mu < +\infty$ and $0 < \sigma < \infty$ is a statistical model. (As mentioned in the preceding section, it is a location-scale family, μ is the location parameter and σ is the scale parameter.) Because the model has two parameters, the parameter space is a subset of $\mathbb{R}^2$

$$\Theta = \{\, (\mu, \sigma) \in \mathbb{R}^2 : \sigma > 0 \,\}.$$

Example 9.2.3 (Normal Distributions, Unit Variance).
The family of $\mathcal{N}(\mu, 1)$ distributions for $-\infty < \mu < +\infty$ is a also statistical model. It is a submodel of the family of all normal distributions in the preceding example. The model has one parameter μ, so the parameter space is a subset of $\mathbb{R}$. In fact, since μ is unrestricted, the parameter space is the whole real line: $\Theta = \mathbb{R}$.

It is important to realize that the examples describe two different models. It is not enough to say we are talking about a normal model. Different parameter spaces make different models.

Example 9.2.4 (Translation Families).
If we take a location-scale family and fix the scale parameter, then we have a one-parameter family. Example 9.2.3 is an example of this. Such a family is called a *location family* or a *translation family*, the latter name arising because the different random variables in the family are related by *translations*, which are changes of variables of the form

$$Y = \mu + X.$$

The distributions in the family differ only in location. They all have the same shape and scale.

Example 9.2.5 (Scale Families).
Conversely, if we take a location-scale family and fix the location parameter, then we also have a one-parameter family. But now the varying parameter is the scale parameter, so the family is called a *scale family.* The distributions in the family differ only in scale. They all have the same shape. They may also differ in location, because a scale transformation of the form

$$Y = \sigma X$$

changes both location and scale. For example, if X has a variance, then

$$\begin{aligned} E(Y) &= \sigma E(X) \\ \operatorname{var}(Y) &= \sigma^2 \operatorname{var}(X) \end{aligned}$$

9.2.2 Nonparametric Models

Some families of distributions are too big to specify in parametric form. No finite set of real parameters can serve to describe the family.

Example 9.2.6 (All Distributions with Finite Variance).
The family of all distributions with finite variance is a statistical model.

At the level of mathematics used in this course, it is hard to see that this model cannot be parameterized, but that does not really matter. The important point is that this is a statistical model even though we do not specify it using a parametric family of densities.

It is important to realize that probabilities and expectations depend on the actual probability distribution of the data, in nonparametric models just as in parametric models. Thus it is still ambiguous to write $P(A)$ or $E(X)$. The probabilities and expectations depend on the actual distribution of X. Since the model is not parametric, we cannot write $P_\theta(A)$ or $E_\theta(X)$ to remind us of the dependence. But it is still there and must be kept in mind.

9.2.3 Semiparametric Models

Sometimes, rather confusingly, we speak of parameters of nonparametric distributions. In this usage a *parameter* is any quantity that is determined by a distribution. In Example 9.2.6 we can still speak of the mean μ as being a parameter of the family. Every distribution in the family has a mean (because it has a variance and this implies existence of moments of lower order). Many different distributions in the family have the same mean, so the mean doesn't determine the distribution, and hence we don't have a parametric family with the mean as its parameter. But we still speak of the mean as being a parameter (rather than *the* parameter) of the family.

Models of this sort are sometimes called *semiparametric*, meaning they have a parametric part of the specification and a nonparametric part. In the example, the parametric part of the specification of the distribution is the mean μ and the nonparametric part is the rest of the description of the distribution (whatever that may be).

9.2.4 Interest and Nuisance Parameters

In multiparameter models, we divide parameters into two categories: *parameters of interest* (also called *interest parameters* though that is not idiomatic English) and *nuisance parameters*. The parameter or parameters of interest are the ones we want to know something about, the nuisance parameters are just complications. We have to deal with the nuisance parameters, but they are not interesting in themselves (in the particular application at hand).

In semiparametric models, the parametric part is typically the parameter of interest, the nonparametric part is the "nuisance" part of the model specification, although we can no longer call it the "nuisance parameter" when it is nonparametric.

Example 9.2.7 (All Distributions with Finite Variance).
The family of all distributions with finite variance is a semiparametric statistical model when we consider the mean μ the parameter of interest.

9.2.5 Statistics

Also somewhat confusingly, the term "statistic" is used as a technical term in this subject. Please do not confuse it with the name of the subject, "statistics." A *statistic* is a function of the data of a random experiment. It cannot involve any parameters or otherwise depend on the true distribution of the data. Since the data make up a random vector and a function of a random variable is a random variable, statistics are random variables. But a random variable can depend on a parameter, and a statistic cannot.

> *All statistics are random variables, but some random variables are not statistics.*

For each fixed μ, the function $X - \mu$ is (if X is the data) a random variable. But if μ is a parameter of the statistical model under consideration, $X - \mu$ is not a statistic. The reason for the distinction is that assuming a statistical model doesn't completely specify the distribution. It only says that X has some distribution in the family, but doesn't say which one. Hence we don't know what μ is. So we can't calculate $X - \mu$ having observed X. But we can calculate any statistic, because a statistic is a function of the observed data only.

9.3 Point Estimation

One form of statistical inference is called *point estimation.* Given data X_1, ..., X_n that are a random sample from some population or that are i. i. d. having a distribution in some statistical model, the problem is to say something about a parameter θ of the population or model, as the case may be. A *point estimate* (also called *point estimator*) of the parameter is a function of the data

$$\hat{\theta}_n(\mathbf{X}) = \hat{\theta}_n(X_1, \ldots, X_n) \tag{9.3}$$

that we use as an estimate of the true unknown parameter value.

This is our first example of "hat" notation. The symbol $\hat{\theta}$ is read "theta hat," the symbol on top of the letter being universally called a "hat" in mathematical contexts (outside math it is called a "caret" or "circumflex accent"). It changes the convention that parameters are Greek letters (like θ) and random variables are Roman letters (like X). Now we are adding Greek letters with hats to the random variables. Since $\hat{\theta}_n$ given by (9.3) is a function of the data, it is a random variable (and not just a random variable, more precisely, it is a *statistic*). The reason we do this is to make the connection between the two clear: $\hat{\theta}_n$ is a point estimator of θ. We often denote a point estimator of a parameter by putting a hat on the parameter. Remember that this puts $\hat{\theta}_n$ in a different conceptual category from θ. The point estimate $\hat{\theta}_n$ is a random variable. The parameter θ is a nonrandom constant.

Example 9.3.1 (Estimating the Mean).
Given i. i. d. data X_1, ..., X_n from a distribution having a mean μ, one point estimator of μ is the sample mean $\overline{X}_n$ defined by (7.15). Another point estimator of μ is the sample median of the empirical distribution of the data $\widetilde{X}_n$ defined in Definition 7.1.4. Yet another point estimator of μ is the constant estimator $\hat{\mu}_n \equiv 42$ that completely ignores the data, producing the estimate 42 for any data.

It is important to understand that any function whatsoever of the data is a point estimate of θ as long as it is a statistic (a function of data values only, not parameters). Even really dumb functions, such as the constant function $\hat{\theta}_n \equiv 42$ that completely ignores the data, are point estimates. Thus calling a function of the data a "point estimate" doesn't say anything at all about the properties of the function except that it is a statistic. The only point of calling a statistic a point estimate of θ is to establish a context for subsequent discussion. It only says we *intend to use* the statistic as a point estimate.

As we just said, every statistic whatsoever is a point estimate of θ, but that doesn't end the story. Some will be better than others. The ones of actual interest, the ones that get used in statistical practice, are the good ones. Much of the theory of point estimation is about which estimators are good. In order to characterize good estimators, we need some criterion of goodness. In fact, there are several different criteria in common use, and different criteria judge estimators differently. An estimator might be good under some criteria and bad under others. But at least if we use an estimator that is good according to some particular criterion, that says something.

The most obvious criterion, how often an estimator is correct, is unfortunately worthless.

> *With continuous data, every continuous estimator is wrong with probability one.*

The true parameter value θ is just a point in the parameter space. We don't know which point, but it is some point. If $\hat{\theta}(\mathbf{X})$ is a continuous random variable, then $P_\theta\{\hat{\theta}(\mathbf{X}) = \theta\}$ is zero, because the probability of every point is zero, as is true for any continuous random variable.

9.3.1 Bias

An estimator T of a parameter θ is *unbiased* if $E_\theta(T) = \theta$, that is, if θ is the mean of the sampling distribution of T when θ is the true parameter value. An estimator that is not unbiased is said to be *biased*, and the difference

$$b(\theta) = E_\theta(T) - \theta$$

is called the *bias* of the estimator.

Example 9.3.2 (Estimating σ^2).
By equations 7.22a and 7.22b, S_n^2 is an unbiased estimator of σ^2 and V_n is a biased estimator of σ^2.

The bias of S_n^2 as an estimator of σ^2 is zero (zero bias is the same as unbiased).

The bias of V_n as an estimator of σ^2 is

$$E(V_n) - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

Note that the bias is negative, meaning that the estimator V_n is below the parameter σ^2 on average.

In a multiparameter problem with a single parameter of interest φ and a nuisance parameter ψ, an estimator T of φ is *unbiased* if $E_\theta(T) = \varphi$, where $\theta = (\varphi, \psi)$ is the parameter vector, and the difference

$$b(\theta) = b(\varphi, \psi) = E_\theta(T) - \varphi$$

is called the *bias* of the estimator. Note that the bias generally depends on both interest and nuisance parameters.

It is generally a bad idea to give mathematical concepts emotionally charged names, and this concept is particularly badly named. Naturally we all want to be unbiased, so we should avoid biased estimators right? Wrong! It is important to remember that this mathematical concept has nothing whatsoever to do with what we call bias in everyday life.

In fact, it can mean the exact opposite! Psychologists call an achievement test unbiased if it satisfies the statistical definition, if it has the correct expectation. The tests are supposed to predict grades in school, and the test is unbiased if it is wrong high and wrong low about equally often so that it is right on average. But grades in school are themselves biased in the everyday sense (unless teachers all turned into saints when I wasn't looking). So in order to be unbiased in the statistical sense the tests must accurately track whatever bias in the everyday sense there is in the grades.

Note that this argument has nothing whatsoever to do with whether the questions on the tests appear to be biased, which is what the argument about "culturally biased" tests usually revolves around. Whatever the appearances, enough cultural bias (or other kinds of bias) must be somehow built into the tests, perhaps without any effort on the part of the people constructing the tests, perhaps even despite efforts to avoid it, to exactly match whatever cultural bias (or whatever) is in grades.

There are also technical arguments against unbiased estimation. I once had a friend who claimed he was ambidextrous because he did equally poorly with both hands. That's the idea behind unbiased estimation, doing equally poorly high and low.

Example 9.3.3 (Constrained Estimation).
Suppose the parameter satisfies a constraint, for example, θ might be a variance, in which case $\theta \ge 0$. Any sensible estimator should take values in the parameter space. Hence we should have $T(\mathbf{X}) \ge 0$ for all values of the data $\mathbf{X}$. Suppose also that our statistical model consists of probability distributions with the same support, hence the same events of probability zero. Then $P_\theta\bigl\{T(\mathbf{X}) > 0\bigr\}$ is either zero for all θ or nonzero for all θ, and by Theorem 5 of Chapter 4 in Lindgren, this implies

$$E_\theta(T) = 0, \qquad \theta \in \Theta \tag{9.4a}$$

or

$$E_\theta(T) > 0, \qquad \theta \in \Theta \tag{9.4b}$$

If (9.4a) holds, then the estimator is biased because $E_\theta(T) = \theta$ does not hold when $\theta \neq 0$. If (9.4b) holds, then the estimator is also biased, because $E_\theta(T) = \theta$ does not hold when $\theta = 0$.

Hence either way we get a biased estimator. The only way to get an unbiased estimator is to sometimes estimate ridiculous (that is, negative) values of the parameter.

So why specifically do I call the principle of unbiasedness the principle of "doing equally poorly with both hands" in this situation? The constraint allows you to do much better on the low side than the high side. Take any estimator $T(\mathbf{X})$, perhaps an unbiased one that sometimes estimates negative parameter values. The estimator is clearly improved by setting all the negative estimates to zero, that is,

$$T_{\text{improved}}(\mathbf{X}) = \begin{cases} T(\mathbf{X}), & T(\mathbf{X}) > 0 \\ 0, & \text{otherwise} \end{cases}$$

is a better estimator, because when $T(\mathbf{X})$ is not ridiculous (i. e., not negative) $T_{\text{improved}}(\mathbf{X})$ has the same value and when when $T(\mathbf{X})$ is ridiculous (negative) $T_{\text{improved}}(\mathbf{X})$ is not (is zero). But $T_{\text{improved}}(\mathbf{X})$ is *biased* whereas $T(\mathbf{X})$ may be *unbiased*. Adopting the principle of unbiasedness here means accepting that one should, as a matter of principle, increase the errors of estimation on the low side to make them as large as the inevitable errors on the high side. Stated that way, it is a mystery why anyone thinks unbiasedness is a good thing. (The solution to the mystery is that people who think unbiasedness is a good thing have never seen this example or other examples where unbiasedness is clearly a bad thing.)

Lest you think the example contrived, let me assure you that it does arise in practice, and I have actually seen real scientists using ridiculous estimators in order to achieve unbiasedness. They must have had a bad statistics course that gave them the idea that unbiasedness is a good thing.

Even though it is not a particularly good thing, unbiasedness is an important theoretical concept. We will meet several situations in which we can prove something about unbiased estimators, but can't do anything with estimators in general. For example, there are theorems that say under certain circumstances that a particular estimator is uniformly minimum variance unbiased (UMVU). It is easy to misinterpret the theorem to say that the best estimator is unbiased, but it doesn't say that at all. In fact, it implicitly says the opposite. It says the particular estimator is better than any other *unbiased* estimator. It says nothing about *biased* estimators, presumably some of them are better still, otherwise we could prove a stronger theorem.

Another issue about bias is that nonlinear functions of unbiased estimators are usually not unbiased. For example, suppose T is an unbiased estimator of θ. Is T^2 also an unbiased estimator of θ^2? No! By the parallel axis theorem

$$E_\theta(T^2) = \operatorname{var}_\theta(T) + E_\theta(T)^2.$$

Unless the distribution of T is concentrated at one point, $\operatorname{var}_\theta(T)$ is strictly greater than zero, and T^2 is biased high, that is, $E_\theta(T^2) > \theta^2$, when T is unbiased for θ. Conversely, if T^2 is unbiased for θ^2, then T is biased low for θ, that is $E_\theta(T) < \theta$.

Example 9.3.4 (Estimating σ).
Since S_n^2 is an unbiased estimator of σ^2, it follows that S_n itself is a biased estimator of σ, in fact $E(S_n) < \sigma$ always holds.

9.3.2 Mean Squared Error

The *mean squared error* (m. s. e.) of an estimator T of a parameter θ is

$$\operatorname{mse}_\theta(T) = E_\theta\bigl\{(T - \theta)^2\bigr\}$$

By the parallel axis theorem

$$\operatorname{mse}_\theta(T) = \operatorname{var}_\theta(T) + b(\theta)^2$$

So mean squared error is variance plus bias squared, and for unbiased estimators m. s. e. is just variance.

Mean squared error provides one sensible criterion for goodness of point estimators. If T_1 and T_2 are estimators of the same parameter θ, then we can say that T_1 is better than T_2 if $\operatorname{mse}(T_1) < \operatorname{mse}(T_2)$. It goes without saying that if we choose a different criterion, the order could come out differently.

An example of another criterion is mean absolute error $E_\theta\{|T - \theta|\}$, but that one doesn't work so well theoretically, because the parallel axis theorem doesn't apply, so there is less we can say about this criterion than about m. s. e.

Example 9.3.5.
Consider the class of estimators of σ^2 of the form kV_n, where $k > 0$ is some constant. The choice $k = 1$ gives V_n itself. The choice $k = n/(n-1)$ gives S_n^2. It turns out that neither of these estimators is the best in this class when we use mean squared error as the criterion. The best in the class is given by the choice $k = n/(n+1)$. No proof is given here. It is a homework problem (Problem 8-7 in Lindgren).

Note that the optimal estimator is *biased.* This gives us yet another example showing that unbiasedness is not necessarily a good thing. The same sort of calculation that shows the choice $k = n/(n+1)$ is optimal, also shows that $\operatorname{mse}(V_n) < \operatorname{mse}(S_n^2)$. So among the two more familiar estimators, the unbiased one is worse (when mean square error is the criterion).

9.3.3 Consistency

Definition 9.3.1 (Consistency).
A sequence of point estimators $\{T_n\}$ of a parameter θ is consistent *if*

$$T_n \xrightarrow{P} \theta, \qquad \text{as } n \to \infty.$$

Generally we aren't so pedantic as to emphasize that consistency is really a property of a sequence. We usually just say T_n is a consistent estimator of θ.

Consistency is not a very strong property, since it doesn't say anything about how fast the errors go to zero nor does it say anything about the distribution of the errors. So we generally aren't interested in estimators that are merely consistent unless for some reason consistency is all we want. We will see the most important such reason in the following section. For now we just list a few consistent estimators.

By the law of large numbers, if X_1, X_2, $\ldots$ are i. i. d. from a distribution with mean μ, then the sample mean $\overline{X}_n$ is a consistent estimator of μ. The only requirement is that the expectation defining μ exist.

Similarly, by Theorem 7.15 every sample moment (ordinary or central) is a consistent estimator of the corresponding population moment. The only requirement is that the population moment exist. Here we have fallen into the sloppy terminology of referring to i. i. d. random variables as a "sample" from a hypothetical infinite "population." What is meant, of course, is that if X_1, X_2, $\ldots$ are i. i. d. from a distribution having ordinary moment α_k or central moment μ_k and if the corresponding sample moments are $A_{k,n}$ and $M_{k,n}$ in the notation of Section 7.3.1, then

$$\begin{aligned} A_{k,n} &\xrightarrow{P} \alpha_k \\ M_{k,n} &\xrightarrow{P} \mu_k \end{aligned}$$

provided only that the moments α_k and μ_k exist.

That doesn't give us a lot of consistent estimators, but we can get a lot more with the following.

Theorem 9.1. *Any continuous function of consistent estimators is consistent. Specifically, if $T_{i,n} \xrightarrow{P} \theta_i$, as $n \to \infty$ for $i = 1, \ldots, m$, then*

$$g(T_{1,n}, \ldots, T_{m,n}) \xrightarrow{P} g(\theta_1, \ldots, \theta_m), \qquad \text{as } n \to \infty$$

if g is jointly continuous at the point $(\theta_1, \ldots, \theta_m)$.

This is just the multivariate version of the continuous mapping theorem for convergence in probability (Theorem 8.7).

9.3.4 Asymptotic Normality

As we said in the preceding section, mere consistency is a fairly uninteresting property, unless it just happens to be all we want. A much more important property is asymptotic normality. Another way to restate the definition of consistency is

$$T_n - \theta \xrightarrow{P} 0.$$

The estimator T_n is supposed to estimate the parameter θ, so $T_n - \theta$ is the error of estimation. Consistency says the error goes to zero. We would like to know more than that. We would like to know how about big the error is, more specifically we would like an approximation of its sampling distribution.

It turns out that almost all estimators of practical interest are not just consistent but also *asymptotically normal*, that is,

$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2) \tag{9.5}$$

holds for some constant σ^2, which may depend on the true distribution of the data. We say an estimator T_n that satisfies (9.5) is *consistent and asymptotically normal* (that is, asymptotically normal when centered at θ). It may be very hard, even impossible, to work out theoretically what the constant σ^2 actually is, although these days one can often use computer simulations to calculate it when pencil and paper analysis fails. Examples of consistent and asymptotically normal estimators are ordinary and central sample moments (Theorems 7.15 and 7.16).

The property (9.5) is not much help by itself, because if σ^2 actually depends on the true distribution of the data (that is, σ^2 is actually a function of the parameter θ, although the notation doesn't indicate this), then we don't know what it actually is because we don't know the true distribution (or the true value of θ). Then the following theorem is useful.

Theorem 9.2 (Plug-In). *Suppose* (9.5) *holds and S_n is any consistent estimator of σ, then*

$$\frac{T_n - \theta}{S_n / \sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

This was proved as a homework problem last semester, and we repeated the proof in Example 8.1.5. It is just (9.5) and Slutsky's theorem. We call this the "plug-in" theorem, because it says asymptotics still works when you plug in S_n for σ.

9.3.5 Method of Moments Estimators

A function of sample moments is called a *method of moments estimator* of a parameter θ if the function evaluated at the corresponding population moments is equal to θ.

Example 9.3.6.
Trivially, sample moments are "method of moments estimators" of the corresponding population moments.

Example 9.3.7 (The Two-Parameter Gamma Model).
The mean and variance of the $\text{Gam}(\alpha, \lambda)$ distribution are

$$\begin{aligned} \mu &= \frac{\alpha}{\lambda} \\ \sigma^2 &= \frac{\alpha}{\lambda^2} \end{aligned}$$

Solving for α and λ gives

$$\begin{aligned} \alpha &= \frac{\mu^2}{\sigma^2} \\ \lambda &= \frac{\mu}{\sigma^2} \end{aligned}$$

Plugging in the corresponding sample moments gives

$$\hat{\alpha}_n = \frac{\overline{X}_n^2}{V_n} \tag{9.6a}$$

$$\hat{\lambda}_n = \frac{\overline{X}_n}{V_n} \tag{9.6b}$$

These are method of moments estimators because they are functions of sample moments, for example

$$\hat{\alpha}_n = g(\overline{X}_n, V_n),$$

where

$$g(u, v) = \frac{u^2}{v}, \tag{9.7}$$

and the function evaluated at the population moments is the parameter to be estimated, for example

$$g(\mu, \sigma^2) = \frac{\mu^2}{\sigma^2} = \alpha.$$

Method of moments estimators are always consistent and asymptotically normal if enough population moments exist and they are nice functions of the sample moments.

Theorem 9.3. *A method of moments estimator involving sample moments of order k or less is consistent provided population moments of order k exist and provided it is a continuous function of the sample moments.*

This is just Theorem 7.15 combined with Theorem 9.1.

Theorem 9.4. *A method of moments estimator involving sample moments of order k or less is asymptotically normal provided population moments of order $2k$ exist and provided it is a differentiable function of the sample moments.*

The proof of this theorem is just the multivariate delta method (Theorem 8.9) applied to the multivariate convergence in distribution of sample moments, for which we have not stated a completely general theorem. What is needed is the multivariate analog of Theorem 7.16 which would give the asymptotic *joint* distribution of several sample moments, rather than the asymptotic *marginal* of just one. Rather than state such a general theorem, we will be content by giving the specific case for the first two moments.

Theorem 9.5. *If X_1, X_2, $\ldots$ are i. i. d. from a distribution having fourth moments, then*

$$\sqrt{n}\begin{pmatrix} \overline{X}_n - \mu \\ A_{2,n} - \alpha_2 \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M}_1), \tag{9.8}$$

where $A_{2,n}$ is the sample ordinary second moment and

$$\mathbf{M}_1 = \begin{pmatrix} \alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1\alpha_2 \\ \alpha_3 - \alpha_1\alpha_2 & \alpha_4 - \alpha_2^2 \end{pmatrix} \tag{9.9}$$

and

$$\sqrt{n}\begin{pmatrix} \overline{X}_n - \mu \\ V_n - \sigma^2 \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M}_2), \tag{9.10}$$

where

$$\mathbf{M}_2 = \begin{pmatrix} \mu_2 & \mu_3 \\ \mu_3 & \mu_4 - \mu_2^2 \end{pmatrix} \tag{9.11}$$

Proof. The first assertion of the theorem was proved in Examples 5.1.1 and 8.1.2. The second assertion was almost, but not quite, proved while we were proving Theorem 7.16. In that theorem, we obtained the asymptotic marginal distribution of V_n, but not the asymptotic joint distribution of $\overline{X}_n$ and V_n. However in the unlabeled displayed equation just below (8.22) we determined

$$\sqrt{n}\begin{pmatrix} \overline{X}_n - \mu \\ M'_{2,n} - \mu_2 \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \mathbf{M}_2)$$

where by the empirical central axis theorem

$$M'_{2,n} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 = V_n + (\overline{X}_n - \mu)^2$$

Hence

$$\sqrt{n}\begin{pmatrix} \overline{X}_n - \mu \\ M'_{2,n} - \mu_2 \end{pmatrix} = \sqrt{n}\begin{pmatrix} \overline{X}_n - \mu \\ V_n - \mu_2 \end{pmatrix} + \sqrt{n}\begin{pmatrix} 0 \\ (\overline{X}_n - \mu)^2 \end{pmatrix}$$

and by Problem 8-6 the second term converges in probability to zero, hence Slutsky's theorem gives the asserted result. □

Example 9.3.8.

In Example 8.2.1 we essentially did a method of moments estimator problem. We just didn't know at the time that "method of moments" is what statisticians call that sort of problem. There we looked at the asymptotic behavior of $1/\overline{X}_n$ for an i. i. d. sample from an $\text{Exp}(\lambda)$ distribution. From the fact that

$$E(X) = \frac{1}{\lambda}$$

we see that

$$\hat{\lambda}_n = \frac{1}{\overline{X}_n}$$

is the obvious method of moments estimator of λ. In Example 8.2.1 we calculated its asymptotic distribution

$$\hat{\lambda}_n \approx \mathcal{N}\left(\lambda, \frac{\lambda^2}{n}\right)$$

Example 9.3.9.
For the method of moments estimators $\hat{\alpha}_n$ and $\hat{\lambda}_n$ defined in Example 9.3.7, these theorems imply that these estimators are consistent and asymptotically normal, because the gamma distribution has moments of all orders and both estimators are differentiable functions of the sample moments they involve.

Getting the actual asymptotic distribution of the estimators is more work. We have to apply the multivariate delta method to the result of Theorem 9.5. We'll just do $\hat{\alpha}_n$ As was pointed out in Example 9.3.7 $\hat{\alpha}_n = g(\overline{X}_n, V_n)$, where the function g is given by (9.7), which has derivative

$$\mathbf{G} = \nabla g(\mu, \sigma^2) = \begin{pmatrix} \frac{2\mu}{\sigma^2} & -\frac{\mu^2}{\sigma^4} \end{pmatrix} \tag{9.12}$$

The specific form of the asymptotic variance matrix of $\overline{X}_n$ and V_n is given by (9.11) with the specific moments of the gamma distribution plugged in. Of course, we already know

$$\mu = \frac{\alpha}{\lambda}$$

and

$$\sigma^2 = \mu_2 = \frac{\alpha}{\lambda^2}$$

Plugging these into (9.12) gives

$$\mathbf{G} = \begin{pmatrix} 2\lambda & -\lambda^2 \end{pmatrix} \tag{9.13}$$

To calculate $\mathbf{M}_2$, we need to also calculate μ_3 and μ_4.

$$\begin{aligned}
\mu_3 &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \left(x - \frac{\alpha}{\lambda}\right)^3 x^{\alpha-1} e^{-\lambda x}\, dx \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \left(x^3 - \frac{3\alpha x^2}{\lambda} + \frac{3\alpha^2 x}{\lambda^2} - \frac{\alpha^3}{\lambda^3}\right) x^{\alpha-1} e^{-\lambda x}\, dx \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha+3-1} e^{-\lambda x}\, dx - \frac{3\alpha\lambda^{\alpha-1}}{\Gamma(\alpha)} \int_0^\infty x^{\alpha+2-1} e^{-\lambda x}\, dx \\
&\qquad + \frac{3\alpha^2\lambda^{\alpha-2}}{\Gamma(\alpha)} \int_0^\infty x^{\alpha+1-1} e^{-\lambda x}\, dx - \frac{\alpha^3\lambda^{\alpha-3}}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-\lambda x}\, dx \\
&= \frac{\lambda^\alpha \Gamma(\alpha+3)}{\lambda^{\alpha+3}\Gamma(\alpha)} - \frac{3\alpha\lambda^{\alpha-1}\Gamma(\alpha+2)}{\lambda^{\alpha+2}\Gamma(\alpha)} + \frac{3\alpha^2\lambda^{\alpha-2}\Gamma(\alpha+1)}{\lambda^{\alpha+1}\Gamma(\alpha)} - \frac{\alpha^3\lambda^{\alpha-3}\Gamma(\alpha)}{\lambda^\alpha\Gamma(\alpha)} \\
&= \frac{(\alpha+2)(\alpha+1)\alpha}{\lambda^3} - \frac{3\alpha(\alpha+1)\alpha}{\lambda^3} + \frac{3\alpha^2\alpha}{\lambda^3} - \frac{\alpha^3}{\lambda^3} \\
&= \frac{(\alpha+2)(\alpha+1)\alpha}{\lambda^3} - \frac{3\alpha(\alpha+1)\alpha}{\lambda^3} + \frac{3\alpha^2\alpha}{\lambda^3} - \frac{\alpha^3}{\lambda^3} \\
&= \frac{2\alpha}{\lambda^3}
\end{aligned}$$

A fairly horrible calculation, but we can check it in Mathematica. In fact we did this problem in the lab as the last question on the quiz. Let's do μ_4 in Mathematica.

```
In[1]:= <<Statistics`ContinuousDistributions`

In[2]:= dist = GammaDistribution[alpha, 1 / lambda]

                                   1
Out[2]= GammaDistribution[alpha, ------]
                                 lambda

In[3]:= f[x_] = PDF[dist, x]

                     -1 + alpha
                    x
Out[3]= ------------------------------------
         lambda x    1    alpha
        E          (------)      Gamma[alpha]
                    lambda

In[4]:= mu = Integrate[ x f[x], {x, 0, Infinity},
        Assumptions -> {alpha > 0 && lambda > 0} ]

              -1 - alpha
        lambda           Gamma[1 + alpha]
Out[4]= ---------------------------------
              1    alpha
           (------)      Gamma[alpha]
            lambda

In[5]:= mu = FullSimplify[mu]

                    -1 - alpha
        alpha lambda
Out[5]= ----------------------
               1    alpha
            (------)
             lambda

In[6]:= mu = PowerExpand[mu]

        alpha
Out[6]= ------
        lambda
```

```
In[7]:= mu4 = Integrate[ (x - mu)^4 f[x], {x, 0, Infinity},
        Assumptions -> {alpha > 0 && lambda > 0} ]

               -4 - alpha          4
Out[7]= (lambda           (-3 alpha  Gamma[alpha] +

                2
>        6 alpha  Gamma[2 + alpha] - 4 alpha Gamma[3 + alpha] +

                                  1    alpha
>        Gamma[4 + alpha])) / ((------)      Gamma[alpha])
                                lambda

In[8]:= mu4 = PowerExpand[FullSimplify[mu4]]

        3 alpha (2 + alpha)
Out[8]= -------------------
                    4
              lambda
```

Thus we finally obtain

$$\mu_4 - \mu_2^2 = \frac{2\alpha(3+\alpha)}{\lambda^4}$$

and

$$\mathbf{M}_2 = \begin{pmatrix} \frac{\alpha}{\lambda^2} & \frac{2\alpha}{\lambda^3} \\ \frac{2\alpha}{\lambda^3} & \frac{2\alpha(3+\alpha)}{\lambda^4} \end{pmatrix} \tag{9.14}$$

So the asymptotic variance of $\hat{\alpha}_n$ is

$$\begin{aligned} \mathbf{G}\mathbf{M}_2\mathbf{G}' &= \begin{pmatrix} 2\lambda & -\lambda^2 \end{pmatrix} \begin{pmatrix} \frac{\alpha}{\lambda^2} & \frac{2\alpha}{\lambda^3} \\ \frac{2\alpha}{\lambda^3} & \frac{2\alpha(3+\alpha)}{\lambda^4} \end{pmatrix} \begin{pmatrix} 2\lambda \\ -\lambda^2 \end{pmatrix} \\ &= 2\alpha(1+\alpha) \end{aligned}$$

and

$$\hat{\alpha}_n \approx \mathcal{N}\left(\alpha, \frac{2\alpha(1+\alpha)}{n}\right) \tag{9.15}$$

9.3.6 Relative Efficiency

Definition 9.3.2 (Relative Efficiency).
The relative efficiency *of two estimators of the same parameter is the ratio of their mean squared errors.*

Lindgren (p. 260) adds an additional proviso that the ratio must not depend on the parameter, but this needlessly restricts the concept. Of course, if the relative efficiency does depend on the parameter then in actual practice you don't know exactly what it is because you don't know the true parameter value. However, you can still make useful statements comparing the estimators. It might be, for example, that the relative efficiency is large for all likely values of the parameter.

Unfortunately, this criterion is almost useless except in toy problems because it is often impossible to calculate mean squared errors of complicated estimators. A much more useful criterion is given in the following section.

9.3.7 Asymptotic Relative Efficiency (ARE)

Definition 9.3.3 (Asymptotic Relative Efficiency).
The asymptotic relative efficiency *of two consistent and asymptotically normal estimators of the same parameter is the ratio of their asymptotic variances.*

Expressed in symbols, if S_n and T_n are two estimators of θ and

$$\sqrt{n}(S_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$
$$\sqrt{n}(T_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \tau^2)$$

then the asymptotic relative efficiency is the ratio of σ^2 to τ^2.

It is unimportant whether you say the ARE is σ^2/τ^2 or τ^2/σ^2. No one can remember which way is up anyway. It is much clearer if you say something like S_n is better than T_n and the ARE is 0.95. It is then clear that σ^2 is smaller than τ^2, because "better" means smaller asymptotic variance. Hence it is clear that the ARE is in this case σ^2/τ^2. Typically the ARE depends on the true distribution of the data.

Example 9.3.10 (Mean versus Median).
Whenever all the distributions in a statistical model are symmetric and have means, the center of symmetry is both the mean and the median. Hence both $\overline{X}_n$ and $\widetilde{X}_n$ are sensible estimators. Which is better?

Generally, it depends on the shape of the population distribution. For a concrete example, we will do the normal distribution. Then the asymptotic variance of $\widetilde{X}_n$ is given in Example 7.4.1, and the asymptotic variance of $\overline{X}_n$ is of course σ^2/n. Hence the sample mean is the better estimator and the ARE is $2/\pi = 0.6366$. Thus the sample median is only about 64% as efficient as the sample mean for normal populations.

For other population distributions the conclusion can be reversed and the sample median may be much better than the mean (Problem 9-12).

Why is ARE interesting? Why ratio of variances? Why not ratio of standard deviations for example? The reason is that ARE has a direct relation to actual costs. To get the same accuracy, we need the same variance. The asymptotic variances are σ^2/m and τ^2/n if we choose sample sizes m and n for the two estimators. So in order to have the same accuracy, we must have

$$m = \frac{\sigma^2}{\tau^2} n = \text{ARE} \times n$$

A large part of the costs of any random experiment will be proportional to the sample size. Hence ARE is the right scale, the one proportional to costs.

9.4 Interval Estimation

Since point estimators are never right, at least when the statistical model is continuous, it makes sense to introduce some procedure that is right some of the time. An *interval estimate* of a real-valued parameter θ is a random interval having endpoints that are statistics, call them $\hat{\theta}_L(\mathbf{X})$ and $\hat{\theta}_R(\mathbf{X})$, the L and R being for "left" and "right." The idea is that the interval estimate says the true parameter value is somewhere between these endpoints. The event $\hat{\theta}_L(\mathbf{X}) < \theta < \hat{\theta}_R(\mathbf{X})$ that the true parameter value is actually in the interval is described as saying the interval *covers* the parameter, and the probability of this event

$$P_\theta\bigl\{\hat{\theta}_L(\mathbf{X}) < \theta < \hat{\theta}_R(\mathbf{X})\bigr\} \tag{9.16}$$

is called the *coverage probability* of the interval estimator. This terminology, "interval estimate," "covers," and "coverage probability," is not widely used, appearing only in fairly technical statistical literature, but the same concepts are widely known under different names. The interval $\bigl(\hat{\theta}_L(\mathbf{X}), \hat{\theta}_R(\mathbf{X})\bigr)$ is called a *confidence interval* and the probability (9.16) is called the *confidence level*, conventionally expressed as a percentage. If the coverage probability is 0.95, then the interval is said to be a "95 percent confidence interval."

The careful reader may have noticed that an important issue was passed over silently in defining the coverage probability (9.16). As the notation indicates, the coverage probability depends on θ, but we don't know what the value of θ is. The whole point of the exercise is to estimate θ. If we knew what θ was, we wouldn't care about a confidence interval.

There are three solutions to this problem.

- Sometimes, by everything working out nicely, the coverage probability (9.16) does not actually depend on θ, so the issue goes away.

- Often, we can't calculate (9.16) exactly anyway and are using the central limit theorem or other asymptotic approximation to approximate the coverage probability. If the asymptotic approximation doesn't depend on θ, the issue goes away.

- Rarely, we can get a lower bound on the coverage probability, say

  $$P_\theta\bigl\{\hat{\theta}_L(\mathbf{X}) < \theta < \hat{\theta}_R(\mathbf{X})\bigr\} \ge p, \qquad \theta \in \Theta$$

  then we are entitled to call p expressed as a percentage the confidence level of the interval. This procedure is conservative but honest. It understates the actual coverage, but guarantees a certain minimum coverage.

9.4.1 Exact Confidence Intervals for Means

Theorem 9.6. *If X_1, …, X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$ where σ is known, then*

$$\overline{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{9.17}$$

is a $100(1 - \alpha)\%$ confidence interval for μ, where $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

The notation in (9.17) means that the confidence interval is

$$\overline{X}_n - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X}_n + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \tag{9.18}$$

Typically, confidence levels of 90%, 95%, or 99% are used. The following little table gives the corresponding $z_{\alpha/2}$ values.

confidence level	z critical value
90%	1.645
95%	1.960
99%	2.576

As the heading of the second column says, the $z_{\alpha/2}$ values are often called "z critical values" for a reason that will become apparent when we get to tests of significance.

The proof of the theorem is trivial. Equation (9.18) holds if and only if

$$\left| \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \right| < z_{\alpha/2}$$

The fraction in the absolute value signs is a standard normal random variable by Theorem 7.12. Hence the confidence level is

$$P(|Z| < z_{\alpha/2}) = 1 - 2P(Z > z_{\alpha/2}) = 1 - \alpha$$

by the symmetry of the normal distribution and the definition of $z_{\alpha/2}$.

This theorem is not much use in practical problems because σ is almost never known. Hence the following.

Theorem 9.7. *If X_1, …, X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$, then*

$$\overline{X}_n \pm t_{\alpha/2} \frac{S_n}{\sqrt{n}} \tag{9.19}$$

is a $100(1 - \alpha)\%$ confidence interval for μ, where $t_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the $t(n - 1)$ distribution.

The proof is again trivial, and follows exactly the same pattern as the previous theorem. Now, however, we can't use our little table of z critical values. We must use a big table of t critical values. The columns labeled 95, 97.5, and 99.5 in Table IIIb in the Appendix of Lindgren give the critical values $t_{\alpha/2}$ for 90%, 95%, and 99% confidence intervals respectively. (Why? Figure it out.) The bottom row of the table, labeled ∞ degrees of freedom gives the corresponding z critical values, so if you forget which column you need to use but can remember what the z critical value would be, that will tell you the right column. Also keep in mind that the degrees of freedom are $n - 1$, not n.

9.4.2 Pivotal Quantities

The random variables

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \tag{9.20a}$$

used in the proof of Theorem 9.6 and

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n - 1) \tag{9.20b}$$

used in the proof of Theorem 9.7 are called "pivotal quantities."

More generally, a random variable is called a *pivotal quantity* or *pivot* if its distribution does not depend on the true distribution of the data so long as that distribution remains in the statistical model under consideration. For a parametric model, this means the distribution of the the pivot does not depend on the parameters of the model. For a nonparametric model, we need the more general definition.

Any pivotal quantity that involves only one parameter can always be used to make confidence intervals for that parameter if its sampling distribution is known. If $g(\mathbf{X}, \theta)$ is a pivotal quantity with known sampling distribution, then we can find numbers a and b such that

$$P\{a < g(\mathbf{X}, \theta) < b\}$$

is our desired confidence level. Then

$$\{\, \theta \in \Theta : a < g(\mathbf{X}, \theta) < b \,\} \tag{9.21}$$

is the desired confidence interval, or perhaps to be precise we should say "confidence set" because the set (9.21) is not necessarily an interval, though it is in cases of practical interest.

Both theorems of the preceding section are examples of intervals derived by the method of pivotal quantities. Another interesting example is Example 8.8b in Lindgren which gives a confidence interval for the parameter of the $\text{Exp}(1/\theta)$ distribution (that is, θ is the mean) using an i. i. d. sample. The pivotal quantity is

$$\frac{2}{\theta} \sum_{i=1}^{n} X_i = \frac{2n\overline{X}_n}{\theta} \sim \text{chi}^2(2n)$$

That this random variable has the asserted distribution comes from Lemma 7.10. The exact assertion we need is given in the unnumbered displayed equation below (7.19) in the example following the lemma (recalling that $1/\theta = \lambda$).

Generally it is not clear how to choose a and b in (9.21) unless the sampling distribution of the pivot is symmetric, as it is for the confidence intervals in the preceding section. Lindgren in Example 8.8b chooses a so-called equal-tailed interval with a and b satisfying

$$P\{g(\mathbf{X}, \theta) < a\} = P\{b < g(\mathbf{X}, \theta)\}$$

but the only reason for doing this is the limitations of the chi-square table used to find a and b. With better tables or a computer, we find that if instead of the 5th and 95th percentiles of the $\text{chi}^2(20)$ used by Lindgren, we use the 0.086 and 0.986 quantiles we get the interval

$$\frac{2 \sum_i X_i}{36.35} < \theta < \frac{2 \sum_i X_i}{12.06}$$

which is about 8% shorter than the equal-tailed interval. In fact, this is the shortest possible 90% confidence interval based on this pivot.

9.4.3 Approximate Confidence Intervals for Means

If no pivotal quantity is known, there still may be an asymptotically pivotal quantity, a function $g_n(X_1, \ldots, X_n, \theta)$ satisfying

$$g_n(X_1, \ldots, X_n, \theta) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \tag{9.22}$$

regardless of the true distribution of the data so long as that distribution remains in the statistical model under consideration. Then

$$\{\, \theta \in \Theta : |g_n(X_1, \ldots, X_n, \theta)| < z_{\alpha/2} \,\} \tag{9.23}$$

is an asymptotic $100(1 - \alpha)\%$ confidence interval for θ, meaning the coverage probability converges to $1 - \alpha$ as $n \to \infty$, and where, as before, $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

Theorem 9.8. *If X_1, …, X_n are i. i. d. from a distribution having mean μ and finite variance σ^2 and S_n is any consistent estimator of σ, then*

$$\overline{X}_n \pm z_{\alpha/2} \frac{S_n}{\sqrt{n}} \tag{9.24}$$

is an asymptotic $100(1-\alpha)\%$ confidence interval for μ, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

This is Theorem 9.2 restated in confidence interval jargon.

Example 9.4.1 (I. I. D. Exponential).
In his Example 8.8b Lindgren gave the following exact equal-tailed 90% confidence interval for the mean θ of an exponential model

$$\frac{2 \sum_i X_i}{31.4} < \theta < \frac{2 \sum_i X_i}{10.9} \tag{9.25}$$

Here we compare this with the asymptotic confidence interval. Since the variance of the exponential is $\sigma^2 = \frac{1}{\lambda^2} = \theta^2$, $\overline{X}_n$ is a consistent estimator of both θ and σ. Hence by the theorem

$$\overline{X}_n \pm 1.645 \frac{\overline{X}_n}{\sqrt{n}}$$

is an asymptotic 90% confidence interval for θ. In Lindgren's example $n = 10$ so the asymptotic interval works out to be

$$0.48\overline{X}_n < \theta < 1.52\overline{X}_n \tag{9.26}$$

For comparison with the exact interval (9.25), we rewrite this as

$$\frac{2 \sum_i X_i}{30.4} < \theta < \frac{2 \sum_i X_i}{9.60}$$

Since $2 \sum_i X_i / \theta$ has a $\text{chi}^2(20)$ distribution (see Example 8.8b in Lindgren), the exact confidence level of this interval is

$$P(9.60 < \chi^2_{20} < 30.4) = 0.911$$

Not perfect, but not too shabby, especially since $n = 10$ is not a very large sample size. We don't usually expect agreement this good.

A useful bit of terminology for discussing asymptotic confidence intervals is the following. In the example, the approximate confidence interval (9.26) has a *nominal* confidence level of 90%, meaning only that we are calling it a 90% interval ("nominal" meaning having a certain name). The *actual* confidence level turns out to be 91.1%. In most applications we have no idea what the actual confidence level of an asymptotic interval really is. The CLT assures us that the actual level is close to the nominal if the sample size is large enough. But we rarely know how large is large enough.

There is in general no reason to expect that the actual level will be greater than the nominal level. It just happened to turn out that way in this example. In another application, actual might be less than nominal.

Most people would find the performance of the asymptotic interval satisfactory in this example and would not bother with figuring out the exact interval. In fact, with the single exception of intervals based on the t distribution, very few exact intervals are widely known or used. None (except t) are mentioned in most introductory statistics courses.

9.4.4 Paired Comparisons

In this section and in the next several sections we cover what is probably the single most useful application of statistics: comparison of the means of two populations. These can be divided into two kinds: paired comparisons and comparisons using independent samples.

In this section we deal with paired comparisons, leaving the other to following sections. The message of this section is that paired comparison problems naturally transform to the one-sample problems already studied. Hence they involve no new theory, just a new application of old theory.

In a paired comparisons problem we observe i. i. d. bivariate data (X_i, Y_i), $i = 1$, …, n. The parameter of interest is

$$\mu_X - \mu_Y = E(X_i) - E(Y_i) \tag{9.27}$$

The standard solution to this problem is to reduce the data to the random variables

$$Z_i = X_i - Y_i, \qquad i = 1, \ldots, n,$$

which are i. i. d. and have mean

$$\mu_Z = \mu_X - \mu_Y,$$

which is the parameter of interest. Hence standard one-sample procedures applied to the Z_i provide point estimates and confidence intervals in this case.

It is an important point that X_i and Y_i do not have to be independent. In fact it is sometimes better, in the sense of getting more accurate estimates of the parameter of interest, if they are dependent. The typical paired comparison situation has X_i and Y_i being different measurements on the same individual, say arm strength of left and right arms or MCAT scores before and after taking a cram course. When X_i and Y_i are measurements on the same individual, they are usually correlated.

The procedure recommended here that reduces the original data to the differences Z_i and then uses one-sample procedures is the only widely used methodology for analyzing paired comparisons. We will study other procedures for paired comparisons when we come to nonparametrics (Chapter 13 in Lindgren), but those procedures also use the same trick of reducing the data to the differences Z_i and then applying one-sample procedures to the Z_i. The only difference between the nonparametric procedures and those described here is that the nonparametric one-sample procedures are not based on the normal or t distributions and do not require normality of the population distribution.

9.4.5 Independent Samples

A more complicated situation where the paired difference trick is not appropriate arises when we have X_1, …, X_m i. i. d. from one population and Y_1, …, Y_n i. i. d. from another population. We assume the samples are independent, that is, $\mathbf{X} = (X_1, \ldots, X_m)$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$ are independent random vectors. The parameter of interest is still (9.27).

Now there is no obvious way to do pairing if $m \neq n$. Even if $m = n$, the pairing is arbitrary and unnatural when X_i and Y_i are measurements on independent randomly chosen individuals.

Asymptotic Confidence Intervals

The obvious estimate of $\mu_X - \mu_Y$ is $\overline{X}_m - \overline{Y}_n$, which has variance

$$\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n} \tag{9.28}$$

An obvious estimator of (9.28) is

$$\frac{S_{X,m}^2}{m} + \frac{S_{Y,n}^2}{n}$$

where $S_{X,m}$ and $S_{Y,n}$ are any consistent estimators of σ_X and σ_Y, such as the usual sample standard deviations. We can use this to construct asymptotic confidence intervals for the parameter of interest as follows.

Theorem 9.9. *Suppose X_1, X_2, ... are i. i. d. with mean μ_X and variance σ_X^2 and Y_1, Y_2, ... are i. i. d. with mean μ_Y and variance σ_Y^2 and $(X_1, \ldots, X_m)$ and $(Y_1, \ldots, Y_n)$ are independent random vectors for each m and n and*

$$S_{X,m} \xrightarrow{P} \sigma_X, \qquad \text{as } m \to \infty \tag{9.29a}$$

$$S_{Y,n} \xrightarrow{P} \sigma_Y, \qquad \text{as } n \to \infty \tag{9.29b}$$

Then

$$\frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\frac{S_{X,m}^2}{m} + \frac{S_{Y,n}^2}{n}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1), \qquad \text{as } m \to \infty \text{ and } n \to \infty.$$

Partial Proof. We will prove the assertion of the theorem under the additional condition that m and n go to infinity in a certain special way, that they are given by sequences m_k and n_k such that

$$\frac{\frac{\sigma_X^2}{m_k}}{\frac{\sigma_X^2}{m_k} + \frac{\sigma_Y^2}{n_k}} \to \alpha \tag{9.30}$$

where α is some constant, necessarily satisfying $0 \le \alpha \le 1$, since the left hand side of (9.30) is always between zero and one.

Then the CLT says

$$\frac{\overline{X}_{m_k} - \mu_X}{\sigma_X / \sqrt{m_k}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

$$\frac{\overline{Y}_{n_k} - \mu_Y}{\sigma_Y / \sqrt{n_k}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

If we let α_k denote the left hand side of (9.30), then by Corollary 8.11 and Slutsky's theorem

$$\begin{aligned} \frac{(\overline{X}_{m_k} - \overline{Y}_{n_k}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{m_k} + \frac{\sigma_Y^2}{n_k}}} &= \sqrt{\alpha_k} \frac{\overline{X}_{m_k} - \mu_X}{\sigma_X / \sqrt{m_k}} + \sqrt{1 - \alpha_k} \frac{\overline{Y}_{n_k} - \mu_Y}{\sigma_Y / \sqrt{n_k}} \\ &\xrightarrow{\mathcal{D}} \sqrt{\alpha} Z_1 + \sqrt{1 - \alpha} Z_2, \end{aligned}$$

where Z_1 and Z_2 are independent standard normal random variables. The limit is a linear combination of independent normal random variables, hence is normal. It has mean zero by linearity of expectation and variance $\alpha + (1 - \alpha) = 1$. Hence it is standard normal. Thus we have established

$$\frac{(\overline{X}_{m_k} - \overline{Y}_{n_k}) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{m_k} + \frac{\sigma_Y^2}{n_k}}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \tag{9.31}$$

Similarly

$$\frac{\frac{S_{X,m_k}^2}{m_k} + \frac{S_{Y,n_k}^2}{n_k}}{\frac{\sigma_X^2}{m_k} + \frac{\sigma_Y^2}{n_k}} = \alpha_k \frac{S_{X,m_k}^2}{\sigma_X^2} + (1 - \alpha_k) \frac{S_{Y,n_k}^2}{\sigma_Y^2} \xrightarrow{P} 1 \tag{9.32}$$

Combining (9.31) and (9.32) and using Slutsky's theorem gives the assertion of the theorem in the presence of our additional assumption (9.30).

The fact that the limit does not depend on α actually implies the theorem as stated (without the additional assumption) but this involves a fair amount of advanced calculus (no more probability) that is beyond the prerequisites for this course, so we will punt on the rest of the proof. □

Corollary 9.10.

$$\overline{X}_m - \overline{Y}_n \pm z_{\alpha/2} \sqrt{\frac{S_{X,m}^2}{m} + \frac{S_{Y,n}^2}{n}}$$

is an asymptotic $100(1 - \alpha)\%$ *confidence interval for* $\mu_X - \mu_Y$*, where* $z_{\alpha/2}$ *is the* $1 - \alpha/2$ *quantile of the standard normal distribution.*

Exact Confidence Intervals

Exact confidence intervals are more problematic. If we assume both populations are normal, then

$$\frac{(m - 1) S_{X,m}^2}{\sigma_X^2} \sim \text{chi}^2(m - 1) \quad \text{and} \quad \frac{(n - 1) S_{Y,n}^2}{\sigma_Y^2} \sim \text{chi}^2(n - 1) \tag{9.33}$$

and are independent. Hence the sum

$$\frac{(m - 1) S_{X,m}^2}{\sigma_X^2} + \frac{(n - 1) S_{Y,n}^2}{\sigma_Y^2} \tag{9.34}$$

is $\text{chi}^2(m+n-2)$. But this doesn't help much. Since it involves the population variances, which are unknown parameters, we can't use (9.34) to make a t distributed pivotal quantity that contains only the parameter of interest. Hence we can't use it to make an exact confidence interval.

In order to make progress, we need to add an additional assumption $\sigma_X = \sigma_Y = \sigma$. Then the variance of the point estimator $\overline{X}_m - \overline{Y}_n$ (9.28) becomes

$$\sigma^2 \left(\frac{1}{m} + \frac{1}{n} \right) \tag{9.35}$$

and (9.34) becomes

$$\frac{(m-1)S^2_{X,m} + (n-1)S^2_{Y,n}}{\sigma^2} \tag{9.36}$$

This gives us a useful pivot. Dividing the standard normal random variable

$$\frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

by the square root of (9.36) divided by its degrees of freedom gives

$$\frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{S_{p,m,n} \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t(m+n-2) \tag{9.37}$$

where

$$S^2_{p,m,n} = \frac{(m-1)S^2_{X,m} + (n-1)S^2_{Y,n}}{m+n-2}$$

It is clear from the fact that (9.36) is $\text{chi}^2(m+n-2)$ that $S^2_{p,m,n}$ is an unbiased estimator of σ^2, but that is not the reason we use it. Rather we use it because of the way the t distribution is defined. $S^2_{p,m,n}$ is called the "pooled" estimator of variance (hence the subscript p).

Thus under the assumptions that

- both samples are i. i. d.
- the samples are independent of each other
- both populations are exactly normal
- both populations have exactly the same variance

an exact $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is given by

$$\overline{X}_m - \overline{Y}_n \pm t_{\alpha/2} S_{p,m,n} \sqrt{\frac{1}{m} + \frac{1}{n}}$$

where $t_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the $t(m+n-2)$ distribution.

The sad truth about this procedure is that, although it is taught in many introductory statistics textbooks, it has (or should have) little practical application. The assumption $\sigma_X = \sigma_Y$ is the archetype of an assumption made for "reasons of mathematical convenience" rather than practical or scientific reasons. If we make the assumption, then we get an exact confidence interval. If we do not make the assumption, then we don't. But the assumption is almost never justified. If you don't know the true population means, how are you to know the population variances are the same?

Welch's Approximation

A better procedure was proposed by Welch in 1937. Unlike the the procedure of the preceding section, there is no set of assumptions that make it "exact." But it is correct for large m and n under any assumptions (like the asymptotic interval) and is a good approximation for small m and n. Welch's procedure uses the same asymptotically pivotal quantity

$$T = \frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}} \tag{9.38}$$

as the one used to make the asymptotic confidence interval. It just uses a better approximation to its sampling distribution than the $\mathcal{N}(0,1)$ approximation appropriate for large m and n.

The key idea goes as follows. The numerator of (9.38) is normal. When standardized, it becomes

$$Z = \frac{(\overline{X}_m - \overline{Y}_n) - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}}} \tag{9.39}$$

Recall that a random variable has a t distribution if it is a standard normal divided by the square root of a chi-square divided by its degrees of freedom. The quantity (9.38) can be rewritten $T = Z/\sqrt{W}$, where Z is given by (9.39) and

$$W = \frac{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}{\frac{\sigma^2_X}{m} + \frac{\sigma^2_Y}{n}} \tag{9.40}$$

Unfortunately, W is not a chi-square divided by its degrees of freedom, so (9.38) is not exactly t distributed. Welch's idea is that although W does not have exactly the desired distribution it will typically have approximately this distribution, so if we figure out the degrees of freedom ν for which a $\text{chi}^2(\nu)$ random variable divided by ν best approximates W, then the distribution of T will be approximately $t(\nu)$.

There could be several definitions of "best approximates." Welch's choice was to match moments. Rewrite W as

$$W = \lambda \frac{U}{m-1} + (1-\lambda)\frac{V}{n-1} \tag{9.41}$$

where

$$\begin{aligned} U &= \frac{(m-1)S_{X,m}^2}{\sigma_X^2} \\ V &= \frac{(n-1)S_{Y,n}^2}{\sigma_Y^2} \end{aligned}$$

and

$$\lambda = \frac{\frac{\sigma_X^2}{m}}{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}$$

Since $U \sim \text{chi}^2(m-1)$ and $V \sim \text{chi}^2(n-1)$ and λ is a constant, we can easily calculate moments.

$$E(W) = \lambda \frac{m-1}{m-1} + (1-\lambda)\frac{n-1}{n-1} = 1$$

which is the right expectation for a chi-square divided by its degrees of freedom, and

$$\begin{aligned} \operatorname{var}(W) &= \left(\frac{\lambda}{m-1}\right)^2 2(m-1) + \left(\frac{1-\lambda}{n-1}\right)^2 2(n-1) \\ &= 2\left[\frac{\lambda^2}{m-1} + \frac{(1-\lambda)^2}{n-1}\right] \end{aligned}$$

Since if $Y \sim \text{chi}^2(\nu)$, then

$$\operatorname{var}\left(\frac{Y}{\nu}\right) = \frac{1}{\nu^2}\operatorname{var}(Y) = \frac{2}{\nu}$$

the Y/ν that gives the best approximation to W in the sense of having the right mean and variance is the one with

$$\frac{1}{\nu} = \frac{\lambda^2}{m-1} + \frac{(1-\lambda)^2}{n-1} \tag{9.42}$$

Thus we arrive at Welch's approximation. The distribution of (9.38) is approximated by a $t(\nu)$ distribution where ν is defined by (9.42).

There are two problems with this approximation. First, we have no t tables for noninteger degrees of freedom and must use computers to look up probabilities. Second, we don't know know ν and must estimate it, using

$$\hat{\nu} = \frac{\left(\frac{S_{X,m}^2}{m} + \frac{S_{Y,n}^2}{n}\right)^2}{\frac{1}{m-1}\left(\frac{S_{X,m}^2}{m}\right)^2 + \frac{1}{n-1}\left(\frac{S_{Y,n}^2}{n}\right)^2} \tag{9.43}$$

Thus (finally) we arrive at the approximate confidence interval based on Welch's approximation. An approximate $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$ is given by

$$\overline{X}_m - \overline{Y}_n \pm t_{\alpha/2}\sqrt{\frac{S^2_{X,m}}{m} + \frac{S^2_{Y,n}}{n}}$$

where $t_{\alpha/2}$ is the $1-\alpha/2$ quantile of the $t(\hat{\nu})$ distribution, $\hat{\nu}$ being given by (9.43).

Example 9.4.2.
We make a confidence interval for the difference of means using the data in Example 9.7b in Lindgren. The statistics are (p. 327 in Lindgren)

	n	$\overline{X}$	S^2
Soil A	6	24	18.000
Soil B	8	29	23.714

Our estimate of the degrees of freedom is

$$\hat{\nu} = \frac{\left(\frac{18.000}{6} + \frac{23.714}{8}\right)^2}{\frac{1}{5}\left(\frac{18.000}{6}\right)^2 + \frac{1}{7}\left(\frac{23.714}{8}\right)^2} = 11.643$$

To get the critical value for this noninteger degrees of freedom we must interpolate in Table IIIb in the Appendix of Lindgren. The critical values for a 95% confidence interval are 2.20 for 11 d. f. and 2.18 for 12 d. f. Interpolating gives 2.19 for 11.6 d. f. R gives

```
> qt(0.975, 11.643)
[1] 2.186245
```

but 2.19 is good enough for all practical purposes.

The 95% confidence interval is thus

$$24 - 29 \pm 2.1862\sqrt{\frac{18.000}{6} + \frac{23.714}{8}}$$

which is $(-10.34, 0.34)$.

For comparison, the procedure of the preceding section gives an interval

$$24 - 29 \pm 2.18\sqrt{\frac{5 \cdot 18.000 + 7 \cdot 23.714}{12}\left(\frac{1}{6} + \frac{1}{8}\right)}$$

which is $(-10.435, 0.435)$.

The two confidence intervals are very similar. Why bother with the more complicated procedure? Because the "exact" procedure makes an assumption which is almost certainly false and is hence indefensible. If we had only done the exact procedure we would have no idea how wrong it was. It is only after we have also used Welch's procedure that we see that *in this particular case* the simpler procedure worked fairly well. In other cases, when the variances or sample sizes are more uneven, there will be an unacceptably large difference between the two answers. For this reason, many statistical computing packages now use Welch's approximation as the primary method of analysis of data like this or at least provide it as an option. Several introductory statistics texts (including the one I use for Statistics 3011) now explain Welch's approximation and recommend its use, although this is still a minority view in textbooks. Textbooks are slow to catch on, and it's only been 60 years.

Example 9.4.3 (Worse Examples).
This analyzes two artificial examples where the standard deviations and sample sizes vary by a factor of 3. First consider

n	S
10	1
30	3

Then

	standard error	d. f.
pooled	0.9733	38
Welch	0.6325	37.96

here "standard error" is the estimated standard deviation of the point estimate, the thing you multiply by the critical value to get the "plus or minus" of the confidence interval. The degrees of freedom, hence the critical values are almost the same, but the standard error using the "pooled" estimator of variance is way too big. Thus the interval is way too wide, needlessly wide, because the only reason it is so wide is that is based on an assumption $\sigma_X = \sigma_Y$ that is obviously false.

Now consider

n	S
10	3
30	1

Then

	standard error	d. f.
pooled	0.6213	38
Welch	0.9661	9.67

Here the "exact" procedure is more dangerous. It gives confidence intervals that are too narrow, not only wrong but wrong in the wrong direction, having far less than their nominal coverage probability.

For example, consider a difference of sample means of 2.0. Then the "exact" procedure gives a 95% confidence interval $2 \pm 2.0244 \cdot 0.6213$ which is

$(0.742, 3.258)$, whereas Welch's procedure gives a a 95% confidence interval $2 \pm 2.2383 \cdot 0.96609$ which is $(-0.162, 4.162)$. Using Welch's approximation to calculate probabilities, the coverage probability of the "exact" interval is about 77.7%. Of course, its coverage probability would be the nominal level 95% if the assumption of equal population variances were true, but here it is obviously false. Welch's approximation isn't exact, so we don't know what the true coverage probability actually is, but it is surely far below nominal.

9.4.6 Confidence Intervals for Variances

Sometimes confidence intervals for variances are wanted. As usual, these come in two kinds, asymptotic and exact.

Asymptotic Intervals

An asymptotic interval for σ^2 can be derived from asymptotic distribution for V_n given by Theorem 7.16 and the "plug-in" theorem (Theorem 9.2).

Theorem 9.11. *If X_1, $\ldots$, X_n are i. i. d. from a distribution having finite fourth moments, V_n is given by* (7.16) *and $M_{4,n}$ is any consistent estimator of μ_4, for example, the fourth sample central moment*

$$M_{4,n} = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^4$$

then

$$V_n \pm z_{\alpha/2} \sqrt{\frac{M_{4,n} - V_n^2}{n}} \tag{9.44}$$

is an asymptotic $100(1-\alpha)\%$ *confidence interval for σ^2, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.*

Example 9.4.4.
Suppose we have i. i. d. data, the sample size is $n = 200$, the sample second central moment is $V_n = 1.7314$, and the sample fourth central moment is $M_{4,n} = 14.2728$.

Plugging this into (9.44), we get

$$1.96 \sqrt{\frac{14.2728 - 1.7314^2}{200}} = 0.46537$$

for the half-width of the asymptotic 95% confidence interval, that is, the interval is 1.73 ± 0.47 or $(1.27, 2.20)$.

Exact Intervals

If the data are assumed to be exactly normally distributed, then by Theorem 7.24

$$\frac{nV_n}{\sigma^2} = \frac{(n-1)S_n^2}{\sigma^2} \sim \text{chi}^2(n-1)$$

and this is a pivotal quantity that can be used to make an exact confidence interval for σ^2 or σ. The calculations are almost exactly like those for the mean of the exponential distribution discussed in Section 9.4.2 that also involved a chi-square distributed pivotal quantity.

Theorem 9.12. *If $X_1, \ldots, X_n$ are i. i. d. $\mathcal{N}(\mu, \sigma^2)$ random variables, V_n is given by* (7.16)*, and $0 < \beta < \alpha < 1$, then*

$$\frac{nV_n}{\chi^2_{1-\alpha+\beta}} < \sigma^2 < \frac{nV_n}{\chi^2_\beta} \tag{9.45}$$

is an exact $100(1-\alpha)\%$ confidence interval for σ^2, where χ^2_β is the β-th quantile of the chi-square distribution with $n-1$ degrees of freedom.

Of course, one can replace nV_n by $(n-1)S_n^2$ both places it appears in (9.45), if one pleases. Usually, one uses $\beta = \alpha/2$ in (9.45), giving a so-called "equal-tailed" interval (equal chance of missing high or low), but other choices of β also give valid confidence intervals, and such intervals may be shorter than the equal-tailed interval.

Example 9.4.5.
Consider the data in Example 8.9a in Lindgren for which $n = 14$ and $S_n^2 = 85.912$. Suppose we want a 95% confidence interval for σ^2. To get an equal-tailed interval, we look up the 0.025 and 0.975 quantiles of the $\text{chi}^2(13)$ distribution in Table Vb of Lindgren. They are 5.01 and 24.7. Hence an exact 95% confidence interval is given by

$$\frac{(n-1)S_n^2}{24.7} < \sigma^2 < \frac{(n-1)S_n^2}{5.01}$$

which after plugging the values of n and S_n^2 becomes

$$45.15 < \sigma^2 < 222.98.$$

Taking square roots gives us a 95% confidence interval for σ

$$6.72 < \sigma < 14.93.$$

As always, we can find a shorter interval if we give up on the equal-tailed idea and use a computer search to find the β that gives the shortest interval for the desired confidence level. The β will depend on whether we are getting an interval for σ^2 or for σ. For σ^2, the optimal β is 0.0465 and the corresponding 95% confidence interval

$$36.12 < \sigma^2 < 192.90.$$

For σ, the optimal β is 0.0414 and the corresponding 95% confidence interval is

$$6.30 < \sigma < 14.08.$$

The Ratio of Variances (Two Samples)

We now go back to the situation of two independent samples from two populations studied in Section 9.4.5. The samples are $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ and the sample variances are denoted $S^2_{X,m}$ and $S^2_{Y,n}$, respectively. Then (9.33) gives the sampling distributions of these sample variances. If we divide the chi-square random variables by their degrees of freedom and form the ratio, we get an F random variable

$$\frac{S^2_{X,m}}{S^2_{Y,n}} \cdot \frac{\sigma^2_Y}{\sigma^2_X} \sim F(m-1, n-1).$$

Hence if a and b are numbers such that $P(a < X < b) = 1 - \alpha$ when $X \sim F(m-1, n-1)$, then

$$a \frac{S^2_{Y,n}}{S^2_{X,m}} < \frac{\sigma^2_Y}{\sigma^2_X} < b \frac{S^2_{Y,n}}{S^2_{X,m}}$$

is a $100(1-\alpha)\%$ confidence interval for the ratio of variances. Taking square roots gives a confidence interval for the ratio of standard deviations.

Of course, there are asymptotic confidence intervals for ratios of variances (or differences of variances) that do not require normal data (Problem 9-19).

9.4.7 The Role of Asymptotics

Why do asymptotics as the sample size goes to infinity matter? Real data have a sample size that is not going anywhere. It just is what it is. Why should anyone draw comfort from the fact that if the sample size were very large, perhaps billions of times larger than the actual sample size, the asymptotics would give a good approximation to the correct sampling distribution of the estimator?

The answer is, of course, that no one does draw comfort from that. What they draw comfort from is that asymptotics actually seem to work, to provide good approximations, at relatively small sample sizes, at least in simple well-behaved situations. Hence the rules of thumb promulgated in introductory statistics books that $n > 30$ is enough to apply "large sample theory" in i. i. d. sampling, except for the binomial distribution and contingency tables,[2] where the rule is the expected value in each cell of the table should be at least five. These rules are known to be simplistic. For skewed distributions n must be larger than 30 for good approximation, much larger if the distribution is highly skewed. Similarly, there are cases where the contingency table rule holds but the distribution of the chi-square statistic is not well approximated by the chi-square distribution. But the rules of thumb are good enough so that textbook authors do not feel negligent in teaching them.

[2] The chi-square test for contingency tables gets us ahead of ourselves. This is the subject of much of Chapter 10 in Lindgren, which we will get to eventually. But as long as we are discussing rules of thumb, we might as well mention all of them, and this is all there are.

If one is worried about the validity of asymptotics, the standard cure is to look at computer simulations. Sometimes simulations show that asymptotic approximations are bad, but asymptotics look good in simulations often enough that people keep on using asymptotics.

So people don't use asymptotics because of the theorems. They use asymptotics because most of the time they actually work in practice. That they work is not something explained by asymptotic theory, because theory only says they work for sufficiently large n. There is no guarantee for the actual n in an actual application.

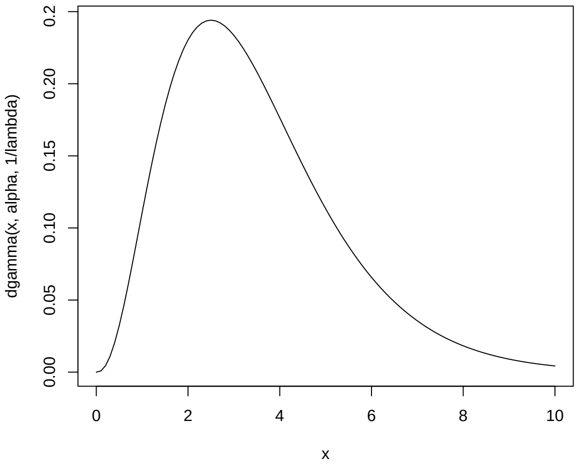
> *Asymptotic theory is only a heuristic. It is a device for producing approximations that may or may not be any good.*

Whether they are any good in an actual application, is something on which the theory is silent.

> *If you are worried the validity of asymptotics, you do simulations. Theory is no help.*

Example 9.4.6 (A Simulation Study).
In Example 9.3.9 and Problem 9-11 we derived the asymptotic distributions of the method of moments estimators of the two parameters of the gamma distribution. What if we are curious whether the asymptotics are good for sample size $n = 25$? Whether the asymptotics are valid also will depend on the shape of the distribution. Skewed distributions require larger sample sizes. Hence it will depend on the shape parameter α, but not on the scale parameter λ. Let's check the case $\alpha = 3.5$. Which a plot shows is moderately skewed.

```
> alpha <- 3.5
> lambda <- 1       # irrelevant, choose something
> curve(dgamma(x, alpha, 1 / lambda), from=0, to=10)
```



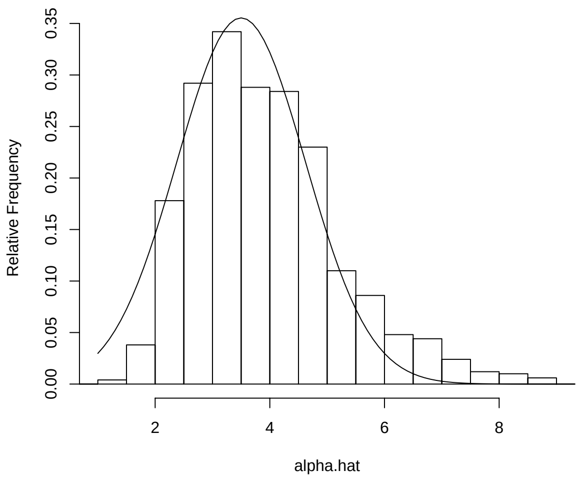
Now what we want to do is simulate lots of random samples of size n and see what the distribution of one of these statistics is. The following code does the job.

```
> alpha <- 3.5
> lambda <- 1       # irrelevant, choose something
> n <- 25           # sample size
> nsim <- 1000      # number of simulations to do
> alpha.hat <- double(nsim)
> for (i in 1:nsim) {
+    x <- rgamma(n, alpha, 1 / lambda)
+    xbar <- mean(x)
+    v <- var(x)
+    alpha.hat[i] <- xbar^2 / v
+ }
```

This only takes a few seconds of computer time. Now we look at a histogram of the data and compare it to the normal density of the asymptotic distribution. Actually, a histogram of all the `alpha.hat` values (not shown) when I did this simulation (of course, you would get different results because the results are random) was clearly nonnormal because it contained two "outliers" very far from the rest of the data. Below is a histogram of all but those two outliers produced by the following R statements

```
> hist(alpha.hat, probability=TRUE, xlim=c(1,9),
+    breaks=seq(0,20,.5))
> curve(dnorm(x, alpha, sqrt(2 * alpha * (1 + alpha) / n)),
+    from=1, to=9, add=TRUE)
```

where in the last line the standard deviations comes from (9.15).



As can be clearly seen from the histogram, the middle of the distribution of $\hat{\alpha}_n$ is clearly skewed and so not very normal (because the normal distribution is symmetric, not skewed). The outliers in the distribution of $\hat{\alpha}_n$ are not a big problem. We are not usually interested in the distribution far out in the tails. The skewness is a big problem. Because of it we should have asymmetric confidence intervals (not $\hat{\alpha}_n$ plus or minus the same thing), and asymptotics doesn't do that. (At least the kind of asymptotics we've studied doesn't do that. So-called "higher order" asymptotics, does correct for skewness, but that subject is beyond the scope of this course.)

Hence the simulation study shows that asymptotics doesn't work, at least for $n = 25$ and $\alpha = 3.5$. For large enough n, it will work, regardless of the value of α. For larger α it will work better for the same n, because the distribution of the X_i will be less skewed. For example, another simulation study (not shown), which I did to check my algebra in deriving the asymptotics, showed that the asymptotics worked fine for $n = 100$ and $\alpha = 2.3$ (there was only a little skewness visible in the histogram, the fit was pretty good).

9.4.8 Robustness

"Robustness" is a word with a lot of meanings in statistics. All of the meanings have something to do with a procedure being insensitive to violation of its assumptions. The differences have to do with what kind of violations are envisaged and what effects are considered important.

Asymptotic Robustness

A confidence interval is *asymptotically robust* or *asymptotically distribution free* for a specified statistical model if it is based on an asymptotically pivotal quantity (the property of being asymptotically pivotal depending, of course, on the model).

Example 9.4.7 (One-Sample t Intervals).
The "exact" confidence interval for the population mean given by Theorem 9.7, which uses the t distribution and assumes normal data, is asymptotically robust (more precisely asymptotically distribution free within the class of all distributions with finite variance) because it is based on the same pivotal quantity as the "large sample" interval given by Theorem 9.8, which needs no assumption about the data except finite variance. Thus we say these one-sample t confidence intervals are robust against departures from the normality assumptions.

Of course asymptotic robustness is inherently a "large sample" result. It may not say much about small sample sizes. Hence the analysis above does not really justify use of the t distribution for small sample sizes when we are worried that the population may be nonnormal. However, one can make the following argument. While it is true that the t confidence intervals are no longer exact if we do not assume exact normality, it is clear that we should make *some* adjustment of the critical value for small sample sizes. Just using the asymptotic interval based on the z critical value is obviously wrong. While the t critical value may not be exactly right, at least it will be a lot closer to the right thing than using the z critical value. If we are really worried we could do some simulations, which would show us that so long as the population distribution is symmetric or close to symmetric the distribution of $\sqrt{n}(\overline{X}_n - \mu)/S_n$ is very well approximated by the $t(n-1)$ distribution even when the population distribution has much heavier tails than normal. When the population is highly skewed, then the distribution of $\sqrt{n}(\overline{X}_n - \mu)/S_n$ is also skewed when n is small and hence cannot be well approximated by a t distribution, which, of course, is symmetric.

Example 9.4.8 (Two-Sample t Intervals).
The "exact" confidence interval for the difference of population means using the pooled estimator of variance is not asymptotically robust within the class of all distributions with finite variances, because the asymptotic distribution of the pivotal quantity (9.37) depends on the ratio of variances σ_X^2/σ_Y^2. Welch's approximate confidence interval is asymptotically robust within this class because it uses the same pivotal quantity as the asymptotic interval.

Both intervals are robust against departures from normality, but the "exact" interval is not robust against departures from its extra assumption $\sigma_X^2 = \sigma_Y^2$.

If we wanted to be pedantic we could say that the "exact" two-sample interval is asymptotically robust within the class of all distributions with finite variances and satisfying the additional assumption $\sigma_X^2 = \sigma_Y^2$, but this would only satisfy a pedant. It only emphasizes that the critical assumption of equality of population variances cannot be violated without destroying any nice properties the procedure is supposed to have. Calling that "robust" is perverse, although it does satisfy the technical condition. Robustness is defined relative to a statistical model. You can always make up a statistical model that makes a procedure robust with respect to that model. The question is whether that model is interesting.

Example 9.4.9 (Variances).
The "exact" confidence interval for the variance using the chi-square distribution is not asymptotically robust with the class of all distributions with finite fourth moments. This is clear because it is not based on the same pivotal quantity as the asymptotic confidence interval given by Theorem 9.11. Hence the "exact" interval is not robust against departures for normality. It critically depends on the property $\mu_4 = 3\sigma^4$ of the normal distribution.

Consider the data from Example 9.4.4 which were computer simulated from a Laplace (double exponential) distribution. The reason for using this distribution as an example is that it has heavier tails than the normal distribution, but not too heavy (the distribution still has moments of all orders, for example). The so-called exact 95% confidence interval for the variance using Theorem 9.12 is

$$1.44 < \sigma^2 < 2.14$$

but here this interval is inappropriate because the normality assumption is incorrect.

In Example 9.4.4 we calculated the correct asymptotic interval using the sample fourth central moment to be

$$1.27 < \sigma^2 < 2.20.$$

Comparing the two, we see that the correct interval is longer than the so-called exact interval based on what is in this case an incorrect assumption (of normality of the population distribution). Hence the so-called exact but in fact incorrect interval will have insufficient coverage. How bad this interval will be, whether it will have 90% coverage or 80% coverage or what instead of its nominal 95% coverage only a simulation study can tell. But we are sure it won't have its nominal coverage, because its assumptions do not hold.

Although we haven't worked out the correct asymptotic interval for the ratio of variances, we can easily believe the "exact" interval for the ratio of variances is also not robust and depends critically on the normality assumption.

These robustness considerations are important. You can find in various textbooks strong recommendations that the "exact" procedures that we have just found to be nonrobust should never be used. Simulations show that they are so critically dependent on their assumptions that even small violations lead to large errors. The robustness analysis shows us why.

Breakdown Point

This section takes up a quite different notion of robustness. The *breakdown point* of a point estimator is the limiting fraction of the data that can be dragged off to infinity without taking the estimator to infinity. More precisely if for each n we have an estimator $T_n(x_1, \ldots, x_n)$ which is a function of the n data values and k_n is the largest integer such that k_n of the x_i can be taken to infinity (with the other $n - k_n$ remaining fixed) and $T_n(x_1, \ldots, x_n)$ remain bounded, then $\lim_{n \to \infty} k_n / n$ is the breakdown point of the sequence of estimators T_n.[3]

The idea behind this technical concept is how resistant the estimator is to junk data. Roughly speaking, the breakdown point is the fraction of junk the estimator can tolerate. Here "junk" is generally considered to consist of gross errors, copying mistakes and the like, where recorded data has nothing whatsoever to do with the actual properties supposedly measured. It can also model rare disturbances of the measurement process or individuals that wind up in a sample though they weren't supposed to be and similar situations.

Example 9.4.10 (The Mean).
The sample mean has breakdown point zero, because

$$\frac{x_1 + \cdots + x_n}{n} \to \infty, \qquad \text{as } x_i \to \infty$$

[3] Some authorities would call this the *asymptotic* breakdown point, since they only use "breakdown point" to describe finite sample properties, that is, they say k_n/n is the breakdown point of T_n. But that needlessly complicates discussions of estimators since k_n is typically a complicated function of n, but the limit is simple.

with x_j, $i \neq j$ fixed. Hence $k_n = 0$ for all n.

Thus the sample mean tolerates zero junk and should only be used with perfect data.

Example 9.4.11 (The Median).
The sample median has breakdown point one-half. If n is odd, say $n = 2m + 1$, then we can drag off to infinity m data points and the sample median will remain bounded (in fact it will be one of the other $m + 1$ data points left fixed). Thus $k_n = \lfloor n/2 \rfloor$ when n is odd. When n is even, say $n = 2m$, then we can drag off to infinity $m - 1$ data points and the sample median will remain bounded, since we are leaving fixed $m + 1$ points and two of them will be the two points we average to calculate the sample median. Thus $k_n = n/2 - 1$ when n is even. In either case k_n is nearly $n/2$ and clearly $k_n/n \to 1/2$ as $n \to \infty$.

This example shows why the finite-sample notion of breakdown points is not so interesting.

Thus we see that the while the mean tolerates only perfect data, the median happily accepts any old junk and still gets decent answers. This is not to say that junk doesn't affect the median at all, only that any amount of junk up to 50% doesn't make the median completely useless. That wouldn't seem like a very strong recommendation until we remember that the mean is completely useless when there is any junk at all no matter how little.

9.5 Tests of Significance

In one sense we are now done with this chapter. This section is just a rehash of what has gone before, looking at the same stuff from a different angle. In another sense we are only half done. Tests of significance (also called hypothesis tests) are as important as confidence intervals, if not more important. So we have to redo everything, this time learning how it all relates to tests of significance. Fortunately, the redo won't take as much time and effort as the first time through, because all of the sampling theory is the same.

The simple story on tests of significance is that they are essentially the same thing as confidence intervals looked from a slightly different angle.

Example 9.5.1 (Difference of Population Proportions).
Suppose two public opinion polls are taken, one four weeks before an election and the other two weeks before. Both polls have sample size 1000. The results in percents were

	1st poll	2nd poll
Jones	36.1	39.9
Smith	30.1	33.0
Miller	22.9	17.4
Undecided	10.9	9.7

The typical reporter looks at something like this and says something like "Jones and Smith both gained ground since the last poll two weeks ago, Jones picking up 4 percent and Smith 3 percent, while Miller lost ground, losing 5 percentage points." This is followed by "news analysis" which reports that Jones is picking up support among college educated voters or whatever. Somewhere down toward the end of the article there may be some mention of sampling variability, a statement like "the polls have a margin of error of 3 percentage points," but it's not clear what anyone is supposed to make of this. Certainly the reporter ignored it in his analysis.

A skeptic might ask the question: has *anything* really changed in two weeks? We know the poll results are random. They are not the true population proportions but only estimates of them (the sample is not the population). Maybe the population proportions haven't changed at all and the apparent change is just chance variation. We don't yet know how to analyze the question of whether anything at all has changed in two weeks—we will get to this in Section 10.5 in Lindgren—but we do know how to analyze whether anything has changed in regard to one candidate, say Jones. The number of people in the samples who expressed a preference for Jones, 361 two weeks ago and 399 now, are binomial random variables with success probabilities p and q (the population proportions). These are estimated by the sample proportions $\hat{p} = 0.361$ and $\hat{q} = 0.399$. An asymptotic confidence interval for $q - p$ is (9.66). Plugging in the numbers gives the 95% confidence interval

$$0.399 - 0.361 \pm 1.960 \sqrt{\frac{0.361 \times 0.639}{1000} + \frac{0.399 \times 0.601}{1000}}$$

or 0.038 ± 0.0425, which is $(-0.0045, 0.0805)$. Multiplying by 100 gives the interval expressed in percent $(-0.45, 8.05)$.

Since the confidence interval contains zero, it is not clear that there has been any increase in voter preference for Jones. No change in preference corresponds to $q - p = 0$, which is a point in the confidence interval, hence is a parameter value included in the interval estimate. It is true that the confidence interval includes a much wider range of positive values than negative values, so the confidence interval includes big increases but only small decreases, but decreases or no change are not ruled out.

Thus it is a bit premature for reporters to be bleating about an increase in the support for Jones. Maybe there wasn't any and the apparent increase is just chance variation in poll results.

The argument carried out in the example is called a test of significance or a statistical hypothesis test. The hypothesis being tested is that there is no real change in the population proportions, in symbols $p = q$. Alternatively, we could say we are testing the complementary hypothesis $p \neq q$, because if we decide that $p = q$ is true this is equivalent to deciding that $p \neq q$ is false and vice versa. We need general names for these two hypotheses, and the rather colorless names that are generally used in the statistical literature are the *null hypothesis* and the *alternative hypothesis*. Lindgren denotes them H_0 and H_A, respectively. These are always two complementary hypotheses, each the negation of the other, so we could do with just one, but we usually wind up mentioning both in discussions, hence the names are handy.

Summarizing what was just said, there are two possible decisions a test of significance can make. It can decide in favor of the null or the alternative. Since exactly one of the two hypotheses is true, deciding that one is true is tantamount to deciding that the other is false. Hence one possible decision is that the null hypothesis is true and the alternative hypothesis is false, which is described as *accepting the null hypothesis* or *rejecting the alternative hypothesis*. The other possible decision, that the alternative hypothesis is true and the null hypothesis is false, is described as described as *accepting the alternative hypothesis* or *rejecting the null hypothesis*.

In the opinion poll example, the null hypothesis is $p = q$ and the alternative hypothesis is $p \neq q$. We accept the null hypothesis if the confidence interval covers the parameter value $q - p = 0$ hypothesized by the null hypothesis. Otherwise we reject the null and accept the alternative. Since the confidence interval $(-0.45, 8.05)$ covers the hypothesized value, we accept the null. We conclude that the hypothesis may well be true, there being no strong evidence against it. In less technical language we conclude that the apparent change in the poll results may be just chance variation. Hence there is no real evidence that Jones is gaining, and hence there is no point in any news analysis of the reasons the gain has occurred.

Of course, the result of the test depends on which confidence interval we use, in particular, on the confidence level. A 90% confidence interval for $q - p$ expressed in percent is $(0.23, 7.37)$. Since this interval does not contain zero, we now reject the null hypothesis and accept the alternative. Thus we come to the opposite conclusion, Jones really is gaining support.

Thus it isn't enough to simply state whether the null hypothesis is accepted or rejected. We must also give the confidence level. For reasons of tradition we actually give something a little bit different. If the test involves a $100(1 - \alpha)\%$ confidence interval, we say we did a test with *significance level* α. People who think it is cool to talk in jargon rather than plain words often call the significance level the "α level," making α a frozen letter in this context and thus violating the principle of "mathematics is invariant under changes of notation." "Significance level" is a much better name.

The significance level is the probability that the confidence interval fails to cover. When the null hypothesis is true and the confidence interval fails to cover, we reject the null hypothesis erroneously. Thus another way of describing the significance level that does not mention confidence intervals is that it is *the probability of erroneously rejecting the null hypothesis*.

You may now be wondering what tests of significance are worth if one can always make a test come out either way by simply choosing to use a higher or lower significance level. The answer is they are worthless if you only pay attention to the decisions ("accept" or "reject") and ignore the significance levels, but when the decision and the significance level are considered together a test does provide useful information. A test using the 0.05 level of significance will erroneously reject the null hypothesis 5% of the time. A test using the 0.10

level of significance will erroneously reject the null hypothesis 10% of the time. That is a weaker evidentiary standard. The 0.10 level test rejects the null, and it may be right in doing so, but it may also be wrong, and the probability of its being wrong is twice that of the 0.05 level test.

That is the basic story on tests of significance. We now begin a systematic development of the theory.

9.5.1 Interest and Nuisance Parameters Revisited

Recall that in Section 9.2.4 we divided parameters into *parameters of interest* and *nuisance parameters*. The parameter or parameters of interest are the ones we want to know something about, the parameter the confidence interval is for, or the parameters involved in the null and alternative hypotheses of a test of significance.

In Example 9.5.1 there are two parameters p and q. Neither is the parameter of interest. The parameter of interest is $q - p$. Thus we see that sometimes we have to reparameterize the model in order to make the parameter of interest one of the parameters of the model. For example, we could choose new parameters α and δ defined by

$$\begin{aligned} \alpha &= p + q \\ \delta &= q - p \end{aligned}$$

Then δ is the parameter of interest, and α is a nuisance parameter.

In general, we write $\theta = (\varphi, \psi)$, where φ is the parameter of interest and ψ is the nuisance parameter. Either or both can be vectors, so

$$\theta = (\theta_1, \ldots, \theta_{k+m}) = (\varphi_1, \ldots, \varphi_k, \psi_1, \ldots, \psi_m)$$

if there are k parameters of interest and m nuisance parameters.

In dealing with confidence intervals there is always exactly one parameter of interest. The confidence interval is an interval estimate of that parameter. There may be many nuisance parameters. When we are estimating a difference of means from independent samples (Section 9.4.5) the parameter of interest is $\mu_X - \mu_Y$. Everything else is a nuisance parameter. As in Example 9.5.1, the parameter of interest is not one of the original parameters. The reparameterization

$$\begin{aligned} \alpha &= \mu_X + \mu_Y \\ \delta &= \mu_X - \mu_Y \end{aligned}$$

makes δ the parameter of interest and α, σ_X^2 and σ_Y^2 the nuisance parameters.

9.5.2 Statistical Hypotheses

In general, a statistical hypothesis can be any statement at all about the parameters of interest. In actual practice, almost all tests involve two kinds of hypotheses.

- The null hypothesis specifies a value of the parameter of interest. Thus it can be written $\varphi = \varphi_0$, where φ_0 is a fixed known value.
- There is a single parameter of interest φ and the null hypothesis is of the form $\varphi \le \varphi_0$ or $\varphi \ge \varphi_0$, where φ_0 is a fixed known value.

When there is a single parameter of interest these have widely used names. A test of

$$\begin{aligned} H_0 &: \varphi \le \varphi_0 \\ H_A &: \varphi > \varphi_0 \end{aligned} \tag{9.46a}$$

is called a *one-tailed test* and H_A is called a *one-sided alternative* (and the same names are used if both inequalities are reversed). A test of

$$\begin{aligned} H_0 &: \varphi = \varphi_0 \\ H_A &: \varphi \ne \varphi_0 \end{aligned} \tag{9.46b}$$

is called a *two-tailed test* and H_A is called a *two-sided alternative.*

When there are several parameters of interest only (9.46b) makes sense so there is usually no need of distinguishing terminology, but in order to discuss the cases of one or several parameters of interest together we will call null hypotheses form in (9.46b) *equality-constrained null hypotheses.*

We also will need notation for the sets of parameter values corresponding to the hypotheses

$$\begin{aligned} \Theta_0 &= \{\, (\varphi, \psi) \in \Theta : \varphi \le \varphi_0 \,\} \\ \Theta_A &= \{\, (\varphi, \psi) \in \Theta : \varphi > \varphi_0 \,\} \end{aligned} \tag{9.47a}$$

in the case of a one-sided alternative and

$$\begin{aligned} \Theta_0 &= \{\, (\varphi, \psi) \in \Theta : \varphi = \varphi_0 \,\} \\ \Theta_A &= \{\, (\varphi, \psi) \in \Theta : \varphi \ne \varphi_0 \,\} \end{aligned} \tag{9.47b}$$

in the case of an equality-constrained null.

As we said above, one can in principle test any hypothesis, but hypotheses other than the two types just described lead to complexities far beyond the scope of this course (in fact beyond the scope of PhD level theoretical statistics courses). So these two kinds of tests are all we will cover. For now we will concentrate on tests of equality-constrained null hypotheses and leave one-tailed tests for a later section (they require only minor changes in the theory).

9.5.3 Tests of Equality-Constrained Null Hypotheses

Most tests of significance (all tests we will consider) are determined by a *test statistic* $T(\mathbf{X})$. The null hypothesis is rejected for large values of the test statistic and accepted for small values. More precisely, there is a number c called the *critical value* for the test such that the decision rule for the test is the following

$$\begin{array}{ll} T(\mathbf{X}) \ge c & \text{reject } H_0 \\ T(\mathbf{X}) < c & \text{accept } H_0 \end{array}$$

Exact Tests

The *significance level* of the test is the probability of rejecting H_0 when H_0 is in fact true, when this probability does not depend on the parameter so long as the parameter remains in Θ_0, that is,

$$\alpha = P_\theta(\text{reject } H_0) = P_\theta\bigl(T(\mathbf{X}) \geq c\bigr), \qquad \theta \in \Theta_0. \tag{9.48}$$

Note that, since the null hypothesis fixes the value of the parameter of interest (for tests we are considering in this section), this means

$$P_\theta(\text{reject } H_0) = P_{(\varphi_0, \psi)}(\text{reject } H_0)$$

does not depend on the value of the nuisance parameter ψ.

How can we arrange for this probability to not depend on the nuisance parameter? We use a pivotal quantity $g(\mathbf{X}, \varphi)$ that only contains the parameter of interest. By definition, its distribution does not depend on the parameter. More precisely, the c. d. f. of the pivotal quantity

$$F(x) = P_{(\varphi,\psi)}\bigl(g(\mathbf{X}, \varphi) \leq x\bigr), \qquad (\varphi, \psi) \in \Theta. \tag{9.49}$$

does not depend on the parameter when the parameter value that is the argument of the pivotal quantity and the true parameter value (i. e., both φ's on the right hand side of the equation) are the same. Of course, $g(\mathbf{X}, \varphi)$ is not a statistic, since it depends on a parameter, but we are only interested right now in parameter values in Θ_0, which means $\varphi = \varphi_0$. Plugging in the hypothesised value φ_0 for φ does give us a statistic $g(\mathbf{X}, \varphi_0)$, and from (9.49) we see that its distribution does not depend on θ for $\theta \in \Theta_0$, which is what is required. Since any function of a statistic is a statistic, any function of $g(\mathbf{X}, \varphi_0)$ can be used as a test statistic.

Example 9.5.2 (t Tests).
Suppose X_1, $\ldots$, X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$ and we wish to test

$$\begin{aligned} H_0 &: \mu = 0 \\ H_A &: \mu \neq 0 \end{aligned}$$

(μ is the parameter of interest and σ^2 is a nuisance parameter.) We know that

$$\frac{\overline{X}_n - \mu}{S_n / \sqrt{n}} \sim t(n-1)$$

is a pivotal quantity. Plugging in μ_0 for μ gives a statistic

$$T = \frac{\overline{X}_n - \mu_0}{S_n / \sqrt{n}}$$

the distribution of which is $t(n-1)$ when the null hypothesis it true (and μ_0 is the true value of the parameter of interest). Which function of T do we use as our test statistic? The analogy with confidence intervals and the symmetry of the t distribution suggest the absolute value $|T|$. Thus we determine the critical value by solving

$$P\bigl(|T| > c\bigr) = \alpha$$

or what is equivalent

$$P\bigl(T > c\bigr) = \alpha/2. \tag{9.50}$$

We denote the solution of (9.50) $c = t_{\alpha/2}$. It is, as we defined it throughout Section 9.4, the $1 - \alpha/2$ quantile of the $t(n-1)$ distribution.

Why would we ever want to test whether μ is zero? Remember paired comparisons, where the test for no difference of population means reduces to a one sample test of $\mu = 0$ after the data are reduced to the differences of the pair values.

Asymptotic Tests

Asymptotic tests work much the same way as exact tests. We just substitute an asymptotically pivotal quantity for an exactly pivotal quantity and substitute asymptotic approximations for exact probabilities.

Sometimes there is a choice of pivotal quantity, as the following examples show.

Example 9.5.3 (Binomial, One Sample).
Suppose X is $\text{Bin}(n, p)$ and we wish to test

$$\begin{aligned} H_0 &: p = p_0 \\ H_A &: p \neq p_0 \end{aligned}$$

where p_0 is a particular number between zero and one. There are two asymptotically pivotal quantities we used to make confidence intervals in this situation

$$g_{1,n}(X, p) = \frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}}$$

and

$$g_{2,n}(X, p) = \frac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)/n}}$$

where, as usual, $\hat{p}_n = X/n$. Both are asymptotically standard normal. Both can be used to make confidence intervals, although the latter is much easier to use for confidence intervals. When it comes to tests, both are easy to use. Plugging in the hypothesized value of the parameter gives test statistics

$$Z_1 = \frac{\hat{p}_n - p_0}{\sqrt{p_0(1-p_0)/n}} \tag{9.51a}$$

and

$$Z_2 = \frac{\hat{p}_n - p_0}{\sqrt{\hat{p}_n(1 - \hat{p}_n)/n}} \tag{9.51b}$$

The two-tailed test with significance level α rejects H_0 when $|Z_i| \ge z_{\alpha/2}$ where, as usual, $z_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

For large n the two test statistics will be very close to each other and the two tests agree with very high probability. Asymptotics gives no reason to choose one over the other. The vast majority of statistics textbooks, however, recommend the test using Z_1. There is a sense in which Z_1 is closer to the standard normal than Z_2. The variance of Z_1 is exactly one, whereas Z_2 does not even have a variance because the denominator is zero when $X = 0$ or $X = n$. Still, neither test is exact, and when n is large enough so that the asymptotics are working well both tests give similar answers. So there is no real reason other than convention for using one or the other. But convention is a good enough reason. Why get into fights with people whose introductory statistics course insisted that the test using Z_1 was the only right way to do it?

Example 9.5.4 (Binomial, Two Sample).
Suppose $X \sim \text{Bin}(m, p)$ and $Y \sim \text{Bin}(n, q)$ and we wish to test

$$\begin{aligned} H_0 &: p = q \\ H_A &: p \ne q \end{aligned}$$

The asymptotically pivotal quantity we used to make confidence intervals in this situation was

$$g_{1,n}(X, Y, p - q) = \frac{(\hat{p}_m - \hat{q}_n) - (p - q)}{\sqrt{\frac{\hat{p}_m(1-\hat{p}_m)}{m} + \frac{\hat{q}_n(1-\hat{q}_n)}{n}}}$$

where, as usual, $\hat{p}_m = X/m$ and $\hat{q}_n = Y/n$. Plugging in the hypothesized value under the null hypothesis $(p - q = 0)$ gives the test statistic

$$Z_1 = \frac{\hat{p}_m - \hat{q}_n}{\sqrt{\frac{\hat{p}_m(1-\hat{p}_m)}{m} + \frac{\hat{q}_n(1-\hat{q}_n)}{n}}} \tag{9.52}$$

A different quantity that is pivotal under the null hypothesis uses a "pooled" estimator of p similar to the pooled estimator of variance used in the two-sample t confidence interval based on the assumption of equality of variances. Here there is nothing controversial or nonrobust about the assumption $p = q$. The theory of tests of significance requires a probability calculation assuming H_0, so that's what we do. Under the null hypothesis $X + Y \sim \text{Bin}(m + n, p)$, hence

$$\hat{r}_{m,n} = \frac{X + Y}{m + n} = \frac{m\hat{p}_m + n\hat{q}_n}{m + n}$$

is the sensible estimator of p (and q). Also, still assuming $p = q$, the variance of the numerator of Z_q is

$$\operatorname{var}(\hat{p}_m - \hat{q}_n) = p(1-p)\left(\frac{1}{m} + \frac{1}{n}\right)$$

Hence under H_0

$$Z_2 = \frac{\hat{p}_m - \hat{q}_n}{\sqrt{\hat{r}_{m,n}(1-\hat{r}_{m,n})\left(\frac{1}{m} + \frac{1}{n}\right)}} \tag{9.53}$$

is also asymptotically standard normal. Thus we again have two test statistics. Either is asymptotically correct. Both will agree with very high probability when m and n are large. Neither is exact. Convention (meaning the vast majority of introductory statistics courses) recommends the test based on Z_2.

9.5.4 P-values

Tests of significance are often done when they do not decide a course of action. This is almost always the case when the issue is a scientific inference. Data are collected, a test of significance is done, a paper is written, and readers are left to judge what it all means. Even the readers make no decisions in the statistical sense, accepting or rejecting some hypothesis solely on the basis of the data reported in the paper.

In such situations it is absurd to simply report that the test of significance rejected H_0 at a particular level of significance chosen by the authors (or accepted H_0 if that is what happened). What if a reader wants a different level of significance?

When the test is based on a test statistic, there is a much more sensible procedure. Suppose our test statistic is $T(\mathbf{X})$ and the value for the actual data being analyzed is $T(\mathbf{x})$. This is the usual "big $\mathbf{X}$" for random vectors and "little $\mathbf{x}$" for possible values (in this case the actual observed value). The significance level of the test corresponding to the critical value c is

$$\alpha(c) = P_\theta\bigl(T(\mathbf{X}) \ge c\bigr),$$

which we assume is the same for all $\theta \in \Theta_0$. Note that as c increases the event $\{\, \mathbf{X} : T(\mathbf{X}) \ge c \,\}$ decreases, and hence $\alpha(c)$ is a decreasing function of c by monotonicity of probability. Since the test rejects H_0 if $T(\mathbf{x}) \ge c$ and otherwise accepts H_0, the null hypothesis is rejected for all critical values c such that $c \le T(\mathbf{x})$ and hence for all significance levels α greater than or equal to

$$\alpha\bigl(T(\mathbf{x})\bigr) = P_\theta\bigl\{T(\mathbf{X}) \ge T(\mathbf{x})\bigr\}.$$

Definition 9.5.1 (P-value).
*The P-*value *of a test based on a test statistic $T(\mathbf{X})$ is*

$$P_\theta\bigl\{T(\mathbf{X}) \ge T(\mathbf{x})\bigr\}$$

provided this does not depend on θ for $\theta \in \Theta_0$. (This definition will later be generalized in Definition 9.5.3).

Summarizing the argument preceding the definition, the relationship between P-values (P), significance levels (α), and decisions is

$$\begin{array}{ll} P \le \alpha & \text{reject } H_0 \\ P > \alpha & \text{accept } H_0 \end{array} \tag{9.54}$$

The P-value (according to textbooks also called "observed level of significance," but I have never seen that outside of textbooks) is the answer to the quandary about different readers wanting different levels of significance. If the scientists report the P-value, then every reader can choose his or her own individual α and apply the rule (9.54) to determine the appropriate decision.

Example 9.5.5 (Example 9.5.1 Continued).
The hypothesis test in Example 9.5.1 is a two-tailed test based on the test statistic $|Z_1|$ where Z_1 is given by (9.52), which has the observed value

$$\frac{0.399 - 0.361}{\sqrt{\frac{0.361 \times 0.639}{1000} + \frac{0.399 \times 0.601}{1000}}} = 1.752$$

Under the null hypothesis, Z_1 is asymptotically standard normal. Hence the P-value is

$$P(|Z_1| > z) \approx 1 - 2\Phi(z) = 2\Phi(-z)$$

where $z = 1.752$ is the observed value of the test statistic, and Φ is, as usual, the standard normal c. d. f. Thus the P-value is $2 \times 0.0399 = 0.0798$.

Interpretation of P-values

P-values have two different interpretations, one trivial and the other controversial. Both interpretations support the following slogan.

> *The lower the P-value, the stronger the evidence against H_0.*

What is controversial is what "strength of evidence" is supposed to mean.

The trivial sense in which the slogan is true is that, if we consider a group of readers with a range of individual significance levels, a lower P-value will convince more readers. So there is no question that a lower P-value is more evidence against H_0. What is controversial is the question: "How much more?"

The controversial interpretation of P-values is the following. When we reject H_0 there are two possibilities

- H_0 actually is false.

- H_0 actually is true, in which case the P-value measures the probability of an actual event $T(\mathbf{X}) \ge T(\mathbf{x})$, which can be stated in words as the probability of seeing data at least as extreme as the data actually observed, where "extreme" is defined by largeness of $T(\mathbf{X})$.

Thus smallness of the P-value is a measure of the improbability of the second possibility.

This "controversial" argument is perfectly correct and strongly disliked by many statisticians, who claim that most people cannot or will not understand it and will confuse the P-value with a quite different notion: the probability of H_0. So let us be very clear about this. Since in frequentist statistics the parameter is not considered a random quantity, H_0 is not an event and $P(H_0)$ is a meaningless expression. Thus not only is the P-value *not* the "probability of the null hypothesis" neither is anything else.

Bayesians do consider parameters to be random quantities and hence do consider $P(H_0)$ to be meaningful. Dogmatic Bayesians consider all non-Bayesian statistics to be bogus, hence they particularly dislike P-values (though they like tests as decision procedures). They write papers[4] with titles like "the irreconcilability of P-values and evidence" by which they mean irreconcilability with *Bayesian notions of evidence.* This paper shows that a P-value always overstates the evidence against H_0 as compared to standard Bayesian notions of evidence.

Bayesian versus frequentist arguments aside, there is nothing controversial about P-values, as the "trivial" argument above makes clear. Lindgren in Section 9.3 gives a long discussion of tests as evidence raising a number of troubling issues, but only the very last paragraph, which mentions the Bayesian view involves P-values *per se.* The rest are troubling issues involving all tests of significance, whether or not P-values are used. We will return to our own discussion of interpretation of tests of significance after we discuss one-tailed tests, themselves one of the most controversial issues.

9.5.5 One-Tailed Tests

Theory

One-tailed tests require changes of the definitions of significance level and P-value. For one-tailed tests, we can no longer arrange for (9.48) to hold. Since the null hypothesis no longer fixes the value of the parameter of interest (only asserts an inequality), the probability we previously used to define the significance level will now depend on the parameter. This leads to the following definition

Definition 9.5.2 (Significance Level).
The significance level *of a test of significance based on a test statistic* $T(\mathbf{X})$ *is*

$$\alpha = \sup_{\theta \in \Theta_0} P_\theta(\textit{reject } H_0) = \sup_{\theta \in \Theta_0} P_\theta\bigl(T(\mathbf{X}) \ge c\bigr). \tag{9.55}$$

In words, the significance level is *the maximum probability of erroneously rejecting the null hypothesis.* Note that (9.55) reduces to our former definition (9.48) in the special case where $P_\theta(\text{reject } H_0)$ is actually the same for all $\theta \in \Theta_0$.

[4] Berger and Sellke, "Testing a point null hypothesis: The irreconcilability of P values and evidence" (with discussion), *Journal of the American Statistical Association*, 82:112-122, 1987

When we carry this new definition through the argument about P-values we obtain the promised generalization of Definition 9.5.1

Definition 9.5.3 (P-value).
The P-value of a test based on a test statistic $T(\mathbf{X})$ is

$$\sup_{\theta \in \Theta_0} P_\theta\bigl\{T(\mathbf{X}) \ge T(\mathbf{x})\bigr\}$$

Suppose the two-tailed test of the hypotheses (9.46b) is based on a pivotal quantity

$$g(\mathbf{X}, \varphi) = \frac{a(\mathbf{X}) - \varphi}{b(\mathbf{X})} \tag{9.56}$$

having a symmetric distribution that does not depend on the true parameter value. The primary examples are the one-sample z and t tests based on the pivotal quantities (9.20a) and (9.20b) and the two-sample test with the pooled estimate of variance based on the pivotal quantity (9.37). Other examples are the sign test and the two Wilcoxon tests that we will meet when we get to nonparametrics (Chapter 13 in Lindgren).

If we follow what we did with two-tailed tests and plug in φ_0 for φ in (9.56), we obtain

$$T(\mathbf{X}) = \frac{a(\mathbf{X}) - \varphi_0}{b(\mathbf{X})} \tag{9.57}$$

The idea of a one-tailed test is to use $T(\mathbf{X})$ itself as the test statistic, rather than $|T(\mathbf{X})|$, which is what we used for two-tailed z and t tests.

Lemma 9.13. *For the test with test statistic* (9.57) *based on the pivotal quantity* (9.56)*, the significance level corresponding to the critical value c is*

$$\alpha = P_{\varphi_0}\bigl\{T(\mathbf{X}) \ge c\bigr\},$$

and the P-value is

$$P_{\varphi_0}\bigl\{T(\mathbf{X}) \ge T(\mathbf{x})\bigr\}.$$

Proof. What we must show is that

$$P_{\varphi_0}\bigl\{T(\mathbf{X}) \ge c\bigr\} = \sup_{\theta \in \Theta_0} P_\theta\bigl\{T(\mathbf{X}) \ge c\bigr\}. \tag{9.58}$$

The assumption that (9.56) is a pivotal quantity means

$$\begin{aligned} P_\theta\bigl\{g(\mathbf{X}, \varphi) \ge c\bigr\} &= P_\theta\left\{\frac{a(\mathbf{X}) - \varphi}{b(\mathbf{X})} \ge c\right\} \\ &= P_{(\varphi,\psi)}\bigl\{a(\mathbf{X}) - cb(\mathbf{X}) \ge \varphi\bigr\} \end{aligned} \tag{9.59a}$$

does not depend on φ and ψ. Now

$$P_\theta\bigl\{T(\mathbf{X}) \ge c\bigr\} = P_{(\varphi,\psi)}\bigl\{a(\mathbf{X}) - cb(\mathbf{X}) \ge \varphi_0\bigr\} \tag{9.59b}$$

does depend on φ and ψ, but for parameter values in the null hypothesis $\varphi \le \varphi_0$ monotonicity of probability implies that (9.59b) is less than or equal to (9.59a), that is,

$$\begin{aligned} P_{\varphi_0}\bigl\{T(\mathbf{X}) \ge c\bigr\} &= P_{(\varphi_0,\psi)}\bigl\{a(\mathbf{X}) - cb(\mathbf{X}) \ge \varphi_0\bigr\} \\ &= P_{(\varphi,\psi)}\bigl\{a(\mathbf{X}) - cb(\mathbf{X}) \ge \varphi\bigr\} \\ &\ge P_{(\varphi,\psi)}\bigl\{a(\mathbf{X}) - cb(\mathbf{X}) \ge \varphi_0\bigr\} \\ &= P_{\varphi}\bigl\{T(\mathbf{X}) \ge c\bigr\} \end{aligned}$$

whenever $\varphi \le \varphi_0$. And this implies (9.58). □

There is an entirely analogous lemma for asymptotic tests, which we won't bother to prove, since the proof is so similar.

Lemma 9.14. *Suppose*

$$g_n(\mathbf{X}, \varphi) = \frac{a_n(\mathbf{X}) - \varphi}{b_n(\mathbf{X})/\sqrt{n}}$$

is an asymptotically pivotal quantity converging to a standard normal distribution as $n \to \infty$, and let

$$T_n(\mathbf{X}) = g_n(\mathbf{X}, \varphi_0).$$

Then a one-tailed test of (9.46a) *rejects H_0 when $T_n \ge z_\alpha$, where z_α is the $1-\alpha$ quantile of the standard normal distribution. The P-value is*

$$P\bigl\{Z > T_n(\mathbf{x})\bigr\},$$

where Z is standard normal and $T_n(\mathbf{x})$ is the observed value of the test statistic.

Practice

The theory in the preceding section is complicated but in practice one-tailed tests are simple as falling off a log. We will just give an example.

Example 9.5.6 (t Tests).
The following data appeared in "Student's" original paper on the t distribution

	X	Y	Z
	0.7	1.9	1.2
	−1.6	0.8	2.4
	−0.2	1.1	1.3
	−1.2	0.1	1.3
	−0.1	−0.1	0.0
	3.4	4.4	1.0
	3.7	5.5	1.8
	0.8	1.6	0.8
	0.0	4.6	4.6
	2.0	3.4	1.4
mean	0.75	2.33	1.58
s. d.			1.23

In each row, X and Y are the additional hours of sleep gained by one patient while using two different soporific drugs (there are 10 patients in the study). The third column is $Z = Y - X$, our usual trick of reducing data for paired comparison to differences. We want to test whether there is a significant difference between the two drugs. The null hypothesis is $\mu_Z = 0$ for a two-tailed test or $\mu_Z \le 0$ for a one-tailed test. The test statistic is

$$t = \frac{1.58}{1.23/\sqrt{10}} = 4.062$$

For a one-tailed test, the P-value is

$$P(T > 4.062) = 0.0014.$$

For a two-tailed test, the P-value is

$$P(|T| > 4.062) = 0.0028.$$

(All we could get from Table IIIa in Lindgren is that the one-tailed P-value is between 0.001 and 0.002 and hence the two-tailed P-value between 0.002 and 0.004. I used a computer to get exact P-values.)

Note that by the symmetry of the t distribution

$$P(T > t) = 2P(|T| > t)$$

so long as $t > 0$. Hence

A two-tailed P-value is twice the one-tailed P-value

so long as the one-tailed P-value is less than one-half. This has nothing to do with the t distribution in particular. It holds whenever the sampling distribution of the test statistic under H_0 is symmetric. By symmetry, two tails have twice the probability of one.

9.5.6 The Duality of Tests and Confidence Intervals

With all this theory we have somewhat lost track of the simple notion we started out with, that tests are just confidence intervals viewed from another angle. This section ties up the loose ends of that notion.

For this section only, forget P-values. Think of a test as a decision procedure that either accepts or rejects H_0 and has a specified significance level. That is the notion of tests that has a simple relationship to confidence intervals.

The word "duality" in the section heading is a fancy mathematical word for the relation between two concepts that are basically two sides of the same coin. Either can be used to define the other. Tests of equality-constrained null hypotheses have exactly that relation to confidence intervals.

> For any $100(1-\alpha)\%$ confidence interval for a parameter θ, the test that rejects $H_0 : \theta = \theta_0$ if and only if the confidence interval does not contain θ_0 has significance level α.
>
> Conversely, for any test with significance level α, the set of parameter values θ_0 such that $H_0 : \theta = \theta_0$ is accepted is a $100(1-\alpha)\%$ confidence interval for θ.

Example 9.5.7 (t Tests Yet Again).

$$\overline{X}_n - t_{\alpha/2} \frac{S_n}{\sqrt{n}} < \mu < \overline{X}_n + t_{\alpha/2} \frac{S_n}{\sqrt{n}}$$

is a $100(1-\alpha)\%$ confidence interval for μ assuming normal data. This confidence interval contains μ_0 if and only if

$$\left| \frac{\overline{X}_n - \mu_0}{S_n / \sqrt{n}} \right| < t_{\alpha/2}$$

which is the criterion for accepting H_0 in the usual two-tailed t test of $H_0 : \mu = \mu_0$. And it works both ways. If we start with the test and work backwards we get the confidence interval.

The duality is not exact if we use different asymptotic approximations for the test and the confidence interval. For example, the standard way to do a confidence interval for the binomial distribution involves the pivotal quantity Z_2 given by (9.51b) but the standard way to do a test involves the pivotal quantity Z_1 given by (9.51a). Z_1 and Z_2 are very close when n is large, but they are not identical. Thus the test and confidence interval will not have exact duality (they would if both were based on the same asymptotically pivotal quantity). However, we can say they have "approximate duality."

One-tailed tests do not seem at first sight to have such a simple duality relationship, but they do. In order to see it we have to change our view of one-tailed tests. All of the one-tailed tests we have considered can also be considered as equality-constrained tests. A test having a test statistic of the form (9.57) can be considered either a test of the hypotheses

$$\begin{aligned} H_0 &: \theta \le \theta_0 \\ H_A &: \theta > \theta_0 \end{aligned}$$

(which is the way we usually consider it) or a test of the hypotheses

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_A &: \theta > \theta_0 \end{aligned}$$

The latter way of thinking about the test changes the statistical model. Since H_0 and H_A partition the parameter space, the parameter space for the first test is $\Theta = \mathbb{R}$ and the parameter space for the second test is $\Theta = \{\, \theta \in \mathbb{R} : \theta \ge 0 \,\}$. But this is the only difference between the two procedures. They have the same test statistic, the same P-value for the same data, and the same decision for the same significance level and same data.

Now we can apply duality of tests and confidence intervals. Consider the t test yet again. The one-tailed t test accepts $H_0 : \mu = \mu_0$ at significance level α when

$$\frac{\overline{X}_n - \mu_0}{S_n / \sqrt{n}} < t_\alpha$$

The set of μ_0 values for which H_0 is accepted, the μ_0 such that

$$\overline{X}_n - t_\alpha \frac{S_n}{\sqrt{n}} < \mu_0 < +\infty,$$

is thus a $100(1 - \alpha)\%$ confidence interval for the true parameter value μ.

Thus we get "one-tailed" confidence intervals dual to one-tailed tests. Such intervals are not widely used, but there is nothing wrong with them. They are perfectly valid confidence intervals. Occasionally they are wanted in real applications.

9.5.7 Sample Size Calculations

All of the problems in this section and the preceding section (tests and confidence intervals), at least those that involve numbers, emulate the most common kind of data analysis, that which is done *after* the data have been collected. In this section we discuss the other kind, done *before* data have been collected.

We're not talking about some sort of magic. What we're talking about is part of most grant proposals and other preliminary work done before any large expensive scientific experiment is done. Of course you can't really analyze data that haven't been collected yet. But you can do *some* calculations that at least give *some* idea how they will come out.

The main issue of interest is whether the proposed sample size is large enough. We know that statistical precision varies as the square root of the sample size (the so-called square root law). So we can always get as precise a result as we please if only we expend enough time, money, and effort. The trouble is that the square root law means that twice the precision costs four times as much, ten times the precision costs a hundred times as much, and so forth. So generally, you must settle for less precision than you would like.

So suppose you are asking for several hundred thousand dollars to do an experiment with sample size 200. Before the funding agency gives you the money, one issue (among many others) that they will want to carefully consider is whether the precision you will get with $n = 200$ is worth the money. After all, if an experiment with $n = 200$ is unlikely to answer any questions of scientific interest because of lack of precision, they should fund some other projects with more promise.

These calculations *before the data are collected* look very different for tests and confidence intervals, so we will look at them separately. We'll do the simpler of the two first.

Confidence Intervals

A large sample confidence interval for the population mean is

$$\overline{X}_n \pm z_{\alpha/2} \frac{S_n}{\sqrt{n}}$$

which is just (9.24) repeated. Before we collect data we will know neither $\overline{X}_n$ nor S_n. We will know the n we propose to use.

Not knowing $\overline{X}_n$ is not a problem here. It will be close to the true population mean μ if n is large, but we don't know μ. In fact the whole point of the experiment (if this sort of confidence interval is of interest) is to estimate μ. So in a way it's a good thing we don't know μ. It gives us something to do.

The question we want to answer, or at least get some idea about, is how wide our confidence interval will be. If we expect it to be too wide to be of any scientific value, then we need a larger sample size or a completely different sort of experiment. So what we need to do is get some idea of the likely size of the plus-or-minus (also called "half-width" because it is half the width of the confidence interval).

Now the half-width depends on three things

- The confidence level (through α), which we know.
- The sample size n, which we know.
- The sample standard deviation S_n, which we do not know.

So in order to make progress we need a guess about the likely size of S_n. We might have some data from a similar experiment that will give us the likely size. Or we might have to just guess. Depending on who does the guessing and how much they know, we might call the guess anything from "expert opinion" to a "wild guess." But no matter where it comes from, we need some number to plug in for S_n.

Questions like this are often phrased backwards. Rather than what half-width will we likely get for a specified confidence level and sample size, one asks what sample size is necessary to get a specified half-width.

Example 9.5.8.
Suppose we are going to do an experiment to measure the expected life time of a new type of light bulb. The old light bulbs had a mean life of 700 hours with a standard deviation of 500 hours. The new light bulbs are supposed to last a lot longer, about 1000 hours, but let's use the same standard deviation in our sample size calculation. With no data yet on the new light bulbs you can call $s = 500$ a guess (we're guessing the s. d. of the new will be the same as the old) or you can call it an estimate based on preliminary data even though the data isn't about the exact same process.

So what sample size do we need to get a half-width of 100 hours for a 95% confidence interval? That's saying we want

$$100 = 1.96 \frac{500}{\sqrt{n}}$$

Solving for n gives

$$n = \left(\frac{1.96 \times 500}{100} \right)^2 = 96.04$$

Of course, a sample size must be a round number. Typically, one rounds up to be conservative, giving $n = 97$ as the answer.

This looks a lot more precise than it really is. Don't forget that we plugged in a guess for S_n. The actual experiment won't produce a confidence interval with a half-width of exactly 100 hours, because S_n won't come out to be exactly 500. We have, however, done the best we could with what we had to work with. Certainly, $n = 97$ is a lot better than complete cluelessness.

The Power of a Hypothesis Test

We will do a similar sort of calculation for a hypothesis test presently, but before we can even discuss such a calculation we need to learn a new term. This new term is called the *power* of a test. It is closely related to the *significance level* of a test, but not exactly the same thing.

First we explain the concepts in words to make the similarities and differences clear.

- The *significance level* of a test is the probability of rejecting the null hypothesis when it is in fact *true*.
- The *power* of a test is the probability of rejecting the null hypothesis when it is in fact *false*, that is when the alternative hypothesis is true.

Thus both level and power are probabilities of the same event "reject H_0" but probabilities under different assumed parameter values.

In symbols the *significance level* is

$$\alpha = P_\theta(\text{reject } H_0), \qquad \theta \in \Theta_0 \tag{9.60a}$$

This is our simpler definition of significance level given by (9.48), which assumes that the probability in (9.60a) does not actually depend on θ for θ in the null hypothesis. Our more general definition (Definition 9.5.2) is more complicated, but we won't worry about that here.

$$\pi(\theta) = P_\theta(\text{reject } H_0), \qquad \theta \in \Theta_A \tag{9.60b}$$

Note that the left hand side in (9.60b) is a function of θ, which we call the *power function* of the test. That's one important difference between level and power. Ideally, the level does not depend on the parameter, that's what the notation in (9.60a) indicates (as the following comment says, if it *did* depend on the parameter we would have to use a more complicated definition). And this is the case in simple situations.

Why don't we arrange for power to be constant too? Well that would defeat the whole purpose of the test. We want

- Low significance level, the lower α the better.
- High power, the higher $\pi(\theta)$ the better.

Why is that? Both level and power refer to the same event "reject H_0" but under different conditions. The level is the probability of a bad thing, erroneously rejecting the null hypothesis when it is true. The power is the probability of a good thing, correctly rejecting the null hypothesis when it is false. So we want low probability of the bad thing (level) and high probability of the good thing (power).

But as probabilities level and power are special cases of the same thing $P_\theta(\text{reject } H_0)$. So the only way power could be constant is if it were the same as the level, which is no good at all. That's not the way to get low level and high power.

What do power functions look like? Lindgren Section 9.10 gives some examples. Power functions are not all alike. It is not the case "seen one, you've seen them all." But fortunately, it is the case "seen two, you've seen them all." Power functions of upper-tailed tests look like Figure 9-10 in Lindgren and those of lower-tailed tests look like mirror images of that figure. Power functions of two-tailed tests look like Figure 9-11 in Lindgren.

Tests of Significance

We now return to sample size calculations. These too are usually phrased as "backwards" questions. What sample size n do we need to achieve a specified power for a specified level test?

Example 9.5.9.
Suppose in the situation explained in Example 9.5.8 we want to do a one-tailed test of whether the new light bulbs are better than the old in the sense of having longer life. We assume the mean life, 700 hours, of the old light bulbs is a known constant (it isn't really, but it is based on much more data than we intend to collect about the new light bulbs, so this isn't too bad an approximation to the right thing, which would be a two-sample test). We will also assume $S_n = 500$, as in Example 9.5.8 even this is only a guess (we need that here too as will be seen presently).

So what sample size is needed for $\alpha = 0.05$ and power 0.99 at the alternative hypothesis of interest, which is 1000 hours mean life for the new bulbs? Just to be perfectly clear, the hypotheses being tested are

$$\begin{aligned} H_0 &: \mu = 700 \text{ hours} \\ H_A &: \mu > 700 \text{ hours} \end{aligned}$$

and the alternative at which we want to calculate the power is $\mu = 1000$ hours.

The event "reject H_0" is in other notation $\overline{X}_n > c$, where c is the critical value for the test. So the first thing we must do is determine c. The test is based on the asymptotically standard normal quantity

$$Z = \frac{\overline{X}_n - \mu}{S_n / \sqrt{n}} \tag{9.61}$$

In power calculations we must be very, very careful with this object. The random variable (9.61) is asymptotically standard normal when μ is the true population mean and *this differs when calculating level and power!* Calculation of the level of a test is done *assuming the null hypothesis*. So in the level calculation we use $\mu = 700$. But calculation of the power is done *assuming the alternative hypothesis*. In particular, our power calculation here will be based on the particular alternative of interest $\mu = 1000$.

As usual, to get a one-tailed "z test" with level $\alpha = 0.05$ we use the critical value on the z scale 1.645. That is we reject H_0 when

$$Z = \frac{\overline{X}_n - 700}{500 / \sqrt{n}} > 1.645$$

Solving for $\overline{X}_n$, this is the same as

$$\overline{X}_n > 700 + 1.645 \frac{500}{\sqrt{n}} = c \tag{9.62}$$

So that finishes our calculation about level. All subsequent calculation assumes $\mu = 1000$. Notice that the critical value c depends on n. We don't get a number. We get a formula.

Now we need to calculate the power at $\mu = 1000$. What is $P(\overline{X}_n > c)$ when $\mu = 1000$. This is just like any other normal probability calculation. The main difficulty is not to get anything confused with the previous calculation. It only requires care and attention to detail. To calculate this probability using a normal table, we need to standardize the number being looked up, which is c.

$$P(\overline{X}_n > c) \approx 1 - \Phi\left(\frac{c - \mu}{S_n / \sqrt{n}}\right)$$

where, as usual, Φ is the standard normal c. d. f.We want this to be 0.99, hence we want the Φ term itself to be 0.01, and from the bottom row of Table IIIb in Lindgren (or from R or Mathematica) we see that we need the argument of Φ to be -2.33 to achieve that. Thus we get another equation

$$\frac{c - \mu}{S_n / \sqrt{n}} = \frac{c - 1000}{500 / \sqrt{n}} = -2.33 \tag{9.63}$$

(Note again, and this is the last time I'll say it, we are using $\mu = 1000$ here).

Plugging (9.62) in here gives

$$\frac{700 + 1.645 \frac{500}{\sqrt{n}} - 1000}{500 / \sqrt{n}} = -2.33$$

or

$$700 + 1.645 \frac{500}{\sqrt{n}} - 1000 = -2.33 \frac{500}{\sqrt{n}}$$

or

$$3.975 \frac{500}{\sqrt{n}} = 300$$

which comes to $n = 43.89$ or (rounded up) $n = 44$.

I concede that a messy calculation like this leaves me in doubt. Let's check that we actually got the right level and size with this critical value

```
> n <- 44
> crit <- 700 + 1.645 * 500 / sqrt(n)
> 1 - pnorm(crit, 700, 500 / sqrt(n))
[1] 0.04998491
> 1 - pnorm(crit, 1000, 500 / sqrt(n))
[1] 0.990227
```

The definition of `crit` is taken from (9.62). We must call it something other than c, because c is an R function name. The last two lines calculate $P(\overline{X} >$ `crit`$)$ under the null and the alternative. In both cases $\overline{X}_n$ is normal with standard deviation $\sigma/\sqrt{n}$, which is approximately $S_n/\sqrt{n}$. The difference between the two is the mean (oh, excuse me, I said I wasn't going to repeat this again, but here I go again), which is assumed to be $\mu = 700$ under the null and $\mu = 1000$ under the alternative.

If we were going to do a lot of these, we could clean up this calculation a bit and make a theorem out of it. It is clear from the way the calculation simplified at the end that some clean up is possible. But that wouldn't help us much. It would only apply to power calculation for tests involving means. Often one does power calculations for chi-square tests or F tests, which are much more complicated. We won't go into the details. We will let this subject go with these examples, which do illustrate the basic idea.

9.5.8 Multiple Tests and Confidence Intervals

All of Section 9.5 so far deals with the situation in which exactly one test is done on a data set. What if we want to do more than one test? Is the theory still valid?

No! If you take any complicated data set and keep doing different tests until one of them rejects the null hypothesis, this will eventually happen, but it proves absolutely nothing because this will always happen. If you keep going until you manage to think up a test that happens to reject H_0, then you will always eventually get a test to reject H_0.

What about situations between the ideal of just doing one test and doing a potentially infinite sequence of tests? What if you have several tests you want to do and will do no more even if none of them rejects H_0? Is there a valid way to do that?

Yes. In fact, many different ways have been proposed in the statistical literature.[5] Most only work with a specific kind of test. However, there is one procedure that is always applicable. It is the only one we will study.

Every known procedure for valid multiple testing conceptually combines the multiple tests into one big test. So you really do only one test. The null hypothesis for the combined test is that *all* the null hypotheses for the separate tests are true. The decision rule for the combined test is to reject the combined null hypothesis if any of the separate tests rejects its null hypothesis.

Suppose we have k tests with null hypotheses H_1, ..., H_k (we can't call them all H_0). The null hypothesis for the combined test is

$$H_0 = H_1 \text{ and } H_2 \text{ and} \cdots \text{and } H_k.$$

In terms of parameter sets, the logical "and" operation corresponds to set intersection. So if Θ_i is the set of parameter values corresponding to the null hypothesis H_i for the i-th separate test, then the parameter values corresponding to H_0 are

$$\Theta_0 = \Theta_1 \cap \Theta_2 \cap \cdots \cap \Theta_k.$$

The significance level is

$$P(\text{reject } H_0) = P(\text{reject } H_1 \text{ or reject } H_2 \text{ or } \ldots \text{ or reject } H_k)$$

assuming this does not depend on the parameter value (otherwise we would have to "sup" over Θ_0). Let E_i denote the event that the i-th test rejects its null hypothesis. If the i-th test is determined by a test statistic $T_i(\mathbf{X})$ and critical value c_i, then

$$E_i = \{\, \mathbf{X} : T_i(\mathbf{X}) \ge c_i \,\}$$

but the exact form of E_i doesn't matter, it is just the set of data values for which the i-th test rejects its null. Since the logical "or" operation corresponds to set union,

$$P(\text{reject } H_0) = P(E_1 \cup E_2 \cup \cdots \cup E_k). \tag{9.64}$$

But now we are stuck. In general we have no way to calculate (9.64). It is the correct significance level for the combined test. If we are to do the test properly, we must calculate it. But in general, especially when we have done a lot of tests with no simple pattern, there is no way to do this calculation.

The right hand side of (9.64) should be familiar. It appears in the addition rule for probability, which is (10) of Theorem 2 of Chapter 2 in Lindgren. But that rule has a condition, the E_i must be mutually exclusive, which never holds in multiple testing situations. So the addition rule is no help. However there is a rule with no conditions, *subadditivity of probability*

$$P(E_1 \cup E_2 \cup \cdots \cup E_k) \le P(E_1) + P(E_2) + \cdots + P(E_k)$$

[5] There are whole books focused on this subject. A good one is *Simultaneous Statistical Inference* by Rupert G. Miller, Jr. (2nd ed., McGraw-Hill, 1981).

that holds for any events $E_1, \ldots, E_k$. This is (b) of Problem 2-22 in Lindgren. We can always apply this rule. Thus we get

$$\alpha = P(\text{reject } H_0) \le \sum_{i=1}^{k} P(\text{test } i \text{ rejects } H_i).$$

This at least provides an upper bound on the significance level. If we use the right hand side instead of α we at least get a conservative procedure. The true error rate, the probability of erroneously rejecting H_0 will be less that our upper bound.

If we adjust the separate tests so they all have the same individual significance level we get the following rules.

> *To do a combined test with significance level at most α, choose level α/k for the k separate tests.*

When we consider this in terms of P-values, the rule becomes

> *When you do k tests, multiply all P-values by k.*

This procedure is usually referred to as a *Bonferroni correction*, because a closely related inequality to subadditivity of probability is sometimes called Bonferroni's inequality.

Using the duality of tests and confidence intervals we immediately get the analogous procedure for multiple confidence intervals. There are two views we can take of multiple confidence intervals. If we have several confidence intervals, all with the same confidence level, for specificity say 95%, that does *not* mean there is 95% probability that they will all simultaneously cover. In fact if there are many intervals, there may be an extremely small probability of simultaneous coverage. Simultaneous confidence intervals is the dual concept of multiple tests. Bonferroni correction applied to confidence intervals says

> *To get simultaneous* $100(1-\alpha)\%$ *coverage for k confidence intervals choose confidence level* $100(1-\alpha/k)\%$ *for the separate intervals.*

Stargazing

Many scientific papers avoid P-values. They only indicate whether certain results are "statistically significant" or not at the 0.05 level and perhaps also at the 0.01 level. Such papers are full of tables like this one

1.13	-1.12	-1.30	1.16	-0.19
-1.18	0.12	0.02	-1.11	0.35
-0.49	-0.11	-0.45	-0.17	-1.66
2.70^{**}	0.03	0.14	-1.64	0.61
-0.35	1.80^{*}	2.65^{**}	-0.73	-1.32

* $P < 0.05$, ** $P < 0.01$

Although the asterisks are just footnote symbols, tables like this are so common that no one familiar with the literature needs to look at the footnote. One star means "significant" (statistically significant at the 0.05 level), and two stars means "highly significant" (statistically significant at the 0.01 level). The stars are supposed to indicate the interesting results. The unstarred numbers are garbage (uninteresting random noise).

Most such tables are completely bogus because *no correction was done for multiple testing.* If a Bonferroni correction (or some other correction for multiple testing) were done, there would be a lot fewer stars. And this would mean a lot fewer so-called significant results for scientists to woof about. Doing the tests honestly, with correction for multiple testing would take all the fun out of the game.

This practice has been disparagingly called "stargazing" by a sociologist (L. Guttman). It should have no place in real science. Yet it is widespread. In many scientific disciplines a paper is unusual if it *doesn't* have tables like this. Scientists being sheep, just like other people, they feel pressure to conform and use the stars. In many disciplines, tables like this are a form of what I call "honest cheating." The tests are bogus, but this is clearly admitted in the paper, so no one should be fooled. Actually, scientists never say anything so harsh as calling the procedure "bogus." That would offend their peers. The emit some academic weasel wording like "no correction was done for multiple testing." If you are statistically astute, you catch the implication of bogosity. Of course the naive reader completely misses the point, but the scientific literature isn't written to be readable by nonexperts.

To give the "honest cheaters" fair credit, they do have an argument for their failure to correct for multiple testing. If they did a Bonferroni correction, that would not be the exactly right thing to do, rather it would be the *conservative* thing to do. No correction is too liberal, Bonferroni is too conservative. The right thing would be somewhere in between, but we usually do not know how to do it. The "honest cheaters" admit they are making a mistake, but they assert that Bonferroni would *also* be a mistake (in the other direction). The trouble with this argument is that the right thing is a lot closer to Bonferroni correction than no correction. Doing many tests with no correction is always bogus.

Of course, tables full of stars are fine if Bonferroni or some other correction for multiple testing was done. Since this is so rare, all authors who do proper correction for multiple testing make it very clear that they did so. They don't want readers to assume they are as clueless and their results as meaningless as in the typical paper in their discipline.

Problems

9-1. The Laplace distribution defined in (9.1) does not have mean zero and variance one. Hence is not the *standard* Laplace distribution. What is the mean and variance of (9.1), and what would be the standard Laplace density (the one with mean zero and variance one)? If we use the standard Laplace density as the reference density of the Laplace location-scale family so that the parameters μ and σ would be the mean and standard deviation, what form would the densities have instead of (9.2)?

9-2. Show that the family of $\text{Gam}(\alpha, \lambda)$ distributions with α fixed and λ varying, taking on all values with $\lambda > 0$ is a scale family.

9-3. Suppose S_n^2 is the sample variance calculated from an i. i. d. normal random sample of size n.

(a) Calculate the bias of S_n as an estimator of the population variance σ.

(b) Find the constant k such that kS_n has the smallest mean square error as an estimator of σ.

9-4. Suppose U and V are statistics that are stochastically independent and are both unbiased estimators of a parameter θ. Write $\text{var}(U) = \sigma_U^2$ and $\text{var}(V) = \sigma_V^2$, and define another statistic $T = aU + (1 - a)V$ where a is an arbitrary but known constant.

(a) Show that T is an unbiased estimator of θ.

(b) Find the a that gives T the smallest mean square error.

9-5. The notes don't give any examples of estimators that are *not* consistent. Give an example of an inconsistent estimator of the population mean.

9-6. If $X \sim \text{Bin}(n, p)$, show that $\hat{p}_n = X/n$ is a consistent and asymptotically normal estimator of p, and give the asymptotic distribution of $\hat{p}_n$.

9-7. If X_1, X_2, $\ldots$ are i. i. d. from a distribution having a variance σ^2, show that both V_n and S_n^2 are consistent estimators of σ^2.

9-8. Suppose X_1, X_2, $\ldots$ are i. i. d. $\text{Geo}(p)$. Find a method of moments estimator for p.

9-9. Suppose X_1, X_2, $\ldots$ are i. i. d. $\text{Beta}(\alpha, 2)$.

(a) Find a method of moments estimator for α.

(b) Find the asymptotic normal distribution of your estimator.

9-10. Let X_1, X_2, $\ldots$, X_n be an i. i. d. sample from a $\text{Beta}(\theta, \theta)$ model, where θ is an unknown parameter. Find a method of moments estimator of θ.

9-11. Suppose X_1, X_2, $\ldots$ are i. i. d. $\text{Gam}(\alpha, \lambda)$. Find the asymptotic normal distribution of the method of moments estimator $\hat{\lambda}_n$ defined in (9.6b).

9-12. Calculate the ARE of $\overline{X}_n$ versus $\widetilde{X}_n$ as an estimator of the center of symmetry for

(a) The double exponential location-scale family having density given in Problem 7-6(b) of these notes. (Note that σ in in the formula for the densities given in that problem is *not* the standard deviation.)

(b) The $t(\nu)$ location-scale family, with densities given by

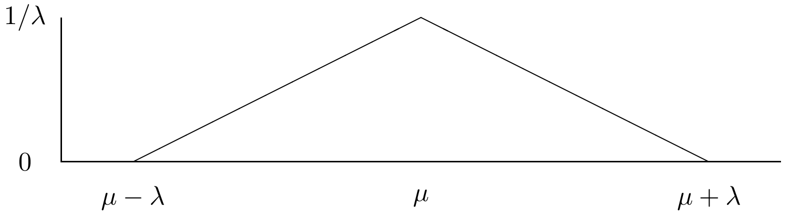
$$f_{\nu,\mu,\sigma}(x) = \frac{1}{\sigma} f_\nu\left(\frac{x-\mu}{\sigma}\right)$$

where f_ν is the $t(\nu)$ density given by (7.32). (Be careful to say things that make sense even considering that the $t(\nu)$ distribution does not have moments of all orders. Again σ is *not* the standard deviation.)

(c) The family of distributions called $\text{Tri}(\mu, \lambda)$ (for triangle) with densities

$$f_{\mu,\lambda}(x) = \frac{1}{\lambda}\left(1 - \frac{|x-\mu|}{\lambda}\right), \qquad |x-\mu| < \lambda$$

shown below



The parameter μ can be any real number, λ must be positive.

9-13. Let X_1, X_2, ..., X_n be an i. i. d. sample from a $\mathcal{N}(\mu, \sigma^2)$ model, where μ and σ^2 are unknown parameters, and let S_n^2 denote the sample variance (defined as usual with $n-1$ in the denominator). Suppose $n = 5$ and $S_n^2 = 53.3$. Give an exact (not asymptotic) 95% confidence interval for σ^2.

9-14. In an experimental weight loss program five subjects were weighed before and after the 15 week treatment. The weights in pounds were as follows

	Subject				
	A	B	C	D	E
Before	225	216	215	225	186
After	193	206	171	223	156

If you want to use R on this problem, the data are in the file

```
http://www.stat.umn.edu/geyer/5102/prob9-14.dat
```

(a) Calculate a 95% confidence interval for the expected weight loss under the program.

(b) Describe the assumptions required to make this a valid confidence interval.

9-15. Suppose we have a sample with replacement of size n from a population and we are interested in the fraction p of the population having a certain property. For concreteness, say the property is that they intend to vote for Jones in an upcoming election. Let $\hat{p}_n$ denote the fraction of the sample having the property (intending to vote for Jones in the example).

(a) Show that

$$\frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1). \tag{9.65a}$$

(b) Show that $\hat{p}_n(1-\hat{p}_n)$ is a consistent estimator of $p(1-p)$.

(c) Show that

$$\frac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)/n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1). \tag{9.65b}$$

(d) Show that

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

is an asymptotic $100(1-\alpha)\%$ confidence interval for p, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

(e) Find another asymptotic confidence interval for p based on the pivotal quantity in (9.65a) rather than the pivotal quantity in (9.65b).

9-16. Suppose we have two independent samples of size m and n from two different populations. We are interested in the fractions p and q of the populations that have a certain property (note: we are not using the $q = 1-p$ convention here, p is the proportion of the first population having the property, and q is the proportion of the second population). We estimate these proportions by the sample proportions $\hat{p}_m$ and $\hat{q}_n$ which are the fractions of the first and second samples having the property. Show that

$$\hat{p}_m - \hat{q}_n \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_m(1-\hat{p}_m)}{m} + \frac{\hat{q}_n(1-\hat{q}_n)}{n}} \tag{9.66}$$

is an asymptotic $100(1-\alpha)\%$ confidence interval for $p-q$, where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

9-17. A physics lab is divided into 20 teams. Each team performs a measurement of the speed of light. Ten teams use one method and the other ten use another method. The average and standard deviation for the teams using each method was given in the following table (units are meters per second times 10^8).

	mean	standard deviation
Method 1	3.00013	0.00395
Method 2	2.99019	0.00853

If you want to use R on this problem, the data are in the file

```
http://www.stat.umn.edu/geyer/5102/prob9-17.dat
```

(a) Assuming that the measurements within each group of ten teams are independent and identically distributed around some unknown mean value (the speed of light as measured by that method), calculate a 95% confidence interval for the difference in the mean values for the two methods using Welch's approximation.

(b) Redo part (a) using the "pooled variance" t confidence interval that assumes both measurement methods have the same variance.

9-18. Suppose a sample of size 100 is assumed to be i. i. d. from a $\text{Gam}(\alpha, \lambda)$ model and the method of moments estimators of the parameters are $\hat{\alpha}_n = 5.23$ and $\hat{\lambda}_n = 21.3$. Find an asymptotic 95% confidence interval for α.

9-19. Suppose $V_{X,m}$ and $V_{Y,n}$ are sample variances and $M_{4,X,m}$ and $M_{4,Y,n}$ are the sample fourth central moments of independent samples from two populations having variances σ_X^2 and σ_Y^2, respectively. Find an asymptotic confidence interval for $\sigma_X^2 - \sigma_Y^2$.

9-20. Show that the "exact" confidence interval for the variance based on the chi-square distribution is asymptotically robust within the class of all distributions having fourth moments and satisfying $\mu_4 = 3\sigma^4$. That is, show that the assumption

$$\frac{nV_n}{\sigma_2} \sim \text{chi}^2(n-1)$$

implies a certain asymptotic limit for V_n and that this limit matches the *correct* asymptotic limit given by Theorem 7.17 only if $\mu_4 = 3\sigma^4$.

9-21. A *trimmed mean* is a point estimator of the form

$$\frac{X_{(k+1)} + \cdots + X_{(n-k)}}{n - 2k} \tag{9.67}$$

that is, the average of the data after the k lowest and k highest order statistics have been thrown away. If $0 \le \alpha < 1/2$ we say that (9.67) is a $100\alpha\%$ trimmed mean if $k = \lfloor n\alpha \rfloor$. We say that the median is a 50% trimmed mean.

For $0 < \alpha < 1/2$, find the breakdown point of the $100\alpha\%$ trimmed mean.

9-22. Given data X_1, ..., X_n, the *Walsh averages* consist of the n data items X_i and the $\binom{n}{2}$ averages $(X_i + X_j)/2$ for distinct pairs of indices i and j. The *Hodges-Lehmann estimator* of the center of symmetry of a symmetric distribution is the empirical median of the vector of Walsh averages.[6] Find the breakdown point of this estimator.

[6] Actually, there are lots of different Hodges-Lehmann estimators. This is the one associated with the Wilcoxon signed rank test.

9-23. Calculate the breakdown point of the MAD (median absolute deviation from the median) defined in Problem 7-1.

9-24. Assume X_1, $\ldots$, X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$. The observed value of $\overline{X}_n$ is 23.6, the observed value of S_n^2 is 103.2, and the sample size is $n = 20$. Perform a one-tailed test of the hypotheses $H_0 : \mu = 20$ versus $H_A : \mu > 20$, finding the P-value.

9-25. Two groups in physics lab have been measuring the density of aluminum at room temperature (20° C). They got the following summary statistics

	n	$\overline{X}_n$	S_n
Group I	10	2.792	0.241
Group II	8	2.538	0.313

(Units are grams per cubic centimeter.) Assume the measurements for group I are i. i. d. $\mathcal{N}(\mu_1, \sigma_1^2)$ and the measurements for group II are i. i. d. $\mathcal{N}(\mu_2, \sigma_2^2)$. We want to perform a test of $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$. Perform Welch's approximate test, come as close as you can to the P-value.

If you want to use R on this problem, the data are in the file

```
http://www.stat.umn.edu/geyer/5102/prob9-25.dat
```

9-26. Suppose I have taken a random sample of size 100 of ears of corn from a field. My sample has mean ear length of 6.13 inches and standard deviation 1.44 inches. This gives me a 95% confidence interval for the true mean ear length all the corn in the field of 6.13 ± 0.28 inches.

Suppose I want a more accurate 95% confidence interval with a half-width (plus-or-minus) of 0.10 inches. What sample size do I need to get that?

9-27. Suppose I intend to collect data about the effect of coaching on SAT scores. The data will be SAT scores for individuals before and after taking a cram course. Suppose the test-retest variability without coaching is known to be about 50 points. How large a sample size do I need to have a power of 0.95 of detecting a true mean difference due to coaching as small as 10 points (the null hypothesis being no difference) at the 0.05 significance level? The test will be an upper-tailed test, since we expect that coaching cannot hurt.

9-28. For the data in Example 9.5.1 compute four confidence intervals, one for difference in each of the four rows of the table, so that your four intervals have 95% probability of *simultaneous* coverage.

Note: This problem can be done in R using the `prop.test` function, but getting the right confidence level is tricky. Be careful.

9-29. A problem on "stargazing." Suppose the twenty-five numbers in the table on p. 294 are all z-scores for different one-tailed, upper-tailed tests. The stars in the table do not reflect any correction for multiple testing. That is a z-score is declared "significant" (gets a star) if $z \geq 1.645$ and is declared "highly significant" (gets two stars) if $z \geq 2.326$. Here 1.645 and 2.326 are the one tailed 0.05 and 0.01 z critical values.

(a) What critical values should replace 1.645 and 2.326 in order to apply a Bonferroni correction to this multiple testing situation?

(b) What would the result of the Bonferroni correction be in terms of stars?

Chapter 10

Likelihood Inference

10.1 Likelihood

“Likelihood” is used as a technical term in statistics. It is not just a vague synonym for probability, as it is in everyday language. It is, however, closely related to probability.

Recall that we use *density* as a term that covers two of Lindgren’s terms: p. f. (probability function) and p. d. f. (probability density function). In this chapter we will see one of the main reasons for our usage. The density will be used in exactly the same way, regardless of whether the data are discrete (so Lindgren would call it a p. f.) or continuous (so Lindgren would call it a p. d. f.). Also recall that a *statistical model* can be described by giving a parametric family of densities $\{ f_\theta : \theta \in \Theta \}$. This means that for each fixed parameter value θ in the parameter space Θ, there is a function $f_\theta(x)$ defined on the sample space that is nonnegative and sums or integrates to one, depending on whether the model is discrete or continuous.

A *likelihood* for the statistical model is defined by the same formula as the density, but the roles of x and θ are interchanged

$$L_x(\theta) = f_\theta(x). \tag{10.1}$$

Thus the likelihood is a different function of the parameter θ for each fixed value of the data x, whereas the density is a different function of x for each fixed value of θ. Likelihood is actually a slightly more general concept, we also call

$$L_x(\theta) = h(x) f_\theta(x) \tag{10.2}$$

a likelihood for the model when $h(x)$ is any nonzero function of x that does not contain the parameter θ. The reason for this extension of the notion is that all of the uses we make of the likelihood function will not be affected in any way by the presence or absence of $h(x)$. The way we make use of the extended definition is to simply drop terms in the density that do not contain the parameter.

Example 10.1.1 (Binomial Likelihood).
If $X \sim \text{Bin}(n, p)$, then

$$f_p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

This is also a likelihood $L_x(p)$. However, we are also free to drop the binomial coefficient, which does not contain the parameter p, writing

$$L_x(p) = p^x (1-p)^{n-x}. \tag{10.3}$$

When the data are an i. i. d. sample from a distribution with density $f_\theta(x)$, the joint density is

$$f_\theta(\mathbf{x}) = \prod_{i=1}^n f_\theta(x_i)$$

Hence this, thought of as a function of the parameter θ rather than the data $\mathbf{x}$, is also a likelihood. As usual we are allowed to drop multiplicative terms not containing the parameter.

When there are several parameters, the likelihood is a function of several variables (the parameters). Or, if we prefer, we can think of the likelihood as a function of a vector variable (the vector of parameters).

Example 10.1.2 (Normal Likelihood).
Suppose X_1, $\ldots$, X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$, then the joint density of the data is

$$f_{\mu,\sigma}(\mathbf{x}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2}$$

We can drop the $\sqrt{2\pi}$ terms. This gives

$$\begin{aligned} L_{\mathbf{x}}(\mu, \sigma) &= \prod_{i=1}^n \frac{1}{\sigma} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\ &= \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2\right) \qquad (10.4) \\ &= \frac{1}{\sigma^n} \exp\left(-\frac{nv_n + n(\bar{x}_n-\mu)^2}{2\sigma^2}\right) \end{aligned}$$

where $\bar{x}_n$ is the empirical mean and v_n is the empirical variance, the last step using the empirical parallel axis theorem. Of course, we are free to use whichever form seems most convenient.

Example 10.1.3 (Normal Likelihood, Known Variance).
This is the same as the preceding example except now we assume σ^2 is a known constant, so μ is the only parameter. Now we are free to drop multiplicative terms not containing μ. Hence we can drop the σ^n term. We can also write

$$\exp\left(-\frac{nv_n + n(\bar{x}_n-\mu)^2}{2\sigma^2}\right) = \exp\left(-\frac{nv_n}{2\sigma^2}\right) \exp\left(-\frac{n(\bar{x}_n-\mu)^2}{2\sigma^2}\right)$$

and the first term on the right does not contain μ, hence

$$L_{\mathbf{x}}(\mu) = \exp\left(-\frac{n(\bar{x}_n - \mu)^2}{2\sigma^2}\right) \tag{10.5}$$

is a likelihood for this model. Note that this is a *different* statistical model from preceding problem because the parameter space is different. Hence it has a *different* likelihood function.

For the next several sections, we will only look at models with one parameter. We will return to multiparameter models in Section 10.4.

10.2 Maximum Likelihood

So far the only general method we have seen for constructing estimators is the method of moments. Now we are going to learn another method, even more general than the method of moments. It has a number of very desirable properties that will be developed as we go along. It is called the method of maximum likelihood. Roughly, the maximum likelihood estimator is the parameter value that maximizes the likelihood. For observed data $\mathbf{x}$ and likelihood $L_{\mathbf{x}}$ the *maximum likelihood estimator* (MLE) is defined to be the parameter value that maximizes the function $L_{\mathbf{x}}$ if the global maximum exists and is unique. If the global maximum does not exist or is not unique, then we have a problem defining the MLE. Mostly we will just deal with situations where there is a unique global maximizer. The MLE is denoted $\hat{\theta}(\mathbf{x})$ or sometimes just $\hat{\theta}$ when we want to leave the dependence on the data out of the notation. When we discuss asymptotics, we will often write it $\hat{\theta}_n(\mathbf{x})$ or $\hat{\theta}_n$ in order to indicate the dependence on the sample size n.

The *log likelihood* is the (natural) logarithm of the likelihood. It is denoted

$$l_{\mathbf{x}}(\theta) = \log L_{\mathbf{x}}(\theta).$$

We define $\log(0) = -\infty$. This makes sense because $\log(x) \to -\infty$ as $x \downarrow 0$.

Because the logarithm function is strictly increasing, a point maximizes the likelihood if and only if it maximizes the log likelihood. It is often simpler to maximize the log likelihood rather than the likelihood.

Example 10.2.1 (Binomial Model).
The log likelihood for the binomial distribution is found by taking logs in (10.3) giving

$$l_x(p) = x \log p + (n - x)\log(1 - p). \tag{10.6a}$$

Calculating derivatives at interior points p of the parameter space gives

$$\begin{aligned} l'_x(p) &= \frac{x}{p} - \frac{n-x}{1-p} \\ &= \frac{x - np}{p(1-p)} \qquad (10.6b) \\ l''_x(p) &= -\frac{x}{p^2} - \frac{n-x}{(1-p)^2}. \qquad (10.6c) \end{aligned}$$

The last equation shows that the log likelihood is a strictly concave function (Definition G.2.1 in Appendix G).

In the general case $0 < x < n$, the first derivative is zero at $\hat{p} = x/n$. By strict concavity, $\hat{p}$ is the unique global maximum. In the special cases $x = 0$ and $x = n$, there is no zero of the derivative, but the endpoints of the parameter space ($0 \le p \le 1$), are local maxima. When $x = 0$

$$l'_x(p) = -\frac{n}{1-p}$$

so $l'_x(0) = -n$ which satisfies (G.2a). Similarly, when $x = n$ we have $l'_x(1) = n$ which satisfies the sufficient condition for $p = 1$ to be a local maximum. Thus in all three cases $\hat{p} = x/n$ is a local maximum of the log likelihood, hence the unique global maximum by strict concavity.

In this case, the MLE is the obvious estimator. By definition X is the sum of i. i. d. Bernoulli random variables, and $\hat{p}$ is the sample mean of these variables. It is also a method of moments estimator and an unbiased estimator, since $E(\hat{p}) = p$.

Example 10.2.2 (Normal Model, Known Variance).
The likelihood for this model is given by (10.5), hence the log likelihood is

$$l_n(\mu) = -\frac{n(\bar{x}_n - \mu)^2}{2\sigma^2} \qquad (10.7)$$

which is clearly maximized at

$$\hat{\mu}_n = \bar{x}_n$$

(since the log likelihood is zero there and negative elsewhere), so that is the MLE.

Example 10.2.3 (Cauchy Location Model).
The Cauchy location model has densities

$$f_\theta(x) = \frac{1}{\pi} \cdot \frac{1}{1 + (x - \theta)^2}$$

Hence the log likelihood for an i. i. d. sample of size n is

$$l_n(\theta) = -\sum_{i=1}^{n} \log\left(1 + (x_i - \theta)^2\right)$$

(we can drop the constant terms $1/\pi$). We can differentiate this

$$l_n'(\theta) = \sum_{i=1}^{n} \frac{2(x_i - \theta)}{1 + (x_i - \theta)^2} \tag{10.8}$$

but the result is horrible. We won't get anywhere by setting that equal to zero and trying to solve for θ.

Fortunately, computers can help. They can't give us a formula expressing the MLE as a function of θ. There isn't any such formula in terms of well-known elementary functions. But for any particular data set, the computer can maximize the likelihood and find the MLE. That's good enough in most applications. R, for example, has a function `nlm` that does nonlinear minimization of a function of several variables. We can use that to find the MLE (minimizing $-f$ maximizes f). First we make up some data

```
Rweb:> n <- 40
Rweb:> theta0 <- 0
Rweb:> x <- theta0 + rcauchy(n)  # make up data
Rweb:> summary(x)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 -5.9590  -0.4615   0.1203  10.0100   1.1990 172.0000
```

In a real problem, of course, we would read in data obtained from the data collectors (though, to be honest, the Cauchy has such heavy tails that it's not used for real data). We have also run the `summary` command that gives the estimators we already know about, the sample mean and median, as well as some other interesting statistics. We know the sample mean is *not* a consistent estimator of θ because the expectation of the Cauchy distribution does not exist. We know the sample median is consistent and asymptotically normal [Problem 7-6(a)].

Then the following code finds the MLE. First it defines an R function `l` that evaluates minus the log likelihood. Then it hands that function to `nlm` to minimize. The `nlm` function uses an iterative algorithm and needs a starting point, which we supply as `median(x)`, the best estimator we know that we have a simple expression for (ignore the third argument to `nlm`, it's helpful but not necessary).

```
Rweb:> l <- function(theta) sum(log(1 + (x - theta)^2))
Rweb:> out <- nlm(l, median(x), fscale=n)
Rweb:> out$estimate
[1] 0.00276767
```

The result is the MLE. Notice that it is a lot closer to the true parameter value (which we know to be zero because we made up the data) than the median. This is no accident. We will eventually see that the MLE is a much better estimator here than the sample median.

10.3 Sampling Theory

10.3.1 Derivatives of the Log Likelihood

When we write $l_{\mathbf{X}}(\theta)$ rather than $l_{\mathbf{x}}(\theta)$, considering the subscript a random vector "big $\mathbf{X}$" rather than a fixed value "little $\mathbf{x}$" the log likelihood becomes a random function, and everything about it, including its derivatives, is also random. Specifically, $l_{\mathbf{X}}$ and its derivatives, $l'_{\mathbf{X}}$, $l''_{\mathbf{X}}$, and so forth, are random functions. The values of these functions at a specified point θ, that is, $l_{\mathbf{X}}(\theta)$, $l'_{\mathbf{X}}(\theta)$, $l''_{\mathbf{X}}(\theta)$, and so forth, are random variables. Note that each different value of θ gives a *different* random variable. Like every other random variable, they have probability distributions. As usual, when we are talking about randomness arising from or mimicking random sampling, we call these the sampling distributions of the random variables, in this case, of the log likelihood and its derivatives.

We are now going to change notation, suppressing the dependence of the log likelihood on the data $\mathbf{X}$ and emphasizing the dependence on the sample size n. As always, this is useful when we discuss asymptotics, and all of the distribution theory in likelihood inference is asymptotic theory. Thus we will write l_n, l'_n, and so forth rather than $l_{\mathbf{X}}$, $l'_{\mathbf{X}}$, and so forth.

The Score

The first derivative of the log likelihood

$$l'_n(\theta) = \frac{d}{d\theta} l_n(\theta)$$

is often called the *score function* or just the *score*. When we consider the score function to be random, $l'_n(\theta)$ is a random variable, a different random variable for each different value of θ.

The score function is important in maximum likelihood. We usually find the MLE by solving the equation $l'_n(\theta) = 0$. Of course, this doesn't work when the MLE is on the boundary of the parameter space or when the MLE doesn't exist, but it does work in the usual case and we have $l'_n(\hat{\theta}_n) = 0$. Note that this does *not* imply that $l'_n(\theta) = 0$ when θ is the true parameter value. Just the opposite! "The sample is not the population" implies that $\hat{\theta}_n$ is *not* θ. In fact, $l'_n(\theta)$ is a random variable and hence doesn't have any constant value.

Expected Fisher Information

The *Fisher information* for a statistical model is the variance of the score $l'_n(\theta)$

$$I_n(\theta) = \operatorname{var}_\theta\{l'_n(\theta)\}. \tag{10.9}$$

It is named after R. A. Fisher, who invented maximum likelihood and discovered many of the properties of maximum likelihood estimators and first called this concept "information." Lindgren calls this concept just "information" instead of “Fisher information,” but the latter is standard terminology because more than one notion of “information” has been used in statistics (although Fisher information is by far the most important and the only one we will consider in this course).

Lindgren uses the notation $I_X(\theta)$ rather than $I_n(\theta)$ but admits this “could be misleading” because the Fisher information does *not* depend on the data X but rather on the *model*, that is on *which* variables we consider “data” rather than on the *values* of those variables. Note that the Fisher information is *not* a random quantity (because *no* unconditional expectation is a random quantity), another reason why the notation $I_X(\theta)$ is very misleading. Since Lindgren's notation is misleading, we will not use it.

Differentiating Under the Integral Sign

Any probability density satisfies

$$\int f_\theta(x)\, dx = 1 \tag{10.10}$$

(or the analogous equation with summation replacing integration if the data are discrete). Usually, although not always,[1] it is possible to take derivatives inside the integral sign, that is,

$$\frac{\partial^k}{\partial \theta^k} \int f_\theta(x)\, dx = \int \frac{\partial^k}{\partial \theta^k} f_\theta(x)\, dx. \tag{10.11}$$

Looking back at the right hand side of (10.10), we see that because the derivative of a constant is zero, that all of the derivatives in (10.11) are zero, that is,

$$\int \frac{\partial^k}{\partial \theta^k} f_\theta(x)\, dx = 0 \tag{10.12}$$

provided that differentiation under the integral sign is valid, that is, provided (10.11) holds.

The partial derivative notation will become unwieldy in the following proof, so we are going to introduce the following shorthand for (10.12)

$$\int f' = 0$$

and

$$\int f'' = 0$$

using primes to indicate partial derivatives with respect to θ (in likelihood theory we always differentiate with respect to parameters, never with respect to data)

[1] We will not worry about the precise technical conditions under which this operation is permitted. They can be found in advanced calculus books. The only condition we will mention is that the limits of integration in (10.11) must not contain the variable of differentiation θ. This will hold in all of the examples we consider.

and also suppressing the variable x entirely (although the integration is still with respect to the data x).

Now we write the log likelihood $l = \log f$, and using the chain rule we obtain

$$l' = \frac{f'}{f} \tag{10.13a}$$

$$l'' = \frac{f''}{f} - \left(\frac{f'}{f}\right)^2 \tag{10.13b}$$

Now note that for any random variable $g(X)$

$$E\{g(X)\} = \int g(x) f(x) \, dx$$

or in the shorthand we are using in this section $E(g) = \int g f$. What this says is that in order to change an expectation to an integral we need an extra f in the integrand. Thus taking expectations of (10.13a) and (10.13b) gives

$$\begin{aligned} E(l') &= \int f' \\ E(l'') &= \int f'' - E\left\{\left(\frac{f'}{f}\right)^2\right\} \end{aligned}$$

and we know from our previous discussion that (still assuming differentiability under the integral sign is valid) that the integrals here are zero, thus

$$E(l') = 0 \tag{10.13c}$$

$$E(l'') = -E\left\{\left(\frac{f'}{f}\right)^2\right\} \tag{10.13d}$$

Finally we note that we can use (10.13a) and (10.13c) to simplify (10.13d). $l' = f'/f$ is a random variable. It has mean zero by (10.13c). For any random variable having mean zero ordinary and central moments are the same, hence the variance is also the ordinary second moment. Thus the second term on the right hand side of (10.13d) is $\operatorname{var}(l')$. Thus we can rewrite (10.13d) as

$$E(l'') = -\operatorname{var}(l') \tag{10.13e}$$

Now we want to get rid of the shorthand, and restate our conclusions as a theorem using ordinary mathematical notation.

Theorem 10.1. *Provided* (10.10) *can be differentiated twice with respect to θ under the integral sign, that is* (10.12) *holds for $k = 1$ and $k = 2$,*

$$E_\theta\{l_n'(\theta)\} = 0 \tag{10.14a}$$

and

$$E_\theta\{l_n''(\theta)\} = -\operatorname{var}_\theta\{l_n'(\theta)\} \tag{10.14b}$$

for all values of θ for which the differentiation under the integral sign is permitted.

This is just what we proved in the preceding discussion.

Note that the variance on the right hand side of (10.14b) is Fisher information. This says that we can calculate Fisher information in two different ways, either the variance of the first derivative of the log likelihood or minus the expectation of the second derivative. You may use whichever seems simpler, your choice. First derivatives are sometimes simpler than second derivatives and sometimes not. Expectations are usually simpler than variances. My experience is that a majority of problems the second derivative calculation is simpler. But in a sizable minority of problems the first derivative calculation is simpler. Don't entirely ignore the first derivative method.

Example 10.3.1 (Binomial Model).
In Example 10.2.1 we found the second derivative of the log likelihood to be

$$l''_X(p) = -\frac{X}{p^2} - \frac{n - X}{(1-p)^2}.$$

This is (10.6c) with "little x" changed to "big X" because we are now considering it a random quantity. Taking expectations using $E(X) = np$ gives

$$\begin{aligned} I_n(p) &= -E\{l''_X(p)\} \\ &= \frac{np}{p^2} + \frac{n - np}{(1-p)^2} \\ &= \frac{n}{p(1-p)} \end{aligned}$$

Example 10.3.2 (Normal Model, Known Variance).
In Example 10.2.2 we found the log likelihood (10.7) for this model Differentiating, we find

$$l''_n(\mu) = -\frac{n}{\sigma^2}$$

Since this happens not to depend on the data, it is nonrandom, hence is its own expectation. Thus it is minus the Fisher information, that is

$$I_n(\mu) = \frac{n}{\sigma^2}$$

The subscripts θ on the expectation and variance operators in (10.14a) and (10.14b) are important. If omitted, it would be possible to give these equations a reading that is false. The point is that there are two θ's involved. When they are different, the statement is simply false.

$$E_{\theta_1}\{l'_n(\theta_2)\} \neq 0$$

when $\theta_1 \neq \theta_2$. If (10.14a) is written with no subscript on the expectation operator

$$E\{l'_n(\theta)\} = 0,$$

then it is not clear what parameter value is meant and under the wrong assumption about what parameter is meant the equation is simply false.

The CLT for the Score

The log likelihood and its derivatives for an i. i. d. sample are sums of i. i. d. terms.

$$l_n(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i) \tag{10.15a}$$

$$l'_n(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f_\theta(X_i) \tag{10.15b}$$

$$l''_n(\theta) = \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \log f_\theta(X_i) \tag{10.15c}$$

and so forth. The middle equation (10.15b) is the score. Also note that each term on the right hand side of (10.15a) is the log likelihood for a model having only X_i as data, that is, a log likelihood for sample size one.

Theorem 10.2.

$$I_n(\theta) = nI_1(\theta)$$

Proof. Because the X_i are assumed independent, the terms on the right hand side of (10.15b) are independent. Hence the variance of the sum is the sum of the variances. By the preceding comment, each term on the right hand side is a score for a sample of size one. □

In words, the theorem says the Fisher information for a sample of size n is equal to the Fisher information for a sample of size 1 multiplied by n.

Example 10.3.3 (Cauchy Location Model).
In Example 10.2.3 we found the first derivative of the log likelihood (the score) for this model in (10.8). The second derivative is

$$l''_n(\theta) = \sum_{i=1}^{n} \left(-\frac{2}{1 + (x_i - \theta)^2} + \frac{4(x_i - \theta)^2}{[1 + (x_i - \theta)^2]^2} \right) \tag{10.16}$$

We called the first derivative “horrible.” This is even more of a mess, but we are only trying to integrate this, and there is a nice analytical (though fairly messy) indefinite integral. Note by Theorem 10.2 that to find the Fisher information, we only need to do the integral for sample size one. Mathematica has no trouble

```
In[1]:= <<Statistics‘ContinuousDistributions‘

In[2]:= dist = CauchyDistribution[theta, 1]

Out[2]= CauchyDistribution[theta, 1]

In[3]:= f[x_, theta_] = PDF[dist, x]
```

```
                    1
Out[3]= ----------------------
                             2
        Pi (1 + (-theta + x) )

In[4]:= Integrate[ D[ Log[f[x, theta]], {theta, 2} ] f[x, theta],
        {x, -Infinity, Infinity} ]

          1
Out[4]= -(-)
          2
```

We don't even need to do the differentiation ourselves. Let Mathematica do both differentiation and integration. The Fisher information is minus this

$$I_1(\theta) = \frac{1}{2} \tag{10.17}$$

And of course, by Theorem 10.2 the Fisher information for sample size n is just n times this.

In the proof of Theorem 10.2 we established that the right hand side of (10.15b) is the sum of i. i. d. terms, each of which is the score for a sample of size one and hence has mean zero by (10.14a) and variance $I_1(\theta)$ by (10.9), the definition of Fisher information.

Thus $l_n'(\theta)/n$ is the average of i. i. d. terms and the central limit theorem applies. Being precise, it says

$$\frac{1}{\sqrt{n}} l_n'(\theta) \xrightarrow{\mathcal{D}} \mathcal{N}\bigl(0, I_1(\theta)\bigr). \tag{10.18}$$

The $1/\sqrt{n}$ arises because we divide $l_n'(\theta)$ by n to get an average then we multiply by $\sqrt{n}$ as usual in the CLT. There is no mean subtracted off on the left hand side because the score has mean zero. The sloppy "double squiggle" version says

$$l_n'(\theta) \approx \mathcal{N}\bigl(0, I_n(\theta)\bigr).$$

Here we wrote $I_n(\theta)$ rather than $nI_1(\theta)$ for the variance (they are, of course, equivalent by Theorem 10.2). Note that the asymptotic mean and variance are no surprise, since by (10.14a) and the definition of Fisher information (10.9) these are the *exact* mean and variance of $l_n'(\theta)$. The only surprise (and it should be no surprise by now) is that the large sample distribution is normal (by the CLT).

Observed Fisher Information

The *observed Fisher information* is

$$J_n(\theta) = -l_n''(\theta). \tag{10.19}$$

For contrast $I_n(\theta)$ is sometimes called *expected* Fisher information to distinguish it from $J_n(\theta)$, although, strictly speaking, the "expected" is redundant.

Note that $J_n(\theta)$ is a random quantity, even though the notation does not explicitly indicate this. In contrast, expected Fisher information, like any other expected value is constant (nonrandom). The connection between observed and expected Fisher information is given by (10.14b), which says, using the notation (10.19) for observed Fisher information

$$E\{J_n(\theta)\} = I_n(\theta). \tag{10.20}$$

Example 10.3.4 (Cauchy Location Model).
In Example 10.3.3 we found the second derivative of the log likelihood for this model in (10.16). The observed Fisher information is just minus this.

$$J_n(\theta) = \sum_{i=1}^{n} \left(\frac{2}{1 + (x_i - \theta)^2} - \frac{4(x_i - \theta)^2}{[1 + (x_i - \theta)^2]^2} \right) \tag{10.21}$$

The LLN for Observed Fisher Information

Equation 10.20 gives us the expectation of the observed Fisher information. Generally, we do not know anything about its variance or any higher moments. Not knowing the variance, the CLT is of no use. But the LLN is still informative.

The analysis is just like the analysis of the sampling distribution of $l'_n(\theta)$ two sections back (but simpler because the LLN is simpler than the CLT). The right hand side of (10.15c) is the sum of i. i. d. terms, each of which is the second derivative of the log likelihood for a sample of size one and hence has mean $I_1(\theta)$ by (10.20).

Thus $J_n(\theta)/n$ is the average of i. i. d. terms and the law of large numbers applies. Being precise, it says

$$\frac{1}{n} J_n(\theta) \xrightarrow{P} I_1(\theta). \tag{10.22}$$

The sloppy "double squiggle" version would be

$$J_n(\theta) \approx I_n(\theta)$$

Note that this doesn't describe an *asymptotic distribution* for $J_n(\theta)$ because the right hand side is *constant* (as always the LLN gives less information than the CLT).

10.3.2 The Sampling Distribution of the MLE

If we expand l'_n using a Taylor series with remainder about the true parameter value θ_0, we get

$$l'_n(\theta) = l'_n(\theta_0) + l''_n(\theta_0)(\theta - \theta_0) + \tfrac{1}{2} l'''_n(\theta^*)(\theta - \theta_0)^2,$$

where θ^* in the remainder term is some point between θ and θ_0.

Using $l''_n(\theta) = -J_n(\theta)$, we get

$$l'_n(\theta) = l'_n(\theta_0) - J_n(\theta_0)(\theta - \theta_0) + \tfrac{1}{2} l'''_n(\theta^*)(\theta - \theta_0)^2. \tag{10.23}$$

Now we assume that the MLE is in the interior of the parameter space, so it satisfies the "likelihood equation" $l'_n(\hat{\theta}_n) = 0$. Then if we plug in $\hat{\theta}_n$ for θ in (10.23), the left hand side is zero, and we get

$$0 = l'_n(\theta_0) - J_n(\theta_0)(\hat{\theta}_n - \theta_0) + \tfrac{1}{2} l'''_n(\theta^*_n)(\hat{\theta}_n - \theta_0)^2, \tag{10.24}$$

where now θ^*_n is some point between θ_0 and $\hat{\theta}_n$.

Now we want to multiply (10.24) by the appropriate constant so that the various terms converge to a nontrivial distribution. Looking at the CLT for $l'_n(\theta)$, equation (10.18) we see that the right constant is $1/\sqrt{n}$ ("constant" here means nonrandom, this is, of course, a function of n). That gives

$$0 = \frac{1}{\sqrt{n}} l'_n(\theta_0) - \frac{1}{n} J_n(\theta_0) \cdot \sqrt{n}(\hat{\theta}_n - \theta_0) + \frac{1}{2\sqrt{n}} l'''_n(\theta^*_n)(\hat{\theta}_n - \theta_0)^2. \tag{10.25}$$

In the middle term we wrote $1/\sqrt{n} = \sqrt{n}/n$ and put each piece with a different factor. We know the behavior of $J_n(\theta)/n$. It's given by (10.22). And we expect from our general experience with asymptotics so far that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ will have a nontrivial asymptotic distribution.

The last term in (10.25) is a mess unlike anything we have ever seen. In order to make progress, we need to make an assumption that gets rid of that messy term. The appropriate assumption is the following.

$$\frac{1}{n} l'''_n(\theta^*_n)(\hat{\theta}_n - \theta) \xrightarrow{P} 0 \tag{10.26}$$

Then rearranging (10.25) gives

$$\sqrt{n}(\hat{\theta}_n - \theta)\left[1 - \frac{\frac{1}{2n} l'''_n(\theta^*_n)(\hat{\theta}_n - \theta)}{\frac{1}{n} J_n(\theta_0)}\right] = \frac{\frac{1}{\sqrt{n}} l'_n(\theta_0)}{\frac{1}{n} J_n(\theta_0)}.$$

Combining our assumption (10.26) with (10.22) and Slutsky's theorem, the messy second term in the square brackets converges in probability to zero (leaving only the unit term). Thus by another use of Slutsky's theorem $\sqrt{n}(\hat{\theta}_n - \theta)$ has the same asymptotic behavior as the right hand side, that is

$$\sqrt{n}\big(\hat{\theta}_n - \theta_0\big) \approx \frac{\frac{1}{\sqrt{n}} l'_n(\theta_0)}{\frac{1}{n} J_n(\theta_0)}.$$

Using (10.18) and (10.22) and Slutsky's theorem yet again, the right hand side converges in distribution to $Z/I_1(\theta)$ where $Z \sim \mathcal{N}\big(0, I_1(\theta)\big)$. Since a linear transformation of a normal is normal, $Z/I_1(\theta)$ is normal with mean

$$E\left\{\frac{Z}{I_1(\theta)}\right\} = \frac{E(Z)}{I_1(\theta)} = 0$$

and variance

$$\operatorname{var}\left\{ \frac{Z}{I_1(\theta)} \right\} = \frac{\operatorname{var}(Z)}{I_1(\theta)^2} = \frac{1}{I_1(\theta)}$$

Thus

$$\sqrt{n}\bigl(\hat{\theta}_n - \theta_0\bigr) \xrightarrow{\mathcal{D}} \mathcal{N}\bigl(0, I_1(\theta_0)^{-1}\bigr) \tag{10.27}$$

This completes a proof of the asymptotic normality of the MLE.

Theorem 10.3. *Suppose the true parameter value θ_0 is in the interior of the parameter space, and suppose the assumptions about differentiability under the integral sign in Theorem 10.1 hold. Suppose we have i. i. d. sampling. And finally suppose that assumption* (10.26) *holds. Then* (10.27) *holds.*

Example 10.3.5 (Cauchy Location Model).
In Example 10.3.3 we found the expected Fisher information for this model (10.17). Inserting that into (10.27) we get

$$\sqrt{n}\bigl(\hat{\theta}_n - \theta_0\bigr) \xrightarrow{\mathcal{D}} \mathcal{N}\bigl(0, 2\bigr)$$

I hope you are suitably impressed at the magic of theoretical statistics here. The other examples we have done in this chapter don't need to use the theory. When the MLE turns out to be $\overline{X}_n$, we already know its asymptotic distribution. In fact, whenever the MLE turns out to be a simple function of any sample moments we could use the delta method to find its asymptotic distribution (though calculating Fisher information is usually easier than applying the delta method). Here we do not have an analytic expression for the MLE as a function of the data. Thus we cannot use the delta method or any other method we have covered (or for that matter any method we *haven't* covered). Fisher information gives us the asymptotic distribution of a random variable we can't even describe (except for the implicit description that it maximizes the likelihood).

Real-life applications of the method of maximum likelihood are more like this Cauchy example than any of the other examples or homework problems. For complicated data (and it seems that real scientific data sets get ever more complicated every year) often the only thing you can write down for the model is the likelihood function. You can't calculate anything else analytically. However, the computer can calculate the MLE and the observed Fisher information (See Example 10.3.6 for more on this), and you're in business. Nothing else from theoretical statistics works, just likelihood theory.

The difficulty with applying Theorem 10.3 is that it is rather hard to verify the conditions. Why should (10.26) hold? The truth is that for some models it does, and for some models it doesn't. The assumption is not as weird as it looks. If the MLE is consistent, then $\hat{\theta}_n - \theta \xrightarrow{P} 0$. Also $\frac{1}{n} l_n'''(\theta_0)$ converges in probability to some constant (its expectation) by the law of large numbers (assuming the expectation exists), because it too is the sum of i. i. d. terms. Then by Slutsky's theorem $\frac{1}{n} l_n'''(\theta_0)(\hat{\theta}_n - \theta)$ converges in probability to zero. Our assumption (10.26) differs only in having θ_n^* in place of θ_0 as the argument of l_n'''. If $\hat{\theta}_n$ converges to θ_0, then so does θ_n^*. Hence we might expect $\frac{1}{n} l_n'''(\theta_n^*)(\hat{\theta}_n - \theta)$

to also converge in probability to zero, which would imply (10.26). Thus the assumption is plausible. But actually showing it holds for any particular model is difficult mathematics, beyond the scope of this course.

What we are left with a rather annoying situation. The "standard asymptotics" of the MLE given by (10.27) usually holds. It holds for "nice" models. But we don't have any definition of "nice" that we can understand intuitively. In fact a half century of research by a lot of smart people has not found any simple definition of "nice" that will do the job here. So though the usual asymptotics usually hold, they don't always, so we are always beset with vague anxiety when using this stuff, or at least we *should* be anxious in order to be proper statisticians. The Alfred E. Newman philosophy "What, me worry?" just isn't appropriate, though to be honest, it does about as much good as the official philosophy that you can't use (10.27) until you have somehow verified the conditions of the theorem (using a lot of math far beyond the scope of this course) or had someone else (a good theoretical statistician) do it for you. Very few users of applied statistics actually do that. Recalling our slogan that asymptotics only produce heuristics and that if you are worried you simulate, one can see why. Even if you managed to verify the conditions of the theorem it still wouldn't tell you how large n would have to be to use the theorem on real data. You would still be in the position of having to simulate if worried.

10.3.3 Asymptotic Relative Efficiency

The MLE usually has the best possible asymptotic variance of any estimator, but a precise statement of this result is tricky, requiring a new concept. We say an estimator $\hat{\theta}_n$, whether or not it is the MLE, is *asymptotically efficient* if it satisfies (10.27). If it does better than that, if its asymptotic variance is less than $I_1(\theta_0)^{-1}$, we say it is *superefficient.* Since the true parameter value θ_0 is unknown, we also insist that it be efficient (at least) at every θ.

Using this new terminology, proving the MLE to be the best possible estimator is the same thing as proving that superefficient estimators do not exist. The trouble with that is that they *do* exist, although they are quite crazy. Here is an example of a superefficient estimator.

Suppose X_1, X_2, are i. i. d. normal with known variance σ^2. In Example 10.2.2 we found that the MLE of the mean μ is $\overline{X}_n$. Consider the estimator

$$X_n^* = \begin{cases} \overline{X}_n, & |\overline{X}_n| > n^{-1/4} \\ 0, & \text{otherwise} \end{cases}$$

If the true μ is not exactly zero, then $\overline{X}_n \xrightarrow{P} \mu$ by the LLN, and $P(\overline{X}_n \le n^{-1/4})$ converges to zero. Thus by Slutsky's theorem X_n^* and $\overline{X}_n$ have the same asymptotics.

But if the true μ is exactly zero, then the CLT says

$$\overline{X}_n \approx \frac{n^{-1/2} Z}{\sigma}$$

where Z is standard normal. Thus $P(\overline{X}_n \le n^{-1/4})$ converges to one because $n^{-1/2}$ is much smaller than $n^{-1/4}$ for large n. Thus in this case $P(X_n^* = 0)$ converges in probability to one. Thus we have two cases, our estimator obeys the usual asymptotics at almost all points

$$\sqrt{n}\,(X_n^* - \mu) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$

but at just one point $\mu = 0$ it has superefficient asymptotics

$$\sqrt{n}\,(X_n^* - \mu) \xrightarrow{\mathcal{D}} 0.$$

Please don't complain about this example. It may seem to be a stupid theoretician trick, but that's the point. All superefficient estimators are stupid in this fashion.

Suppose we have an estimator θ_n^* of some parameter θ that is consistent and asymptotically normal, that is,

$$\sqrt{n}\,(\theta_n^* - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}\bigl(0, \tau^2(\theta)\bigr),$$

where we have written the asymptotic variance as $\tau^2(\theta)$ to indicate that, in general, it is a function of the true parameter value θ. If $\tau^2(\theta)$ is a continuous function of θ, the estimator cannot be superefficient. We must have

$$\tau^2(\theta) \ge I_1(\theta)^{-1}, \qquad \text{for all } \theta$$

Superefficiency can only occur discontinuously, as in the example. Thus superefficient estimators are a pathological phenomenon. They are no use in real applications, because you can never know whether such an estimator is superefficient at the true parameter value.

Ignoring superefficient estimators, the theorem says that it is not possible asymptotically to do better than the MLE. Of course you can do better for finite sample size n, but at least if n is moderately large you can't do much better. This fact creates a strong argument for using maximum likelihood estimators.

Example 10.3.6 (Cauchy Location Model).
In Example 10.3.5 we asymptotic distribution for the MLE in the Cauchy Location Model. The only other sensible estimator we know about is the sample median, whose asymptotic distribution was found in Problem 7-6(a). The asymptotic variance of the MLE is 2. The asymptotic variance of the sample median is $\pi^2/4$. The ARE is 0.81 or 1.23 depending on which way you write it. The MLE is more efficient (the MLE is always more efficient).

Example 10.3.7 (Normal Location Model).
In Example 10.2.2 we found that the sample mean is the MLE for the normal location model (normal, known variance). In Example 7.4.1 we found that the sample median was asymptotically less efficient than the sample mean for this model (asymptotic variance σ^2 for the mean and $\pi\sigma^2/2$ for the median). Now we find out why. The sample mean, being the MLE is better than *any* other estimator (barring weird superefficient estimators).

Example 10.3.8 (Laplace Location Model).
The Laplace (double exponential) location model has densities

$$f_\mu(x) = \frac{1}{2} e^{-|x-\mu|}$$

hence log likelihood

$$l_n(\mu) = -\sum_{i=1}^{n} |x_i - \mu| \tag{10.28}$$

Now this is a problem for which the tools we usually apply are no help. We can't take derivatives, because the absolute value function is nondifferentiable (having a kink at zero). However, we can use Corollary 7.7 (the characterization of the empirical median), which says, phrased in terms of the current context, that the maximizer of (10.28) is the sample median. Thus the sample median is the MLE. In Problem 7-6(b) we found the asymptotic variance of the sample median (now discovered also to be the MLE) to be one (in this parameterization). In problem 9-1 we found the variance of X to be two (in this parameterization). Hence the ARE of the mean to the median is either 1/2 or 2, depending on how you write it. Now we find out that it is no surprise the sample median is better. It's the MLE so it's better that any other estimator (not just better than the mean).

As an aside, note that we can't use Fisher information to calculate the asymptotic variance because it isn't defined, the log likelihood not being differentiable. The theorem that the MLE is more efficient, still holds though. So when we find out that the sample median is the MLE, we can use the theorem about the asymptotics of the sample median (Corollary 7.28) to calculate the asymptotic variance.

We summarize the results of the preceding three examples in a little table of asymptotic variances.

	MLE	median	mean
Cauchy	2	2.47	∞
Normal	1	1.57	1
Laplace	1	1	2

In the first line, all three estimators are different. The sample mean is useless, not even consistent (the LLN doesn't hold because the expectation of the Cauchy distribution doesn't exist). In the other two lines, the MLE is the same as one of the other estimators. In all three cases the MLE is best (by the theorem, the MLE is best in every case, not just these).

10.3.4 Estimating the Variance

One remaining problem with Theorem 10.3 is that the asymptotic variance $I_1(\theta_0)^{-1}$ is unknown because the true parameter value θ_0 is unknown. (If we knew θ_0, we wouldn't be bothered with estimating it!) But this is a minor problem. We just estimate the asymptotic variance using the "plug-in" principle by $I_1(\hat{\theta}_n)$. If I_1 is a continuous function of θ, then

$$I_1(\hat{\theta}_n) \xrightarrow{P} I_1(\theta_0), \qquad \text{as } n \to \infty$$

by the continuous mapping theorem. So we can use the left hand side as an approximation to the right hand side. Being a bit sloppy, we can write

$$\hat{\theta}_n \approx \mathcal{N}\left(\theta_0, \frac{1}{n I_1(\hat{\theta}_n)}\right).$$

Caution: There is a natural tendency, exhibited by many students, to get confused about whether one uses $I_1(\theta)$ or $I_n(\theta)$. They get either too many or too few n's or root n's in their standard errors. The error variance is

$$\frac{1}{n I_1(\hat{\theta}_n)} = \frac{1}{I_n(\hat{\theta}_n)}$$

In words, this can be called "inverse Fisher information" *if* (big if) one remembers which Fisher information is meant. It is the inverse of the Fisher information for the *actual problem at hand* (sample size n). Another way to remember which is the correct standard error is that the standard error must obey the "square root law," that is, it must decrease like $1/\sqrt{n}$. If one gets confused about the standard error, one gets ridiculous confidence intervals, too wide or too narrow by a factor of n or $\sqrt{n}$.

A second problem with the theorem is that the Fisher information $I_1(\theta)$ is defined by an expectation, which may be difficult or impossible to derive in closed form. In that case, we can use the observed Fisher information, substituting $J_n(\hat{\theta}_n)$ for $I_n(\hat{\theta}_n)$. These will typically be close to each other. Assumptions similar to those of Theorem 10.3 (ignoring the same sort of remainder term) imply

$$\frac{1}{n} J_n(\hat{\theta}_n) \xrightarrow{P} I_1(\theta_0), \qquad \text{as } n \to \infty.$$

Since $J_n(\theta)$ involves no expectations, only derivatives, it can be calculated whenever the likelihood itself can be calculated, and hence can almost always be used in calculating standard errors. Of course, one can use observed Fisher information even when the expected Fisher information can also be calculated. One can use whichever seems more convenient.

10.3.5 Tests and Confidence Intervals

These variance estimates are combined using the "plug-in" theorem to construct asymptotically pivotal quantities for tests and confidence intervals. If z_α is the $1 - \alpha$ quantile of the standard normal distribution, then

$$\hat{\theta}_n \pm z_{\alpha/2} \frac{1}{\sqrt{I_n(\hat{\theta}_n)}}$$

or

$$\hat{\theta}_n \pm z_{\alpha/2} \frac{1}{\sqrt{J_n(\hat{\theta}_n)}}$$

is an asymptotic $100(1 - \alpha)\%$ confidence interval for θ. We can use whichever is convenient. If we are doing a hypothesis test, then

$$T_n = \left(\hat{\theta}_n - \theta_0\right) \sqrt{I_n(\hat{\theta}_n)} \tag{10.29a}$$

or

$$T_n = \left(\hat{\theta}_n - \theta_0\right) \sqrt{J_n(\hat{\theta}_n)} \tag{10.29b}$$

is an asymptotically pivotal quantity (either is asymptotically standard normal) that can be used to construct a test. A two-tailed test of

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_A &: \theta \neq \theta_0 \end{aligned}$$

rejects H_0 when $|T_n| \geq z_{\alpha/2}$, a one-tailed test of

$$\begin{aligned} H_0 &: \theta \leq \theta_0 \\ H_A &: \theta > \theta_0 \end{aligned}$$

rejects H_0 when $T_n \geq z_\alpha$, and a one-tailed test of

$$\begin{aligned} H_0 &: \theta \geq \theta_0 \\ H_A &: \theta < \theta_0 \end{aligned}$$

rejects H_0 when $T_n \leq -z_\alpha$.

This should all seem very familiar. It is just like all other asymptotic confidence intervals and tests. The only novelty is using observed or expected Fisher information to calculate the asymptotic standard error.

Example 10.3.9 (A Problem in Genetics).
In his influential monograph *Statistical Methods for Research Workers*, first published in 1925, R. A. Fisher described the following problem in genetics. The data are counts of seedling plants classified according to their values of two traits, green or which leaf color and starchy or sugary carbohydrate content

	green	white
starchy	1997	904
sugary	906	32

The probability model for data such as this is the *multinomial distribution* (Section 5.4 in these notes). Data are assumed to be observations on an i. i. d. sample of individuals classified into k categories (here $k = 4$, the number of cells in the table). Because of the assumption the individuals are identically distributed, each has the same probability of falling in the i-th cell of the table, denote it p_i. Because probabilities sum to one, we must have

$$p_1 + \cdots + p_k = 1 \tag{10.30}$$

If the entries in the table (the category counts are denoted X_i), then the joint density of these random variables is given by equation (3) on p. 187 in Lindgren

$$f_{\mathbf{p}}(\mathbf{x}) = \binom{n}{x_1, \ldots, x_k} \prod_{i=1}^{k} p_i^{x_i} \tag{10.31}$$

As the boldface type in the notation on the left hand side indicates, this describes the distribution of the random vector $\mathbf{X} = (X_1, \ldots, X_k)$ which depends on the vector parameter $\mathbf{p} = (p_1, \ldots, p_k)$. This is the multivariate analog of the binomial distribution.

The components X_i of this random vector are dependent. In fact, if n is the sample size, they must add to n because each individual falls in some cell of the table

$$X_1 + \cdots + X_k = n \tag{10.32}$$

Thus there are "really" only $n - 1$ random variables, because one can be eliminated using (10.32), and only $n - 1$ parameters, because one can be eliminated using (10.30). But doing this elimination of variables and parameters spoils the symmetry of (10.31). It does not simplify the formulas but complicates them. The log likelihood for the multinomial model is

$$l_n(\mathbf{p}) = \sum_{i=1}^{k} x_i \log(p_i)$$

(we can drop the multinomial coefficient because it does not contain parameters).

Returning to the genetics, Fisher was actually interested in a one-parameter submodel of the multinomial model, having cell probabilities specified by a single parameter θ, given by

	green	white
starchy	$\frac{1}{4}(2 + \theta)$	$\frac{1}{4}(1 - \theta)$
sugary	$\frac{1}{4}(1 - \theta)$	$\frac{1}{4}\theta$

In order for this to make sense as a submodel of the multinomial model, the cell probabilities must add to one (10.30), which is easily checked. They must also all be greater than zero, which requires $0 \le \theta \le 1$. The parameter θ has a scientific interpretation. Under a specific genetic model $\sqrt{\theta}$ is the *recombination fraction*, which is a measure of the distance along the chromosome between the two genes controlling these two traits (assuming they are on the same chromosome, if not then $\theta = 1/4$).

Numbering the cells of the table from one to four, going across rows starting at the upper left corner, the log likelihood becomes

$$l_n(\theta) = x_1 \log(2+\theta) + (x_2 + x_3)\log(1-\theta) + x_4 \log(\theta) \tag{10.33}$$

(where we dropped some more terms involving $\log(\frac{1}{4})$ but not containing the parameter). And the score is

$$l_n'(\theta) = \frac{x_1}{2+\theta} - \frac{x_2+x_3}{1-\theta} + \frac{x_4}{\theta} \tag{10.34}$$

In order to find the maximum likelihood estimate, we need to solve the equation $l_n'(\theta) = 0$. Multiplying through by the product of the denominators gives

$$x_1(1-\theta)\theta - (x_2+x_3)(2+\theta)\theta + x_4(2+\theta)(1-\theta) = 0$$

or simplifying a bit

$$2x_4 + (x_1 - 2x_2 - 2x_3 - x_4)\theta - (x_1 + x_2 + x_3 + x_4)\theta^2 = 0$$

or using (10.32)

$$n\theta^2 - (x_1 - 2x_2 - 2x_3 - x_4)\theta - 2x_4 = 0.$$

This is a quadratic equation with solutions

$$\hat{\theta}_n = \frac{x_1 - 2x_2 - 2x_3 - x_4 \pm \sqrt{(x_1 - 2x_2 - 2x_3 - x_4)^2 + 8nx_4}}{2n}$$

Since the square root is larger than the first term of the numerator, choosing the minus sign always gives a negative solution, which is not in the set of allowed parameter values. Hence the only "solution of the likelihood equation" is

$$\hat{\theta}_n = \frac{x_1 - 2x_2 - 2x_3 - x_4 + \sqrt{(x_1 - 2x_2 - 2x_3 - x_4)^2 + 8nx_4}}{2n} \tag{10.35}$$

In order to check whether this is a local or global maximum we need to look at the second derivative of the log likelihood, which also be needed to calculate Fisher information,

$$l_n''(\theta) = -\frac{x_1}{(2+\theta)^2} - \frac{x_2+x_3}{(1-\theta)^2} - \frac{x_4}{\theta^2} \tag{10.36}$$

Since this is negative for all θ in the interior of the parameter space ($0 < \theta < 1$), the log likelihood is strictly concave and (10.35) is the unique global maximum of the log likelihood.

Plugging in the actual data from the table

$$x_1 - 2x_2 - 2x_3 - x_4 = 1997 - 2(904 + 906) - 32 = -1655$$

and

$$n = 1997 + 904 + 906 + 32 = 3839$$

$$\begin{aligned}
\hat{\theta}_n &= \frac{-1655 + \sqrt{(-1655)^2 + 8 \cdot 3839 \cdot 32}}{2 \cdot 3839} \\
&= \frac{-1655 + \sqrt{1655^2 + 982784}}{7678} \\
&= 0.0357123
\end{aligned}$$

To make a confidence interval we need the Fisher information, either observed or expected (we'll do both to show how it's done, but in a real application you would choose one or the other). Finding the variance of the score (10.34) is a bit tricky because the x_i are correlated. Thus in this example the calculation of expected Fisher information using the second derivative is a good deal easier than the calculation using the first derivative. The observed Fisher information is just minus (10.36)

$$J_n(\theta) = \frac{x_1}{(2+\theta)^2} + \frac{x_2 + x_3}{(1-\theta)^2} + \frac{x_4}{\theta^2}$$

and the expected Fisher information is its expectation. Since the marginal distribution of X_i is $\text{Bin}(n, p_i)$ (Section 5.4.5 of these notes) $E(X_i) = np_i$ and

$$\begin{aligned}
I_n(\theta) &= \frac{np_1}{(2+\theta)^2} + \frac{n(p_2 + p_3)}{(1-\theta)^2} + \frac{np_4}{\theta^2} \\
&= n \left(\frac{\frac{1}{4}(2+\theta)}{(2+\theta)^2} + \frac{\frac{1}{2}(1-\theta)}{(1-\theta)^2} + \frac{\frac{1}{4}\theta}{\theta^2} \right) \\
&= \frac{n}{4} \left(\frac{1}{2+\theta} + \frac{2}{1-\theta} + \frac{1}{\theta} \right)
\end{aligned}$$

Plugging the data into these formulas gives

$$\begin{aligned}
J_n(\hat{\theta}_n) &= \frac{1997}{(2 + 0.0357123)^2} + \frac{904 + 906}{(1 - 0.0357123)^2} + \frac{32}{0.0357123^2} \\
&= 27519.2
\end{aligned}$$

and

$$\begin{aligned}
I_n(\hat{\theta}_n) &= \frac{3839}{4} \left(\frac{1}{2 + 0.0357123} + \frac{2}{1 - 0.0357123} + \frac{1}{0.0357123} \right) \\
&= 29336.5
\end{aligned}$$

We may use either to construct confidence intervals. If we use the observed Fisher information, we get

$$\hat{\theta}_n \pm \frac{1.96}{\sqrt{J_n(\hat{\theta}_n)}}$$

as a 95% confidence interval for the unknown true θ. The "plus or minus" is $1.96/\sqrt{27519.2} = 0.011815$, so our 95% confidence interval is 0.036 ± 0.012. If we use expected Fisher information instead the "plus or minus" would be 0.011443, almost the same.

To illustrate a hypothesis test, the natural null hypothesis to test is that the genes are *unlinked* (not on the same chromosome), in which case $\theta = 1/4$. Thus we test

$$\begin{aligned} H_0 &: \theta = 1/4 \\ H_A &: \theta \neq 1/4 \end{aligned}$$

Under certain genetic models a one-tailed test would be appropriate, but Fisher doesn't give us enough information about the data to know which tail to test, we will do a two-tailed test. The test statistic is either (10.29a) or (10.29b) with $\theta_0 = 1/4$, depending on whether we want to use observed or expected Fisher information. These give

$$\begin{aligned} \left(\hat{\theta}_n - \theta_0\right) \sqrt{I_n(\hat{\theta}_n)} &= (0.0357123 - 0.25)\sqrt{27519.2} = -35.548 \\ \left(\hat{\theta}_n - \theta_0\right) \sqrt{J_n(\hat{\theta}_n)} &= (0.0357123 - 0.25)\sqrt{29336.5} = -36.703 \end{aligned}$$

In either case the test statistics are so large that we get zero for the P-value (not exactly zero, R gives 4×10^{-277} for one and 3×10^{-295} for the other, but the normal approximation has no validity whatsoever this far out in the tail, so $P \approx 0$ is a more sensible answer). What this says, is that there is clear evidence of linkage (genes on the same chromosome). Here the evidence is so strong that there's almost no doubt remaining.

One moral of the story is that you can use either observed or expected Fisher information, whichever you prefer, whichever seems easier. They won't always give *exactly* the same confidence interval or P-value. But they both give valid asymptotic approximations (neither is *exactly* right but both are *approximately* right for large n and neither is preferred over the other).

Another moral of the story is a lesson about what hypothesis tests say. The test above says there is no question that $\theta \neq 1/4$, and hence the genes are linked. It does not say anything else. It does not tell you anything about the value of θ other than that it is not the value hypothesized under the null hypothesis. If you want to know more, you must look at a confidence interval.

10.4 Multiparameter Models

All of the preceding theory goes through to the multiparameter case. It just becomes more complicated. In particular, the MLE $\hat{\boldsymbol{\theta}}_n$ is a vector (obviously we need a vector estimator of a vector parameter). Hence Fisher information, to describe the variance of a vector, must become a matrix. Basically, that's the whole story. We just need to fill in the details.

10.4.1 Maximum Likelihood

Example 10.4.1 (Two-Parameter Normal Model).
The log likelihood for the two-parameter normal model is found by taking logs in (10.4) giving

$$l_n(\mu, \sigma) = -n \log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \tag{10.37}$$

Actually, the variance σ^2 is a more sensible parameter to estimate than the standard deviation σ. So define $\varphi = \sigma^2$, which means $\sigma = \varphi^{1/2}$. Plugging this into (10.37) gives

$$l_n(\mu, \varphi) = -\frac{n}{2} \log(\varphi) - \frac{1}{2\varphi} \sum_{i=1}^n (x_i - \mu)^2 \tag{10.38}$$

Differentiating gives

$$\begin{aligned} \frac{\partial l_n(\mu, \varphi)}{\partial \mu} &= \frac{1}{\varphi} \sum_{i=1}^n (x_i - \mu) \\ &= \frac{n(\bar{x}_n - \mu)}{\varphi} \qquad (10.39a) \\ \frac{\partial l_n(\mu, \varphi)}{\partial \varphi} &= -\frac{n}{2\varphi} + \frac{1}{2\varphi^2} \sum_{i=1}^n (x_i - \mu)^2 \qquad (10.39b) \end{aligned}$$

We have two partial derivatives. They are the two components of the score vector. We find a point where the first derivative *vector* is zero by setting them both to zero and solving the simultaneous equations. In general, this is hard. Here it is actually easy, because it is clear that (10.39a) is zero when and only when $\mu = \bar{x}_n$. So that is the MLE of μ (just as we found when we assumed the variance was known). Plugging that solution into (10.39b) gives

$$-\frac{n}{2\varphi} + \frac{1}{2\varphi^2} \sum_{i=1}^n (x_i - \bar{x}_n)^2 = -\frac{n}{2\varphi} + \frac{n v_n}{2\varphi^2}$$

where v_n is the usual variance of the empirical distribution (note not s_n^2). And this is clearly zero when $\varphi = v_n$. So that is the MLE of the variance φ.

Since we have two parameters, it is tempting to say "the MLE's are ...," but we can also think of *the* parameter as a vector $\boldsymbol{\theta} = (\mu, \varphi)$ and the MLE of this vector parameter is

$$\hat{\boldsymbol{\theta}}_n = \begin{pmatrix} \hat{\mu}_n \\ \hat{\varphi}_n \end{pmatrix} = \begin{pmatrix} \bar{x}_n \\ v_n \end{pmatrix} \tag{10.40}$$

Actually, we are being a bit overconfident here. What we have found is a zero of the first derivative, in fact, the only zero. But we haven't shown yet that this is even a local maximum, much less a global maximum.

So we look at the second derivative matrix. This has components

$$\frac{\partial^2 l_n(\mu, \varphi)}{\partial \mu^2} = -\frac{n}{\varphi} \tag{10.41a}$$

$$\frac{\partial^2 l_n(\mu, \varphi)}{\partial \mu \partial \varphi} = -\frac{n(\bar{x}_n - \mu)}{\varphi^2} \tag{10.41b}$$

$$\frac{\partial^2 l_n(\mu, \varphi)}{\partial \varphi^2} = \frac{n}{2\varphi^2} - \frac{1}{\varphi^3} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{10.41c}$$

or, more precisely, the second derivative is a 2×2 matrix having (10.41a) and (10.41c) as diagonal elements and off-diagonal elements (10.41b) (both the same because a second derivative matrix is symmetric). Unfortunately, this is not a negative definite matrix for all values of parameters and data, because (10.41c) is not always negative. Thus the log likelihood is not strictly concave, and the theory developed in Appendix G does not guarantee (10.40) is a global maximum.[2]

If we evaluate the second derivative matrix at the MLE, we get considerable simplification. When we plug in $\mu = \bar{x}_n$ (10.41b) is zero. Thus we get a diagonal matrix

$$\nabla^2 l_n(\hat{\boldsymbol{\theta}}_n) = \begin{pmatrix} -\frac{n}{\hat{\varphi}_n} & 0 \\ 0 & -\frac{n}{2\hat{\varphi}_n^2} \end{pmatrix} \tag{10.42}$$

the 1,1 component being (10.41a) with the MLE plugged in for the parameter, and the 2,2 component being (10.41c) simplified by using the fact that the sum in (10.41c) is $n v_n = n\hat{\varphi}_n$ when the MLE is plugged in for the parameter. Now (10.42) is negative definite (a diagonal matrix is negative definite if and only if each element on the diagonal is negative). So the theory we know does establish that (10.40) is a local maximum of the log likelihood.

Problems that work out so simply are quite rare. Usually there are no obvious solutions of the "likelihood equations" (first partial derivatives set equal to zero). In most problems the only way to find MLE's is ask a competent computer.

Example 10.4.2 (Cauchy Location-Scale Model).
The Cauchy Location-Scale model has densities

$$f_{\mu,\sigma}(x) = \frac{1}{\pi} \cdot \frac{\sigma}{\sigma^2 + (x - \mu)^2}$$

(p. 191 in Lindgren). Here μ is the location parameter and σ is the scale parameter (of course we could use any other Greek letters for these parameters). We know that the Cauchy distribution is symmetric and hence μ is the population median. We used the sample median as an estimator of μ in Problem 7-6(a).

[2] It actually is a global maximum, but we won't develop the theory needed to show that.

The log likelihood for an i. i. d. sample of size n is

$$l_n(\mu, \sigma) = n \log(\sigma) - \sum_{i=1}^{n} \log\left(\sigma^2 + (x_i - \mu)^2\right)$$

There's no point in differentiating. If we had to use the computer for the one-parameter Cauchy model, the two-parameter model certainly isn't going to be simple enough to do by hand. We proceed just as in the one-parameter problem (Example 10.2.3).

Define an R function that evaluates minus the log likelihood (minus because the `nlm` function minimizes rather than maximizes)

```
> l <- function(theta) {
+     mu <- theta[1]
+     sigma <- theta[2]
+     return(- n * log(sigma) + sum(log(sigma^2 + (x - mu)^2)))
+ }
```

Then we hand this to the `nlm` function as the function to be minimized. But first we have to figure out what to use as a starting point. The better starting point we have, the better chance that we find the right local minimum if more than one exists. We know a good estimator of μ, the sample median. What might be a good estimator of scale? Variance is no good. The Cauchy distribution doesn't even have a mean, much less a variance. The only other general estimator of scale that has even been mentioned is IQR (p. 202 in Lindgren) and we've never used it in any examples, nor do we know anything about its properties. However, it's the only idea we have, so let's try it. The reason that IQR works as a scale estimator is that the IQR of a general Cauchy distribution is just σ times the IQR of a standard Cauchy distribution (obvious from a picture of the density). The IQR of the standard Cauchy happens to be simple

```
> qcauchy(0.75) - qcauchy(0.25)
[1] 2
```

Thus half the IQR of the data is a sensible estimate of scale, and a good starting point for the optimization algorithm.

In real life, we would use real data. Here we just make up the data.

```
> n <- 80
> mu <- 0
> sigma <- 1
> x <- mu + sigma * rcauchy(n)  # make up data
> summary(x)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-38.2900  -0.6647   0.1869   0.6895   1.0630  54.1200
```

and then hand the data to the optimizer

```
out <- nlm(l, c(median(x), IQR(x) / 2), fscale=n, hessian=TRUE)
```

The result returned by the optimizer, saved in the variable `out`, which is a list with several components, only the interesting ones of which we show below

```
> out
$minimum
[1] 123.7644

$estimate
[1] 0.2162629 0.9229725

$gradient
[1] -6.470202e-05  5.941558e-05

$hessian
          [,1]      [,2]
[1,] 50.320659  2.273821
[2,]  2.273821 43.575535
```

- `estimate` is the optimum parameter value, that is

$$\hat{\boldsymbol{\theta}}_n = \begin{pmatrix} \hat{\mu}_n \\ \hat{\sigma}_n \end{pmatrix} = \begin{pmatrix} 0.2162629 \\ 0.9229725 \end{pmatrix}$$

- `minimum` is the optimal value of the objective function, that is

$$l_n(\hat{\boldsymbol{\theta}}_n) = -123.7644$$

  (recall that the objective function handed to `nlm` is *minus* the log likelihood).

- `gradient` is $-\nabla l_n(\hat{\boldsymbol{\theta}}_n)$ the first derivative vector of the objective function at the MLE. It should be zero to convergence tolerance of the optimization algorithm (and it is).

- `hessian` is $-\nabla^2 l_n(\hat{\boldsymbol{\theta}}_n)$, the second derivative matrix of the objective function at the MLE.

If we want to check that this is a local maximum of the log likelihood (hence a local *minimum* of the objective function passed to `nlm`) we check whether the `hessian` component is positive definite

```
> eigen(out$hessian)
$values
[1] 51.01558 42.88061

$vectors
           [,1]       [,2]
[1,] -0.9563346  0.2922741
[2,] -0.2922741 -0.9563346
```

Since both eigenvalues are positive, the Hessian is positive definite and the solution is a local maximum of the log likelihood.

Whether done by hand or done by computer, the process is much the same. Find a zero of the first derivative, and the second derivative tells you whether it is a local maximum or not (since the Hessian returned by the computer is the Hessian at the reported optimal value, it is of no use in checking whether you have a global maximum). As we will see in the next section, the Hessian is also observed Fisher information, and hence tells us important things about the asymptotic sampling distribution of the MLE.

10.4.2 Sampling Theory

We now have to repeat all of Section 10.3 giving the multiparameter analogs of everything in there. We will omit details that are basically the same as in the uniparameter case and concentrate on the differences.

In the multiparameter case the score is a vector $\nabla l_n(\boldsymbol{\theta})$. The *Fisher information* is, as in the uniparameter case, its variance

$$\mathbf{I}_n(\boldsymbol{\theta}) = \operatorname{var}_{\boldsymbol{\theta}}\{\nabla l_n(\boldsymbol{\theta})\}, \tag{10.43}$$

but now, since the variance of a random vector is a (nonrandom) matrix, the Fisher information is a matrix. Like any variance matrix, it is symmetric square and positive semi-definite. If $\nabla l_n(\boldsymbol{\theta})$ is not concentrated on a hyperplane (see Section 5.1.9 in these notes), then the Fisher information is actually positive definite.

As in the uniparameter case we sometimes call $\mathbf{I}_n(\boldsymbol{\theta})$ the "expected" Fisher information for contrast with "observed" Fisher information, but, strictly speaking the "expected" is redundant. "Fisher information" with no qualifying adjective always means $\mathbf{I}_n(\boldsymbol{\theta})$.

We still have the multiparameter analogs of the two important theorems about the score and Fisher information (Theorems 10.1 and 10.2).

Theorem 10.4. *Provided first and second order partial derivatives of* (10.10) *with respect to components of* θ *can be taken under the integral sign,*

$$E_{\boldsymbol{\theta}}\{\nabla l_n(\boldsymbol{\theta})\} = 0 \tag{10.44a}$$

and

$$E_{\boldsymbol{\theta}}\{\nabla^2 l_n(\boldsymbol{\theta})\} = -\operatorname{var}_{\boldsymbol{\theta}}\{\nabla l_n(\boldsymbol{\theta})\} \tag{10.44b}$$

for all values of $\boldsymbol{\theta}$ *for which the differentiation under the integral sign is permitted.*

Theorem 10.5.

$$\mathbf{I}_n(\boldsymbol{\theta}) = n\mathbf{I}_1(\boldsymbol{\theta})$$

The proofs of these two theorems are exactly the same as in the uniparameter case. One only has to use partial derivatives instead of ordinary derivatives to prove Theorem 10.4. Theorem 10.5 follows just as in the uniparameter case from the fact that the variance of a sum is the sum of the variance when the terms are independent, the multivariate version of which is (5.10) in Chapter 5 of these notes.

As in the uniparameter case, (10.44b) tells us we can calculate Fisher information in two quite different ways, either (10.43) or

$$\mathbf{I}_n(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}\{\nabla^2 l_n(\boldsymbol{\theta})\}. \tag{10.45}$$

It's your choice. Use whichever seems easier.

Also as in the uniparameter case, Theorem 10.5 tells us we can calculate Fisher information using sample size one (which won't have any summations) and multiply by n.

Example 10.4.3 (Two-Parameter Normal Model).
The log likelihood and derivatives for this model were figured out in Example 10.4.1. The components of the (expected) Fisher information are the negative expectations of (10.41a), (10.41b), and (10.41c), that is,

$$\begin{aligned}
-E\left(\frac{\partial^2 l_n(\mu,\varphi)}{\partial\mu^2}\right) &= \frac{n}{\varphi} \\
-E\left(\frac{\partial^2 l_n(\mu,\varphi)}{\partial\mu\partial\varphi}\right) &= \frac{n[E(\overline{X}_n)-\mu]}{\varphi^2} \\
&= 0 \\
-E\left(\frac{\partial^2 l_n(\mu,\varphi)}{\partial\varphi^2}\right) &= -\frac{n}{2\varphi^2} + \frac{1}{\varphi^3}\sum_{i=1}^n E\{(x_i-\mu)^2\} \\
&= -\frac{n}{2\varphi^2} + \frac{n\varphi}{\varphi^3} \\
&= \frac{n}{2\varphi^2}
\end{aligned}$$

Thus the Fisher information is the diagonal matrix

$$\mathbf{I}_n(\boldsymbol{\theta}) = \begin{pmatrix} \frac{n}{\varphi} & 0 \\ 0 & \frac{n}{2\varphi^2} \end{pmatrix} \tag{10.46}$$

Observed Fisher information is also defined in the same way as in the uniparameter case, as minus the second derivative of the log likelihood

$$\mathbf{J}_n(\boldsymbol{\theta}) = -\nabla^2 l_n(\boldsymbol{\theta}). \tag{10.47}$$

Note that the second derivative is a matrix here, which is a good thing, because the expected Fisher information is also a matrix.

The CLT and the LLN apply in exactly the same way to the score and observed Fisher information. The analog of (10.18) is

$$\frac{1}{\sqrt{n}} \nabla l_n(\boldsymbol{\theta}) \xrightarrow{\mathcal{D}} \mathcal{N}\bigl(0, \mathbf{I}_1(\boldsymbol{\theta})\bigr) \tag{10.48}$$

(just the same, except for some boldface type). Of course, this is a multivariate CLT because $\nabla l_n(\boldsymbol{\theta})$ is a random vector rather than a random scalar. The sloppy "double squiggle" version is

$$\nabla l_n(\boldsymbol{\theta}) \approx \mathcal{N}\bigl(0, \mathbf{I}_n(\boldsymbol{\theta})\bigr).$$

The analog of (10.22) is

$$\frac{1}{n} \mathbf{J}_n(\boldsymbol{\theta}) \xrightarrow{P} \mathbf{I}_1(\boldsymbol{\theta}). \tag{10.49}$$

(again, just the same, except for some boldface type). It is, of course, a multivariate convergence in probability statement. The sloppy "double squiggle" version would be

$$\mathbf{J}_n(\boldsymbol{\theta}) \approx \mathbf{I}_n(\boldsymbol{\theta})$$

Still proceeding as in the univariate case, expanding ∇l_n using a Taylor series with remainder about the true parameter value $\boldsymbol{\theta}_0$, we get

$$\nabla l_n(\boldsymbol{\theta}) = \nabla l_n(\boldsymbol{\theta}_0) + \nabla^2 l_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \text{remainder}$$

This is a bit harder to interpret than the uniparameter analog. First, it is a vector equation, each term having k components if there are k parameters. As in the uniparameter case, we can use $\nabla^2 l_n(\boldsymbol{\theta}) = -\mathbf{J}_n(\boldsymbol{\theta})$ to rewrite this as

$$\nabla l_n(\boldsymbol{\theta}) = \nabla l_n(\boldsymbol{\theta}_0) - \mathbf{J}_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \text{remainder}, \tag{10.50}$$

but that still doesn't make it obvious what k-dimensional vector the middle term on the right hand side is supposed to be. Since $\mathbf{J}_n(\boldsymbol{\theta}_0)$ is a $k \times k$ matrix and $\boldsymbol{\theta} - \boldsymbol{\theta}_0$ is a k vector, this must be a matrix multiplication, which does indeed produce a k vector. We won't bother to write out the "remainder" term in detail.

Now we apply the same sort of argument we used in the uniparameter case. If the MLE is in the interior of the parameter space, the first derivative of the log likelihood will be zero at the MLE. So if we plug in the MLE for $\boldsymbol{\theta}$, the left hand side of (10.50) is zero, and we get

$$\nabla l_n(\boldsymbol{\theta}_0) \approx \mathbf{J}_n(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \tag{10.51}$$

(we dropped the remainder term, assuming as in the uniparameter case that it is asymptotically negligible, and replaced the equals sign with a "double squiggle" to indicate this is not an exact inequality). Again we divide through by $\sqrt{n}$ to make both sides the right size to converge in distribution to a nontrivial random variable

$$\frac{1}{\sqrt{n}} \nabla l_n(\boldsymbol{\theta}_0) \approx \frac{1}{n} \mathbf{J}_n(\boldsymbol{\theta}_0) \cdot \sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \tag{10.52}$$

This is almost, but not quite, what we need. We need $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$ by itself on one side. The way to do that is multiply through by the inverse of the matrix multiplying it.

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \approx \left(\frac{1}{n} \mathbf{J}_n(\boldsymbol{\theta}_0) \right)^{-1} \frac{1}{\sqrt{n}} \nabla l_n(\boldsymbol{\theta}_0) \tag{10.53}$$

Because matrix inversion is a continuous operation, the continuous mapping theorem and (10.49) imply that the first factor on the right hand side converges in probability to $\mathbf{I}_1(\boldsymbol{\theta}_0)^{-1}$. The convergence in distribution for second factor on the right hand side is given by (10.48). By Slutsky's theorem, the right hand side converges to the product, that is

$$\left(\frac{1}{n} \mathbf{J}_n(\boldsymbol{\theta}_0) \right)^{-1} \frac{1}{\sqrt{n}} \nabla l_n(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1} \mathbf{Y}$$

where

$$\mathbf{Y} \sim \mathcal{N}(0, \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1}).$$

Since a linear transformation of a multivariate normal is multivariate normal (Theorem 12 of Chapter 12 in Lindgren), $\mathbf{I}_1(\boldsymbol{\theta}_0)^{-1}\mathbf{Y}$ is multivariate normal with mean vector

$$E\left\{ \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1} \mathbf{Y} \right\} = \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1} E(\mathbf{Y}) = 0$$

and variance matrix

$$\begin{aligned} \operatorname{var}\left\{ \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1} \mathbf{Y} \right\} &= \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1} \operatorname{var}(\mathbf{Y}) \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1} \\ &= \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1} \mathbf{I}_1(\boldsymbol{\theta}_0) \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1} \\ &= \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1} \end{aligned}$$

the middle equality being the formula for the variance of a (multivariate) linear transformation, (5.18b) in Chapter 5 of these notes.

Thus we have arrived at the multiparameter version of the "usual asymptotics" of maximum likelihood.

Theorem 10.6. *If the true parameter value $\boldsymbol{\theta}_0$ is in the interior of the parameter space, first and second order partial derivatives of* (10.10) *with respect to components of θ can be taken under the integral sign, and the difference of the two sides of* (10.52) *converges in probability to zero, then*

$$\sqrt{n}\big(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \mathbf{I}_1(\boldsymbol{\theta}_0)^{-1}\big) \tag{10.54}$$

This looks just like the uniparameter version (10.27), except for some boldface type.

As in the uniparameter case, the conditions of the theorem are hard to verify. We often use it without any attempt to verify the conditions. As we remarked in regard to the uniparameter case, even if you have verified the conditions, that still doesn't prove that the normal approximation given by the theorem is good at the actual n in which you are interested (the sample size of the data in some real application). Hence our slogan about asymptotics only providing a heuristic applies. If you are worried about the validity of the asymptotics, check it with computer simulations.

One final point, just like in the uniparameter case, Theorem 10.6 must be combined with the plug-in theorem to get a useful result. Since we don't know the true parameter value θ_0, we don't know the asymptotic variance $\mathbf{I}_1(\boldsymbol{\theta}_0)^{-1}$ and must estimate it. Plugging either observed or expected Fisher information evaluated at the MLE gives

$$\hat{\boldsymbol{\theta}}_n \approx \mathcal{N}\big(\boldsymbol{\theta}_0, \mathbf{I}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\big)$$

or

$$\hat{\boldsymbol{\theta}}_n \approx \mathcal{N}\big(\boldsymbol{\theta}_0, \mathbf{J}_n(\hat{\boldsymbol{\theta}}_n)^{-1}\big).$$

Example 10.4.4 (Two-Parameter Normal Model).
The MLE for this model is given by (10.40) in Example 10.4.1. The observed Fisher information evaluated at $\hat{\boldsymbol{\theta}}_n$ is given by minus (10.42) in the same example. The expected Fisher information is given by (10.46) in Example 10.4.3. When evaluated at the MLE, observed and expected Fisher information are the same

$$\mathbf{I}_n(\hat{\boldsymbol{\theta}}_n) = \mathbf{J}_n(\hat{\boldsymbol{\theta}}_n) = \begin{pmatrix} \frac{n}{\hat{\varphi}_n} & 0 \\ 0 & \frac{n}{2\hat{\varphi}_n^2} \end{pmatrix}$$

Inverting a diagonal matrix is easy. Just invert each term on the diagonal.

$$\mathbf{I}_n(\hat{\boldsymbol{\theta}}_n)^{-1} = \mathbf{J}_n(\hat{\boldsymbol{\theta}}_n)^{-1} = \begin{pmatrix} \frac{\hat{\varphi}_n}{n} & 0 \\ 0 & \frac{2\hat{\varphi}_n^2}{n} \end{pmatrix}$$

That's our estimate of the asymptotic variance of the MLE (vector).

This example doesn't tell us anything we didn't already know. It says the MLE's of μ and of $\varphi = \sigma^2$ are asymptotically independent, because the covariance is zero and uncorrelated jointly multivariate normal random variables are independent (Theorem 4 of Chapter 12 in Lindgren). This is no surprise, because we know that the MLE's $\overline{X}_n$ and V_n are actually independent (not just asymptotically) by the corollary to Theorem 10 of Chapter 7 in Lindgren. Since they are independent, their joint distribution is uninteresting (just the product of the marginals), and what this says about the marginals we also have long known

$$\hat{\mu}_n = \overline{X}_n \approx \mathcal{N}\left(\mu, \frac{V_n}{n}\right)$$

(which is just the CLT plus the plug-in theorem) and

$$\hat{\varphi}_n = V_n \approx \mathcal{N}\left(\sigma^2, \frac{2V_n^2}{n}\right)$$

This may not ring a bell right away. It is the asymptotic distribution of V_n worked out in Example 7.3.6 in these notes plus the plug-in theorem.

Example 10.4.5 (Cauchy Location-Scale Model).
The MLE and observed Fisher information for this model were found in Example 10.4.5. Of course they were not found by hand calculation, just by numerical optimization using the computer. To repeat the relevant bits of the computer output,

```
$estimate
[1] 0.2162629 0.9229725
```

is the MLE $\hat{\boldsymbol{\theta}}_n$, and

```
$hessian
          [,1]      [,2]
[1,] 50.320659  2.273821
[2,]  2.273821 43.575535
```

is the observed Fisher information evaluated at the MLE $\mathbf{J}_n(\hat{\boldsymbol{\theta}}_n)$.

Now to make confidence intervals for the parameters we need to calculate *inverse* Fisher information, because that is the asymptotic variance of the MLE. The R function that inverts matrices has the totally unintuitive name `solve` (because it also solves linear equations). Hence the inverse Fisher information $\mathbf{J}_n(\hat{\boldsymbol{\theta}}_n)^{-1}$ is given by

```
> solve(out$hessian)
             [,1]         [,2]
[1,]  0.019919520 -0.001039426
[2,] -0.001039426  0.023002887
```

The numbers on the diagonal are the asymptotic variances of the first component of the MLE ($\hat{\mu}_n$) and of the second component ($\hat{\sigma}_n$). So

```
> avar <- solve(out$hessian)
> out$estimate[1] + c(-1,1) * qnorm(0.975) * sqrt(avar[1,1])
[1] -0.0603596  0.4928854
> out$estimate[2] + c(-1,1) * qnorm(0.975) * sqrt(avar[2,2])
[1] 0.6257106 1.2202344
```

give asymptotic 95% confidence intervals for μ and σ, respectively. Note these are not *simultaneous* confidence intervals, because we did not do Bonferroni correction.

There are two important points illustrated by the last example.

- If you can write down the log likelihood for a model, you can do likelihood inference, without doing any derivatives or expectations. (In this example, the computer found the MLE and calculated the second derivative matrix by finite differences. We didn't do any differentiation. And because we used observed rather than expected Fisher information, we didn't need to do any integrals either.)

- In order to calculate asymptotic variances, you do need to invert the Fisher information matrix, but the computer easily does that too.

Honesty compels me to admit that this example is not as convincing as it might be, because Mathematica easily calculates the expected Fisher information and it is diagonal and so trivially invertible. In fact, all location-scale models with symmetric densities have diagonal Fisher information. You will just have to take my word for it that there are many interesting complicated models that arise in applied statistics (too complicated to discuss here) for which you can't do anything but write down the log likelihood and hand it to the computer to do the rest, just like in this example.

10.4.3 Likelihood Ratio Tests

One last subject before we are done with sampling theory: In this section we learn about a completely different kind of hypothesis test. All of the tests we studied in the last chapter (and in this chapter up to here) had null hypotheses that fixed the value of *one* parameter (the so-called *parameter of interest*) and said nothing about any other parameters (the so-called *nuisance parameters*). Now we are going to learn about tests with multiple parameters of interest.

A likelihood ratio test compares two models, which we will call the *little model* and the *big model*. The little model is a *submodel* of the big model. Another term commonly used to describe this situation is *nested* models (one is a submodel of the other).

Roughly speaking, the little and big models correspond to the null and alternative hypotheses for the test. To be precise, let Θ be the whole parameter space of the problem, which is the parameter space of the big model, and let Θ_0 be the parameter space of the little model. These are nested models if $\Theta_0 \subset \Theta$. The null hypothesis corresponds to the little model,

$$H_0 : \theta \in \Theta_0, \tag{10.55}$$

but the alternative hypothesis is not supposed to include the null and hence must correspond to the part of the big model that is not in the little model

$$H_0 : \theta \in \Theta_A, \tag{10.56}$$

where $\Theta_A = \Theta \setminus \Theta_0$ (the operation indicated here is called "set difference" and says that Θ_A consists of points in Θ that are not in Θ_0, that is, are in the big model but not in the little model).

This is almost enough in the way of preliminary discussion to describe the likelihood ratio test, but not quite. What we need is for the little model to be a *smooth* submodel of the big model. The simplest way to describe that is as follows. Suppose that the big model has m parameters, which we can also think of as a single vector parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, so $\boldsymbol{\theta}$ is a point in m-dimensional Euclidean space $\mathbb{R}^m$, and Θ is a subset of $\mathbb{R}^m$. We need the little model to be a smooth k-dimensional surface in Θ. The simplest way to describe such a surface is by a differentiable map $\mathbf{g}$ from a subset Φ of k-dimensional Euclidean space $\mathbb{R}^k$ into Θ. We let Θ_0 be the *range* of the function $\mathbf{g}$, that is,

$$\Theta_0 = g(\Phi) = \{\, \mathbf{g}(\boldsymbol{\varphi}) : \boldsymbol{\varphi} \in \Phi \,\}$$

This gives us two ways to think of the little model. When we think of it as a submodel of the big model, it has parameter $\boldsymbol{\theta} \in \Theta_0$, which is an m-dimensional parameter, just like the parameter of the big model. In fact, since the models are nested, each parameter value $\boldsymbol{\theta}$ in the little model is also a possible parameter value of the big model. But this parameter cannot be varied freely. In order to do maximum likelihood, we must use the parameterization $\boldsymbol{\varphi}$. We say the big model has m free parameters, but the little model only has k free parameters.

To do the likelihood ratio test, we first find the MLE's for the two models. We denote the MLE in the big model $\hat{\boldsymbol{\theta}}_n$, and we denote the MLE in the little model by $\boldsymbol{\theta}^*_n = \mathbf{g}(\hat{\boldsymbol{\varphi}}_n)$. (We can't call them both "theta hat." We wouldn't be able to tell them apart.)

The *likelihood ratio test statistic* for comparing these two models, that is, for testing the hypotheses (10.55) and (10.56) is

$$2[l_n(\hat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}^*_n)]. \tag{10.57}$$

It is twice the log of the maximized likelihood ratios for the two models

$$2 \log \left(\frac{\max_{\boldsymbol{\theta} \in \Theta_A} L_n(\boldsymbol{\theta})}{\max_{\boldsymbol{\theta} \in \Theta_0} L_n(\boldsymbol{\theta})} \right), \tag{10.58}$$

where we are using the convention that big L is the likelihood and little l the log likelihood: $l_n(\boldsymbol{\theta}) = \log L_n(\theta)$. To see why these are the same, the parameter value at which the maximum in the numerator is achieved is by definition $\hat{\boldsymbol{\theta}}_n$, and the parameter value at which the maximum in the denominator is achieved is by definition $\boldsymbol{\theta}^*_n$, so (10.58) is equal to

$$2 \log \left(\frac{L_n(\hat{\boldsymbol{\theta}}_n)}{L_n(\boldsymbol{\theta}^*_n)} \right)$$

and by rule for the the log of a ratio, this is the same as (10.57).

Why this is interesting is the following, which for once we do not state as a formal theorem. If we assume

(i) the null hypothesis is correct, that is, the true parameter value has the form $\boldsymbol{\theta}_0 = \mathbf{g}(\boldsymbol{\varphi}_0)$ for some point $\boldsymbol{\varphi}_0$ in Φ, and

(ii) $\boldsymbol{\varphi}_0$ is an interior point of Φ, and $\boldsymbol{\theta}_0$ is an interior point of Θ,

then under all of the conditions required for the usual asymptotics of maximum likelihood (Theorem 10.6) plus a little bit more (we for once omit the gory details) the asymptotic distribution of (10.57) or (10.58) is $\text{chi}^2(m-k)$.

This is quite a remarkable property of maximum likelihood. When doing a *likelihood ratio test*, one using (10.57) or (10.58) as the test statistic, the asymptotic distribution of the test statistic does not depend on any details of the model. You simply calculate the MLE's in the big and little model, calculate the difference in log likelihoods, multiply by two, and compare to the chi-square distribution with the appropriate number of degrees of freedom.

Example 10.4.6 (Equality of Poisson Means).
Consider the following data, which are counts in regions of equal area in what is assumed to be a Poisson process, which makes the counts independent Poisson random variables.

$$26 \quad 37 \quad 31 \quad 30 \quad 26 \quad 42$$

The question we want to examine is whether the Poisson process is *homogeneous* or *inhomogeneous.* If homogeneous, the counts have mean $\mu = \lambda A$, and since the area A is assumed to be the same for each, the counts all have the same mean, and since the mean is the parameter of the Poisson distribution, that means they all have the same distribution. This is our null hypothesis. The counts X_i are i. i. d. $\text{Poi}(\mu)$. This is the little model. The big model allows unequal means $\mu_i = \lambda_i A$. So in this model the X_i are independent but not identically distributed $X_i \sim \text{Poi}(\mu_i)$. The little model has one free parameter (dimension $k = 1$, and the big model has m parameters if there are m counts in the data, here $m = 6$).

The MLE for the little model, data i. i. d. $\text{Poi}(\mu)$ has already been found in Problem 7-42(c) in Lindgren. It is the sample mean, which here we write $\hat{\mu} = \bar{x}_m$. (This is the $\hat{\varphi}$ parameter estimate in the general discussion.) The corresponding parameter in the big model is m-dimensional with all components equal

$$\boldsymbol{\mu}^* = \begin{pmatrix} \bar{x}_m \\ \vdots \\ \bar{x}_m \end{pmatrix}$$

(This is the $\boldsymbol{\theta}^*$ parameter estimate in the general discussion).

The MLE in the big model also seems obvious. The only data relevant to the mean μ_i is the count x_i, so we "really" have m separate problems, and the MLE is given by the special case $m = 1$ of the solution for the little model, that is, $\hat{\mu}_i = x_i$ and the vector MLE is just the data vector

$$\hat{\boldsymbol{\mu}} = \begin{pmatrix} \hat{\mu}_1 \\ \vdots \\ \hat{\mu}_m \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}$$

Plausible though this may be, it is not completely convincing. We should wade through the gory details to be sure this is really the MLE in the big model. The density for x_i is

$$f_{\mu_i}(x_i) = \frac{\mu_i^{x_i}}{x_i!} e^{-\mu_i}$$

The joint density is the product (because the x_i are independent)

$$f_{\boldsymbol{\mu}}(\mathbf{x}) = \prod_{i=1}^{m} \frac{\mu_i^{x_i}}{x_i!} e^{-\mu_i}$$

The $x_i!$ terms do not contain parameters and can be dropped from the likelihood

$$L(\boldsymbol{\mu}) = \prod_{i=1}^{m} \mu_i^{x_i} e^{-\mu_i}$$

So the log likelihood is

$$l(\boldsymbol{\mu}) = \sum_{i=1}^{m} \bigl(x_i \log(\mu_i) - \mu_i \bigr) \tag{10.59}$$

The derivatives are

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\mu})}{\partial \mu_i} &= \frac{x_i}{\mu_i} - 1 \\
\frac{\partial^2 l(\boldsymbol{\mu})}{\partial \mu_i^2} &= -\frac{x_i}{\mu_i^2} \\
\frac{\partial^2 l(\boldsymbol{\mu})}{\partial \mu_i \partial \mu_j} &= 0, \qquad i \neq j
\end{aligned}$$

Since the second derivative is diagonal with negative diagonal elements, it is negative definite and the log likelihood is a strictly convex function. So the MLE is found by setting the first derivatives equal to zero and solving, which does indeed give $\hat{\mu}_i = x_i$.

The likelihood ratio test statistic is found by plugging $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{\mu}^*$ into the log likelihood (10.59), subtracting, and multiplying by two. Here's how to do it in R

```
> x <- c(26, 37, 31, 30, 26, 42)
> m <- length(x)
> mu.star <- mean(x)
> mu.hat <- x
> l <- function(mu, x) sum(x * log(mu) - mu)
> lrt <- 2 * (l(mu.hat, x) - l(mu.star, x))
> 1 - pchisq(lrt, m - 1)
[1] 0.2918324
```

There are $m - 1$ degrees of freedom in the chi-square distribution because the little model has one parameter and the big model has $m = 6$ parameters. The P-value $P = 0.29$ obviously is not close to statistical significance by any reasonable criterion. Thus we “accept” the little model, and conclude that the data may well be from a homogeneous Poisson process.

There are a couple issues about this example that may be bother you. First, what happened to n? How can we do a "large n" analysis, when there is no n? If m is supposed to be n, it certainly isn't large. This has to do with a special property of the Poisson distribution which we saw in getting a normal approximation. As it says in Appendix F,

$$\text{Poi}(\mu) \approx \mathcal{N}(\mu, \mu)$$

if μ *is large*. There doesn't have to be any n for the CLT to apply. So asymptotics work here not because m is large, but because the means of the X_i are large.

As usual though, we have our dictum that asymptotics only provides a heuristic. If you are worried about the validity of the asymptotics, you simulate. This your author did, and the asymptotics provide a very good approximation here (details not shown, you'll have to take my word for it).

A test of this sort in which the whole point is to accept the little model is often called a *goodness of fit* test. When the little model is accepted, we say it seems to fit the data well. At least the test gives no evidence that the big model fits any better. Thus the principle of parsimony (other things being equal, simpler is better) says to choose the little model.

There is no difference between goodness of fit tests and any other kind of tests except which hypothesis you are rooting for. When you like the simpler model, you call the test a goodness of fit test and are happy when the null hypothesis is accepted, and you conclude that the little model fits just as well as the bigger, more complicated model to which it was compared. When you like the more complicated model, there is no special term for that situation, because that describes most tests. But then you are happy when the null hypothesis is rejected, and you conclude that the complexity of the big model is necessary to fit the data well.

Example 10.4.7 (A Problem in Genetics).
This revisits Example 10.3.9. Here we want to do a goodness of fit test. The little model is the model fit in Example 10.3.9. The big model to which we compare it is the general multinomial model. The log likelihood for the big model is given by the unnumbered displayed equation on p. 10.3.9

$$l_n(\mathbf{p}) = \sum_{i=1}^{k} x_i \log(p_i)$$

because of the constraint (10.30) there are actually only $k - 1$ free parameters in the big model, and in order to fit the model by maximum likelihood we must eliminate on of the parameters by writing it in terms of the others

$$p_k = 1 - p_1 + \cdots + p_{k=1} \tag{10.60}$$

Then we get

$$\begin{aligned}
\frac{\partial l_n(\mathbf{p})}{\partial p_i} &= \frac{x_i}{p_i} + \frac{x_k}{p_k} \cdot \frac{\partial p_k}{\partial p_i} \\
&= \frac{x_i}{p_i} - \frac{x_k}{p_k} \\
\frac{\partial^2 l_n(\mathbf{p})}{\partial p_i^2} &= -\frac{x_i}{p_i^2} - \frac{x_k}{p_k^2} \\
\frac{\partial^2 l_n(\mathbf{p})}{\partial p_i \partial p_j} &= -\frac{x_k}{p_k^2}, \qquad i \neq j
\end{aligned}$$

The second derivative matrix is rather messy. It is in fact negative definite, but we won't go through the details of showing this. Hence the log likelihood is strictly concave and the unique global maximum is found by setting the first derivative equal to zero and solving for the parameter vector $\mathbf{p}$. This gives us $k-1$ equations

$$\frac{x_i}{p_i} = \frac{x_k}{p_k} = \frac{x_k}{1 - p_1 + \cdots + p_{k-1}}$$

in the $k-1$ unknowns p_1, $\ldots$, p_{k-1}. It turns out that the expression on the right hand side here is not helpful. Rewriting the left hand equality gives

$$p_i = \frac{x_i p_k}{x_k}, \qquad i = 1, \ldots, k-1. \tag{10.61}$$

Now use the fact that probabilities sum to one and the x_i sum to n (10.30) and (10.32)

$$\begin{aligned}
1 &= \sum_{i=1}^{k} p_i \\
&= p_k + \sum_{i=1}^{k-1} \frac{x_i p_k}{x_k} \\
&= p_k + \frac{p_k}{x_k} \sum_{i=1}^{k-1} x_i \\
&= p_k + \frac{p_k}{x_k}(n - x_k) \\
&= \frac{n p_k}{x_k}
\end{aligned}$$

Solving for p_k gives

$$p_k = \frac{x_k}{n}$$

and plugging this back into (10.61) gives

$$p_i = \frac{x_i}{n}, \qquad i = 1, \ldots, k-1.$$

Putting both together gives $\hat{p}_i = x_i/n$ for all i, so that is the MLE in the big model.

Now we are ready to do the likelihood ratio test,

```
> x <- c(1997, 904, 906, 32)
> p.hat <- x / sum(x)
> theta <- 0.0357123
> p.star <- c((2 + theta) / 4, (1 - theta) / 4,
+    (1 - theta) / 4, theta / 4)
> l <- function(p) sum(x * log(p))
> lrt <- 2 * (l(p.hat) - l(p.star))
> 1 - pchisq(lrt, 2)
[1] 0.3644519
```

The MLE in the little model ($\hat{\theta} = 0.0357123$) was found in Example 10.3.9. The P-value for $P = 0.36$ shows no significant lack of fit of the small model (that is, we accept H_0 which is the small model, and this is tantamount to saying it fits the data well).

10.5 Change of Parameters

This section is more careful than the proceeding ones. It is so formal that it may be hard to see the forest for the trees. Hence before starting we present the "cartoon guide," which consists of two simple ideas

- a change of parameters does not affect likelihood inference, except that
- in calculating Fisher information some extra terms arise from the chain rule.

10.5.1 Invariance of Likelihood

What happens to the likelihood and log likelihood under a change of parameters? The answer to this question seems so obvious, that we have done the right thing in one of the preceding examples without making a point of it: in Example 10.4.1 we changed parameters from the standard deviation σ to the variance $\varphi = \sigma^2$. Here is an even simpler example.

Example 10.5.1 (Exponential Model).
Suppose X_1, X_2, $\ldots$ are i. i. d. exponential. If we take the parameter to be the usual parameter λ, the likelihood was figured out in Problem 7-38(b) in Lindgren

$$\begin{aligned} L_n(\lambda) &= \lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i \right) \\ &= \lambda^n e^{-\lambda n \bar{x}_n} \end{aligned}$$

So the log likelihood is

$$l_n(\lambda) = n \log(\lambda) - \lambda n \bar{x}_n \tag{10.62}$$

Now suppose that we are not interested in the parameter λ but in the mean $\mu = 1/\lambda$. Then what is the log likelihood? The answer is that we do the obvious: plug $\lambda = 1/\mu$ into (10.62). Giving

$$\tilde{l}_n(\mu) = -n \log(\mu) - \frac{n \bar{x}_n}{\mu} \tag{10.63}$$

for the log likelihood in this parameterization.

Because we are being very careful in this section, we have used different notation for the log likelihood. Recognizing that (10.62) and (10.63) define different functions, we denote one l_n and the other $\tilde{l}_n$. Formally, if we have two parameterizations related by an invertible transformation $\varphi = g(\theta)$ and $\theta = g^{-1}(\varphi)$, then the two log likelihoods are related by *invariance*

$$\tilde{l}_n(\varphi) = l_n(\theta), \qquad \text{when } \varphi = g(\theta). \tag{10.64}$$

The log likelihood has the same values at parameters representing the same probability distribution. In order to clearly show the effect of the change of parameter, we need to plug the condition in (10.64) into the invariance relation giving

$$\tilde{l}_n[g(\theta)] = l_n(\theta) \tag{10.65}$$

This clearly shows that $\tilde{l}_n$ and l_n are not the same function. Note that if we write $h = g^{-1}$ then an equivalent way to write (10.65) is

$$\tilde{l}_n(\varphi) = l_n[h(\varphi)]. \tag{10.66}$$

Also note that exactly the same formulas would hold in the multiparameter case, except that we would use some boldface type.

10.5.2 Invariance of the MLE

What happens to maximum likelihood estimates under a change of parameters?

Theorem 10.7 (Invariance of MLE's). *Suppose that $\varphi = g(\theta)$ is an invertible change of parameter. If $\hat{\theta}$ is the MLE for θ, then $\hat{\varphi} = g(\hat{\theta})$ is the MLE for φ.*

This is obvious from (10.65). If $\hat{\theta}_n$ maximizes l_n, then $\hat{\varphi}_n = g(\hat{\theta}_n)$ maximizes $\tilde{l}_n$. And vice versa: if $\hat{\varphi}_n$ maximizes $\tilde{l}_n$, then $\hat{\theta}_n = g^{-1}(\hat{\varphi}_n)$ maximizes l_n.

This theorem on invariance of maximum likelihood estimates seems obvious, and it is, but it shouldn't be ignored on that account. Other estimates do not possess such an invariance property. Method of moments estimates don't (at least not necessarily), and, as we shall see when we get to them, Bayes estimates don't either.

Example 10.5.2 (Exponential Model).
In Problem 7-42(b) in Lindgren we found $\hat{\lambda}_n = 1/\bar{x}_n$ as the MLE for the i. i. d. Exp(λ) model. By the theorem, $1/\hat{\lambda}_n = \bar{x}_n$ is the MLE of the parameter $\mu = 1/\lambda$. We don't have to maximize (10.63) to find the MLE. We can get it from the theorem.

Example 10.5.3 (Normal Location-Scale Model).
In Example 10.4.1 we found $\hat{\varphi}_n = v_n$ for the MLE of the variance $\varphi = \sigma^2$. By the invariance theorem, the MLE of the standard deviation $\sigma = \sqrt{\varphi}$ is $\hat{\sigma}_n = \sqrt{v_n}$.

We also make the same remark here as in the preceding section, that exactly the same phenomenon holds in the multiparameter case. The formulas would even be exactly the same, except for some boldface type.

10.5.3 Invariance of Likelihood Ratio Tests

Theorem 10.8 (Invariance of the Likelihood Ratio Test). *The likelihood ratio test statistic* (10.57) *or* (10.58) *is unchanged by an invertible change of parameter.*

If g is an invertible change of parameter and $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}^*_n$ are the MLE's in the big model and the little model, respectively, then by Theorem 10.7 $\hat{\boldsymbol{\varphi}}_n = g(\hat{\boldsymbol{\theta}}_n)$ and $\boldsymbol{\varphi}^*_n = g(\boldsymbol{\theta}^*_n)$ are the MLE's in the transformed coordinates, and Theorem 10.8 asserts

$$2[l_n(\hat{\boldsymbol{\theta}}_n) - l_n(\boldsymbol{\theta}^*_n)] = 2[l_n(\hat{\boldsymbol{\varphi}}_n) - l_n(\boldsymbol{\varphi}^*_n)].$$

This is clear from the invariance of likelihood (10.64).

Again this seems obvious, and it is, but it is an important property not shared by other forms of inference. The value of the likelihood ratio test statistic, and hence the P-value for the test, does not depend on the parameterization.

10.5.4 Covariance of Fisher Information

What happens to observed and expected Fisher information under a change of parameters is a bit more complicated.

Theorem 10.9. *Suppose that $\boldsymbol{\varphi} = \mathbf{g}(\boldsymbol{\theta})$ is an invertible change of parameter with differentiable inverse $\boldsymbol{\theta} = \mathbf{h}(\boldsymbol{\varphi})$, and write*

$$\mathbf{H}(\boldsymbol{\varphi}) = \nabla \mathbf{h}(\boldsymbol{\varphi}).$$

Then $\mathbf{I}_n(\boldsymbol{\theta})$, the expected Fisher information for $\boldsymbol{\theta}$, and $\widetilde{\mathbf{I}}_n(\boldsymbol{\varphi})$, the expected Fisher information for $\boldsymbol{\varphi}$, are related by

$$\widetilde{\mathbf{I}}_n(\boldsymbol{\varphi}) = \mathbf{H}(\boldsymbol{\varphi}) \cdot \mathbf{I}_n[\mathbf{h}(\boldsymbol{\varphi})] \cdot \mathbf{H}(\boldsymbol{\varphi})'. \tag{10.67}$$

If $\hat{\boldsymbol{\theta}}_n = \mathbf{h}(\hat{\boldsymbol{\varphi}}_n)$ is an interior point of the parameter space, then $\mathbf{J}_n(\hat{\boldsymbol{\theta}}_n)$, the observed Fisher information for $\boldsymbol{\theta}$ evaluated at the MLE, and $\widetilde{\mathbf{J}}_n(\hat{\boldsymbol{\varphi}}_n)$, the observed Fisher information for $\boldsymbol{\varphi}$ evaluated at the MLE, are related by

$$\widetilde{\mathbf{J}}_n(\hat{\boldsymbol{\varphi}}_n) = \mathbf{H}(\hat{\boldsymbol{\varphi}}_n) \cdot \mathbf{J}_n(\hat{\boldsymbol{\theta}}_n) \cdot \mathbf{H}(\hat{\boldsymbol{\varphi}}_n)'. \tag{10.68}$$

Each of the terms in (10.67) and (10.68) is an $m \times m$ matrix if there are m parameters. Hence this theorem is not so easy to use, and we won't give any examples.

The uniparameter case is much simpler, and we record as in explicit corollary.

Corollary 10.10. *Suppose that $\varphi = g(\theta)$ is an invertible change of parameter with differentiable inverse $\theta = h(\varphi)$. Then $I_n(\theta)$, the expected Fisher information for θ, and $\widetilde{I}_n(\varphi)$, the expected Fisher information for φ, are related by*

$$\widetilde{I}_n(\varphi) = I_n[h(\varphi)] \cdot [h'(\varphi)]^2 \tag{10.69}$$

If $\hat{\theta}_n = h(\hat{\varphi}_n)$ is an interior point of the parameter space, then of the parameter space, then $J_n(\hat{\theta}_n)$, the observed Fisher information for θ evaluated at the MLE, and $\widetilde{J}_n(\hat{\varphi}_n)$, the observed Fisher information for φ evaluated at the MLE, are related by

$$\widetilde{J}_n(\hat{\varphi}_n) = J_n(\hat{\theta}_n) \cdot [h'(\hat{\varphi}_n)]^2 \tag{10.70}$$

Note the difference between (10.67) and (10.68). The transformation rule for expected Fisher information holds for all parameter values. The transformation rule for observed Fisher information holds only when it is evaluated at the MLE and the MLE is in the interior of the parameter space, not on the boundary.

Example 10.5.4 (Exponential Model).
Consider again the Exp(λ) model we looked at in Examples 10.5.1 and 10.5.2. In those examples we found the MLE's of λ and $\mu = 1/\lambda$ to be $\hat{\mu}_n = \bar{x}_n$ and $\hat{\lambda}_n = 1/\bar{x}_n$.

We also found in Problem 10-1(a) that the Fisher information for λ is

$$I_n(\lambda) = \frac{n}{\lambda^2}$$

Let us apply the corollary to find the Fisher information for μ.

The inverse transformation is

$$\lambda = h(\mu) = \frac{1}{\mu}$$

and the derivative is

$$h'(\mu) = -\frac{1}{\mu^2}$$

Thus (10.69) gives

$$\begin{aligned}
\widetilde{I}_n(\mu) &= I_n[h(\mu)] \cdot [h'(\mu)]^2 \\
&= I_n(1/\mu) \cdot \left(-\frac{1}{\mu^2}\right)^2 \\
&= \frac{1}{(1/\mu)^2} \cdot \frac{1}{\mu^4} \\
&= \frac{1}{\mu^2}
\end{aligned}$$

Thus we get for the asymptotic distribution

$$\hat{\mu}_n \approx \mathcal{N}\left(\mu, \frac{\mu^2}{n}\right)$$

Of course, we didn't need to do this calculation to find the asymptotic distribution. Since $\hat{mu}_n = \bar{x}_n$ the CLT gives its asymptotic distribution directly

$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

We see that these are indeed the same because we know that for the exponential distribution $\mu = 1/\lambda$ and $\sigma^2 = 1/\lambda^2$ so $\sigma^2 = \mu^2$.

Problems

10-1. Suppose X_1, X_2, ... are i. i. d. $\text{Exp}(\lambda)$. We found in Problem 7-42(b) in Lindgren that the MLE is $\hat{\lambda}_n = 1/\bar{x}_n$.

(a) Find the asymptotic distribution of $\hat{\lambda}_n$ using expected Fisher information, and check that this gives the same answer as the delta method (which was done in Example 8.2.1 in these notes).

(b) Find an asymptotic 95% confidence interval for λ, again using Fisher information (either observed or expected, your choice).

10-2. Suppose (X_i, Y_i), $i = 1$, ..., n are i. i. d. with joint density

$$f(x, y) = e^{-\theta x - y/\theta}, \qquad x > 0,\ y > 0.$$

(a) Find the MLE of θ.

(b) Find the observed and expected Fisher information (both) and asymptotic standard errors for the MLE based on each. Are they the same?

10-3. Let X_1, X_2, ..., X_n be an i. i. d. sample from a model having densities

$$f_\theta(x) = (\theta - 1)x^{-\theta}, \qquad 1 < x < \infty,$$

where $\theta > 1$ is an unknown parameter.

(a) Find the MLE of θ and prove that it is the global maximizer of the likelihood.

(b) Find the expected Fisher information for θ.

(c) Give an asymptotic 95% confidence interval for θ.

(d) Show that

$$\hat{\theta}_n = \frac{2\overline{X}_n - 1}{\overline{X}_n - 1}$$

is a method of moments estimator of θ.

(e) Use the delta method to calculate the asymptotic distribution of this method of moments estimator.

(f) Calculate the ARE for the two estimators.

10-4. Show that for data that are i. i. d. $\mathcal{U}(0, \theta)$ the MLE is $\hat{\theta}_n = X_{(n)}$, the maximum data value, the asymptotic distribution of which was found in Problem 7-7. (**Hint:** Careful! The solution involves the boundary of the sample space. Also Example 8.6a in Lindgren is similar and may give you some ideas.)

This shows that Theorem 10.3 doesn't always hold. The asymptotics

$$n\bigl(\theta - X_{(n)}\bigr) \xrightarrow{\mathcal{D}} \operatorname{Exp}(1/\theta)$$

found in Problem 7-7 don't even remotely resemble the "usual asymptotics" of maximum likelihood.

10-5. Suppose x_1, $\ldots$, x_n are known numbers (not random), and we observe random variables $Y_1, \ldots, Y_n$ that are independent but *not* identically distributed random variables having distributions

$$Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2),$$

where α, β, and σ^2 are unknown parameters.

(a) Write down the log likelihood for the parameters α, β, and $\varphi = \sigma^2$.

(b) Find the maximum likelihood estimates of these parameters.

(c) Find the expected Fisher information matrix for these parameters.

(**Caution:** In taking expectations remember only the Y_i are random. The x_i are known constants.)

10-6. Find the maximum likelihood estimates for the two-parameter gamma model with densities

$$f_{\alpha,\lambda}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$$

This cannot be done in closed form, you will have to use R. Use the method of moments estimates derived in (9.6a) and (9.6b) in these notes the starting point supplied to the `nlm` function in R.

One way to write an R function that evaluates minus the log likelihood for this model, assuming the data are a vector `x` is

```
l <- function(theta) {
      alpha <- theta[1]
      lambda <- theta[2]
      return(- sum(log(dgamma(x, alpha, 1 / lambda))))
  }
```

Data for the problem are at the URL

`http://www.stat.umn.edu/geyer/5102/prob10-6.dat`

It helps the `nlm` function to add the optional argument `fscale=length(x)` to give it an idea of the approximate size of the log likelihood.

(a) Find the MLE (vector) for the data at this URL.

(b) Find the observed Fisher information at the MLE and show that it is a positive definite matrix.

(c) Find asymptotic 95% confidence intervals for the parameters α and λ. (They do not need to have simultaneous coverage, that is, you need not use Bonferroni correction).

10-7. Prove Corollary 10.10 directly. (It is just the one-parameter case of Theorem 10.9, so the corollary follows trivially from the theorem, but we didn't prove the theorem. So prove the corollary without using the theorem.)

Hint: Start with (10.66) and use the chain rule.

10-8. Suppose X_1, $\ldots$, X_n are i. i. d. Poi(μ). The probability that X_i is zero is $p = e^{-\mu}$. Note that the transformation $p = g(\mu) = e^{-\mu}$ is one-to-one because the exponential function is monotone. Hence we can also consider p a parameter of this distribution. It's just not the usual one.

(a) Find the MLE for μ and for p.

(b) Find the (expected) Fisher information for μ.

(c) Find the (expected) Fisher information for p. Corollary 10.10 may be helpful.

(d) Suppose we observe data $\overline{X}_n = 5.7$ and $n = 30$. Find a 95% confidence interval for the parameter p.

Chapter 11

Bayesian Inference

A *Bayesian* is a person who treats parameters as random variables, and a "frequentist" is a person who doesn't. The "frequentist" slogan that expresses this is "parameters are unknown constants, not random variables." This is supposed to explain why Bayesian inference is wrong. But it is a cheap rhetorical trick. Bayesians think that probability theory is a way to express lack of knowledge, so they *agree* that "parameters are unknown constants" and continue with "hence we describe our uncertainty about them with a probability model."

Slogans can be tossed back and forth forever with no change in positions. To see what the argument is about, we have to learn Bayesian inference.

11.1 Parametric Models and Conditional Probability

The Bayesian notion gives us another view of conditional probability.

> *Conditional probability distributions are no different from parametric families of distributions.*

For each fixed value of y, the conditional density $f(x \mid y)$, considered as a function of x alone, is a probability density. So long as the two properties

$$f(x \mid y) \geq 0, \qquad \text{for all } x \tag{11.1a}$$

and

$$\int f(x \mid y) \, dx = 1 \tag{11.1b}$$

(with the integral replaced by a sum in the discrete case) hold for all y, then this defines a conditional probability model. There is no other requirement. We also made this point when we considered conditional probability in Chapter 3 of these notes. In fact, (11.1a) and (11.1b) just repeat (3.5a) and (3.5b) from that chapter.

Last semester most of our time spent studying conditional probability involved deriving conditional densities from joint densities. When you are doing that, there is another requirement conditional densities must satisfy: "joint equals marginal times conditional"

$$f(x, y) = f(x \mid y) f_Y(y).$$

But that's a very different issue. If we are only interested in whether a formula defines a conditional density and have no interest in whether or not this conditional density is the one associated with a particular joint density, then the only requirements are that (11.1a) and (11.1b) hold for every possible value of y. These are, of course, the same requirements that apply to *any* probability density, conditional or unconditional, that it is nonnegative and integrate or sum, as the case may be, to one.

The point of the slogan about there being no difference between conditional probability and parametric families is that parametric families must satisfy the same two properties with y replaced by θ

$$f(x \mid \theta) \ge 0, \qquad \text{for all } x \tag{11.2a}$$

and

$$\int f(x \mid \theta) \, dx = 1 \tag{11.2b}$$

A frequentist may write a probability density $f_\theta(x)$ to emphasize that θ is not a random variable, but just an adjustable constant. We used this notation ourselves in the chapters on frequentist inference (9 and 10). A Bayesian always writes $f(x \mid \theta)$ to emphasize that θ is a random variable (a Bayesian is a person who treats parameters as random variables), and the density is being considered the conditional density of the random variable X given the random variable θ.

The multivariable or multiparameter case is no different except for some boldface type (and perhaps sums and integrals become multiple too). We still say that conditional probability and parametric models are different ways of looking at the same thing, and that the only requirement for a function to be a conditional density is that it be nonnegative and integrate (or sum) to one, integrating (or summing) with respect to the variable "in front of the bar."

11.2 Prior and Posterior Distributions

11.2.1 Prior Distributions

Bayesians use the same statistical models as frequentists. If X_1, X_2, …, X_n are i. i. d. with density $f(x \mid \theta)$, then the joint distribution of the data is

$$f(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta).$$

When this is thought of as a function of the parameter rather than the data, it becomes the likelihood

$$L_{\mathbf{x}}(\theta) = f(\mathbf{x} \mid \theta) \tag{11.3}$$

or, more generally,

$$f_{\mathbf{X}}(\mathbf{x} \mid \theta) = c(\mathbf{x}) L_{\mathbf{x}}(\theta) \tag{11.4}$$

where $c(\mathbf{x})$ is any nonzero function of $\mathbf{x}$ alone, not a function of θ. These are just (10.1) and (10.2) repeated, the same definition of likelihood as in the preceding chapter. The only difference is that we are using the Bayesian notation $f(\mathbf{x} \mid \theta)$ rather than the frequentist notation $f_\theta(\mathbf{x})$ for densities. The point is that the Bayesian thinks of both $\mathbf{X}$ and θ as random variables and thinks of the density $f(\mathbf{x} \mid \theta)$ as a conditional density of $\mathbf{x}$ given θ.

So far Bayesians and non-Bayesians agree, except for notation. They part company when the Bayesian goes on to put a probability distribution on the parameter θ. In order to specify a joint distribution for $\mathbf{X}$ and θ, we need the marginal for θ. For reasons to be explained later, this is called the *prior* distribution of θ. Since we have already used the letter f for the density of $\mathbf{x}$ given θ, we (following Lindgren) will use g for the prior density.

We should take a brief time-out for a reminder that *mathematics is invariant under changes of notation.* Up to this point random variables have always been Roman letters, never Greek letters. You may have unconsciously made this a rule. If so, you will now have to unlearn it. For Bayesians, the parameters (usually Greek letters) are also random variables.

Example 11.2.1 (Exponential Data, Gamma Prior).
Suppose the data are one observation $X \sim \text{Exp}(\lambda)$ so the conditional density of the data given the parameter is

$$f(x \mid \lambda) = \lambda e^{-\lambda x}, \tag{11.5a}$$

and suppose the prior distribution for λ is $\text{Gam}(a, b)$. We call a and b *hyperparameters* of the prior. We can't use the usual notation α and λ for gamma distribution parameters for the hyperparameters, at least we can't use λ, because we are already using λ for something else, the parameter of the data distribution. Although a and b are parameters (of the prior), it would be too confusing to simply call them "parameters" as in "parameters of the distribution of the parameter." Hence the term "hyperparameter" which indicates parameters "one level up" (in the prior rather than the data distribution). Bayesians treat parameters (here λ) as random variables, but not hyperparameters (here a and b). The hyperparameters are constants chosen to give a particular prior distribution.

What is the prior density of λ? We usually write the gamma density as

$$f(x \mid \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} \tag{11.5b}$$

but as we already noticed we can't use α and λ as the hyperparameters because we are already using λ for the parameter. The problem says we are to use a

and b for those parameters. This means that (11.5b) becomes

$$f(x \mid a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx} \tag{11.5c}$$

That is what the notation $\operatorname{Gam}(a, b)$ means. We also have to make another change. It is not X but λ that has this distribution. This means that (11.5c) becomes

$$f(\lambda \mid a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \tag{11.5d}$$

That is the prior density for λ. While we are making changes using "mathematics is invariant under changes of notation" we might as well make one more, changing the name of the function from f to g because we are already using f for the function in (11.5a). This means that (11.5d) becomes

$$g(\lambda \mid a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda} \tag{11.5e}$$

These sorts of changes of notation, changing α and λ to a and b to get (11.5c), then changing x to λ to get (11.5d), then changing f to g to get (11.5e) should be easy. If they throw you, practice until they become easy. The past experience with this course is that some students never understand these trivial manipulations. Hence they can't even get started right on any Bayesian problem, and hence completely botch all of them. So a word to the wise: if you haven't understood "mathematics is invariant under changes of notation" yet, get it now.

> ***A Sanity Check:*** *The prior density for θ is a function of θ, not some other variable (like x).*

Making this simple sanity check will save you from the worst errors: using (11.5b) or (11.5c) for the prior. Not that there are no other ways to screw up. If you decide to change x to λ first, paying no attention to the fact that there already is a λ in (11.5b), you get

$$f(\lambda \mid \alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\lambda^2} \tag{11.5f}$$

and the problem is now irretrievably botched. There is no way to get from (11.5f) to the right formula (11.5e) except recognizing you've goofed and starting over.

11.2.2 Posterior Distributions

The *joint* distribution of data and parameters, that is, of the pair $(\mathbf{X}, \theta)$, is the conditional times the marginal $f(\mathbf{x} \mid \theta) g(\theta)$. The next step in Bayesian inference is to produce the conditional distribution of the parameter given the data. For this Lindgren uses yet a third letter h. We know how to find a conditional given a joint distribution: conditional = joint/marginal or written out in symbols

$$h(\theta \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \theta) g(\theta)}{p(\mathbf{x})} \tag{11.6}$$

where $p(\mathbf{x})$ is the marginal of $\mathbf{X}$, that is,

$$p(\mathbf{x}) = \int f(\mathbf{x} \mid \theta) g(\theta) \, d\theta \tag{11.7}$$

if θ is a continuous random variable. Of course, if θ is discrete we use the same formula except that the integral is replaced by a sum, but applications with discrete parameter spaces are very rare. We won't consider any in this course.

The conditional distribution $h(\theta \mid \mathbf{x})$ of the parameter given the data is called the *posterior* distribution of the parameter. The idea behind the terminology is that the prior represents knowledge (or conversely uncertainty) about the parameter before the data are observed and the posterior represents knowledge about the parameter after the data are observed. That agrees with our usual notion of conditional probability: $f(x \mid y)$ is what you use for the distribution of X after you observe Y. The only novelty is applying this notation to parameters (Greek letters) rather than to data (Roman letters). To a Bayesian these are both random variables, so there is no novelty. Bayesian inference is just an application of conditional probability. The only novelty is the notion of treating parameters as random variables in the first place.

If we use (11.4) to replace the conditional density $f(\mathbf{x} \mid \theta)$ by a constant times the likelihood, we see that this does not affect the calculation of the posterior, because (11.7) becomes

$$p(\mathbf{x}) = c(\mathbf{x}) \int L_{\mathbf{x}}(\theta) g(\theta) \, d\theta \tag{11.8}$$

so plugging both (11.4) and (11.8) into (11.6) gives

$$h(\theta \mid \mathbf{x}) = \frac{L_{\mathbf{x}}(\theta) g(\theta)}{\int L_{\mathbf{x}}(\theta) g(\theta) \, d\theta} \tag{11.9}$$

the factor $c(\mathbf{x})$ that appears in both the numerator and denominator cancels. Either of the formulas (11.6) or (11.9) is commonly called *Bayes' rule* or *Bayes' theorem*. Calling it a "theorem" seems a bit much, since it is a trivial rearrangement of the definition of conditional probability density. In fact, exactly this formula was introduced in the chapter on conditional probability of these notes (Section 3.4.5) last semester. The only difference is that we used Roman letters for the variables behind the bar, and now we are going to use Greek letters. Same mathematical idea, just different notation.

In the same chapter we also introduced the notion of unnormalized probability densities (Section 3.4.2) and calculation of conditional probabilities as a renormalization process (Sections 3.4.3 and 3.4.4). If you weren't in my section first semester (or if you have forgotten this material), you should review it.

A function is called an *unnormalized density* if it is nonnegative and has a finite nonzero integral, which is called the *normalizing constant* of the function. Using this notion, a simple way to express (11.9) is to say that $L_{\mathbf{x}}(\theta)g(\theta)$, thought of as a function of θ for fixed $\mathbf{x}$, is an unnormalized posterior density. The denominator in (11.9) is its normalizing constant. Another way to say it is the pseudo-mathematical expression

$$\text{posterior} \propto \text{likelihood} \times \text{prior}. \tag{11.10}$$

The symbol $\propto$, read "proportional to," expresses the notion that the right hand side is an unnormalized density.

Similarly, it does no harm if the prior density is unnormalized. Suppose we specify the prior by an unnormalized density $g_u(\theta)$. The normalized prior is then $g(\theta) = c g_u(\theta)$, where c is a nonzero constant. Plugging this into (11.9) gives

$$h(\theta \mid \mathbf{x}) = \frac{L_{\mathbf{x}}(\theta)g_u(\theta)}{\int L_{\mathbf{x}}(\theta)g_u(\theta)\,d\theta}.$$

The factor c appears in both the numerator and denominator and cancels. This is exactly the same as (11.9) except that g is replaced by g_u. Thus it makes no difference whether or not the prior is normalized. So we could re-express our slogan (11.10) as

$$\text{posterior} \propto \text{likelihood} \times \text{possibly unnormalized prior},$$

but that seems too verbose. We'll just use (11.10) with the tacit understanding that the prior density need not be normalized.

Example 11.2.2 (Example 11.2.1 Continued).
Since the hyperparameters a and b are constants, the first factor in the (correct!) prior density (11.5e) is constant and we can drop it, giving the unnormalized prior

$$g_u(\lambda \mid a, b) = \lambda^{a-1} e^{-b\lambda}. \tag{11.11}$$

Multiplying by the data distribution gives the unnormalized posterior

$$h_u(\lambda \mid x) = \lambda e^{-\lambda x} \lambda^{a-1} e^{-b\lambda} = \lambda^a e^{-(b+x)\lambda}. \tag{11.12}$$

Keep in mind that the random variable here is λ, the data x is fixed (because we are conditioning on it) and so are the hyperparameters a and b.

To normalize this density, we use our favorite trick of recognizing the unnormalized density of a brand name distribution. Clearly (11.12) has the same form as (11.11). The only difference is that a has been replaced by $a + 1$ and b has been replaced by $b + x$. Thus the posterior distribution of λ given x is $\text{Gam}(a + 1, b + x)$.

That constitutes a satisfactory answer to the problem. We don't even have to write down the density to specify the posterior distribution. If for some reason we do actually need the density, we can find it from the formula for the gamma density

$$h(\lambda \mid x) = \frac{(b+x)^{a+1}}{\Gamma(a+1)} \lambda^a e^{-(b+x)\lambda}. \tag{11.13}$$

What if we hadn't thought of the trick? Plodding on using (11.9), we would see that in order to find the denominator we would have to evaluate the integral

$$\int_0^\infty \lambda^a e^{-(b+x)\lambda} \, d\lambda$$

(Note well! The variable of integration is λ not x. The variable in a Bayesian problem is the parameter. If you proceed by reflexes rather than thinking during an exam, you are liable to write dx. If you remember this warning, you won't make that mistake.) This integral is rather hard to do unless we recognize its relationship to the gamma function or just find in a book. It is equation (4) on p. 173 in Lindgren. Evaluating the integral and plugging into (11.9) gives us (11.13) again.

Doing the calculation of the integral just rederives the normalizing constant of the gamma distribution, redoing the work on p. 173 in Lindgren. The trick saves you this extra work.

Example 11.2.3 (Binomial Data, Uniform Prior).
This example is the first Bayesian analysis ever done. It was discovered by Thomas Bayes and published posthumously in 1764 and gives Bayesian inference its name. Suppose the data are $X \sim \text{Bin}(n, p)$ and our prior distribution for p is $\mathcal{U}(0, 1)$.

Then the prior is $g(p) = 1$ for $0 < p < 1$, and the likelihood is (10.3). Since the prior is identically equal to one, the likelihood is also the unnormalized posterior

$$h(p \mid x) \propto p^x (1-p)^{n-x} \tag{11.14}$$

To normalize this density, we again use our favorite trick of recognizing the unnormalized density of a brand name distribution. In this case the factors p and $(1-p)$ should recall the beta distribution,[1] which has densities of the form

$$f(x \mid s, t) = \frac{\Gamma(s+t)}{\Gamma(s)\Gamma(t)} x^{s-1} (1-x)^{t-1} \tag{11.15}$$

(p. 175 in Lindgren). Comparing (11.15) with x changed to p with (11.14), we see that they are the same except for constants if $s = x + 1$ and $t = n - x + 1$.

[1] Why not the binomial distribution? That's the one that has p and $1-p$ in the formula! The beta distribution has x and $1-x$. If that's what you are thinking, you have again run afoul of "mathematics is invariant under changes of notation." The letters don't matter. A binomial distribution is a binomial distribution no matter whether you call the parameter p or x, and a beta distribution is a beta distribution no matter whether you call the random variable p or x. What matters is not which letter you use, but the role it plays. Here p is the random variable, the letter in front of the bar in the conditional density $h(p \mid x)$, hence we want to find a distribution having a density with factors p and $1-p$ where p is the *random variable.* The beta distribution is the only one we know with that property.

Thus the posterior distribution of p given x is $\text{Beta}(x + 1, n - x + 1)$. No integration is necessary if we see the trick.

If you don't see the trick, you must integrate the right hand side of (11.14) to find the normalizing constant of the posterior. Again this integral is rather hard unless you recognize it as a beta integral, equation (14) on p. 175 in Lindgren. As in the preceding example, this just redoes the work of deriving the normalizing constant of a brand name distribution. The trick is easier.

Example 11.2.4 (Normal Data, Normal Prior on the Mean).
Assume X_1, $\ldots$, X_n are i. i. d. with unknown mean μ and known variance σ^2, and assume a normal prior distribution for the unknown parameter μ. This example is not very practical, because we rarely know σ^2, but it makes a good example. Inference for the case where both μ and σ^2 are unknown will be covered in Section 11.4.3.

Let us denote the prior distribution for μ by $\mathcal{N}(\mu_0, \sigma_0^2)$. As in Example 11.2.1 we can't use μ and σ for the hyperparameters, because we are already using these letters for parameters of the data distribution. Then an unnormalized prior density is

$$g(\mu) = \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right)$$

(we can drop the constant $\sqrt{2\pi}\sigma_0$). Combining this with the likelihood (10.5) gives the unnormalized posterior

$$\begin{aligned} h_u(\mu \mid x) &= L_{\mathbf{x}}(\mu) g(\mu) \\ &= \exp\left(-\frac{n(\bar{x}_n - \mu)^2}{2\sigma^2}\right) \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) \\ &= \exp\left(-\frac{n(\bar{x}_n - \mu)^2}{2\sigma^2} - \frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) \end{aligned}$$

It will considerably simplify notation in the rest of the problem if we introduce $\lambda = 1/\sigma^2$ and $\lambda_0 = 1/\sigma_0^2$. The technical term for reciprocal variance is *precision*, so λ is the precision of the data, and λ_0 is the precision hyperparameter. Then the unnormalized posterior becomes

$$h_u(\mu \mid x) = \exp\left(-\frac{n\lambda}{2}(\bar{x}_n - \mu)^2 - \frac{\lambda_0}{2}(\mu - \mu_0)^2\right) \tag{11.16}$$

Since the exponent is quadratic in μ, the posterior must be some normal distribution (this is the "e to a quadratic theorem, Theorem 5.10 of Chapter 5 of last semester's notes). To see which normal distribution, we could apply the theorem, but we will just do the calculation from first principles since the theorem is multivariate and we are only interested in the univariate case here (and, to be honest, I don't want to rewrite this, which was written last year before I wrote the theorem this year). To do the calculation we compare the exponent in (11.16) with the exponent of a normal density with mean a and precision b, which is $-b(\mu - a)^2/2$. Matching coefficients of μ and μ^2 gives the posterior mean a and posterior precision b. That is, we must have

$$\begin{aligned} & b\mu^2 - 2ab\mu + \text{a constant} \\ & \qquad = (n\lambda + \lambda_0)\mu^2 - 2(n\lambda\bar{x}_n + \lambda_0\mu_0)\mu + \text{some other constant} \end{aligned}$$

Hence

$$\begin{aligned} b &= n\lambda + \lambda_0 \qquad (11.17\text{a}) \\ ab &= n\lambda\bar{x}_n + \lambda_0\mu_0 \end{aligned}$$

so

$$a = \frac{n\lambda\bar{x}_n + \lambda_0\mu_0}{n\lambda + \lambda_0} \qquad (11.17\text{b})$$

and the posterior distribution of μ is normal with mean (11.17b) and precision (11.17a).

11.3 The Subjective Bayes Philosophy

That's more or less the story on the mechanics of Bayesian inference. There are some bells and whistles that we will add later, but this is the basic story. It's just conditional probability coupled with the notion of treating parameters as random variables. For the most part the calculations are no different from those we did when we studied conditional probability last semester. If you can get used to Greek letters as random variables, the rest is straightforward.

Here we take a time out from learning mechanics to learn enough of the Bayesian philosophy to understand this chapter. The Bayesian philosophy holds that all uncertainty can be described by means of probability distributions. This has far reaching implications. For one thing, it means that, since everyone is uncertain about many things, everyone has probability distributions inside their heads. These are the prior distributions that appear in Bayesian problems. A subjective Bayesian believes that everyone has a different prior distribution for any particular problem. An objective Bayesian believes there are ways in which different people can agree on a common prior distribution (by convention if no other way). We will only explain the subjective Bayesian view.

So in any particular problem, once the probability model for the data (and hence the likelihood) is decided, one then gets a prior by "eliciting" the prior distribution that represents the knowledge (or uncertainty, depending on how you look at it) of some expert. Then you apply Bayes rule, and you're done.

> *Once agreement is reached about being Bayesian and on the likelihood and prior, Bayesian inference is straightforward.*

Frequentist inference involves many technical difficulties, choice of point estimators, test statistics, and so forth. Bayesian inference involves no such difficulties.

Every inference is a straightforward conditional probability calculation, which if not doable with pencil and paper is usually doable by computer.

Getting the agreement mentioned in the slogan may be difficult. Bayesian inference is controversial. For a century roughly from 1860 to 1960, it was considered absurd, obviously completely wrong (and many other pejorative terms were applied). Now the pendulum of fashion has swung the other way, and Bayesian inference, if not the most popular, is at least very trendy in certain circles. But it takes some people a long time to get the word. Textbooks are often decades behind research. Opinions are passed from scientist to scientist without influence by statisticians and statistics courses. So there are still a lot of people out there who think Bayesian inference is a no-no.

Agreement to be Bayesian does not end philosophical arguments. There can be arguments about the appropriateness of the probability model for the data, but exactly the same arguments would arise if one wanted to use the same model for frequentist inference, so those arguments are not peculiar to Bayesian inference. And there can be arguments about the prior. Whose prior (what expert's opinion) is used? How it is elicited? Was it elicited correctly? The elicitation problem is made more difficult (or perhaps simpler, I'm not sure) by the fact that it does not really involve getting a probability distribution from inside someone's head down on paper. All psychological study of people's behavior involving probability and statistics has revealed no evidence for, and a good deal of evidence against, the notion that real people think in accord the rules of Bayesian inference.

What do we mean by "think in accord with the rules of Bayesian inference"? We will explain that, the so-called *Bayesian model of learning*, and that will end our discussion of philosophy.

Suppose data X_1, X_2, $\ldots$ are assumed to have a probability model with likelihood

$$L_{x_1,\ldots,x_n}(\theta) = \prod_{i=1}^n f(x_i \mid \theta).$$

(Note that we have gone back to our original notation of indicating the data by a subscript. The reason for it will become apparent presently.) And suppose we start with a prior $g(\theta)$ that represents our knowledge (or, looked at the other way, uncertainty) about the parameter θ before any data are observed.

Suppose the data arrive over the course of time, and intermediate analyses are done as the data arrive. For simplicity, we only consider two analyses, but it will be clear that everything said here extends to multiple analyses. For concreteness, say the analyses are done after observing m and n data values X_i, respectively, with $m < n$. In the first analysis, we derive a posterior distribution

$$h(\theta \mid x_1, \ldots, x_m) \propto L_{x_1,\ldots,x_m}(\theta) g(\theta) \tag{11.18}$$

that represents our knowledge (or, looked at the other way, uncertainty) about the parameter θ at this time and reflects both the information from the prior and from the data x_1, $\ldots$, x_m.

Now we take a time out for more philosophy. The distribution (11.18) can be thought of as *both* a prior and a posterior. It is a posterior, when we consider that it describes our knowledge *after* x_1, ..., x_m have been observed. It is a prior, when we consider that it describes our knowledge *before* x_{m+1}, ..., x_n are observed. So it should serve as our prior for the subsequent analysis of those data.

The likelihood for those data is

$$L_{x_{m+1},\ldots,x_n}(\theta) = \prod_{i=m+1}^{n} f(x_i \mid \theta).$$

and the posterior after observing those data is

$$h(\theta \mid x_1, \ldots, x_n) \propto L_{x_{m+1},\ldots,x_n}(\theta) h(\theta \mid x_1, \ldots, x_m) \tag{11.19}$$

Great! So what's the point? The point is that there is another way of thinking about this problem. If we ignore the fact that the data arrived in two clumps, we would analyze the whole data set at once, using the likelihood for all the data and the original prior $g(\theta)$. This would give a formula just like (11.18) except with m replaced by n. Now there is no philosophical reason why these two procedures (two-stage analysis and one-stage analysis) should give the same answers, but it is a remarkable fact that they do. No matter how you do a Bayesian analysis, so long as you correctly apply Bayes rule, you get the same answer (starting from the same prior).

Note that frequentist inference does not have this property. There is no way to use the results of a frequentist analysis of part of the data in subsequent analyses. In fact, the very fact of having done an analysis on part of the data *changes the answer* of the analysis of the complete data, because it gives rise to a need for correction for multiple testing. What we learned here is that Bayesian inference has (and needs) no analog of frequentist correction for multiple testing. So long as you apply Bayes rule correctly, you get the correct answer.

This same issue means that Bayesian inference can serve as a model of learning, but frequentist inference can't. The Bayesian notion of learning is just the transformation from prior to posterior via Bayes rule. The prior describes your knowledge before the data are observed, the posterior your knowledge after the data are observed, the difference is what you learned from the data.

11.4 More on Prior and Posterior Distributions

11.4.1 Improper Priors

We saw in the preceding section that it does no harm to use an unnormalized prior. We can go even further and drop the requirement that the prior be integrable. If $g(\theta)$ is a nonnegative function such that

$$\int g(\theta) \, d\theta = \infty$$

but the normalizing constant

$$\int L_{\mathbf{x}}(\theta) g(\theta) \, d\theta$$

is still finite, then (11.9) still defines a probability density.

Then we say we are using an *improper prior* and sometimes we say we are using the *formal Bayes rule*. It isn't really Bayes' rule, because neither the prior nor the joint distribution of parameter and data are proper probability distributions. But we use the same equation (11.9) and so the method has the same form ("formal" here doesn't mean "dressed for a prom," it means "having the same form").

Thus what we are doing here is philosophically bogus from the subjective point of view. An improper prior cannot represent "prior opinion" because it is not a probability distribution. That's why Lindgren, for example, makes a point of deriving the results with improper priors as limits of procedures using proper priors (Problem 7-83, for example). But not all results involving improper priors can be derived as such limits, so the limiting argument really contributes nothing to our understanding of improper priors. Hence our approach will be "just do it" with no worries about consequent philosophical difficulties.

Example 11.4.1 (Normal Data, Improper Prior).
Suppose X_1, X_2, $\ldots$, X_n are i. i. d. $\mathcal{N}(\mu, \sigma^2)$ where μ is a known number (not a parameter) and σ is an unknown parameter. We need a prior distribution for the unknown parameter σ, which we take to be the improper prior[2] with density $g(\sigma) = 1$ for all σ. This is improper because

$$\int_0^\infty d\sigma = \infty.$$

The likelihood is

$$L(\sigma) = \frac{1}{\sigma^n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\}. \tag{11.20}$$

Since we are using a flat prior (11.20) is also the unnormalized posterior.

We now want to use our trick of recognizing the density of a known model, but (11.20) isn't proportional to any of the densities in Chapter 6 in Lindgren. It turns out, however, that a change of variable gives us a known family. Define a new parameter $\lambda = 1/\sigma^2$ (precision again). Then the likelihood becomes

$$L(\lambda) = \lambda^{n/2} \exp\left\{ -\frac{\lambda}{2} \sum_{i=1}^n (x_i - \mu)^2 \right\}. \tag{11.21}$$

[2] The reader should perhaps be warned that aficionados of improper priors consider this the wrong improper prior. The "natural" improper prior is $g(\sigma) = 1/\sigma$. The reasons why, however, are too complicated to explain here, and do not generalize to other problems. So we will use this one.

There is no use of the change-of-variable theorem (Theorem 8 of Chapter 3 in Lindgren) because λ is not a random variable in the *data model* or in the *likelihood*. There the X_i are the random variables.

We do, however, need to apply the change-of-variable theorem to the prior. The inverse transformation is

$$\sigma = h(\lambda) = \lambda^{-1/2},$$

and the change-of-variable theorem says the prior for λ is

$$g_\Lambda(\lambda) = g[h(\lambda)] |h'(\lambda)|$$

where

$$|h'(\lambda)| = \left| -\frac{1}{2} \lambda^{-3/2} \right| = \tfrac{1}{2} \lambda^{-3/2} \tag{11.22}$$

Since g, the prior for σ, is identically equal to 1, the prior for λ is

$$g_\Lambda(\lambda) = \tfrac{1}{2} \lambda^{-3/2}$$

The unnormalized posterior for λ is likelihood (11.21) times prior (11.22)

$$h(\lambda \mid \mathbf{x}) \propto \lambda^{n/2-3/2} \exp\left\{ -\frac{\lambda}{2} \sum_{i=1}^{n} (x_i - \mu)^2 \right\}.$$

Considered as a function of λ (not the x_i) the right hand side must be an unnormalized density. Using the trick of recognizing an unnormalized brand name density, we see that the posterior distribution of λ is $\text{Gam}(a, b)$ with

$$\begin{aligned} a &= \frac{n-1}{2} \\ b &= \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2 \end{aligned}$$

Since the parameters a and b of the gamma distribution must be strictly positive, we need $n > 1$ in order to have a proper posterior ($b > 0$ is satisfied automatically). This check whether the posterior is integrable is always necessary when using an improper prior (and never necessary when using a proper prior). An improper posterior (one that doesn't integrate) is nonsense.

11.4.2 Conjugate Priors

The definition of *conjugate prior family* of distributions given on p. 247 in Lindgren, is fairly cryptic. A family $\mathcal{F}$ of probability distributions is conjugate to a probability model if the posterior is in $\mathcal{F}$ whenever the prior is in $\mathcal{F}$. How does one find such a family? One trivial example is the (nonparametric) family of *all* probability distributions on the parameter space. (The posterior is a probability distribution, hence is trivially in the family of all probability distributions.)

If that were all there was to the notion of conjugate families, it would be useless concept. The idea is to find a parametric conjugate family, one we recognize. Here is a recipe for such families, so-called *natural* conjugate families. If we have independent sampling, the likelihood satisfies

$$L_{x_1,\ldots,x_{m+n}}(\theta) \propto L_{x_1,\ldots,x_m}(\theta) L_{x_{m+1},\ldots,x_{m+n}}(\theta).$$

Thus we see that we can take a likelihood with some "made up" data as the prior and the unnormalized posterior will be the likelihood for the combination of real and "made up" data. In short, likelihoods form a conjugate family, so long as we include all sample sizes and all possible data values.

Usually we take a slightly larger family. If the likelihood is a well-defined positive function for noninteger values of the sample size, then we allow noninteger values. Similarly if the data are discrete, we also allow arbitrary data values so long as the resulting function is still well-defined and positive. It is clear that the result is a conjugate family so long as our possible "made up" data includes all possible actual data values.

Example 11.4.2 (Example 11.2.4 Continued).
In Example 11.2.4, we found a normal prior for μ resulted in a normal posterior. Thus family of normal distributions for the parameter μ is a conjugate prior family for a normal data model when the variance is known and μ is the only parameter.

11.4.3 The Two-Parameter Normal Distribution

The two-parameter normal model has data X_1, X_2, ... i. i. d. $\mathcal{N}(\mu, \sigma^2)$ with both μ and σ considered parameters. The likelihood is given by (10.4). As in Examples 11.2.4 and 11.4.1, the analysis becomes simpler if we use precision $\lambda = 1/\sigma^2$ as one of the parameters, giving

$$\begin{aligned} L_{\mathbf{x}}(\mu, \lambda) &= \lambda^{n/2} \exp\left\{-\tfrac{n}{2}\lambda[v_n + (\bar{x}_n - \mu)^2]\right\} \\ &= \lambda^{n/2} \exp\left\{-\tfrac{n}{2}\lambda[v_n + \bar{x}_n^2 - 2\bar{x}_n\mu + \mu^2]\right\} \end{aligned} \tag{11.23}$$

This has three bits of "made up data" to adjust: n, v_n, and $\bar{x}_n$. Replacing them with Greek letters α, β, and γ gives a conjugate family of priors with unnormalized densities

$$g(\mu, \lambda \mid \alpha, \beta, \gamma) = \lambda^{\alpha/2} \exp\left\{-\tfrac{1}{2}\alpha\lambda(\beta - 2\gamma\mu + \mu^2)\right\}. \tag{11.24}$$

Here α, β, and γ are hyperparameters of the prior, known constants, not random variables. Choosing the hyperparameters chooses a particular prior from the conjugate family to represent prior opinion about the parameters μ and λ.

The next task is to figure out the properties of the conjugate family we just discovered. With a little work we will be able to "factor" these distributions as joint = conditional × marginal and recognize the marginals and conditionals.

The conditional of μ given λ is clearly a normal distribution, because it is "e to a quadratic" (as a function of μ for fixed λ). To figure out which normal, we have to match up coefficients of powers of μ in the exponential. If $\mu \mid \lambda \sim \mathcal{N}(a, b)$, then we must have

$$\begin{aligned}(\mu - a)^2/b &= a^2/b - 2a\mu/b + \mu^2/b \\ &= \alpha\lambda(\beta - 2\gamma\mu + \mu^2) + \text{a constant}\end{aligned}$$

hence we can determine a and b by matching the coefficients of μ and μ^2 giving

$$a = \gamma \tag{11.25a}$$

$$b = \frac{1}{\alpha\lambda} \tag{11.25b}$$

To figure out the marginal of λ we have to do the "factorization" into conditional and marginal. The conditional $\mathcal{N}(a, b)$ with a and b given by (11.25a) and (11.25b) has density proportional to

$$b^{-1/2} \exp\left\{-\frac{1}{2b}(\mu - a)^2\right\} = \alpha^{1/2}\lambda^{1/2} \exp\left\{-\frac{1}{2}\alpha\lambda\left(\gamma^2 - 2\gamma\mu + \mu^2\right)\right\} \tag{11.26}$$

Thus the marginal of λ must have density proportional to (11.24) divided by (11.26), that is,

$$\lambda^{(\alpha-1)/2} \exp\left\{-\tfrac{1}{2}\alpha\lambda(\beta - \gamma^2)\right\}.$$

This is clearly proportional to a $\text{Gam}(c, d)$ density with

$$c = (\alpha + 1)/2 \tag{11.27a}$$

$$d = \alpha\left(\beta - \gamma^2\right)/2 \tag{11.27b}$$

Thus we have discovered that our conjugate family can be "factored" as a product of normal and gamma distributions. The connection between the shape parameter $(\alpha + 1)/2$ of the gamma and the precision $\alpha\lambda$ of the normal seems arbitrary. Thus one usually allows the two to be varied independently, which gives a family with four hyperparameters.

Definition 11.4.1 (The Normal-Gamma Family of Distributions). *If a random variable X has a* $\text{Gam}(\alpha, \beta)$ *distribution, and the conditional distribution of another random variable Y given $X = x$ is a* $\mathcal{N}(\gamma, \delta^{-1}x^{-1})$ *distribution, then we say the joint distribution of X and Y is* normal-gamma *with parameters α, β, γ, and δ. The parameter γ can be any real number, the rest must be strictly positive.*

Following the usual practice of making random variables Roman letters near the end of the alphabet, we have changed (λ, μ) to (X, Y) for this definition only. As we continue the Bayesian analysis we will go back to having the random variables being λ and μ. We have also redefined α, β, and γ and will no longer use the parameterization (11.24) for the normal-gamma family. The new parameterization given in the definition is standard.

Theorem 11.1. *If X_1, X_2, …, X_n are i. i. d. $\mathcal{N}(\mu, \lambda^{-1})$, then the normal-gamma family is a conjugate family of prior distributions for (λ, μ). If the prior distribution is*

$$\begin{aligned} \lambda &\sim \text{Gam}(\alpha_0, \beta_0) \\ \mu \mid \lambda &\sim \mathcal{N}(\gamma_0, \delta_0^{-1}\lambda^{-1}) \end{aligned}$$

then the posterior distribution is

$$\begin{aligned} \lambda &\sim \text{Gam}(\alpha_1, \beta_1) \\ \mu \mid \lambda &\sim \mathcal{N}(\gamma_1, \delta_1^{-1}\lambda^{-1}) \end{aligned}$$

where

$$\alpha_1 = \alpha_0 + \frac{n}{2} \tag{11.28a}$$

$$\beta_1 = \beta_0 + \frac{n}{2}\left(v_n + \frac{\delta_0(\bar{x}_n - \gamma_0)^2}{\delta_0 + n}\right) \tag{11.28b}$$

$$\gamma_1 = \frac{\gamma_0\delta_0 + n\bar{x}_n}{\delta_0 + n} \tag{11.28c}$$

$$\delta_1 = \delta_0 + n \tag{11.28d}$$

where $\bar{x}_n$ is the empirical mean and v_n is the empirical variance.

Proof. If (λ, μ) is normal-gamma with parameters α, β, γ, and δ, the unnormalized density is

$$\lambda^{\alpha-1}\exp\{-\beta\lambda\} \cdot \lambda^{1/2}\exp\left\{-\tfrac{1}{2}\delta\lambda(\mu-\gamma)^2\right\}. \tag{11.29}$$

Putting subscripts of zero on the hyperparameters in (11.29) and multiplying by the likelihood (11.23) gives the unnormalized posterior

$$\lambda^{\alpha_0+n/2-1/2}\exp\left\{-\beta_0\lambda - \tfrac{1}{2}\delta_0\lambda(\mu-\gamma_0)^2 - \tfrac{n}{2}\lambda(v_n + \bar{x}_n^2 - 2\bar{x}_n\mu + \mu^2)\right\} \tag{11.30}$$

To prove the theorem we have to show that this is equal to (11.29) with subscripts of one on the hyperparameters and that the relationship between the hyperparameters of prior and posterior is the one stated.

Comparing the exponent of λ in (11.29) and (11.30) gives (11.28a). The other three relationships between hyperparameters are found by equating the coefficients of λ, of $\lambda\mu$, and of $\lambda\mu^2$ in the exponential terms, which gives

$$-\beta_1 - \tfrac{1}{2}\delta_1\gamma_1^2 = -\beta_0 - \tfrac{1}{2}\delta_0\gamma_0^2 - \tfrac{n}{2}(v_n + \bar{x}_n^2) \tag{11.31a}$$

$$\gamma_1\delta_1 = \gamma_0\delta_0 + n\bar{x}_n \tag{11.31b}$$

$$-\tfrac{1}{2}\delta_1 = -\tfrac{1}{2}\delta_0 - \tfrac{n}{2} \tag{11.31c}$$

Equation (11.31c) immediately implies (11.28d). Plugging (11.28d) into (11.31b) gives (11.28c). Plugging (11.28c) and (11.28d) into (11.31a) gives

$$\beta_1 = \beta_0 + \tfrac{1}{2}\delta_0\gamma_0^2 + \tfrac{n}{2}(v_n + \bar{x}_n^2) - \tfrac{1}{2}\frac{(\gamma_0\delta_0 + n\bar{x}_n)^2}{\delta_0 + n}$$

which with a bit of formula manipulation is (11.28b). □

We also want to learn about the other "factorization" of the normal-gamma family into the marginal of μ times the conditional of λ given μ. This involves Student's t-distribution with noninteger degrees of freedom (Definition 7.3.2 in these notes).

Theorem 11.2. *If*

$$\begin{aligned} \lambda &\sim \operatorname{Gam}(\alpha, \beta) \\ \mu \mid \lambda &\sim \mathcal{N}(\gamma, \delta^{-1}\lambda^{-1}) \end{aligned}$$

then

$$\begin{aligned} (\mu - \gamma)/d &\sim t(\nu) \\ \lambda \mid \mu &\sim \operatorname{Gam}(a, b) \end{aligned}$$

where

$$\begin{aligned} a &= \alpha + \tfrac{1}{2} \\ b &= \beta + \tfrac{1}{2}\delta(\mu - \gamma)^2 \\ \nu &= 2\alpha \\ d &= \sqrt{\frac{\beta}{\alpha\delta}} \end{aligned}$$

Proof. The unnormalized joint density for μ and λ is given by (11.29). The conditional distribution of λ given μ is clearly the $\operatorname{Gam}(a, b)$ asserted by the theorem. This has density

$$\frac{b^a}{\Gamma(a)} \lambda^{a-1} e^{-b\lambda}. \tag{11.32}$$

We may ignore the factor $\Gamma(a)$, which is a constant, but we must keep b^a, which contains μ. The unnormalized marginal for μ is thus (11.29) divided by (11.32), which is b^{-a} or

$$h(\mu) = \frac{1}{[\beta + \frac{1}{2}\delta(\mu - \gamma)^2]^{\alpha+1/2}} \tag{11.33}$$

Hence we see that some linear function of μ has a t distribution with $\nu = 2\alpha$ degrees of freedom. To determine the linear function we must equate coefficients of powers of μ in

$$\beta + \tfrac{1}{2}\delta(\mu - \gamma)^2 = k\left(1 + \frac{(\mu - c)^2}{d^2\nu}\right)$$

which is derived by plugging in $(\mu - c)/d$ for x in (7.32) and matching the $1 + x^2/\nu$ term in the denominator to the term in square brackets in (11.33). In order for $(\mu - c)/d$ to have a $t(\nu)$ distribution, these two terms must be proportional, hence equal when multiplied by k, a yet to be determined constant of proportionality.

Hence

$$\begin{aligned}
\beta + \frac{\delta \gamma^2}{2} &= k + \frac{k c^2}{d^2 \nu} \\
- \delta \gamma &= - \frac{2 k c}{d^2 \nu} \\
\frac{\delta}{2} &= \frac{k}{d^2 \nu}
\end{aligned}$$

Solving for k, c, and d we get $c = \gamma$, $k = \beta$, and

$$d^2 = \frac{2 \beta}{\delta \nu} = \frac{\beta}{\delta \alpha}$$

which finishes the proof of the theorem. □

Example 11.4.3 (Improper Prior for the Normal).
Bayesian inference for the two-parameter normal is quite complicated. Choosing a conjugate prior involves specifying four hyperparameters. The hyperparameters of the posterior are complicated functions of the hyperparameters of the prior and the data.

Here we analyze a simple case. Choose $\beta = \delta = 0$ in (11.29). Then the unnormalized prior is just $\lambda^{\alpha - 1/2}$. This is, of course, an improper prior. The posterior is normal-gamma with parameters

$$\begin{aligned}
\alpha_1 &= \alpha + \frac{n}{2} \\
\beta_1 &= \frac{n v_n}{2} \\
\gamma_1 &= \bar{x}_n \\
\delta_1 &= n
\end{aligned}$$

and is a proper distribution so long as $\alpha > -n/2$.

The posterior marginal distribution of λ is $\text{Gam}(\alpha_1, \beta_1)$ and the posterior marginal distribution of $(\mu - \bar{x}_n)/\sqrt{v_n/\nu}$ is $t(\nu)$, where $\nu = 2\alpha_1 = n + 2\alpha$.

Suppose we decide on the value $\alpha = -\frac{1}{2}$ for the remaining hyperparameter. Then $\nu = n - 1$. And since $v_n/(n-1) = s_n^2/n$, we get a marginal posterior distribution

$$\frac{\mu - \bar{x}_n}{s_n/\sqrt{n}} \sim t(n-1)$$

Thus for this particular improper prior, the marginal posterior distribution of this quantity agrees with its sampling distribution.

The agreement of Bayesian posterior and frequentist sampling distributions leads to numerically identical though philosophically different inferences. But no great message should be read into this. No Bayesian with a proper prior would get the "same" inference as the frequentist. A Bayesian with a subjective prior would not get the same inference, because subjective priors representing prior knowledge about the parameters are supposed to be proper.

11.5 Bayesian Point Estimates

Most Bayesians are not much interested in point estimates of parameters. To them a parameter is a random variable, and what is important is its distribution. A point estimate is a meager bit of information as compared, for example, to a plot of the posterior density.

Frequentists too are not much interested in point estimates for their own sake, but frequentists find many uses for point estimates as tools for constructing tests and confidence intervals. All asymptotic arguments start with calculating the asymptotic distribution of some point estimate. They may also require a point estimate of the asymptotic standard deviation for use in the "plug-in" theorem. Bayesians do not need point estimates for any of these purposes. All Bayesian inference starts with calculating the posterior distribution. To go from there to a point estimate is to throw away most of the information contained in the posterior.

Still, point estimates are easy to calculate (some of them, at least) and easy to discuss. So they are worth some study. Bayesians use three main kinds of point estimates: the posterior mean, median, and mode. The first two we have already met.

Definition 11.5.1 (Mode).
A mode *of a random variable having a continuous density is a local maximum of the density. The variable is* unimodal *if it has one mode,* bimodal *if two, and* multimodal *if more than one.*

When we say the *mode (rather than* a *mode) in reference to a multimodal distribution, we mean the highest mode (if one is higher than the others).*

All of the brand name continuous distributions introduced in Chapter 6 in Lindgren are unimodal. The normal distribution is unimodal, and the mode is the mean. In fact this is obviously true (draw a picture) for any symmetric unimodal distribution.

> *For a symmetric unimodal distribution, the mean (if it exists), the median, and the mode are all equal to the center of symmetry.*

The gamma distribution, and its special cases the exponential and chi-square, are not symmetric, but are unimodal. For them the mean, median, and mode are three different points.

Example 11.5.1 (Mode of the Gamma Distribution).
It does not matter if we use an unnormalized density. Multiplying by a constant changes only the vertical scale, not the position of the mode. An unnormalized $\text{Gam}(\alpha, \lambda)$ density is

$$f(x) = x^{\alpha-1} e^{-\lambda x}.$$

As in maximum likelihood, it is often easier to maximize the log density, which must have the same mode. The log density is

$$g(x) = \log f(x) = (\alpha - 1) \log(x) - \lambda x$$

Differentiating gives

$$g'(x) = \frac{\alpha - 1}{x} - \lambda \tag{11.34}$$

$$g''(x) = -\frac{\alpha - 1}{x^2} \tag{11.35}$$

From (11.35) we see that the log density is strictly concave, hence the local maximum is unique if it exists. Solving $g'(x) = 0$, we get

$$x = (\alpha - 1)/\lambda \tag{11.36}$$

for the mode when $\alpha \geq 1$. When $\alpha < 1$ (11.36) is negative and hence not in the sample space. In that case (11.34) is negative for all x, hence $g(x)$ is strictly decreasing and the only local maximum occurs at $x = 0$. The mode is a bit weird because $f(x) \to \infty$ as $x \to 0$ in this case. But we still call it the mode.

Frequentists are not much interested in the mode as a point estimate of location, because it is very hard to estimate and may be far from the main mass of the distribution, even when the distribution is unimodal (but not symmetric). For example, consider the $\text{Exp}(\lambda)$ distribution. The mean is $1/\lambda$, the median is $\log(2)/\lambda = 0.693/\lambda$ (Problem 6-47 in Lindgren), and the mode is zero.

Bayesians are interested in the posterior mode because of its analogy to maximum likelihood. As we saw in the preceding example, it does not matter if we use an unnormalized objective function in determining the mode, since normalization only changes the vertical scale and does not change the position of the mode. An unnormalized posterior is likelihood times prior. Thus we find the posterior mode by maximizing $L_{\mathbf{x}}(\theta)g(\theta)$, considered as a function of θ for fixed x. If we use a flat prior, this is the same as maximum likelihood. If we do not use a flat prior, then the posterior mode will be different from the MLE. But in either case the posterior mode can be calculated directly from the unnormalized posterior $L_{\mathbf{x}}(\theta)g(\theta)$. There is no need to calculate the normalizing constant, integral in (11.9), if all we want is the posterior mode.

Example 11.5.2 (Example 11.2.1 and Example 11.2.2 Continued).
In Example 11.2.2 we found a $\text{Gam}(a + 1, b + x)$ posterior distribution, where x was the data and a and b hyperparameters of the prior. The mode of this distribution, hence the posterior mode is given by (11.36) with $a+1$ plugged in for α and $b + x$ plugged in for λ. Hence the posterior mode is

$$\lambda^* = \frac{a}{b + x}$$

For comparison the MLE is

$$\hat{\lambda} = \frac{1}{x}$$

and this is the posterior mode for a flat prior (rather than the gamma prior used in Examples 11.2.1 and 11.2.2). The posterior mean is

$$E(\lambda \mid x) = \frac{a + 1}{b + x}$$

The posterior median is hard to calculate. There is no closed form expression for it as a function of a, b, and x. For any fixed values of a, b, and x, we could use tables of the incomplete gamma function (not in Lindgren but in reference books) or a computer statistics package to calculate the posterior median, but we cannot exhibit a formula like those for the other estimators.

11.6 Highest Posterior Density Regions

This section covers the Bayesian competitor to confidence intervals, which go under the name "highest posterior density regions." A *highest posterior density (HPD) region* is a level set of the posterior, that is a set of the form

$$\{\, \theta : h(\theta \mid \mathbf{x}) > \alpha \,\}$$

for some $\alpha > 0$, that has a specified posterior probability, i. e.,

$$P\{h(\theta \mid \mathbf{X}) > \alpha \mid \mathbf{X} = \mathbf{x}\} = \beta$$

Note that, *as always when we are being Bayesians,* we are conditioning on the data $\mathbf{X}$, what is random here is the parameter θ. The idea behind the HPD region is that all of the points included in the region should be more probable (in the sense of higher posterior density) than those not in the region.

Example 11.6.1 (Examples 11.2.4 and 11.4.2 Continued).
In Example 11.2.4 we saw that if the data are i. i. d. normal with mean μ and precision λ with λ known and the prior for μ was normal with mean μ_0 and precision λ_0, then the posterior is normal with mean (11.17b) and precision (11.17a). By the symmetry of the normal distribution, the 95% HPD region is a symmetric interval centered at the posterior mean. The same logic we use to figure out critical values for confidence intervals tells us the half width of the interval is 1.96 posterior standard deviations, that is, the 95% HPD region for μ is

$$\frac{n\lambda\bar{x}_n + \lambda_0\mu_0}{n\lambda + \lambda_0} \pm 1.96\sqrt{\frac{1}{n\lambda + \lambda_0}}$$

(recall that $n\lambda + \lambda_0$ is the *precision* not the variance, so the standard deviation is the square root of its reciprocal).

Comparing this with the frequentist 95% confidence interval, which is

$$\bar{x}_n \pm 1.96\sqrt{\frac{1}{n\lambda}}$$

(recall that $\sigma^2 = 1/\lambda$), we see that in general the two may be rather different, although they do become very close in the limit as $\lambda_0 \to 0$. The case $\lambda_0 = 0$ does not correspond to any normal prior, but is the what results from using a flat, improper prior (Problem 7-82 in Lindgren). Thus the frequentist and the Bayesian produce the same interval, albeit with different philosophical interpretation, when (and only when) the Bayesian uses the flat improper prior.

Otherwise, they disagree. The disagreement will be slight if the Bayesian's prior is very diffuse (this means the prior variance is very large, hence the prior precision λ_0 is very small). If the Bayesian's prior is fairly precise, the disagreement may be substantial and the 95% HPD region much shorter than the 95% confidence interval.

Example 11.6.2 (Marginal t Posterior for μ).
When the data are i. i. d. normal with both mean and variance unknown parameters and we use a conjugate prior, then Theorems 11.1 and 11.2 tell us that the marginal posterior for μ is a location-scale transform of a t distribution with noninteger degrees of freedom. More precisely, Theorem 11.2 says that $(\mu - \gamma)/d$ has a $t(\nu)$ distribution, where γ is a hyperparameter of the posterior and d and ν are defined (in the theorem) in terms of the other hyperparameters of the posterior α, β, and δ, and Theorem 11.1 gives the relation between the hyperparameters of the posterior and the hyperparameters of the prior and the data.

What is new in this example is that we want to figure out the HPD region for μ. This is easily done by the same logic that gives frequentist confidence intervals. By the symmetry of the t distribution, the HPD region is the set of μ values satisfying

$$\left| \frac{\mu - \gamma}{d} \right| < t_{\alpha/2}$$

where $t_{\alpha/2}$ is the $1 - \alpha/2$ quantile of the $t(\nu)$ distribution. So the HPD region is $\gamma \pm t_{\alpha/2} d$.

This resembles the frequentist confidence interval in being

$$\text{something} \pm t_{\alpha/2} \times \text{something else},$$

but the "something" is not $\overline{X}_n$, the "something else" is not $S_n/\sqrt{n}$ and the degrees of freedom ν is not $n - 1$ except for the particular improper prior used in Example 11.4.3.

Calculating HPD regions is not so easy when the posterior is not symmetric. Then it is generally necessary to do a computer search to find the endpoints of the region.

Example 11.6.3.
This continues Example 8.8b in Lindgren, which in class we gave both exact and asymptotic "frequentist" analyses. The data X_1, $\ldots$, X_n are i. i. d. Exp$(1/\theta)$. In order to be Bayesians we need a prior for θ, which we take to be $g(\theta) = 1/\theta$, an improper prior. The likelihood is

$$L(\theta) = \theta^{-n} \exp\left\{ -\frac{1}{\theta} \sum_{i=1}^{n} X_i \right\}$$

so the unnormalized posterior is

$$h(\theta) = L(\theta) g(\theta) = \theta^{-n-1} \exp\left\{ -\frac{1}{\theta} \sum_{i=1}^{n} X_i \right\} \tag{11.37}$$

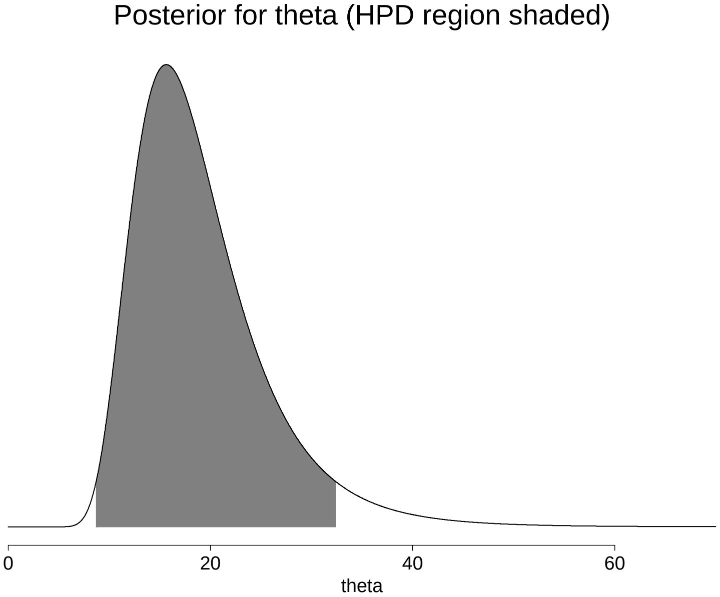
Lindgren would write $h(\theta \mid \mathbf{X})$ but we will temporarily suppress the data in the notation.

This is not a recognizable distribution, but it looks like $\lambda = 1/\theta$ has a gamma distribution. Let's check. Solving for θ the change of variable is $\theta = 1/\lambda = w(\lambda)$. The derivative of this transformation is $w'(\lambda) = -1/\lambda^2$, hence the unnormalized posterior for λ is

$$\begin{aligned} h[w(\lambda)] \cdot |w'(\lambda)| &= \lambda^{n+1} \exp\left\{ -\lambda \sum_{i=1}^{n} X_i \right\} \cdot \frac{1}{\lambda^2} \\ &= \lambda^{n-1} \exp\left\{ -\lambda \sum_{i=1}^{n} X_i \right\} \end{aligned}$$

which we recognize as $\text{Gam}\,(n, \sum_i X_i)$. Thus we can use the known distribution of λ to calculate probabilities about θ.

Finding the HPD region is not so easy. There is no way to calculate the endpoints or look them up in a table. There is a simple method using a computer. Plot the unnormalized posterior (11.37). For a specific numerical example we used before $\sum_i X_i = 172.0$. The unnormalized posterior for θ is the curve plotted below.



The shaded area is the probability of the HPD region. The region itself is the range of θ values covered $(8.70, 32.41)$. The posterior probability of the HPD region is 95% (this is the Bayesian analog of the "confidence" in a confidence interval).

The HPD region was determined from the plot as follows. The curve is actually plotted on a grid of points, spaced .01 apart on the θ axis. The sum of the y-values for these points approximates the integral for the normalizing constant of the unnormalized density $h(\theta)$ given by (11.37). The points with the highest y-values that constitute 95% of the sum are easily found by the computer and give a good approximation to the HPD region. The actual probability of the region, calculated using the gamma distribution of λ, is 94.99% (pretty close), and the heights of the unnormalized density (11.37) at the endpoints are 1.200×10^{-19} and 1.19610^{-19}. So we have come pretty close to a level set of the posterior.

11.7 Bayes Tests

The Bayesian view of one-tailed tests is fairly straightforward. Strangely, two-tailed tests, which the frequentist finds to be a minor variant of one-tailed tests, the Bayesian finds incredibly complicated and somewhat bizarre, so much so that Lindgren just avoids the subject. He blames the problem, calling it a "mathematical idealization," which is a meaningless criticism since so is everything else in statistics and every other mathematical subject.

A Bayesian one-tailed test is simple. The null and alternative hypotheses, being subsets of the parameter space are events, because the Bayesian considers parameters to be random variables. Hence they have probabilities (both prior and posterior). The test is done by calculating the posterior probabilities of the hypotheses and seeing which is bigger.

Example 9.5a in Lindgren provides an example of this. The data $Y \sim \text{Bin}(n, p)$ with $n = 15$ and the observed value of the data $y = 12$. The parameter p is unknown and given a $\mathcal{U}(0, 1)$ prior distribution. The hypotheses are

$$\begin{aligned} H_0 &: p \le \tfrac{1}{2} \\ H_A &: p > \tfrac{1}{2} \end{aligned}$$

Lindgren calculates $P(H_0 \mid Y = 12) = 0.0106$ and hence by the complement rule $P(H_A \mid Y = 12) = 1 - P(H_0 \mid Y = 12) = 0.9894$.

For comparison, Lindgren gives the P-value of the frequentist test, which is $P(Y \ge 12 \mid p = \frac{1}{2}) = 0.0176$. Both the Bayesian and frequentist tests strongly favor the alternative by conventional standards of evidence, and the P-value and Bayesian posterior probability of the null hypothesis are fairly similar, though different in philosophical interpretation. The frequentist says "probability of the null hypothesis" is a meaningless phrase because parameters are not random. The Bayesian says this probability exists and is 0.0106. The P-value is quite a different philosophical animal. It is the probability of seeing data at least as extreme as the actual observed data under the assumption that the null hypothesis is true. As we saw with confidence intervals and HPD regions, the numbers are slightly different, but the philosophical interpretations are wildly different.

The Bayesian two-tailed test runs into a problem. The null and alternative hypotheses are still subsets of the parameter space, hence are still events (to the Bayesian), and hence still have probabilities. The trouble is that when the hypotheses are

$$\begin{aligned} H_0 &: p = \tfrac{1}{2} \\ H_A &: p \neq \tfrac{1}{2} \end{aligned}$$

and the prior is continuous, the null hypothesis, being a single point, has probability zero. Thus if we use the same prior as we used for the one-tailed test, or any continuous prior, the posterior probability will be zero. But this is only because the prior probability is zero. As we saw in Problem 7-84 in Lindgren, whenever the prior probability is zero, the posterior probability will be zero too. So we haven't learned anything by doing such a test. Our mind was made up before we observed any data that H_0 was impossible and no data can change our minds. Might as well not bother to collect data or analyze it.

As Lindgren says on p. 313 a way out of this dilemma is to make the null hypothesis an interval, say

$$\begin{aligned} H_0 &: \tfrac{1}{2} - \epsilon \le p \le \tfrac{1}{2} + \epsilon \\ H_A &: p < \tfrac{1}{2} - \epsilon \text{ or } \tfrac{1}{2} + \epsilon < p \end{aligned}$$

for some $\epsilon > 0$. But this only adds to the problems. True the prior and posterior probabilities are now no longer zero, but where did ϵ come from? This "solution" has raised more questions than it answers. Furthermore, the posterior probability will still converge to zero if we let ϵ go to zero (by continuity of probability, Theorem 4 of Chapter 2 in Lindgren) so our analysis will depend very strongly on the choice of ϵ. We've only added to our troubles in finding a sensible Bayesian analysis.

The choice of ϵ is so problematic that most Bayesians that bother with two-tailed tests at all use a different solution to the dilemma. It is also weird, but less weird. The solution is to choose a prior that is not continuous, and puts some probability on the point null hypothesis $\Theta_0 = \{\frac{1}{2}\}$. For example, continuing the use of a uniform prior as much as possible, consider the prior distribution defined as follows

$$\begin{aligned} P(H_0) &= \alpha \\ P(H_A) &= 1 - \alpha \\ p \mid H_A &\sim \mathcal{U}(0, 1) \end{aligned}$$

This is a mixture of a distribution concentrated at one point ($\frac{1}{2}$) and a uniform distribution.

Allowing such a prior takes us out of the theory we know. If the prior is continuous, we calculate expectations, marginals, etc. by integrating. If it is discrete, by summing. If it is neither discrete nor continuous, we don't know what to do. Fortunately we can describe what to do in this very simple case where the distribution is continuous except for a single atom and avoid the complexity of the general situation.

In order to apply Bayes rule, we need to calculate the marginal probability of the observed data $P(Y = y)$ in the binomial example. We can do the calculation in two parts, using what Lindgren calls the "law of total probability" (Theorem 3 of Chapter 2), which in this case says

$$P(Y = y) = P(Y = y \,\text{and}\, p = \tfrac{1}{2}) + P(Y = y \,\text{and}\, p \neq \tfrac{1}{2}).$$

First

$$\begin{aligned} P(Y = y \,\text{and}\, p = \tfrac{1}{2}) &= P(Y = y \mid p = \tfrac{1}{2}) P(p = \tfrac{1}{2}) \\ &= \alpha \binom{n}{y} \left(\frac{1}{2}\right)^y \left(\frac{1}{2}\right)^{n-y} \\ &= \alpha \binom{n}{y} \left(\frac{1}{2}\right)^n \end{aligned}$$

Second

$$\begin{aligned} P(Y = y \,\text{and}\, p \neq \tfrac{1}{2}) &= P(Y = y \mid p \neq \tfrac{1}{2}) P(p \neq \tfrac{1}{2}) \\ &= (1 - \alpha) \binom{n}{y} \int_0^1 p^y (1 - p)^{n-y} \, dp \\ &= (1 - \alpha) \binom{n}{y} B(y + 1, n - y + 1) \\ &= \frac{1 - \alpha}{n + 1} \end{aligned}$$

Putting these together, the marginal probability is

$$P(Y = y) = \alpha \binom{n}{y} \left(\frac{1}{2}\right)^n + \frac{1 - \alpha}{n + 1}$$

Hence the probability of H_0 is

$$P(p = \tfrac{1}{2} \mid Y = y) = \frac{P(p = \tfrac{1}{2} \,\text{and}\, Y = y)}{P(Y = y)} = \frac{\alpha \binom{n}{y} \left(\frac{1}{2}\right)^n}{\alpha \binom{n}{y} \left(\frac{1}{2}\right)^n + \frac{1-\alpha}{n+1}} \qquad (11.38)$$

That's how a Bayesian two-tailed test is done.

Some practical or philosophical (non-mathematical anyway) issues remain. The probability $P(H_0)$ given by (11.38) still depends strongly on the prior probability of H_0 (that is, α). This means that no two Bayesians will produce the same answer, since each will have a different prior probability (they will agree on the formula but plug in different numbers for α).

In order to eliminate this source of disagreement, we need a new notion, which is called the "Bayes factor" for the test. It is the ratio of the posterior to prior odds. Recall that *odds* are probabilities expressed as a ratio rather than a fraction. If the probability of an event is p, then the odds are $p/(1 - p)$. Here the prior odds of H_0 are $\alpha/(1 - \alpha)$ and the posterior odds are

$$\frac{P(p = \frac{1}{2} \mid Y = y)}{P(p \neq \frac{1}{2} \mid Y = y)} = \frac{\binom{n}{y} \left(\frac{1}{2}\right)^n}{\frac{1}{n+1}} \cdot \frac{\alpha}{1 - \alpha}$$

Hence the Bayes factor is the first term on the right hand side. Notice that it does not depend on α at all, although it still does depend on prior probabilities, since it depends on the choice of a prior that is uniform on the alternative hypothesis. The Bayes factor eliminates some, but not all, of the dependence on the prior.

Now let us plug in a few numbers, to get a concrete example. Continuing the example above with observed data $y = 12$, the Bayes factor is

$$\binom{15}{12} \left(\frac{1}{2}\right)^{15} \cdot (15 + 1) = 0.2221$$

For comparison, the two-tailed P-value is twice the one tailed $P = 0.035$ (because the distribution of the test statistic Y is symmetric under H_0, the binomial distribution is only symmetric when $p = \frac{1}{2}$ but that's what H_0 asserts).

Notice the big difference between the Bayesian and frequentist analyses. Frequentists are impressed with the evidence against H_0, at least those frequentists who think $P < 0.05$ implies "statistical significance." Bayesians are unimpressed. The data only lower the odds in favor of H_0 by a factor between 4 and 5 ($1/0.2221 = 4.5$). If the prior odds in favor of H_0 were even (1 to 1), then the posterior odds in favor of H_0 are now 0.222, and the posterior probability of H_0 is $0.222/(1 + 0.222) = .182$, still almost one chance in 5 that H_0 is true.

It shouldn't be any surprise that the frequentist and Bayesian answers turn out so different. They don't purport to resemble each other in any way. The only connection between the two is that they are competitors, different approaches to the same issue, saying something about whether H_0 or H_A is correct. The situation we saw here is typical, the Bayesian is always less impressed by evidence against H_0 and "accepts" H_0 less often than the frequentist.[3] This gives the Bayesians a problem with selling Bayes factors. Users of tests generally want to reject H_0. They didn't collect their data with the idea that there was nothing interesting in it (which is what H_0 usually says). Thus they are reluctant to switch to a procedure that makes rejecting H_0 even harder. Of course the Bayesian argues that frequentist tests are too lenient, but since frequentist tests are widely accepted and everyone is used to them, this sales job is an uphill battle.

Now let us go back and redo the calculation above abstractly so we get a general formula for the Bayes factor. Suppose we are doing a problem with likelihood $L_x(\theta)$ and put prior probability α on the point null hypothesis $H_0 : \theta = \theta_0$ and $1 - \alpha$ on the alternative hypothesis $H_A : \theta \neq \theta_0$ distributed according to the conditional density $g(\theta)$. Unlike the situation in most Bayesian inference, we must have $g(\theta)$ a proper probability density (not improper, not unnormalized).

As in the example above, the marginal probability of the data is

$$\begin{aligned} P(X = x) &= P(X = x \,\text{and}\, H_0) + P(X = x \,\text{and}\, H_A) \\ &= P(X = x \mid H_0)\alpha + P(X = x \mid H_A)(1 - \alpha). \end{aligned}$$

[3] Berger and Sellke, "Testing a point null hypothesis: The irreconcilability of P values and evidence" (with discussion), *Journal of the American Statistical Association*, 82:112-122, 1987.

The posterior probability of H_0 is

$$\begin{aligned} P(H_0 \mid X = x) &= \frac{P(X = x \,\text{and}\, H_0)}{P(X = x)} \\ &= \frac{P(X = x \mid H_0)\alpha}{P(X = x \mid H_0)\alpha + P(X = x \mid H_A)(1 - \alpha)} \end{aligned}$$

and the posterior odds in favor of H_0 are

$$\frac{P(X = x \mid H_0)}{P(X = x \mid H_A)} \cdot \frac{\alpha}{1 - \alpha}$$

Thus the Bayes factor is the first term above, the ratio of prior to posterior odds

$$\frac{P(X = x \mid H_0)}{P(X = x \mid H_A)}$$

To proceed we need the density of the data, which as always is proportional to the likelihood, $f(x \mid \theta) = c(x) L_x(\theta)$. Then

$$P(X = x \mid H_0) = c(x) L_x(\theta_0)$$

and

$$P(X = x \mid H_A) = c(x) \int L_x(\theta) g(\theta) \, d\theta$$

So the Bayes factor in favor of H_0 is

$$\frac{L_x(\theta_0)}{\int L_x(\theta) g(\theta) \, d\theta}$$

Notice several things. First, the factor $c(x)$ appears in both the numerator and denominator of the Bayes factor and hence cancels, not appearing in the result. Second, the prior on the alternative $g(\theta)$ appears *only* in the denominator. That's why it must be a proper density. If it were unnormalized or improper, that would introduce an arbitrary constant that would not cancel into the Bayes factor, rendering it meaningless.

11.8 Bayesian Asymptotics

For large sample sizes, frequentist and Bayesian procedures (most of them anyway) give approximately the same answers. This is the result of a theorem that we will not state precisely. Under the same conditions required for the usual asymptotics of maximum likelihood plus one additional condition (that usually holds, but we won't describe since it is fairly technical) the asymptotic posterior distribution is "the same" as the asymptotic sampling distribution of the MLE. We put "the same" in quotes because the philosophical interpretation is radically different, but the asymptotic distribution is the same in both cases. Here's what we mean. The asymptotic sampling distribution of the MLE $\hat{\theta}_n$ is

$$\hat{\theta}_n \approx \mathcal{N}\left(\theta, \frac{1}{I_n(\theta)}\right)$$

where θ is the true parameter value. Of course we don't know θ (that's why we are estimating it) so we don't know the asymptotic variance $1/I_n(\theta)$. But we can consistently estimate it, plugging in $\hat{\theta}_n$ for θ giving

$$\hat{\theta}_n \approx \mathcal{N}\left(\theta, \frac{1}{I_n(\hat{\theta}_n)}\right) \tag{11.39}$$

Equation (11.39) is fairly sloppy notation. Strictly speaking, we should write

$$\sqrt{I_n(\hat{\theta}_n)}\bigl(\hat{\theta}_n - \theta\bigr) \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1) \tag{11.40}$$

but it is clear what is meant. The asymptotic posterior distribution of the parameter θ is

$$\theta \approx \mathcal{N}\left(\hat{\theta}_n, \frac{1}{I_n(\hat{\theta}_n)}\right) \tag{11.41}$$

Comparing with (11.39) we see that they differ only in the interchange of θ and $\hat{\theta}_n$. The frequentist considers θ fixed and $\hat{\theta}_n$ random and the asymptotic sampling distribution of $\hat{\theta}_n$ to be a normal distribution centered at the unknown true parameter value θ. The Bayesian considers $\hat{\theta}_n$ fixed (Bayesians *condition* on the data) and θ random and the asymptotic posterior distribution of θ to be a normal distribution centered at the MLE $\hat{\theta}_n$.

It is an important point that the asymptotic posterior distribution does *not* depend on the prior distribution of the parameter so long as the prior density is continuous and nonzero at the true parameter value. The catch phrase that expresses this is that the likelihood "outweighs" the prior for large sample sizes. Thus for large (perhaps very large) sample sizes all Bayesians agree (priors don't matter) and they also agree with the frequentists.

At least they agree about most things. Frequentist asymptotic confidence intervals will also be Bayesian asymptotic HPD regions. Frequentist asymptotic P-values for one-tailed tests will also be Bayesian asymptotic posterior probabilities of the null hypothesis for the same tests. One thing that will stay different is two-tailed tests. For them the posterior probabilities do not go away asymptotically and the frequentist and Bayesian do not get the same results no matter how large the sample size.

Problems

11-1. Suppose we observe $X \sim \text{Poi}(\mu)$ and we want to do a Bayesian analysis with prior distribution $\text{Gam}(\alpha, \beta)$ for μ, where α and β are known numbers expressing our prior opinion about probable values of μ.

(a) Find the posterior distribution of μ.

(b) Find the posterior mean of μ.

11-2. Suppose X is a single observation from a $\text{Gam}(\alpha, \lambda)$ distribution, where α is a known constant. Suppose our prior distribution for λ is $\text{Gam}(\alpha_0, \lambda_0)$, where the hyperparameters α_0 and λ_0 are also known constants.

(a) Find the posterior distribution for λ given X.

(b) Find the posterior mean $E(\lambda \mid X)$.

(c) Find the posterior mode of λ.

11-3. Using the same improper prior as was used in Example 11.4.3, show that the posterior marginal distribution of $(n-1)S_n^2\lambda$ is the same as its sampling distribution. More precisely stated, show that the frequentist sampling distribution of $(n-1)S_n^2\lambda$ with λ considered a nonrandom constant is the same as the Bayesian marginal posterior distribution of $(n-1)S_n^2\lambda$ with λ considered random and $S_n^2 = s_n^2$ fixed at the observed value.

11-4. Find the posterior mean and variance of μ when the data are i. i. d. normal and the prior is a general normal-gamma prior. Say for which values of the hyperparameters the posterior mean and variance of μ exist.

11-5. Suppose X_1, $\ldots$, X_n are i. i. d. $\mathcal{N}(\mu, 4)$, the prior distribution for μ is $\mathcal{N}(10, 9)$, and the sample mean of a sample of size 10 is $\overline{X}_n = 12$. Calculate a 90% HPD region for μ (note not 95%).

11-6. Suppose $X_1, \ldots, X_n$ are i. i. d. $\mathcal{N}(\mu, \lambda^{-1})$, the prior distribution for (μ, λ) is the conjugate normal-gamma prior with

$$\begin{aligned} \lambda &\sim \text{Gam}(3, 3) \\ \mu \mid \lambda &\sim \mathcal{N}(10, 16\lambda^{-1}) \end{aligned}$$

the sample mean of a sample of size 15 is $\overline{X}_n = 12$ and the sample variance is $S_n^2 = 50$ (note not V_n). Calculate a 95% HPD region for μ.

11-7. Suppose $X \sim \text{Bin}(n, p)$, where p is an unknown parameter. Find a formula giving the Bayes factor for the two-tailed test of

$$\begin{aligned} H_0 &: p = p_0 \\ H_A &: p \neq p_0 \end{aligned}$$

when the prior distribution for p given H_A is $\text{Beta}(s, t)$, where s and t are known constants. Hint: this is just like the test worked out in Section 11.7 except for the prior.

11-8. Suppose X_1, …, X_n are i. i. d. $\mathcal{N}(\mu, \lambda^{-1})$, where μ is an unknown parameter and the precision λ is a known constant. Find a formula giving the Bayes factor for the two-tailed test of

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_A &: \mu \neq \mu_0 \end{aligned}$$

when the prior distribution for μ given H_A is $\mathcal{N}(\mu_0, \lambda_0^{-1})$, where μ_0 and λ_0 are known constants.

11-9. Suppose the setup described at the end of Section 11.3. Verify that the posterior (11.19) from the two-stage analysis described in that section is the same as the posterior from analyzing all the data at once, which would be (11.18) with m replaced by n.

11-10. Suppose X_1, X_2, … X_n are i. i. d. from the distribution with density

$$f(x) = \theta x^{-\theta - 1}, \qquad x > 1,$$

where $\theta > 0$ is an unknown parameter. Suppose our prior distribution for the parameter θ is $\text{Exp}(\lambda)$, where λ is a known number (hyperparameter of the prior).

(a) Find the posterior density of θ.

(b) Find the posterior mean of θ.

(c) Find the posterior mode of θ.

Chapter 12

Regression

12.1 The Population Regression Function

12.1.1 Regression and Conditional Expectation

Recall from last semester (Section 3.3.2 of these notes) that "regression function" is another name for conditional expectation. Recall that a conditional expectation *is not* a function of the variable or variables "in front of the bar" and *is* a function of the variable or variables "behind the bar." Thus $E(Y \mid X)$ is not a function of Y and is a function of X, so we can write

$$h(X) = E(Y \mid X).$$

This function h is an ordinary function. When we wish to emphasize this and write it as a function of an ordinary variable, we write

$$h(x) = E(Y \mid x),$$

but the meaning is the same in either case. This function h is called the *regression function of Y on X*, the reason for the long name being that

$$g(Y) = E(X \mid Y)$$

defines a different function, the *regression function of X on Y*.

When we developed this terminology last semester, we had not begun systematic study of random vectors. Now we want to generalize this to allow a vector variable "behind the bar" leaving the variable in "front of the bar" a scalar. Then the regression function is a scalar function of a vector variable

$$h(\mathbf{X}) = E(Y \mid \mathbf{X})$$

which we can also think of as a function of several variables

$$h(X_1, \ldots, X_k) = E(Y \mid X_1, \ldots, X_k).$$

12.1.2 Best Prediction

There is a connection between conditional expectation (or the regression function) and prediction, which is given by the following theorem, which is Theorem 3.6 in last semester's notes improved to have a random vector "behind the bar." The proof is exactly the same as for Theorem 3.6 except for boldface type for $\mathbf{X}$, which does not make any essential difference.

Theorem 12.1 (Best Prediction). *For predicting a random variable Y given the value of a random vector $\mathbf{X}$, the predictor function $a(\mathbf{X})$ that minimizes the expected squared prediction error*

$$E\{[Y - a(\mathbf{X})]^2\}$$

is the conditional expectation $a(\mathbf{X}) = E(Y \mid \mathbf{X})$.

This theorem is analogous to theorem about the characterization of the mean (Corollary 7.2 in these notes). Together these two theorems say

- The best estimate of the the value of a random variable Y, where "best" means minimizing expected squared prediction error, is the mean $E(Y)$, when no other information is available.

- The best estimate of the the value of a random variable Y given the value of a random vector $\mathbf{X}$, where "best" means minimizing expected squared prediction error, is the conditional mean $E(Y \mid \mathbf{X})$.

The theorem gives yet another name for $E(Y \mid \mathbf{X})$. In addition to *conditional expectation* and the *regression function*, we also call it the *best predictor* (BP). Sometimes the best predictor is called the *best unbiased predictor* (BUP) because it is unbiased in the sense that its expectation is the mean of Y. This is a consequence of the iterated expectation property (Axiom CE2 for conditional expectation in Chapter 3 of these notes).

$$E\{E(Y \mid \mathbf{X})\} = E(Y).$$

Since the best predictor is always unbiased, it is irrelevant whether or not you bother to mention that it is unbiased. BP and BUP mean the same thing.

We give no examples because our interest in BP is mostly abstract. If you know the regression function, then you use it to give the best prediction. But when we are doing statistics, we don't know the regression function, because it depends on the true distribution of the data, and that depends on unknown parameters. Thus when doing statistics, the regression function isn't something we calculate, it's something we *estimate*. And often, as we will see in the next section, we don't even try to use the regression function (use best prediction), because it's too hard.

12.1.3 Best Linear Prediction

A widely used simplification of the regression problem restricts the allowable predictions to linear predictions, functions of the form

$$h(\mathbf{X}) = \alpha + \boldsymbol{\beta}'\mathbf{X} = \alpha + \sum_{i=1}^{n} \beta_i X_i. \tag{12.1}$$

(where α and β_1, …, β_n are constants). The function of this form that has the smallest mean square prediction error

$$E\{(Y - \alpha - \boldsymbol{\beta}'\mathbf{X})^2\} \tag{12.2}$$

is called the *best linear predictor* (BLP).

It should be understood, that using BLP is the Wrong Thing (not BP) unless the regression function just happens to be linear. The reason for doing the Wrong Thing is presumably because the Right Thing (using BP) is too hard, or we are too ignorant, or something of the sort.

Estimating the regression function is hard, but can be done. The main reason for the widespread use of linear prediction is that it had a 150 year head start (the development of linear regression theory started around 1800, whereas the development of nonlinear regression theory didn't really take off until the 1970's and is still a very active research area). So people understand linear regression much better, there is a long history of use in the various sciences, and so forth. Hence we will study it because of its popularity. (You should keep in mind though, that it is usually the Wrong Thing, decidedly not "best" despite the name).

We are now going to do a "stupid math trick" that simplifies notation at the expense of some mystification. Expression (12.1) is needlessly complicated (says the mathematician) by having two kinds of coefficients: α and the β_i. Only one kind is actually needed. We can consider (12.1) a special case of

$$h(\mathbf{X}) = \boldsymbol{\beta}'\mathbf{X} = \sum_{i=1}^{n} \beta_i X_i, \tag{12.3}$$

because if we make X_1 the constant random variable $X_1 = 1$, then (12.3) becomes

$$h(\mathbf{X}) = \beta_1 + \sum_{i=2}^{n} \beta_i X_i,$$

and this describes the same family of predictor functions as (12.1). Only the notation has changed (what was α is now β_1, what was β_i is now β_{i+1} for $i > 1$).

Thus we see that, although the simpleminded notion of the relationship between our two expressions for a linear prediction function is that (12.3) is a special case of (12.1) obtained by taking $\alpha = 0$, the really sophisticated notation is just the reverse, that (12.1) is a special case of (12.3) obtained by taking one of the $X_i = 1$. Having seen this, we will just use the mathematically simpler form (12.3) without any assertion that any of the X_i are constant. Understanding the general case tells us about the special case.

Theorem 12.2 (Best Linear Prediction). *For predicting a random variable Y given the value of a random vector $\mathbf{X}$, the linear predictor function* (12.3) *that minimizes the expected squared prediction error*

$$E\{(Y - \boldsymbol{\beta}'\mathbf{X})^2\} \tag{12.4}$$

is defined by

$$\boldsymbol{\beta} = E(\mathbf{X}\mathbf{X}')^{-1}E(Y\mathbf{X}) \tag{12.5}$$

assuming the inverse exists.[1]

Proof. The m. s. p. e. (12.4) is a quadratic function of β_1, ..., β_n. Since it is nonnegative it is a positive semi-definite quadratic function and hence has a global minimum where the first derivative is zero.

The proof is simpler if we rewrite (12.4) in non-matrix notation as

$$Q(\boldsymbol{\beta}) = E\left\{\left(Y - \sum_{i=1}^n \beta_i X_i\right)\left(Y - \sum_{j=1}^n \beta_j X_j\right)\right\}$$

and further simplify using linearity of expectation

$$Q(\boldsymbol{\beta}) = E(Y^2) + \sum_{i=1}^n \sum_{j=1}^n \beta_i \beta_j E(X_i X_j) - 2\sum_{i=1}^n \beta_i E(Y X_i)$$

The first derivative vector has components

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_k} = 2\sum_{i=1}^n \beta_i E(X_i X_k) - 2E(Y X_k) \tag{12.6}$$

These are n linear equations in n unknowns β_1, ..., β_k. They always have a solution (not necessarily unique), but in general there is no nice expression for the solution, except that we can always write the solution of any set of linear equations in terms of a matrix inverse (if the inverse exists, or a generalized inverse if not). Before we can do that, we need to put (12.6) back in matrix notation, using the fact that $E(\mathbf{X}\mathbf{X}')$ is a matrix with components $E(X_i X_j)$ so (12.6) can be rewritten

$$\nabla Q(\boldsymbol{\beta}) = 2E(\mathbf{X}\mathbf{X}')\boldsymbol{\beta} - 2E(Y\mathbf{X}) \tag{12.7}$$

Hence the equations to be solved are (12.7) set to zero, that is

$$E(\mathbf{X}\mathbf{X}')\boldsymbol{\beta} = E(Y\mathbf{X}) \tag{12.8}$$

Multiplying both sides on the left by $E(\mathbf{X}\mathbf{X}')^{-1}$ gives (12.5). □

[1] If the inverse does not exist, it can be replaced by a so-called "generalized inverse" and the same formula still produces a best linear predictor, but a generalized inverse is non-unique, so the $\boldsymbol{\beta}$ produced by the formula is non-unique. However, every such $\boldsymbol{\beta}$ gives the same prediction $\boldsymbol{\beta}'\mathbf{X}$ for the same value of $\mathbf{X}$. The nonuniqueness arises because $\mathbf{X}$ is a degenerate random vector (concentrated on a hyperplane). We will ignore this issue henceforth and assume the inverse exists.

For convenience we give the special case in which there is one constant predictor variable and one non-constant predictor, in which case we write the linear predictor function as $\alpha + \beta X$.

Corollary 12.3. *For predicting a random scalar Y given the value of a random scalar X, the linear predictor function of the form $\alpha + \beta X$ that minimizes the expected squared prediction error is defined by*

$$\alpha = \mu_Y - \beta \mu_X \tag{12.9a}$$

$$\beta = \rho \frac{\sigma_Y}{\sigma_X} \tag{12.9b}$$

This looks so different from the form given in the theorem that it is simpler to just derive it separately, which is done on p. 426 in Lindgren. The reason we have called it a "corollary" is to remind you that it is, despite appearances, a special case of the theorem.

As with BP and BUP, sometimes that BLP is BLUP (best linear unbiased prediction). In general the BLP is not unbiased, but when one of the predictors is constant (when we are using our "stupid math trick") it is.

> *BLP is BLUP when one of the predictors is constant.*

In particular, there is no difference between BLP and BLUP in Corollary 12.3.

The proof of this assertion is a direct consequence of (12.8) in the proof of the theorem. This one vector equation is equivalent to n scalar equations, which are sometimes called the "normal equations." If $X_k = 1$ with probability one, then the k-th normal equation

$$E(X_k \mathbf{X}') \boldsymbol{\beta} = E(Y X_k)$$

becomes

$$E(Y) = E(\mathbf{X})' \boldsymbol{\beta} = \boldsymbol{\beta}' E(\mathbf{X}) = E(\boldsymbol{\beta}' \mathbf{X})$$

and this says that the prediction $\boldsymbol{\beta}'\mathbf{X}$ is unbiased for Y.

Example 12.1.1 (A Pretty Bad "Best" Linear Prediction).
In Example 3.5.1 in Chapter 3 of these notes we considered positive scalar random variables X and Y having joint density

$$f(x, y) = \tfrac{1}{2}(x + y) e^{-x-y}, \qquad x > 0,\ y > 0.$$

There we found the best predictor of X given Y is

$$a(Y) = E(X \mid Y) = \frac{2 + Y}{1 + Y}, \qquad Y > 0.$$

This is a fairly nonlinear function so we don't expect BLP to do very well, and it doesn't.

Direct calculation using gamma integrals gives

$$
\begin{aligned}
E(X) &= \frac{1}{2} \int_0^\infty \int_0^\infty (x^2 + xy) e^{-x-y} \, dx \, dy \\
&= \frac{1}{2} \int_0^\infty (2 + y) e^{-y} \, dy \\
&= \frac{3}{2} \\
E(X^2) &= \frac{1}{2} \int_0^\infty \int_0^\infty (x^3 + x^2 y) e^{-x-y} \, dx \, dy \\
&= \frac{1}{2} \int_0^\infty \int_0^\infty (6 + 2y) e^{-y} \, dy \\
&= 4 \\
E(XY) &= \frac{1}{2} \int_0^\infty \int_0^\infty (x^2 y + x y^2) e^{-x-y} \, dx \, dy \\
&= \frac{1}{2} \int_0^\infty \int_0^\infty (2y + y^2) e^{-y} \, dy \\
&= 2
\end{aligned}
$$

By symmetry $E(X) = E(Y)$ and $E(X^2) = E(Y^2)$. So

$$
\begin{aligned}
\operatorname{var}(X) &= E(X^2) - E(X)^2 = \frac{7}{4} \\
\operatorname{var}(Y) &= \operatorname{var}(X) \\
\operatorname{cov}(X, Y) &= E(XY) - E(X)E(Y) = -\frac{1}{4} \\
\operatorname{cor}(X, Y) &= \frac{\operatorname{cov}(X, Y)}{\sqrt{\operatorname{var}(X) \operatorname{var}(Y)}} = -\frac{1}{7}
\end{aligned}
$$

So the BLP corollary gives

$$
\begin{aligned}
\beta &= -\frac{1}{7} \\
\alpha &= \frac{12}{7}
\end{aligned}
$$

and best linear prediction is

$$
a_{\text{blp}}(Y) = \frac{12 - Y}{7}, \qquad Y > 0.
$$

Note that for $Y > 12$ the BLP is negative, whereas the variable X it is predicting is necessarily positive. So the prediction isn't very good. The theorem asserts that this prediction is the best of all linear predictions. The problem is that no linear prediction is very good, even the best of them.

The BLP isn't the BP. Sometimes the BLP is a very bad predictor.

The theorem describes the BLP (or BLUP when one of the predictors is constant) in the case where you *know* the "population" distribution, the true distribution of $\mathbf{X}$ and Y. But when we are doing statistics, we don't know the true distribution of the data, because it depends on unknown parameters. Thus when doing statistics, the BLP or BLUP isn't something we calculate, it's something we *estimate*. It's the true unknown "population" function that we are trying to estimate from a sample.

12.2 The Sample Regression Function

Recall the empirical distribution introduced in Section 7.1 of these notes. It is the distribution that puts probability $1/n$ at each of n points. There we were interested in the case where the points were scalars. Here we are interested in the case where the points are vectors, but there is no real difference except for boldface. The empirical expectation operator associated with the vector points $\mathbf{x}_1$, $\ldots$, $\mathbf{x}_n$ is defined by

$$E_n\{g(\mathbf{X})\} = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i). \tag{12.10}$$

which is just (7.2) with some type changed to boldface.

In regression, we are interested in vectors of a rather special form, consisting of a scalar "response" variable y and a vector "predictor" variable $\mathbf{x}$. Suppose we have observed a sample of predictor-response pairs $(\mathbf{x}_i, y_i)$, then the corresponding empirical expectation formula is

$$E_n\{g(Y, \mathbf{X})\} = \frac{1}{n}\sum_{i=1}^{n} g(y, \mathbf{x}_i). \tag{12.11}$$

In particular, the empirical mean square prediction error for a linear predictor of the form described in the theorem is

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{\beta}'\mathbf{x}_i)^2 \tag{12.12}$$

Now the theorem applied to the empirical distribution gives the following. The empirical BLP is

$$\hat{\boldsymbol{\beta}} = E_n(\mathbf{X}\mathbf{X}')^{-1}E_n(Y\mathbf{X}) \tag{12.13}$$

where

$$E_n(\mathbf{X}\mathbf{X}') = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i' \tag{12.14}$$

$$E_n(Y\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} y_i\mathbf{x}_i \tag{12.15}$$

However, this is not the usual form in which the empirical analogue of the theorem is stated. The usual form involves yet another "stupid math trick." The formulas above have some explicit sums, those involved in the empirical expectation, and some implicit sums, those involved in matrix multiplications. The "stupid math trick" we are now introducing makes all the sums implicit (matrix multiplications).

To understand the trick, we need a closer look at the predictor variables. The subscripts on the $\mathbf{x}_i$ do not denote components, but different vectors in the sample. Each $\mathbf{x}_i$ is a vector and has components, say

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$$

(Before we were writing n as the dimension of the vectors. Now we are using n for the sample size. So the dimension must be a different letter, here p.) Thus the "predictor" part of the observed data are np variables

$$x_{ij}, \qquad i = 1, \ldots, n \text{ and } j = 1, \ldots, p.$$

which if we like, we can think of as an $n \times p$ matrix $\mathbf{X}$ (we've introduced a new notation here: $\mathbf{X}$ with no subscripts will henceforth be an $n \times p$ matrix). This matrix is very important in the theory of linear regression. It is commonly called the *design matrix*. The reason for the name is that if the data are from a designed experiment, then the design matrix incorporates everything about the design that is involved in linear regression theory. If the data are not from a designed experiment, then the name is inappropriate, but everyone uses it anyway. The relationship between the design matrix $\mathbf{X}$ and the predictor vectors $\mathbf{x}_1$, ..., $\mathbf{x}_p$ is that the predictor vectors are the columns of the design matrix.

Now write $\boldsymbol{\mu}$ as the n-dimensional vector of all the theoretical predictions (the conditional expectation of Y_i given all the $\mathbf{x}$'s), which has components

$$\mu_i = \boldsymbol{\beta}'\mathbf{x}_i = \sum_{j=1}^{p} X_{ij}\beta_j$$

This sum can be written as a matrix multiplication

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}. \tag{12.16}$$

because the dimensions match

$$\underset{n \times 1}{\boldsymbol{\mu}} = \underset{n \times p}{\mathbf{X}} \quad \underset{p \times 1}{\boldsymbol{\beta}}$$

Now we want to also write the sum in (12.12) as a matrix multiplication. The way we do this is to note that for any vector $\mathbf{z} = (z_1, \ldots, z_n)$

$$\mathbf{z}'\mathbf{z} = \sum_{i=1}^{n} z_i^2.$$

Applying this with $\mathbf{z} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$ gives

$$\frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{12.17}$$

as another expression for the empirical m. s. p. e. (12.12).

Now we also want to rewrite (12.14) and (12.15) using this trick. When we write these equations out explicitly using all the subscripts, we see that (12.14) is a matrix with (j, k)-th element

$$\frac{1}{n}\sum_{i=1}^{n} x_{ij}x_{ik}$$

which is seen to be the j, k element of $\mathbf{X}'\mathbf{X}/n$. Similarly, (12.15) is a vector with j-th element

$$\frac{1}{n}\sum_{i=1}^{n} y_i X_{ij}$$

which is seen to be the j-th element of $\mathbf{y}'\mathbf{X}/n$ or of $\mathbf{X}'\mathbf{y}/n$. Putting these together we get the following very compact matrix notation for (12.13)

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

This is the "usual" way the empirical version of the BLP theorem is written

Corollary 12.4 (Multiple Linear Regression). *The $\boldsymbol{\beta}$ that minimizes* (12.17) *is*

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{12.18}$$

For completeness, we also record the empirical analog of Corollary 12.3

Corollary 12.5 (Simple Linear Regression). *The values α and β that minimize the empirical expected squared prediction error*

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$

are

$$\hat{\alpha} = \bar{y} - \beta\bar{x} \tag{12.19a}$$

$$\hat{\beta} = r\frac{s_y}{s_x} \tag{12.19b}$$

Fortunately, we do not have to do the calculations described by these corollaries by hand. Many calculators will do the "simple case" of Corollary 12.5. Any computer statistics package will do the "multiple" case of Corollary 12.4.

Here's an example using R.

Example 12.2.1 (Multiple Regression).
We use the data in the URL

```
http://www.stat.umn.edu/geyer/5102/ex12.2.1.dat
```

The R command that does multiple linear regression is `lm` (for "linear model"). This data set has three variables `x1`, `x2`, and `y`. In R each is a vector, and they all have the same length (in this particular data set $n = 100$). The response is `y`, and the predictor variables are `x1` and `x2`. The specific R commands that do the regression and print the results are

```
out <- lm(y ~ x1 + x2)
summary(out)
```

The first command doesn't print anything (it just returns the dataset `out`), the latter prints the fairly voluminous output

```
Call:
lm(formula = y ~ x1 + x2)

Residuals:
     Min       1Q   Median       3Q      Max
-121.338  -32.564    5.525   35.309  124.846

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  94.4803    10.9616   8.619 1.27e-13 ***
x1            0.8503     0.5606   1.517    0.133
x2            1.3599     0.5492   2.476    0.015 *
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 54.22 on 97 degrees of freedom
Multiple R-Squared: 0.5875,     Adjusted R-squared: 0.579
F-statistic: 69.08 on 2 and 97 degrees of freedom,      p-value:     0
```

most of which we won't explain now (and a fair amount of which we won't explain ever).

The first thing we will explain is what model was fit, and where to find the estimates of the β's in the printout. R always assumes by default that you want a constant predictor. Hence the model fit here has *three* predictors, not just the two explicitly mentioned. Hence it also has three corresponding parameters. We can write the model as

$$h(\mathbf{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2.$$

Information about the parameter estimates is found in the section labeled `Coefficients:` (α and the β_i and their estimates are often called *regression* *coefficients* because they are the coefficients of the predictor variables in the definition of the regression function). The estimates are given in the column labeled `Estimate` in that section, which we repeat here

```
             Estimate
(Intercept)  94.4803
x1            0.8503
x2            1.3599
```

The three estimates are $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ respectively. The coefficients of the non-constant predictors are labeled by the names of the variables they multiply. The coefficient of the constant predictor is labeled `(Intercept)` because α is usually called the "y-intercept" in elementary math.

If you actually wanted to do regression *without* a constant predictor, you would need to know the magic incantation that makes R do this. In fact, it has two[2]

```
out <- lm(y ~ x1 + x2 + 0)
out <- lm(y ~ x1 + x2 - 1)
```

So that covers the mechanics of doing linear regression. Let the computer do it!

12.3 Sampling Theory

12.3.1 The Regression Model

In order to have sampling theory, we need a probability model. The probability model usually adopted assumes that we observe pairs $(Y_i, \mathbf{X}_i)$, $i = 1$, 2, $\ldots$. The Y_i are scalars, and the $\mathbf{X}_i$ are vectors. The $\mathbf{X}_i$ may or may not be random, but if random we condition on them, meaning we *treat* them as if *not* random. Thus we will henceforth write them as lower case $\mathbf{x}_i$.

The *linear regression model* (sometimes just called the *linear model*) is that the means of the Y_i are linear functions of the $\mathbf{x}_i$

$$E(Y_i) = \boldsymbol{\beta}' \mathbf{x}_i. \tag{12.20}$$

Note that this formula assumes the $\mathbf{x}_i$ are constants. If we didn't assume that, we would have to write (12.20) as

$$E(Y_i \mid \mathbf{X}_i) = \boldsymbol{\beta}' \mathbf{X}_i \tag{12.21}$$

(this is the last time we will note the distinction between the two approaches).

We also usually write

$$Y_i = \boldsymbol{\beta}' \mathbf{x}_i + e_i, \qquad i = 1, \ldots, n,$$

[2] The reason for two is that the R team like the first, and the second is provided for backwards compatibility with the S language, which R is more or less a clone of. I find neither very intuitive.

which can be written as a single vector equation

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{12.22}$$

where $\mathbf{e} = (e_1, \ldots, e_n)$.

This equation has no statistical content. We are just defining variables e_i to be the deviations of the Y_i from their means. The e_i are usually called "errors." Despite the lower case letter, they are random variables ("big E" is a frozen letter, reserved for expectation operators). In order for (12.20) to hold, the errors must have mean zero.

To further specify the distribution, we can describe the distribution of either the Y_i or the e_i. The latter is simpler. There are two different types of assumptions that can be made about the errors: strong and weak. The weak assumption, used in the following section, describes only the first two moments. The weak assumption says

$$\begin{aligned} E(\mathbf{e}) &= 0 \\ \operatorname{var}(\mathbf{e}) &= \sigma^2 \mathbf{I} \end{aligned} \tag{12.23}$$

or in words, the errors

- have mean zero,
- are uncorrelated, and
- have constant variance

(that is, they all have the same variance).[3]

The strong assumption actually gives the distribution of the errors

$$\mathbf{e} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}) \tag{12.24}$$

It is a special case of the weak assumption. It says the errors have the mean and variance specified by (12.23), and in addition that they are multivariate normal. Note that by Theorem 5.13 of Chapter 5 in last semester's notes (uncorrelated implies independent for jointly multivariate normal random variables), an equivalent way to state the strong assumption is that the e_i are i. i. d. $\mathcal{N}(0, \sigma^2)$.

Thus the weak assumption only makes the errors uncorrelated (which does *not* imply they are independent if they are not multivariate normal), whereas the strong assumption makes the errors both independent and normally distributed.

Both the weak and strong assumption make the same assumption (12.20) about the means of the Y_i. Another way to describe this part of the model assumptions is by saying that we are assuming that the true population regression function is linear. It is clear from (12.21) that

$$h(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x} \tag{12.25}$$

[3] The assumption of constant variance is so important that some statistician invented a big word to describe it: *homoscedasticity*. Violation of the assumption (different variances) is called *heteroscedasticity*. But we'll just say "constant variance" and "non-constant variance."

is assumed to be the regression function of Y on $\mathbf{X}$. We often call this the *population regression function* in our usual abuse of terminology that talks about i. i. d. variables as a "sample" from an infinite "population." What we mean is only the assertion (12.20) or (12.21) that this is the true unknown regression function.

This is the usual statistical set-up. The $\mathbf{x}_i$ and Y_i are observed data; the β_i are unknown parameters. The best we can do is to *estimate* the unknown parameters with *estimates* $\hat{\beta}_i$ that are functions of the data and study the statistical properties of the estimates (and eventually get confidence intervals, hypothesis tests, and the rest of the paraphernalia of statistical inference).

The estimators we are going to study are the empirical BLUP estimates described by Corollaries 12.4 and 12.5. The name "empirical BLUP" we used for them is nonstandard. They have accumulated a lot of names over the years. One common name for the $\hat{\beta}_i$ is *sample regression coefficients.* (And the analogous term for the β_i is *population regression coefficients.*) Another common name, in use for 200 years, for the $\hat{\beta}_i$ is *least squares estimates* because they minimize the empirical m. s. p. e. (12.12).

When we plug the estimates into (12.25) we get

$$\hat{h}(\mathbf{x}) = \hat{\boldsymbol{\beta}}' \mathbf{x}, \tag{12.26}$$

which is the *sample regression function.*[4]

It is important here, as everywhere else in (frequentist) statistics to keep the slogan about *the sample is not the population* firmly in mind. Even assuming that the population regression function is linear so (12.25) gives best predictions, the sample regression function (12.26) does *not* give best predictions because the sample is not the population. How far off they are is the job of sampling theory to describe.

12.3.2 The Gauss-Markov Theorem

This section uses only the "weak" distributional assumptions (12.23). Normality is not used. The content of the Gauss-Markov theorem is simply stated as the least squares estimates are *best linear unbiased estimates* (BLUE).

Before we can even state the theorem properly we need to explain what "best" means in this context. For unbiased estimates mean square error is the same as variance. So best means smallest variance. The problem is that the estimate $\hat{\boldsymbol{\beta}}$ is a vector, so its variance is a matrix. What does it mean for one matrix to be "smaller" than another?

In general there is no sensible definition of a "less than" relation for matrices. Recall, though that variance matrices have the special property of being positive semi-definite (Corollary 5.5 of Chapter 5 of these notes). There is a natural

[4] You also hear people say a lot of other pairs: *sample regression thingummy* and *population regression thingummy* for instances of "thingummy" other than *function* and *coefficients*, such as *equation*, *line* (thinking of the graph of a linear function being a line in the "simple" case of one non-constant predictor), and so forth.

partial order for positive semi-definite matrices. We say $A \leq B$ if $B - A$ is a positive semi-definite matrix.

To understand what this means, look at the proof of why covariance matrices are positive semi-definite. A matrix $\mathbf{M}$ is positive semi-definite if

$$\mathbf{c}'\mathbf{M}\mathbf{c} \geq 0, \qquad \text{for every vector } \mathbf{c}.$$

We also know that for any random vector $\mathbf{X}$ having variance matrix $\mathbf{M}$, the variance of a scalar linear function is given by

$$\operatorname{var}(a + \mathbf{c}'\mathbf{X}) = \mathbf{c}'\mathbf{M}\mathbf{c} \tag{12.27}$$

by (5.19b) from Chapter 5 of these notes. Since variances are nonnegative, this shows $\mathbf{M}$ is positive semi-definite.

Now consider two random vectors $\mathbf{X}$ and $\mathbf{Y}$ with variance matrices $\mathbf{M}_{\mathbf{X}}$ and $\mathbf{M}_{\mathbf{Y}}$, respectively. We say that $\mathbf{M}_{\mathbf{X}} \leq \mathbf{M}_{\mathbf{Y}}$ if and only if $\mathbf{M}_{\mathbf{Y}} - \mathbf{M}_{\mathbf{X}}$ is a positive semi-definite matrix (that's the definition of the partial order). This means

$$\mathbf{c}'(\mathbf{M}_{\mathbf{Y}} - \mathbf{M}_{\mathbf{X}})\mathbf{c} \geq 0, \qquad \text{for every vector } \mathbf{c},$$

and this is equivalent to

$$\mathbf{c}'\mathbf{M}_{\mathbf{X}}\mathbf{c} \leq \mathbf{c}'\mathbf{M}_{\mathbf{Y}}\mathbf{c}, \qquad \text{for every vector } \mathbf{c},$$

and by (12.27) this is also equivalent to

$$\operatorname{var}(a + \mathbf{c}'\mathbf{X}) \leq \operatorname{var}(a + \mathbf{c}'\mathbf{Y}), \qquad \text{for every vector } \mathbf{c}. \tag{12.28}$$

This characterization tells us what the partial order means. The variance matrices are ordered $\operatorname{var}(\mathbf{X}) \leq \operatorname{var}(\mathbf{Y})$ if and only if the variance of *every scalar-valued linear function* of $\mathbf{X}$ is no greater than the variance of the same function of $\mathbf{Y}$. That's a strong condition!

Now that we have got this rather complicated definition of "best" explained, the theorem itself is very simple.

Theorem 12.6 (Gauss-Markov). *Under the assumptions* (12.22) *and* (12.23), *the least squares estimate* (12.18) *is an unbiased estimate of* $\boldsymbol{\beta}$. *Furthermore it is the best linear unbiased estimate, where "best" means smallest variance.*

In short, the least squares estimate is BLUE.

Proof. The first assertion is that $\hat{\boldsymbol{\beta}}$ given by (12.18) is an unbiased for $\boldsymbol{\beta}$. This is trivial

$$E(\hat{\boldsymbol{\beta}}) = E\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right\} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

(just linearity of expectation and the definition of matrix inverse).

Consider an unbiased but otherwise completely arbitrary estimate, which will have the form $\boldsymbol{\beta}^* = \mathbf{A}\mathbf{Y}$ for some constant matrix $\mathbf{A}$. (Saying $\mathbf{A}$ is constant means $\mathbf{A}$ can depend on $\mathbf{X}$ but not on $\mathbf{Y}$. Saying $\boldsymbol{\beta}^*$ is an estimate means $\mathbf{A}$ cannot depend on the parameters $\boldsymbol{\beta}$ and σ^2.) It simplifies the proof somewhat if we define

$$\mathbf{B} = \mathbf{A} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

so

$$\boldsymbol{\beta}^* = \left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{B}\right]\mathbf{Y} = \hat{\boldsymbol{\beta}} + \mathbf{B}\mathbf{Y}$$

The condition that $\boldsymbol{\beta}^*$ be unbiased is then

$$\boldsymbol{\beta} = E(\boldsymbol{\beta}^*) = E(\hat{\boldsymbol{\beta}}) + \mathbf{B}E(\mathbf{Y}) = \boldsymbol{\beta} + \mathbf{B}\mathbf{X}\boldsymbol{\beta}$$

which simplifies to

$$\mathbf{B}\mathbf{X}\boldsymbol{\beta} = 0$$

Unbiasedness means this must hold for all possible values of $\boldsymbol{\beta}$. Hence $\mathbf{B}\mathbf{X} = 0$.

Now we calculate

$$\operatorname{var}(\boldsymbol{\beta}^*) = \operatorname{var}(\hat{\boldsymbol{\beta}} + \mathbf{B}\mathbf{Y}) = \operatorname{var}(\hat{\boldsymbol{\beta}}) + \operatorname{var}(\mathbf{B}\mathbf{Y}) + 2\operatorname{cov}(\hat{\boldsymbol{\beta}}, \mathbf{B}\mathbf{Y}) \tag{12.29}$$

The formula for the variance of a sum here is (5.9) from Chapter 5.

I now claim that the covariance is zero (meaning we haven't proved that *yet*, but we want to look at its consequences to see why it is worth proving), from which the BLUE assertion follows immediately, because then (12.29) becomes

$$\operatorname{var}(\boldsymbol{\beta}^*) = \operatorname{var}(\hat{\boldsymbol{\beta}}) + \operatorname{var}(\mathbf{B}\mathbf{Y})$$

and, since $\operatorname{var}(\mathbf{B}\mathbf{Y})$ like any variance matrix must be positive semi-definite, this implies that $\operatorname{var}(\boldsymbol{\beta}^*) - \operatorname{var}(\hat{\boldsymbol{\beta}})$ is positive semi-definite, which according to the definition of partial order for matrices is the same as $\operatorname{var}(\hat{\boldsymbol{\beta}}) \le \operatorname{var}(\boldsymbol{\beta}^*)$, which is the "$\hat{\boldsymbol{\beta}}$ is best" assertion of the theorem.

Thus we have a proof that is complete except for the unproved claim that the covariance term in (12.29) is zero. So we now prove that claim. Again we calculate, using

$$\operatorname{cov}(\hat{\boldsymbol{\beta}}, \mathbf{B}\mathbf{Y}) = E\left\{\hat{\boldsymbol{\beta}}(\mathbf{B}\mathbf{Y})'\right\} - E(\hat{\boldsymbol{\beta}})E(\mathbf{B}\mathbf{Y})' = E\left\{\hat{\boldsymbol{\beta}}(\mathbf{B}\mathbf{Y})'\right\}$$

because $E(\mathbf{B}\mathbf{Y}) = \mathbf{B}\mathbf{X}\boldsymbol{\beta} = 0$. And

$$\begin{aligned} E\left\{\hat{\boldsymbol{\beta}}(\mathbf{B}\mathbf{Y})'\right\} &= E\left\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{B}'\right\} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{Y}\mathbf{Y}')\mathbf{B}' \end{aligned}$$

Now

$$E(\mathbf{Y}\mathbf{Y}') = \operatorname{var}(\mathbf{Y}) + E(\mathbf{Y})E(\mathbf{Y})' = \sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{X}'$$

so

$$\begin{aligned} E\left\{\hat{\boldsymbol{\beta}}(\mathbf{B}\mathbf{Y})'\right\} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I} + \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{X}')\mathbf{B}' \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{B}' + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}'\mathbf{X}'\mathbf{B}' \end{aligned}$$

And both terms contain $\mathbf{X}'\mathbf{B}' = (\mathbf{B}\mathbf{X})' = 0$. □

This is a very famous theorem. All students should know about it. But for all its supposed theoretical importance, you can't actually do anything with it. It's only good for theoretical woofing. Saying "least squares estimates are BLUE" intimidates people who haven't had a course like this.

The BLUE assertion of the theorem isn't even that important. After all, it doesn't claim that the least squares estimates are *best*. It only claims that they are *best linear unbiased*, which means best among linear and unbiased estimators. Presumably there are nonlinear or biased estimators that are better. Otherwise we could prove a stronger theorem. You have to read between the lines. It sounds like a claim that least squares estimates are the best, but when you decode the qualifications, it actually suggests that they aren't the best.

Moreover, the whole analysis is based on the assumption that the linear model is correct, that the true unknown population regression function is *linear*, that is, has the form (12.25). If the true unknown population regression function is *not* linear, then the least squares estimates are not even unbiased, much less BLUE.

12.3.3 The Sampling Distribution of the Estimates

The Regression Coefficients

We now turn to a much more straightforward problem: what is the sampling distribution of $\hat{\boldsymbol{\beta}}$. In order to have a sampling distribution, we need to specify the whole distribution of the data (not just two moments like we used in the Gauss-Markov theorem). Thus we now switch to the strong linear regression assumptions (12.24).

The least squares estimates are a linear transformation of the data $\mathbf{Y}$ by (12.18), hence if the data are multivariate normal, so are the estimates. A multivariate normal distribution is determined by its mean vector and variance matrix, so we only need to calculate the mean and variance to figure out the distribution.

Theorem 12.7. *Under the assumptions* (12.22) *and* (12.24)

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right). \tag{12.30}$$

Proof. We already showed that $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ under the weak assumptions in the proof of the Gauss-Markov theorem. Since the strong assumptions are stronger, this holds here too.

Now

$$\begin{aligned}
\operatorname{var}(\hat{\boldsymbol{\beta}}) &= \operatorname{var}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\right) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \operatorname{var}(\mathbf{Y}) \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \operatorname{var}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

because $\text{var}(\mathbf{Y}) = \text{var}(\mathbf{e}) = \sigma^2 \mathbf{I}$. □

Please note the assumptions of the theorem. If we assume

- the true regression function is linear (12.22) and
- the errors are independent, identically distributed, and *exactly* normally distributed (12.24),

then (12.30) gives the *exact* (not asymptotic) sampling distribution of the least squares estimates. If any of these assumptions are not exactly correct, then it doesn't.

The Error Variance

Unfortunately, the theorem by itself is useless for inference because the distribution contains an unknown parameter σ^2. To make progress, we need an estimate of this parameter and knowledge of its sampling distribution.

If we observed the actual errors e_i, the natural estimate of their variance would be

$$\frac{1}{n} \sum_{i=1}^{n} e_i^2$$

We don't subtract off their mean, because we know $E(e_i) = 0$.

Unfortunately, we do not observe the errors, and must estimate them. Since

$$e_i = y_i - \boldsymbol{\beta}' \mathbf{x}_i$$

the natural estimate is

$$\hat{e}_i = y_i - \hat{\boldsymbol{\beta}}' \mathbf{x}_i \tag{12.31}$$

Because the sample is not the population, these are not the right thing. Hence we should call them not "errors" but "estimated errors." The usual name, however, for (12.31) is *residuals*. We often rewrite (12.31) as a vector equation

$$\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}. \tag{12.32}$$

Plugging the residuals in for the errors in our "natural estimate" gives

$$\frac{1}{n} \hat{\mathbf{e}}' \hat{\mathbf{e}} = \frac{1}{n} \sum_{i=1}^{n} \hat{e}_i^2 \tag{12.33}$$

as a sensible estimate of σ^2, and it turns out this is the MLE (p. 491 in Lindgren). However, this is not the estimator commonly used, because it is biased. The commonly used estimator is

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} \hat{e}_i^2 \tag{12.34}$$

where p is the number of predictor variables (and regression coefficients). As we shall see, this estimate turns out to be unbiased.

The sum in either of these estimators referred to often enough that it needs a name. It is called the *sum of squares of the residuals* (SSResid) or the *residual sum of squares*

$$\text{SSResid} = \sum_{i=1}^{n} \hat{e}_i^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}}. \tag{12.35}$$

Theorem 12.8. *Under the assumptions* (12.22) *and* (12.24) SSResid *is independent of $\hat{\boldsymbol{\beta}}$, and*

$$\frac{\text{SSResid}}{\sigma^2} \sim \text{chi}^2(n-p),$$

where n is the number of observations and p the number of regression coefficients.

From (12.34) and (12.35) we see that the theorem is equivalent to

Corollary 12.9. *Under the assumptions* (12.22) *and* (12.24) $\hat{\sigma}^2$ *is independent of $\hat{\boldsymbol{\beta}}$, and*

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \text{chi}^2(n-p),$$

where n and p are as in the theorem.

We will have to defer the proof of the theorem for a bit while we develop a deeper understanding of what linear regression does. First we will look at what we can do with the theorem.

12.3.4 Tests and Confidence Intervals for Regression Coefficients

The main thing we can do with these theorems is make pivotal quantities having Student's t distribution. Recall the definition of Student's t distribution from Section 7.3.5 of these notes: the ratio of a standard normal and the square root of an independent chi-square divided by its degrees of freedom. The vector $\hat{\boldsymbol{\beta}}$ of sample regression coefficients is multivariate normal by Theorem 12.7. To simplify notation define

$$\mathbf{M} = (\mathbf{X}'\mathbf{X})^{-1}.$$

Then the variance of $\hat{\boldsymbol{\beta}}$ is $\sigma^2\mathbf{M}$. Hence a particular sample regression coefficient $\hat{\beta}_k$ has variance $\sigma^2 m_{kk}$ (where, as usual, the elements of $\mathbf{M}$ are denoted m_{ij}). The mean of $\hat{\beta}_k$ is β_k. Thus

$$Z_k = \frac{\hat{\beta}_k - \beta_k}{\sigma\sqrt{m_{kk}}}$$

is standard normal. By Corollary 12.9, $\hat{\sigma}^2/\sigma^2$ is an independent chi-square divided by its degrees of freedom, hence

$$T_k = \frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}\sqrt{m_{kk}}} \sim t(n-p). \tag{12.36}$$

Since the right hand side does not contain unknown parameters, this is a *pivotal quantity*. Hence it can be used for exact confidence intervals and tests about the unknown parameter β_k.

The only difficulty in using this pivotal quantity is calculating the denominator $\hat{\sigma}\sqrt{m_{kk}}$. Because it involves a matrix inverse, there is no simple formula for calculating it. You must use a computer. When using R to do the regression, it always calculates this quantity and prints it out.

Example 12.3.1.
This continues Example 12.2.1. Again we repeat part of the printout from that example

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  94.4803    10.9616   8.619 1.27e-13 ***
x1            0.8503     0.5606   1.517    0.133
x2            1.3599     0.5492   2.476    0.015 *
```

Recall from Example 12.2.1 that we explained that the column labeled `Estimate` from this table gives the sample regression coefficients $\hat{\beta}_k$. Now we explain the rest of the table. The column labeled `Std. Error` gives the denominators $\hat{\sigma}\sqrt{m_{kk}}$ of the t pivotal quantities involving the regression coefficients. The label follows the widespread convention of calling an estimated standard deviation a "standard error." The standard deviation of $\hat{\beta}_k$ is $\sigma\sqrt{m_{kk}}$, which involves an unknown parameter. Estimating it by plugging in $\hat{\sigma}$ for σ gives the *standard error*.

The column labeled `t value` gives the value of the t statistic (12.36) for testing the null hypothesis $\beta_k = 0$. This means that it is the value of (12.36) with zero plugged in for β_k. Let's check this. Looking at the last row, for example $1.3599/0.5492 = 2.476147$, and we see that the third column is indeed the first column divided by the second.

The column labeled `Pr(>|t|)` gives the P-value for the two-tailed test of $\beta_k = 0$ (that is, the alternative is $H_A : \beta_k \neq 0$). Let's also check this. The degrees of freedom of the relevant t distribution are $n - p$, where $n = 100$ and $p = 3$ (there are three regression coefficients including the intercept). Actually, we do not even have to do this subtraction. The degrees of freedom are also given in the R printout in the line

```
Residual standard error: 54.22 on 97 degrees of freedom
```

The P-value corresponding to the t statistic 2.476 in the bottom row of the table is

```
> 2 * (1 - pt(2.476, 97))
[1] 0.01501905
```

and this does indeed agree with the number in the fourth column of the table.

Thus R makes the test with null hypothesis $\beta_k = 0$ easy. It prints the P-value for the two-tailed test and, of course, the P-value for a one-tailed test, if desired, would be half the two-tailed P-value.

The confidence intervals are a bit harder. They have the form

$$\text{estimate} \pm \text{critical value} \times \text{standard error}$$

and all the pieces except the critical value are given in the printout, but that is easily looked up. The critical value for a 95% confidence interval is

```
> qt(0.975, 97)
[1] 1.984723
```

Thus a 95% confidence interval for β_2 (using numbers from the bottom row of the table in the printout) is

```
> 1.3599 + c(-1,1) * 1.984723 * 0.5492
[1] 0.2698901 2.4499099
```

One final warning: with three regression coefficients here, you can do three confidence intervals or three tests. But doing that without correction for multiple testing (Bonferroni correction, for example) is *bogus.* In fact, R's attempt to be helpful by providing the "stars" necessary for "stargazing" is just the bogosity we warned about in Section 9.5.8. So unless there is a strong tradition of stargazing in your scientific subdiscipline, so strong that you just have to do it no matter how bogus, ignore the stars. You can turn off the printing of stars by inserting the command

```
> options(show.signif.stars=FALSE)
```

before the `summary(out)` command.

12.3.5 The Hat Matrix

Also of interest besides the sample regression coefficients is the estimate of the regression function itself

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

The matrix that multiplies $\mathbf{y}$ on the right hand side

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'. \tag{12.37}$$

was dubbed by someone suffering a fit of cuteness the *hat matrix* because it puts the "hat" on $\mathbf{y}$, that is, $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ (actually, I enjoy terminology like this, I just don't care to defend it against stuffed shirts who think scientific terminology should be very serious and boring).

Lemma 12.10. *The hat matrix* (12.37) *is the orthogonal projection onto the subspace spanned by the predictor vectors (the columns of* **X***).*

The notion of an orthogonal projection matrix is defined in Section H.1 in Appendix H. Another name for the subspace mentioned in the theorem is just the range of the linear transformation represented by the matrix $\mathbf{X}$, which we write range($\mathbf{X}$). The theorem asserts that this is also the range of the linear transformation represented by the hat matrix $\mathbf{H}$, that is, range($\mathbf{X}$) = range($\mathbf{H}$).

Proof. That the hat matrix is symmetric is obvious from the formula and the rule that the transpose of a matrix product is the product of the transposes in reverse order. That the hat matrix is idempotent is verified by just looking at the formula for $\mathbf{H}^2$.

So the only thing left to verify is that $\mathbf{H}$ actually maps onto range($\mathbf{X}$). We need to show that an arbitrary element of range($\mathbf{X}$), which has the form $\mathbf{X}\boldsymbol{\beta}$ for an arbitrary vector $\boldsymbol{\beta}$, is equal to $\mathbf{H}\mathbf{y}$ for some vector $\mathbf{y}$. It is easily verified that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ does the job. □

With this lemma we can finally finish the proof of the theorem that gives t statistics.

Proof of Theorem 12.8. First observe that $\mathbf{H}$ and $\mathbf{I} - \mathbf{H}$ are orthogonal projections that are orthogonal to each other (see Section H.1 in Appendix H for definitions). Define $\mathbf{Z} = \mathbf{e}/\sigma$, then $\mathbf{Z}$ is a multivariate standard normal random vector (that is, the components are i. i. d. standard normal). Theorem H.3 in Appendix H says that $\mathbf{HZ}$ and $(\mathbf{I} - \mathbf{H})\mathbf{Z}$ are independent multivariate normal random vectors and their squared lengths are chi-square random variables. The next step is to see what these vectors are in terms of the variables we have been using.

$$\mathbf{HZ} = \frac{1}{\sigma}\mathbf{He} = \frac{1}{\sigma}\mathbf{H}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{1}{\sigma}(\hat{\mathbf{y}} - \boldsymbol{\mu})$$

where $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, originally defined in (12.16), is the vector of means of the response variables (the fact that $\mathbf{HX}\boldsymbol{\beta} = \mathbf{X}\boldsymbol{\beta}$ was verified in the preceding proof). And

$$(\mathbf{I} - \mathbf{H})\mathbf{Z} = \frac{1}{\sigma}(\mathbf{I} - \mathbf{H})\mathbf{e} = \frac{1}{\sigma}(\mathbf{I} - \mathbf{H})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \frac{1}{\sigma}(\mathbf{y} - \hat{\mathbf{y}})$$

Thus we see that

$$\hat{\mathbf{y}} = \boldsymbol{\mu} + \sigma\mathbf{HZ}$$

and

$$\frac{\text{SSResid}}{\sigma^2} = \|(\mathbf{I} - \mathbf{H})\mathbf{Z}\|^2 \tag{12.38}$$

As we said above, Theorem H.3 in Appendix H says that these are independent random variables, and the latter has a chi-square distribution with degrees of freedom rank($\mathbf{I} - \mathbf{H}$).

That almost finishes the proof. There are two loose ends. We were supposed to show that SSResid is independent of $\hat{\boldsymbol{\beta}}$, but what we showed above is that it is independent of $\hat{\mathbf{y}}$. However, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{y}}$, so independence of $\hat{\mathbf{y}}$ implies independence of $\hat{\boldsymbol{\beta}}$.

The last loose end is that we need to calculate the rank of $\mathbf{I}-\mathbf{H}$. Since $\mathbf{I}-\mathbf{H}$ is the projection on the subspace orthogonally complementary to range($\mathbf{H}$), its rank is $n-p$, where n is the dimension of the whole space and $p = \operatorname{rank}(\mathbf{H})$. Lemma 12.10 asserts that $\operatorname{rank}(\mathbf{H}) = \operatorname{rank}(\mathbf{X})$, which is number of predictor variables. So we are done. □

12.3.6 Polynomial Regression

The reader may have been lead from what has been said so far to think that linear regression is only useful for fitting *linear* regression functions. No! It's much more useful than that. Here is a slogan that captures the issue.

> *It's called "linear regression" because it's linear in the β's, not because it's linear in the x's.*

Here's what the slogan means. Suppose I have a function that is linear in the β's but not linear in the x's, for example

$$h(x) = \beta_1 \sin(x) + \beta_2 \log(x) + \beta_3 x^2.$$

We can put this in linear regression form by simply making up new predictor variables

$$\begin{aligned} x_1 &= \sin(x) \\ x_2 &= \log(x) \\ x_3 &= x^2 \end{aligned}$$

It matters not a bit that these new variables are dependent, all functions of the original predictor variable x. In terms of these "made up" predictor variables our regression function is now linear in both the β's and the x's

$$h(x) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

and is in the form required by the assumptions for linear regression.

> *You can make up as many predictors as you please.*

This section describes one way to "make up predictors."

Example 12.3.2 (Polynomial Regression).
Look at Figure 12.1 which is a plot of some regression data found in the data set at the URL

`http://www.stat.umn.edu/geyer/5102/ex12.3.2.dat`

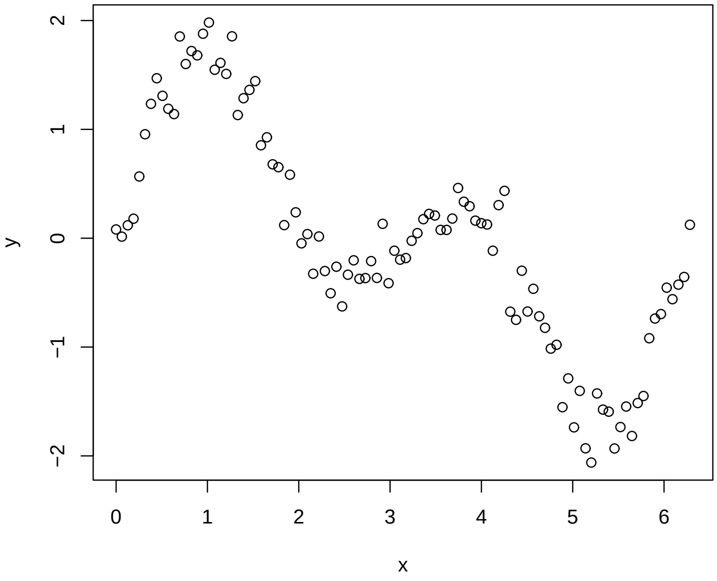


Figure 12.1: Some regression data.

A mere glance at the plot shows that y is decidedly not a linear function of x, not even close. However, the slogan says there is nothing that prevents us from making up more predictor variables. The data set itself has no other variables in it, just x and y. So any new predictor variables we make up must be functions of x. What functions?

Since we are told nothing about the data, we have no guidance as to what functions to make up. In a real application, there might be some guidance from scientific theories that describe the data. Or there might not. Users of linear regression often have no preconceived ideas as to what particular functional form the regression function may have. The title of this section suggests we try a polynomial, that is we want to use a regression function of the form

$$E(Y \mid X) = \sum_{i=0}^{k} \beta_i X^i. \tag{12.39}$$

This is a linear regression function if we consider that we have $k + 1$ different predictor variables $X^0 = 1$, $X^1 = X$, X^2, $\ldots$, X^k. What we have done is "made up" new predictor variables, which are the higher powers of X. Here's how R does it.

```
> out <- lm(y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5) + I(x^6))
> summary(out)

Call:
```

```
lm(formula = y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5) + I(x^6))

Residuals:
      Min        1Q    Median        3Q       Max
-0.577040 -0.192875 -0.004183  0.196342  0.734926

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.654541   0.192106  -3.407 0.000971 ***
x            7.868747   0.868968   9.055 2.04e-14 ***
I(x^2)      -8.408605   1.228159  -6.847 8.00e-10 ***
I(x^3)       3.334579   0.741135   4.499 1.97e-05 ***
I(x^4)      -0.554590   0.215506  -2.573 0.011651 *
I(x^5)       0.029160   0.029832   0.977 0.330884
I(x^6)       0.000576   0.001577   0.365 0.715765
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

Residual standard error: 0.308 on 93 degrees of freedom
Multiple R-Squared: 0.9122,     Adjusted R-squared: 0.9065
F-statistic: 160.9 on 6 and 93 degrees of freedom,      p-value:     0
```

The function `I()` is used to give arithmetic expressions their literal meaning in the model formula. If you leave it out, R doesn't do the right thing.

Why start with a sixth degree polynomial? No particular reason. We'll examine this issue later. For now, just accept it.

How do we interpret this mess? The *naive* way is to pay attention to the stars (of course, you wouldn't be that naive now, after all our harping on the bogosity of stargazing, would you?). They seem to say that the coefficients up the x^4 term are statistically significant, and the coefficients of the two higher powers are not. So we should try a fourth degree polynomial next.

Stargazing violates the "do *one* test" dogma. To do the right thing we must do only one test. The obvious coefficient to test is the one for the highest power of x. Clearly it is not statistically significant. Thus we can accept the null hypothesis of that test and set the corresponding regression coefficient equal to zero. And that is the end of the conclusions we can draw from this regression!

However, that conclusion leaves us with a new model to fit. The part of the printout about the regression coefficients for that model is

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.687409   0.168940  -4.069 9.83e-05 ***
x            8.112834   0.552844  14.675  < 2e-16 ***
I(x^2)      -8.808811   0.552153 -15.954  < 2e-16 ***
I(x^3)       3.592480   0.224068  16.033  < 2e-16 ***
I(x^4)      -0.631962   0.039391 -16.043  < 2e-16 ***
I(x^5)       0.040017   0.002495  16.039  < 2e-16 ***
```

Surprise! The coefficient of x^5 wasn't statistically significant before, but now it is. In fact, it wasn't even close to significance before, and now it is extremely significant. What's happening?

A long answer could get very complicated. All of the β's are correlated with each other, so it is very hard to tell what is going on. A short answer is that we shouldn't have been surprised. None of the theory we have developed so far gives any positive reason why this can't happen. An even shorter answer is the following slogan.

> *If you want to know anything about a model, you* ***must*** *fit that model.*
> *You can't tell* ***anything*** *about one model by looking at the regression output for* ***some other model.***

Guessing that the coefficient of x^5 wouldn't be significant in this model from looking at the output of the other model is a mugs game. If that's not a clear example of the bogosity of stargazing, I don't know what is.

Now let us add the sample regression function to the plot. The vector $\hat{\mathbf{y}}$, which is the sample regression function evaluated at the predictor values in the data set is in the component `fitted.values` of the list returned by `lm` function (what we usually store in the variable `out`). The R commands

```
> out <- lm(y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5))
> plot(x, y)
> lines(x, out$fitted.values)
```

Figure 12.2 shows this plot. The fit isn't bad, but it isn't great either. I think I could draw a better fitting curve by hand. How can that be? Isn't linear regression BLUE? How can I do better than the "best?" Linear regression is BLUE only if the model assumptions are true. In particular, in this case, its BLUE only if the true unknown regression function is a polynomial of degree five. Doesn't appear to be. So much for the optimality of linear regression. It's only optimal in toy problems for which the answer is known. For real data where you don't know the true population regression function, it isn't.

Before leaving this section we should ask and answer the following question.

> *What's so special about polynomials? Nothing!*
> *Made up predictors can be* ***any*** *functions of the original predictors.*

Problem 12-4 explores using sines and cosines instead of polynomials.

We revisit the issue we started with, just to make sure everything is clear: is this linear regression or not? It was certainly done with a linear regression computer program, and looked at abstractly enough, it is linear regression. As we explained at the beginning (12.39) is in the form of a linear population regression function, if we consider it a function of $k+1$ variables, X^0, ..., X^k instead of just the one variable X. But if we consider it a function of just one variable X, the graph of which is the line in Figure 12.2, it isn't linear. Thus we see that linear regression is more versatile than it appears at first sight. It also does nonlinear regression by making it a special case of linear regression (quite a trick, that).

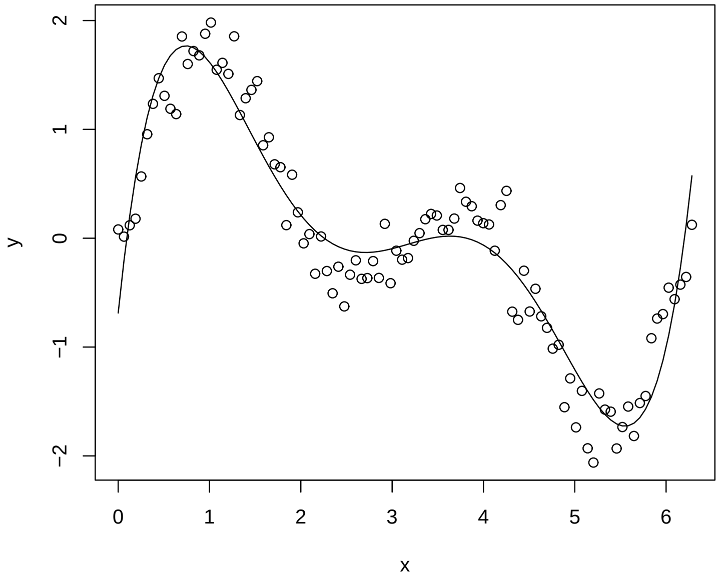


Figure 12.2: The same regression data plotted in Figure 12.1 with the best fitting polynomial of degree five added.

12.3.7 The F-Test for Model Comparison

This section deals with the regression analog of the likelihood ratio test. Suppose we have two nested regression models, and we want do a test comparing them. The null and alternative hypotheses are exactly as extensively discussed in Section 10.4.3 (Likelihood Ratio Tests). We could, in fact, just use the likelihood ratio test exactly as described in that section. It would provide an asymptotically valid test, approximately correct when the sample size is large. However, likelihood ratio tests are not traditionally used in regression. What is used, what we will develop in this section, are exact tests in which the test statistic has an F distribution. For large sample sizes, these F tests give P-values very close to the likelihood ratio test. The difference is with small sample sizes, where the likelihood ratio test is not valid, but the F tests are valid under the "strong" assumption that the errors are i. i. d. normal. (The F tests are, of course, not exact when that assumption is violated.)

Before we can state the theorem we need to see what the condition that models be "nested" means in the regression context. As with many other things about linear regression, there is a simple notion, useful when you are actually doing regression, and a mathematically sophisticated notion, useful in proofs. The simple notion is that the little model is just like the big model except that some of the regression coefficients in the big model are fixed in the little model, usually at zero.

For example, if the models under consideration are the polynomial regression models considered in the preceding section, then the big model might be all sixth degree polynomials, having regression functions of the form (12.39) with $k = 6$ and the little model might be all third degree polynomials having regression functions of the same form with $k = 3$. The little model is clearly obtained from the big model by setting $\beta_4 = \beta_5 = \beta_6 = 0$. Also clearly, the big model has three more parameters than the little model.

The mathematically sophisticated notion is that if $\mathbf{X}_{\text{big}}$ is the design matrix of the big model and $\mathbf{X}_{\text{little}}$ is the design matrix of the little model, then the models are nested if $\text{range}(\mathbf{X}_{\text{little}}) \subset \text{range}(\mathbf{X}_{\text{big}})$, or, what is equivalent because we know the range of the design matrix is the same as the range of the hat matrix, $\text{range}(\mathbf{H}_{\text{little}}) \subset \text{range}(\mathbf{H}_{\text{big}})$, where $\mathbf{H}_{\text{little}}$ and $\mathbf{H}_{\text{big}}$ are the corresponding hat matrices.

While we are at it, we generalize to a sequence of nested models. Suppose $\mathbf{H}_i$, $i = 1, \ldots, k$ are the hat matrices of a sequence of regression models. Then we say the sequence is *nested* if

$$\text{range}(\mathbf{H}_i) \subset \text{range}(\mathbf{H}_{i+1}), \qquad i = 1, \ldots, k-1 \tag{12.40}$$

Theorem 12.11. *Let* SSResid$_i$ *denote the residual sum of squares for the i-th model in a sequence of k nested regression models. Assume the smallest model is true, that is,*

$$E(Y) = \mathbf{X}_1 \boldsymbol{\beta}$$

where $\mathbf{X}_1$ *is the design matrix for the smallest model, and assume the errors satisfy* (12.24)*, then* $\text{SSResid}_k / \sigma^2$ *and*

$$\frac{\text{SSResid}_i - \text{SSResid}_{i+1}}{\sigma^2}, \qquad i = 1, \ldots, k-1$$

are independent random variables, and

$$\frac{\text{SSResid}_k}{\sigma^2} \sim \text{chi}^2(n - p_k) \tag{12.41a}$$

and

$$\frac{\text{SSResid}_i - \text{SSResid}_{i+1}}{\sigma^2} \sim \text{chi}^2(p_{i+1} - p_i), \tag{12.41b}$$

where p_i *is the dimension (number of regression coefficients) of the i-th model.*

Proof. Assertion (12.41a) does not have to be proved, since it is just the assertion of Theorem 12.8 applied to the k-th model. In the proof of Theorem 12.8, in equation (12.38) we derived a formula giving SSResid in terms of the hat matrix. If we add subscripts to make it apply to the current situation, it becomes

$$\frac{\text{SSResid}_i}{\sigma^2} = \|(\mathbf{I} - \mathbf{H}_i)\mathbf{Z}\|^2 \tag{12.42}$$

where as in the proof of Theorem 12.8, we are defining $\mathbf{Z} = \mathbf{e}/\sigma$. Now note that

$$\begin{aligned} \|(\mathbf{H}_{i+1} - \mathbf{H}_i)\mathbf{Z}\|^2 &= \mathbf{Z}'(\mathbf{H}_{i+1} - \mathbf{H}_i)^2\mathbf{Z} \\ &= \mathbf{Z}'(\mathbf{H}_{i+1}^2 - \mathbf{H}_{i+1}\mathbf{H}_i - \mathbf{H}_i\mathbf{H}_{i+1} + \mathbf{H}_i^2)\mathbf{Z} \\ &= \mathbf{Z}'(\mathbf{H}_{i+1} - \mathbf{H}_i)\mathbf{Z} \\ &= \mathbf{Z}'(\mathbf{I} - \mathbf{H}_i)\mathbf{Z} - \mathbf{Z}'(\mathbf{I} - \mathbf{H}_{i+1})\mathbf{Z} \\ &= \frac{\text{SSResid}_i - \text{SSResid}_{i+1}}{\sigma^2} \end{aligned} \tag{12.43}$$

where in the middle we use the fact that the hat matrices are idempotent and $\mathbf{H}_{i+1}\mathbf{H}_i = \mathbf{H}_i\mathbf{H}_{i+1} = \mathbf{H}_i$, which comes from Lemma H.1 in Appendix H.

Now we want to apply Theorem H.3 in the same appendix. This will show both the asserted independence and the chi-square distributions if we can show the following. The matrices

$$\mathbf{I} - \mathbf{H}_k \tag{12.44a}$$

and

$$\mathbf{H}_{i+1} - \mathbf{H}_i, \qquad i = 1, \ldots, k-1 \tag{12.44b}$$

are an orthogonal set of orthogonal projections, and

$$\operatorname{rank}(\mathbf{H}_{i+1} - \mathbf{H}_i) = p_{i+1} - p_i. \tag{12.44c}$$

Note that we can avoid treating (12.44a) as a special case by defining $H_{k+1} = \mathbf{I}$.

First we have to show that $\mathbf{H}_{i+1} - \mathbf{H}_i$ is an orthogonal projection. It is clearly symmetric, and idempotence was already shown in (12.43).

Then we have to show

$$(\mathbf{H}_{i+1} - \mathbf{H}_i)(\mathbf{H}_{j+1} - \mathbf{H}_j) = 0, \qquad i < j.$$

This also follows directly from Lemma H.1 in Appendix H.

$$\begin{aligned} (\mathbf{H}_{i+1} - \mathbf{H}_i)(\mathbf{H}_{j+1} - \mathbf{H}_j) &= \mathbf{H}_{i+1}\mathbf{H}_{j+1} - \mathbf{H}_{i+1}\mathbf{H}_j - \mathbf{H}_i\mathbf{H}_{j+1} + \mathbf{H}_i\mathbf{H}_j \\ &= \mathbf{H}_{i+1} - \mathbf{H}_{i+1} - \mathbf{H}_i + \mathbf{H}_i \end{aligned}$$

The only bit remaining to prove is (12.44c). Note that $p_i = \operatorname{rank}(\mathbf{H}_i)$, so this is the same thing as

$$\operatorname{rank}(\mathbf{H}_{i+1} - \mathbf{H}_i) = \operatorname{rank}(\mathbf{H}_{i+1}) - \operatorname{rank}(\mathbf{H}_i)$$

By definition $\operatorname{rank}(\mathbf{H}_{i+1} - \mathbf{H}_i)$ is the dimension of $\operatorname{range}(\mathbf{H}_{i+1} - \mathbf{H}_i)$. We now claim that this is the orthogonal complement of $\operatorname{range}(\mathbf{H}_i)$ in $\operatorname{range}(\mathbf{H}_{i+1})$. Consider an arbitrary vector $\mathbf{y}$ in $\operatorname{range}(\mathbf{H}_{i+1})$. Then

$$(\mathbf{H}_{i+1} - \mathbf{H}_i)\mathbf{y} = \mathbf{y} - \mathbf{H}_i\mathbf{y}$$

which is a vector orthogonal to $\operatorname{range}(\mathbf{H}_i)$. Since every vector in $\operatorname{range}(\mathbf{H}_{i+1} - \mathbf{H}_i)$ is orthogonal to every vector in $\operatorname{range}(\mathbf{H}_i)$, this implies that a basis for one is orthogonal to a basis for the other. Hence the union of the bases is a basis for $\operatorname{range}(\mathbf{H}_{i+1})$ and the dimensions add, which is what was to be proved. □

Corollary 12.12. *If* SSResid_{little} *and* SSResid_{big} *are the residual sums of squares for nested models of dimension* p_{little} *and* p_{big}*, respectively, and the "strong" model assumptions* (12.22) *and* (12.24) *hold for the little model, then*

$$\frac{\text{SSResid}_{little} - \text{SSResid}_{big}}{p_{big} - p_{little}} \cdot \frac{n - p_{big}}{\text{SSResid}_{big}} \sim F(p_{big} - p_{little}, n - p_{big}).$$

Proof. The theorem says the two random variables involving residual sums of squares are independent chi-square random variables. Dividing each by its degrees of freedom and forming the ratio makes an F random variable. □

Example 12.3.3 (Multivariable Polynomial Regression).
Let us consider whether a quadratic model or higher polynomial would fit the data of Example 12.2.1 better than the linear model used in that example. The most general quadratic model has six terms

$$E(Y \mid x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_1 x_2 + \beta_5 x_2^2$$

The R command to fit this model is

```
out <- lm(y ~ x1 + x2 + I(x1^2) + I(x1 * x2) + I(x2^2))
summary(out)
```

The part of the output that describes the regression coefficients is shown below.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 91.011781  16.984368   5.359 5.95e-07 ***
x1           0.460700   1.540967   0.299    0.766
x2           1.921907   1.399151   1.374    0.173
I(x1^2)      0.013292   0.052642   0.252    0.801
I(x1 * x2)  -0.020873   0.097794  -0.213    0.831
I(x2^2)      0.005785   0.047867   0.121    0.904
```

It says, if we are naive enough to believe the "stars" (which of course we aren't), that none of the regression coefficients except the one for the constant predictor is interesting. Of course this contradicts Example 12.2.1 where we found that the coefficient of x_2 was "significant" (yet another case illustrating how misleading stargazing is).

In order to compare this quadratic model with the linear model fit in Example 12.2.1 we should do the F-test described in this section. R provides a way to do this easily. First fit the two models, saving both results.

```
out.lin <- lm(y ~ x1 + x2)
out.quad <- lm(y ~ x1 + x2 + I(x1^2) + I(x1 * x2) + I(x2^2))
```

Then the function `anova` computes a so-called "analysis of variance" (ANOVA) table for the model comparison.

```
anova(out.lin, out.quad)
```

The output of the `anova` function is

```
Analysis of Variance Table

Model 1: y ~ x1 + x2
Model 2: y ~ x1 + x2 + I(x1^2) + I(x1 * x2) + I(x2^2)
  Res.Df Res.Sum Sq Df Sum Sq F value Pr(>F)
1     97     285119
2     94     284370  3    749  0.0825 0.9694
```

The last three lines of the printout here are a so-called "analysis of variance" table. Back in the stone age, when people calculated this stuff without computers, someone decided it was helpful to lay out the arithmetic in calculating the F statistic this way. Nowadays only the final result $F = 0.0825$ and the P-value of the test $P = 0.9694$ are interesting. But these tables give old timers a warm fuzzy feeling, so computers still print them out.

Since R calculates everything, there is nothing left for you to do except interpreting the P-value. Low P-values are evidence in favor of the alternative, high P-values in favor of the null. This one is certainly high, much higher than one would expect by chance if the null hypothesis is true. Thus we accept the null hypothesis (here, the linear model). We say it fits just as well as the quadratic model. The extra terms in the quadratic model add no predictive or explanatory value.

Thus we should examine the linear model having only x_1, x_2, and the constant predictor. But we already did this in Example 12.3.1. The printout for that example apparently shows that x_1 can also be dropped.

12.3.8 Intervals for the Regression Function

Confidence Intervals

An important problem is estimating the regression function itself, either at some specified $\mathbf{x}$ value, or at all $\mathbf{x}$ values. As everywhere else the sample is not the population. What we want to know is

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta},$$

the population mean vector. But as the Greek letters indicate, we don't know $\boldsymbol{\beta}$ hence don't know $\mathbf{X}\boldsymbol{\beta}$. What we do know is the corresponding sample quantity

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}},$$

the vector of "predicted values." What is the relation between the two?

More generally, we can consider the regression function as a function of the predictor vector $\mathbf{x} = (x_1, \ldots, x_p)$

$$E(Y \mid \mathbf{x}) = \sum_{i=1}^{p} \beta_i x_i = \mathbf{x}'\boldsymbol{\beta}$$

that we can evaluate at arbitrary $\mathbf{x}$ values, not just at so-called "design points" ($\mathbf{x}$ values occurring in the data set being analyzed). If we write this function as

$$h(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}, \tag{12.45}$$

it is obvious that the most sensible estimate is

$$\hat{h}(\mathbf{x}) = \mathbf{x}'\hat{\boldsymbol{\beta}}. \tag{12.46}$$

Before we become bogged down in details, it is worthwhile to get the big picture firmly in mind. The sample regression coefficient vector $\hat{\boldsymbol{\beta}}$ is multivariate normal and independent of the error sum of squares SSResid. The regression function estimate (12.46), being a linear transformation of a multivariate normal random vector, is a normal random scalar. Hence we can combine it with the error sum of squares to make t-confidence intervals and t-tests. All we need to do is work out the details.

We have already said that (12.46) is normal. Clearly its mean is (12.45). Hence it is an unbiased estimate of the population regression function. By Corollary 5.4 in Chapter 5 of these notes, the variance of (12.46) is $\mathbf{x}' \operatorname{var}(\hat{\boldsymbol{\beta}})\mathbf{x}$, and plugging in the variance of the sample regression coefficient vector from Theorem 12.7 gives

$$\operatorname{var}(\mathbf{x}'\hat{\boldsymbol{\beta}}) = \sigma^2 \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}. \tag{12.47}$$

If you are confused about what big $\mathbf{X}$ and little $\mathbf{x}$ are here, $\mathbf{X}$ is the design matrix for the original data set and $\mathbf{x}$ is one possible value of the predictor vector. If $\mathbf{x}$ is a value that occurs in the original data, then it is one row of the design matrix $\mathbf{X}$, otherwise $\mathbf{X}$ and $\mathbf{x}$ are unrelated. The vector $\mathbf{x}$ can be any vector of predictor values for any individual, whether one in the original data set or some other individual.

Of course, we have the usual problem that we don't know σ^2 and have to plug in the estimate $\hat{\sigma}^2$ given by (12.34). This gives a t confidence interval

$$\mathbf{x}'\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\hat{\sigma}\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}$$

where, as usual, the degrees of freedom for the t distribution used to calculate the critical value is $n - p$.

Fortunately, we don't need to do any of these calculations. R has a function `predict` that does them all. Suppose we have a data set with three predictors `x1`, `x2`, and `x3`, and a response `y`, then, as usual,

```
out <- lm(y ~ x1 + x2 + x3)
```

fits the model. Now

```
predict(out, data.frame(x1=1, x2=1, x3=1), interval="confidence")
```

produces a 95% confidence interval for the value of the population regression function at $\mathbf{x} = (1, 1, 1)$. The output is (for data which are not shown)

```
          fit      lwr      upr
[1,] 3.623616 2.022624 5.224608
```

The component labeled `fit` is the estimated value of the population regression function $\mathbf{x}'\hat{\boldsymbol{\beta}}$. The component labeled `lwr` is the lower endpoint of the 95% confidence interval and the component labeled `upr` is the upper endpoint. For different confidence level use the optional argument `level`, for example,

```
predict(out, data.frame(x1=1, x2=1, x3=1), interval="confidence",
level=.90)
```

Prediction Intervals

Actually, one rarely wants a confidence interval of the kind described in the preceding section. One usually wants a very closely related interval called a *prediction interval*. The idea is this. What is the point of knowing the population regression $E(Y \mid \mathbf{x})$? It gives BLUP (best linear unbiased predictions) for the response Y (with the usual proviso, the model must be correct). However, these predictions, even if they used the true population regression function would not be exactly correct, because Y is observed "with error." If we write $\mu = E(Y \mid \mathbf{x})$, then $Y \sim \mathcal{N}(\mu, \sigma^2)$ under the "strong" linear regression assumptions. So our "best" estimate will be wrong by about σ (the error standard deviation), sometimes more, sometimes less, because Y is a random variable.

Of course, we don't know the population regression function and are forced to substitute the sample regression function. We can write

$$Y = \mathbf{x}'\boldsymbol{\beta} + e$$

for the population regression model, but we can't use that. Let us rewrite this

$$Y = \mathbf{x}'\hat{\boldsymbol{\beta}} + \mathbf{x}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + e.$$

The first term on the right, is what we use for prediction. The second two terms are unknown errors (the middle term being unknown because we don't know the true population regression coefficient vector $\boldsymbol{\beta}$). However we do know the sampling distribution of the sum of the two terms on the right. Being a linear transformation of jointly normal random variables, it is normal with mean

$$E\{\mathbf{x}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + e\} = 0$$

and variance

$$\operatorname{var}\{\mathbf{x}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + e\} = \operatorname{var}\{\mathbf{x}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\} + \operatorname{var}(e) = \sigma^2 + \sigma^2 \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x},$$

where we used (12.47) to get the variance of $\mathbf{x}'\hat{\boldsymbol{\beta}}$. Here the first equality assumes that e and $\hat{\boldsymbol{\beta}}$ are independent random variables, which will be the case if Y here refers to a "new" individual, *not* one in the original data set used to calculate the sample regression function. Thus the "prediction" interval is almost exactly like the "confidence" interval with just a slight change in the formula

$$\mathbf{x}'\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\hat{\sigma}\sqrt{1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}$$

The "1+" in the square root is only difference between the two formulas. R also makes this interval convenient to calculate. Just do the same thing as for the confidence interval but use `interval="prediction"` as the argument specifying the type of interval wanted.

Example 12.3.4.
This continues Example 12.3.2. What we want to do here is add lines indicating the prediction intervals to Figure 12.2. The following code

```
out <- lm(y ~ x + I(x^2) + I(x^3) + I(x^4) + I(x^5))
pout <- predict(out, data.frame(x=x), interval="prediction")
plot(x, y, ylim=range(pout))
lines(x, out$fitted.values)
lines(x, pout[ , "lwr"], lty=2)
lines(x, pout[ , "upr"], lty=2)
```

(This uses a bit of magic of optional arguments. The `ylim=range(pout)` argument to the `plot` command leaves room for the confidence intervals. The `lty=2` says to use a line type different from the default. Supplying a whole vector `data.frame(x=x)` to the `predict` function, produces all the prediction intervals in one statement. Using labels as subscripts, as in `pout[ , "lwr"]` is another R idiom we won't try to explain.)

12.4 The Abstract View of Regression

> *Regression coefficients are meaningless. Only regression functions and fitted values are meaningful.*

The idea of a regression problem is to estimate a regression function. When there are several predictors, there is no unique way to express the regression function as a linear function of the predictors.

> *Any linearly independent set of linear combinations of predictor variables makes for an equivalent regression problem.*

Suppose $\mathbf{X}$ is a design matrix for a regression problem. The columns of $\mathbf{X}$ correspond to the predictor variables. Using linear combinations of predictors is like using a design matrix

$$\mathbf{X}^* = \mathbf{X}\mathbf{A},$$

where A is an invertible $p \times p$ matrix (where, as usual, there are p predictors, including the constant predictor, if there is one). The requirement that $\mathbf{A}$ be invertible is necessary so that $\mathbf{X}^*$ will have rank p if $\mathbf{X}$ does. Then

$$(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1} = [(\mathbf{X}\mathbf{A})'(\mathbf{X}\mathbf{A})]^{-1} = [\mathbf{A}'\mathbf{X}'\mathbf{X}\mathbf{A}]^{-1} = \mathbf{A}^{-1}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{A}')^{-1}$$

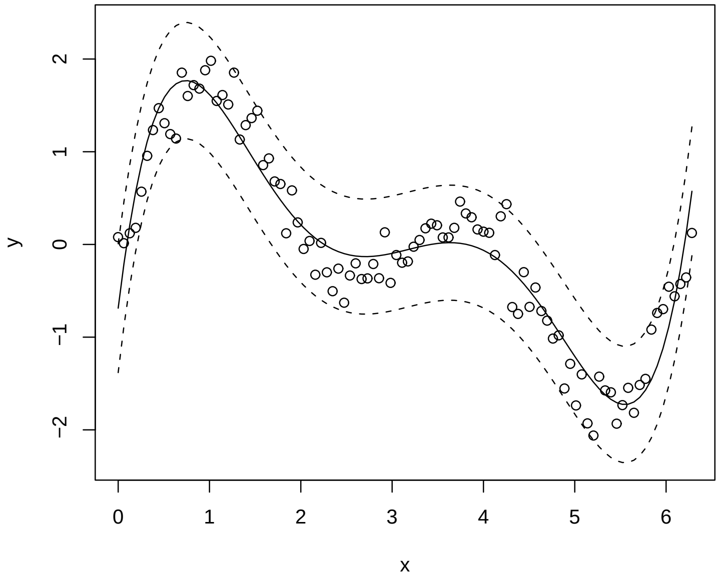


Figure 12.3: The same regression data plotted in Figure 12.1 with the best fitting polynomial of degree five added and pointwise prediction intervals.

because the inverse of a product is the product of the inverses in reverse order. (We can't apply this rule to $\mathbf{X}$ itself because $\mathbf{X}$ is not invertible. Only the product $\mathbf{X}'\mathbf{X}$ is invertible).

The regression coefficients for the "starred problem" are different

$$\begin{aligned}\hat{\boldsymbol{\beta}}^* &= (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{y} \\ &= \mathbf{A}^{-1}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{A}')^{-1}\mathbf{A}'\mathbf{X}'\mathbf{y} \\ &= \mathbf{A}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{A}^{-1}\hat{\boldsymbol{\beta}}\end{aligned}$$

because $(\mathbf{A}')^{-1}\mathbf{A}'$ is the identity and using the definition (12.18) of $\hat{\boldsymbol{\beta}}$.

Although the *regression coefficients* are different, the *fitted values* are not different!

$$\hat{\mathbf{y}}^* = \mathbf{X}^*\hat{\boldsymbol{\beta}}^* = \mathbf{X}\mathbf{A}\mathbf{A}^{-1}\hat{\boldsymbol{\beta}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{y}}$$

This means the hat matrices for the two problems are also the same (as is easily checked).

Since F tests for model comparison depend only on residual sums of squares, which depend only on fitted values, no F test is changed by replacing an "unstarred" problem with a "starred" problem in the manner described above. Nothing of statistical significance changes. There is no such thing as a test of whether the "starred" or the "unstarred" model fit the data better. Both always fit exactly the same, no better and no worse. They have the same residual sum of squares. In a sophisticated abstract sense they are two equivalent descriptions of the *same* model.

But when you look at the regression coefficients, for example, at the table the R `summary` command prints out giving the coefficients, standard errors, t statistics, and P-values, the "starred" and "unstarred" models look *very* different. The regression coefficients seem to have nothing to do with each other (though, of course, they are actually related by the linear relationship $\boldsymbol{\beta}^* = \mathbf{A}^{-1}\boldsymbol{\beta}$, this is impossible to visualize if $\mathbf{A}$ has complicated structure).

So whenever you see someone taking regression coefficients very seriously, remember that they are actually meaningless. The discussion would be better phrased in terms of prediction, predicted (fitted) values, and regression functions.

> *Regression is for prediction, not for explanation.*

Of course, what scientists mostly want from regression is explanation not prediction. But what we are saying that what they want and what regression actually delivers are two different things.

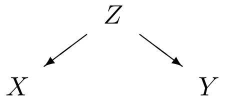
Another related slogan that is a bit off the subject, but worth throwing into the discussion for the sake of completeness, is

> *Correlation is not causation, and regression isn't either.*

What is meant by the slogan "correlation is not causation" is that mere correlation doesn't show a causative relationship between variables. This is clear from the fact that correlation is a symmetric relation (the correlation of x and y is the same as the correlation of y and x), but causal relationships are not symmetric ("x causes y" is not the same as "y causes x"). If we denote causal relationships by arrows, there are two possible causal relationships involving x and y

$$X \longrightarrow Y \qquad \text{or} \qquad X \longleftarrow Y$$

If we admit other variables into consideration, there are many possible causal relationships, for example



Here neither "x causes y" nor "y causes x" holds. Both are controlled by a third variable z. So mere existence of a correlation does not entitle us to say anything about underlying causal relationships.

Now regression is just correlation looked at from a different angle. This is clear in the "simple" case (one non-constant predictor) where the slope of the regression line $\hat{\beta}$ is related to the correlation coefficient r by

$$\hat{\beta} = r \frac{s_y}{s_x}.$$

In general, the relationship between regression and correlation is less transparent, but the regression coefficients and fitted values are functions of the sample first and second moments of the response and predictor variables (including covariances). This is clear from the formulation of least squares estimates as functions of "empirical" second moments given (12.13), (12.14), and (12.15).

Regression does not "explain" the relationship between the variables. There is no way to tell which way the causal arrow goes between the variables, or even if there is any direct causal relationship. What scientists *want* is to find causal relationships. Often, as in many social sciences, there is no way do do controlled experiments and regression is the only tool available to explore relationships between variables. Scientists want so hard to find causal relationships that they often forget the slogans above (or pay lip service to them while ignoring their content).

There is even a school of regression use called *causal modeling* that claims to use regression and similar tools to find causal relationships. But the theorems of that school are of the "ham and eggs" variety (if we had some ham, we'd have ham and eggs, if we had some eggs). First they assume there are *no* unmeasured variables that have any causal relationships with *any* measured variables (predictor or response). That is, they assume there are no variables like Z in the picture above involving three variables. Then they assume that there are non-statistical reasons for deciding which way the arrow goes in the other picture. Then they have a theorem that says causal relationships can be determined. But in the "simple" case (only two variables X and Y) this is a pure tautology

If we assume X causes Y, then we can conclude X causes Y

(well, duh!) When there are more than two variables, so-called causal modeling can yield conclusions that are not purely tautological, but they are always based on exceedingly strong assumptions (no unmeasured variables with causal connection to measured variables) that are always known to be false without a doubt. There is no real escape from "correlation is not causation, and regression isn't either."

12.5 Categorical Predictors (ANOVA)

ANOVA is just regression with all predictor variables categorical.

12.5.1 Categorical Predictors and Dummy Variables

When a predictor variable is categorical, there is no sense in which there can be one regression coefficient that applies to the variable. If x_1 is a variable taking values in the set $\{\text{Buick}, \text{Toyota}, \text{Mercedes}\}$, then $\beta_1 x_1$ makes no sense because the values of x_1 are not numbers. However, categorical predictor variables are easily incorporated into the regression framework using the device of so-called *dummy variables*.

If x is a categorical predictor variable taking values in an arbitrary set S, then the *dummy variables* associated with x are the indicator random variables (zero-one valued)

$$I_{\{s\}}(x), \qquad s \in S.$$

For example, if we have a predictor variable x associated with the predictor vector

$$\mathbf{x} = \begin{pmatrix} \text{Buick} \\ \text{Buick} \\ \text{Mercedes} \\ \text{Buick} \\ \text{Toyota} \\ \text{Mercedes} \end{pmatrix}$$

then this is associated with three dummy variable vectors that make up three columns of the design matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{12.48}$$

the first column is the indicator of the category *Buick*, the second column the indicator of the category *Toyota*, the third column the indicator of the category *Mercedes*.

Suppose we fit a model with this design matrix (no constant predictor). Call the three predictor vectors, the columns of (12.48), $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$. Then the regression model is

$$\mathbf{y} = \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \beta_3 \mathbf{x}_3 + \mathbf{e}$$

Note that exactly one of the three predictor vectors is nonzero, that is, if we write the scalar equations with one more subscript,

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i,$$

then this is the same as

$$y_i = \beta_k + e_i,$$

if the i-th individual is in category k. Thus the regression coefficient β_k is just the population mean for category k.

For technical reasons, to be explained presently, we often drop one of the predictors (it doesn't matter which) and add a constant predictor. This gives us a different design matrix. If we drop the *Mercedes* dummy variable in the example, this gives us a design matrix

$$\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \tag{12.49}$$

Now the first column is the constant predictor, the second column is the *Buick* predictor and the third is the *Toyota* predictor.

Although this seems at first sight to change the model, in the abstract sense discussed in the preceding section, it does not. The constant predictor is the sum of the rows of the original design matrix (12.48). Thus (12.48) and (12.49) are abstractly equivalent design matrices: "starred" and "unstarred" matrices in the notation of the preceding section. The two models both fit three parameters which determine estimates of the population means for the three categories. In the representation using design matrix (12.48) the regression coefficients *are* just the sample means for the three categories. In the other representation, they aren't, but the *fitted value* for each individual *will* be the sample mean for the individual's category.

Categorical predictors are so important that R makes it easy to fit models involving them

```
x <- c("Buick", "Buick", "Mercedes", "Buick", "Toyota", "Mercedes")
y <- c(0.9, 1.0, 1.9, 1.1, 3.0, 2.1)
xf <- factor(x)
out <- lm(y ~ xf)
summary(out)
```

The first two statements define the predictor `x` and the response `y`. The last two are the usual R commands for fitting a regression model and printing out various information about it. The only new wrinkle is the command in the middle that makes a special "factor" variable `xf` that will be dealt with in the right way by the `lm` command. We don't have to set up the dummy variables ourselves. R will do it automagically whenever a "factor" (i. e., categorical) variable appears in the regression formula. The reason why we have to run the original predictor variable `x` through the `factor` function is because if the category labels are numeric (instead of text as in this example) there is no way for R to tell whether we want the variable treated as quantitative or categorical unless we tell it. The `factor` function is the way we tell R we want a variable treated as categorical. We could also have compressed the two lines above involving the `factor` and `lm` functions into one

```
three <- lm(y ~ factor(x))
```

Either way, we get the following table of regression coefficients in the output of the `summary` command. Only the labels of the regression coefficients differ. All the numbers are identical.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.00000    0.06667  15.000 0.000643 ***
xfMercedes   1.00000    0.10541   9.487 0.002483 **
xfToyota     2.00000    0.13333  15.000 0.000643 ***
```

These regression coefficients are not easy to interpret. Their interpretation depends on the actual design matrix R uses, which is neither of the design matrices (12.48) or (12.49) described above. However, we shouldn't let this bother us in the slightest. The slogan at the beginning of the preceding section tells us that regression coefficients are meaningless anyway. They are especially meaningless here. What is important are the fitted values

```
> predict(out)
1 2 3 4 5 6
1 1 2 1 3 2
```

Comparing this with the definition of x. We see that individuals 1, 2, and 4 are in the *Buick* category. All have predicted value 1. Thus that is the sample mean for the *Buick* category. Similarly, the sample means for the *Toytota* and *Mercedes* categories are 3 and 2, respectively.

If we actually wanted to force the regression coefficients to be the sample means, we could do that.

```
  x1 <- as.numeric(x == "Buick")
  x2 <- as.numeric(x == "Toyota")
  x3 <- as.numeric(x == "Mercedes")
  out.too <- lm(y ~ x1 + x2 + x3 + 0)
  summary(out.too)
```

Gives the output

```
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x1  1.00000    0.06667   15.00 0.000643 ***
x2  3.00000    0.11547   25.98 0.000125 ***
x3  2.00000    0.08165   24.50 0.000149 ***
```

The design matrix for this regression is (12.48)

```
> cbind(x1, x2, x3)
     x1 x2 x3
[1,]  1  0  0
[2,]  1  0  0
[3,]  0  0  1
[4,]  1  0  0
[5,]  0  1  0
[6,]  0  0  1
```

But we don't need to worry about this because "regression coefficients are meaningless." Either regression gives the same predicted values. They agree about every statistically meaningful quantity.

Now we return to the promised explanation of the technical reason why a design matrix like (12.49) is preferred to one like (12.48). Suppose we have *two* categorical predictors, say

$$\begin{pmatrix} \text{Buick} & \text{red} \\ \text{Buick} & \text{yellow} \\ \text{Mercedes} & \text{red} \\ \text{Buick} & \text{yellow} \\ \text{Toyota} & \text{red} \\ \text{Mercedes} & \text{yellow} \end{pmatrix}$$

Now there are *five* dummy variables

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix}$$

The first three columns are the same as in (12.48), the fourth column the indicator of the category *red*, and the fifth column the indicator of the category *yellow*. But (and this is a very important "but") this design matrix does not have full rank, because the first three columns add to the predictor vector that is all ones, and so do the last two columns. The rank is only 4, not 5. In order to have uniquely determined regression coefficients, we must have an $n \times 4$ design matrix. The simple way to achieve this is to drop one dummy variable from each set, it doesn't matter which, and add the constant predictor. This gives us something like

$$\begin{pmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

here we have the *constant*, *Buick*, *Toyota*, and *red* dummy variables. We've kept all but one of each set (two cars, one color). R does this automagically, we don't have to do anything special. With `x` and `y` defined as above

```
z <- c("red", "yellow", "red", "yellow", "red", "yellow")
out <- lm(y ~ factor(x) + factor(z))
summary(out)
```

produces the following table of regression coefficients

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.88571    0.04518  19.606  0.00259 **
factor(x)Mercedes  1.02857    0.04949  20.785  0.00231 **
factor(x)Toyota    2.11429    0.06999  30.210  0.00109 **
factor(z)yellow    0.17143    0.04949   3.464  0.07418 .
```

It always does the right thing, provided you remember to tell it that categorical variables are "factors." (Well perhaps we should have said it always does *a* rather than *the* right thing. It didn't keep the same dummy variables, that we suggested. But it did keep two cars and one color, which is all that matters.)

No problem arises in mixing quantitative and categorical random variables. Just do it (remembering to tell R which predictors are categorical)!

Example 12.5.1.
Suppose we have a data set like

```
http://www.stat.umn.edu/geyer/5102/ex12.5.1.dat
```

which has one categorical predictor variable `sex`, one quantitative predictor `x` and a response `y`. Suppose we want to fit parallel regression lines for each of the categories, as in Figure 12.4. We will see how to make such a plot below, but first we need to discuss how to fit the regression model. If we let z denote the dummy variable indicating one of the two category values, the regression model we want has the form

$$\mathbf{y} = \alpha + \beta \mathbf{z} + \gamma \mathbf{x} + \mathbf{e}.$$

Here γ is the slope of both regression lines in the figure, α is the y-intercept of one of the lines, and $\alpha + \beta$ is the y-intercept of the other. Now we see that

```
out <- lm(y ~ factor(sex) + x)
summary(out)
```

fits this regression model. The part of the printout concerning the regression coefficients is

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.9253     0.2352  20.938  < 2e-16 ***
factor(sex)  -2.0633     0.1945 -10.606  < 2e-16 ***
x             1.0688     0.3316   3.223  0.00173 **
```

Figure 12.4 was made with the following commands.

```
f <- sex == "female"
plot(x, y, type="n")
points(x[f], y[f], pch="f")
points(x[!f], y[!f], pch="m")
lines(x[f], predict(out)[f])
lines(x[!f], predict(out)[!f])
```

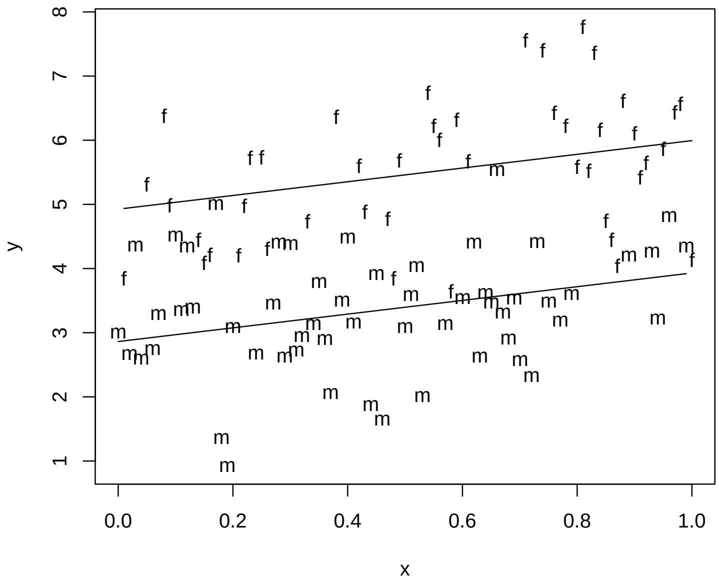


Figure 12.4: Scatter plot with parallel regression lines. The letters "f" and "m" indicate points in the two categories. The upper line is the regression line for the female category, the lower for the male category.

(If this seems a little magical, never mind. Doing fancy things with R graphics is complicated and beyond the scope of this course, not theoretical statistics.)

Great! But how about some statistics, say a test or a confidence interval? One question that is interesting is whether the true population slopes of the regression lines are the same or different. In order to find out about the "big model" that allows different slopes, we need to fit that model.

One obvious way to fit it is to divide the data, and fit a regression line to each category separately. There will be two regression coefficients (slope and intercept) for each category, making four in all. But this won't be useful for doing the test. We need fit a model to all the data that has the same predicted values (is abstractly the same regression). A little thought about dummy variables tells us that the following model will do what we want

$$\mathbf{y} = \alpha + \beta\mathbf{z} + \gamma\mathbf{x} + \delta\mathbf{x}\cdot\mathbf{z} + \mathbf{e}.$$

Here γ is the slope of one regression line and $\gamma + \delta$ is the slope of the other. As before, α is the y-intercept of one of the lines, and $\alpha + \beta$ is the y-intercept of the other. Thus something like

```
out.too <- lm(y ~ factor(sex) + x + I(factor(sex) * x)) # bogus!
```

would seem to be what is wanted. But actually, the much simpler

```
out.too <- lm(y ~ factor(sex) * x)
```

works. R assumes we want the so-called "main effects" `sex` and `x` whenever we specify the "interaction" `sex * x`. Also we do not need to enclose the multiplication in the `I()` function, because the `*` here doesn't really indicate multiplication. Rather it is a magic character indicating "interaction" that R recognizes in model formula and treats specially (just like `+` is magic). In fact, the more complicated form *doesn't* work. One must use the simple form. The part of the printout of the `summary(out.too)` command that is the table of regression coefficients is

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.7479     0.3013  15.757  < 2e-16 ***
factor(sex)    -1.7464     0.3885  -4.495 1.92e-05 ***
x               1.3822     0.4698   2.942  0.00408 **
factor(sex).x  -0.6255     0.6637  -0.943  0.34825
```

The four regression coefficients are α, β, γ, δ in the discussion above (in that order). A test of whether the two lines have the same slope or not, is just a test of $\delta = 0$. Hence we can read the P-value right off the printout: $P = 0.348$ (two-tailed). There is no statistically significant difference in the slopes of the two regression lines. Thus we are free to adopt the simpler model fit before with only three parameters.

If we wished to next ask the question whether a single line would fit the data (a two-parameter model), we could read the P-value for that test off the printout of the three-parameter model: $P < 2 \times 10^{-16}$ (two-tailed, though it doesn't matter for a P-value this low). Hence there is a highly statistically significant difference between the intercepts for the two categories.

12.5.2 ANOVA

Often, regression with all predictors categorical is called *analysis of variance* (ANOVA). Most textbooks, Lindgren (Sections 14.7 through 14.10) give this special case very special treatment. We won't bother, being content with the slogan that began this section.

We will just redo Example 14.7a in Lindgren to show how to do it in R

```
analyst <- c(1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 3, 3, 3, 4, 4, 4, 4)
yield <- c(8, 5, -1, 6, 5, 3, 7, 12, 5, 3, 10, 4, -2, 1, 1, 6, 10, 7)
out <- aov(yield ~ factor(analyst))
summary(out)
```

produces the same table as in Lindgren, the only difference is that R adds the P-value for the F test.

The `aov` function "cuts to the chase." In ANOVA you almost always want to do F tests for models that include all the dummy variables for a given category or none. It just goes straight to the analysis of variance table for such comparisons.

Here where there is just one categorical variable, there is just one test. In so-called two-way ANOVA (Section 14.9 in Lindgren) there are two tests, one for each category. If an interaction is added (p. 533 ff. in Lindgren) that adds another test. And so forth.

12.6 Residual Analysis

An important topic that has been ignored so far is how one checks whether the model assumptions are plausible. There is, of course, never any way to prove they are correct, but it is generally accepted that one should make some effort to show that the model assumptions are not completely ridiculous.

This has not always been the case. Back in the stone age, when computers didn't come with video monitors and statistics programs just produced printout, techniques in this section were not used. People did regression with no useful checks of model assumptions, hence often when it was completely ridiculous, although they had no awareness of the ridiculosity.

Nowadays, you can install R (or similar software) on any computer and easily make diagnostic plots that will reveal some violations of model assumptions. (And miss some. There is no magic that will reveal *all* violations.) We can only scratch the surface of this area. Books on regression have much more.

For a start, we divide the (strong) assumptions into two classes.

- Assumptions about **errors**.
  - independent
  - normal
  - homoscedastic (same variance)
- The assumption about **the regression function**.

These two classes of assumptions are treated quite differently. The assumption of a particular form for the regression function is checked using F tests for model comparison. If a particular model is wrong, a larger model may be right. Presumably, some large enough model will be right. The only problem is to find it. So when we said we had been ignoring model checking, that wasn't quite right. We haven't ignored this part (although we will have a bit more to say about it later.)

To be precise, we should modify the last paragraph to say that F tests check the assumption about the regression function, *if the other assumptions are correct*. If the error assumptions don't hold, then the F statistic doesn't have an F distribution, and there's no way to interpret it.

Thus logically, the error assumptions come first, but there is a slight problem with this. We don't see the errors, so we can't check them. We do have error estimates $\hat{e}_i$, but they depend on the model we fit, which depends on the assumed regression function.

Some misguided souls attempt to avoid this dilemma by applying their checks to the responses y_i, but this is entirely misguided. The responses y_i are not identically distributed, either conditionally or unconditionally, so there is no point in looking at their distribution. Moreover the *marginal* distribution of the responses y_i is not assumed to be normal in regression theory, only the *conditional* distribution given the predictor variables. Hence there is *no* useful conclusion about regression that can be derived from checking whether the responses appear normally distributed. If they appear normal, that doesn't prove anything. If they appear non-normal, that doesn't prove anything either. I stress this because I often see naive users of regression looking at histograms of the response variable, and asking what it means. Then I trot out a slogan.

Normality checks must be applied to ***residuals****, not responses.*

Thus to return to our dilemma. We can't check the assumptions about the regression function until we have checked the error assumptions. But we can't check the error assumptions without knowing the correct regression function. The only way to proceed is to apply checks about error assumptions to residuals from a model that is large enough so that one can reasonably hope it is correct. So always apply these checks to residuals from the *largest* model under consideration, not any smaller model. That doesn't really avoid the dilemma, but it's the best one can do.

So what is the distribution of the residuals? The *errors* are i. i. d. normal (at least, that's the assumption we want check), but the residuals aren't.

Theorem 12.13. *Under the assumptions* (12.22) *and* (12.24)

$$\hat{\mathbf{e}} \sim \mathcal{N}\bigl(0, \sigma^2(\mathbf{I} - \mathbf{H})\bigr),$$

where $\mathbf{H}$ *is the "hat" matrix* (12.37).

The proof is left as an exercise.

The theorem says the residuals are jointly multivariate normal, but are neither independent nor identically distributed. Hence they do have the *normality* property assumed for the errors, but not the *independence* or *constant variance* properties.

It is a sad fact that there is no sensible test for independence of the errors. Even if we observed the errors (which we don't), there would be no test that could, even in principle tell whether they were independent. The problem is that there are too many ways that random variables can be dependent, and no test can rule them all out. If you test for some particular form of dependence, and the test accepts the null hypothesis, that does not prove independence. Some other form of dependence may be there. Thus the independence assumption is usually not checked. We have to proceed on hope here.

In checking the other properties, the lack of identical distribution is a problem. How can we check if the residuals are normal, if each one has to be checked against a different normal distribution? The obvious solution is to standardize the residuals. The random quantities

$$\frac{\hat{e}_i}{\sigma\sqrt{1 - h_{ii}}} \tag{12.50}$$

are not independent because the $\hat{e}_i$ are correlated, but they are (marginally) identically distributed, in fact, standard normal (under the "strong" regression assumptions). Hence, plugging in $\hat{\sigma}$ for σ gives quantities

$$t_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \tag{12.51}$$

that are identically $t(n - p)$ distributed. The quantities (12.51) are called *internally studentized residuals* and are often used to check whether the residuals are normal.

We, following R, are going to ignore them and look at a better idea, so-called *externally studentized residuals.* But the path to get there is long. It will require some patience to follow.

12.6.1 Leave One Out

A problem with residuals and internally studentized residuals is that in the i-th residual

$$\hat{e}_i = y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}$$

the data y_i is used twice because $\hat{\boldsymbol{\beta}}$ depends on all the y's including y_i. A better, more honest, estimate of the error e_i, is

$$\hat{e}_{(i)} = y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}_{(i)}$$

where $\hat{\boldsymbol{\beta}}_{(i)}$ is the regression estimate obtained by dropping the i-th case from the data. This is called a *leave-one-out residual.* Note that subscripts in parentheses do not indicate order statistics (as they did in Chapter 7 of these notes). In this section the indicate various quantities associated with a leave-one-out regression.

It would seem that leave-one-out residuals would be a big pain. It would require doing n regressions rather than just one to calculate these residuals. It is a very interesting fact about regression that this not so. The leave-one-out residuals can be calculated from the original regression using all the data. We will now see how to do this. Our analysis will also derive the distribution of the leave-one-out residuals.

Lemma 12.14. *If* $\mathbf{A}$ *is a symmetric matrix,* $\mathbf{a}$ *is a vector, and* $\mathbf{a}'\mathbf{A}\mathbf{a} \neq 1$*, then*

$$(\mathbf{A} - \mathbf{a}\mathbf{a}')^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}\mathbf{a}\mathbf{a}'\mathbf{A}^{-1}}{1 - \mathbf{a}'\mathbf{A}^{-1}\mathbf{a}}$$

Proof. We merely need to multiply $\mathbf{A} - \mathbf{a}\mathbf{a}'$ by the formula the lemma asserts is its inverse and check that we get the identity.

$$\begin{aligned}(\mathbf{A} - \mathbf{a}\mathbf{a}') \left(\mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}\mathbf{a}\mathbf{a}'\mathbf{A}^{-1}}{1 - \mathbf{a}'\mathbf{A}^{-1}\mathbf{a}} \right) &= I - \mathbf{a}\mathbf{a}'\mathbf{A}^{-1} + \frac{\mathbf{a}\mathbf{a}'\mathbf{A}^{-1} - \mathbf{a}\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}\mathbf{a}'\mathbf{A}^{-1}}{1 - \mathbf{a}'\mathbf{A}^{-1}\mathbf{a}} \\ &= I - \mathbf{a}\mathbf{a}'\mathbf{A}^{-1} + \frac{(1 - \mathbf{a}'\mathbf{A}^{-1}\mathbf{a})\mathbf{a}\mathbf{a}'\mathbf{A}^{-1}}{1 - \mathbf{a}'\mathbf{A}^{-1}\mathbf{a}} \\ &= I - \mathbf{a}\mathbf{a}'\mathbf{A}^{-1} + \mathbf{a}\mathbf{a}'\mathbf{A}^{-1} \\ &= I\end{aligned}$$

The only tricky bit is the second equality, which results from the realization that the factor $\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}$ in $\mathbf{a}\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}\mathbf{a}'\mathbf{A}^{-1}$ is a *scalar* an hence can be factored out. □

Lemma 12.15. *For any matrix* $\mathbf{X}$*, let* $\mathbf{x}_i$ *denote the (column) vector corresponding to the i-th row of* $\mathbf{X}$ *and* $\mathbf{X}_{(i)}$ *the matrix obtained by deleting the i-th row from* $\mathbf{X}$*, then*

$$\mathbf{X}'\mathbf{X} = \mathbf{X}'_{(i)}\mathbf{X}_{(i)} + \mathbf{x}_i\mathbf{x}'_i$$

Proof. Obvious. Just write out in detail what the formulas mean. □

Lemma 12.16. *With the notation in the preceding lemma, if*

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

has elements h_{ij}*, then*

$$h_{ii} = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$$

Proof. Obvious. Just write out in detail what the formulas mean. □

Corollary 12.17.

$$\left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)} \right)^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1 - h_{ii}} \tag{12.52}$$

Theorem 12.18. *Let* $\mathbf{y}$ *have elements* y_i *and let* $\mathbf{y}_{(i)}$ *denote the vector obtained by deleting the i-th element from* $\mathbf{y}$*. Define*

$$\begin{aligned}\hat{y}_i &= \mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ \hat{y}_{(i)} &= \mathbf{x}'_i \left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)} \right)^{-1} \mathbf{X}'_{(i)}\mathbf{y}_{(i)}\end{aligned}$$

and

$$\begin{aligned}\hat{e}_i &= y_i - \hat{y}_i \\ \tilde{e}_i &= y_i - \hat{y}_{(i)}\end{aligned}$$

Then

$$\hat{e}_{(i)} = \frac{\hat{e}_i}{1 - h_{ii}} \tag{12.53}$$

Proof. Using the corollary, and

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'_{(i)}\mathbf{y}_{(i)} + y_i\mathbf{x}_i \tag{12.54}$$

which is proved like Lemma 2.39,

$$\begin{aligned}
\hat{y}_{(i)} &= \mathbf{x}_i'\left((\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}}{1 - h_{ii}}\right)(\mathbf{X}'\mathbf{y} - y_i\mathbf{x}_i) \\
&= \hat{y}_i - h_{ii}y_i + \frac{h_{ii}\hat{y}_i - h_{ii}^2 y_i}{1 - h_{ii}} \\
&= \frac{\hat{y}_i - h_{ii}y_i}{1 - h_{ii}}
\end{aligned}$$

And

$$\hat{e}_{(i)} = \frac{y_i - \hat{y}_i}{1 - h_{ii}} = \frac{\hat{e}_i}{1 - h_{ii}}$$

which is the assertion (12.53) of the theorem. □

Thus we finally arrive at the definition of the leave-one-out residuals in terms of the ordinary residuals (12.53). At first sight, this doesn't seem to do much because the leave-one-out residuals are just a constant times the ordinary residuals (a different constant for each residual, but "constant" here means nonrandom rather than "same") hence when standardized are exactly the same (12.50). However, a bit deeper thought says that the "plug-in" step that follows is different. Instead of (12.51) we should plug in the standard error for the *leave-one-out* regression obtaining

$$t_{(i)} = \frac{\hat{e}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} \tag{12.55}$$

where $\hat{\sigma}_{(i)}$ is the estimate of σ obtained by dropping the i-th case from the data. These residuals (12.55) are called *externally studentized residuals.*

These residuals are identically $t(n - 1 - p)$ distributed, because $\hat{\sigma}_{(i)}$ is based on $n - 1$ data points and p predictors. They are exactly the t statistics for the test of whether y_i is data from the model by whether the prediction interval for y_i based on the other $n-1$ data points covers y_i. (This is not obvious, since we didn't derive them that way.)

We are not quite finished with our theoretical derivation. We still need a formula for $\hat{\sigma}_{(i)}$ that doesn't require a new regression procedure.

Lemma 12.19.

$$\hat{\sigma}^2_{(i)} = \hat{\sigma}^2 \frac{n - p - t_i^2}{n - p - 1}$$

where the t_i are the internally studentized residuals (12.51).

Proof. By definition

$$
\begin{aligned}
\hat{\sigma}^2 &= \frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{n-p} \\
&= \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})^2\mathbf{y}}{n-p} \\
&= \frac{\mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}}{n-p}
\end{aligned}
$$

because $\hat{\mathbf{e}} = (\mathbf{I}-\mathbf{H})\mathbf{y}$ and $\mathbf{I}-\mathbf{H}$ is symmetric and idempotent.

Hence the analogous formula

$$
\hat{\sigma}^2_{(i)} = \frac{\mathbf{y}'_{(i)}\bigl(\mathbf{I}-\mathbf{H}_{(i)}\bigr)\mathbf{y}_{(i)}}{n-p-1}
$$

holds for the leave-one-out regression. Now

$$
\mathbf{y}'_{(i)}\bigl(\mathbf{I}-\mathbf{H}_{(i)}\bigr)\mathbf{y}_{(i)} = \mathbf{y}'_{(i)}\mathbf{y}_{(i)} - \mathbf{y}'_{(i)}\mathbf{H}_{(i)}\mathbf{y}_{(i)}
$$

and the first term is

$$
\mathbf{y}'_{(i)}\mathbf{y}_{(i)} = \mathbf{y}'\mathbf{y} - y_i^2. \tag{12.56}
$$

The second term is more complicated but can be calculated using (12.54) and (12.52).

$$
\begin{aligned}
\mathbf{y}'_{(i)}\mathbf{H}_{(i)}\mathbf{y}_{(i)} &= \mathbf{y}'_{(i)}\mathbf{X}_{(i)}\left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}'_{(i)}\mathbf{y}_{(i)} \\
&= (\mathbf{y}'\mathbf{X} - y_i\mathbf{x}'_i)\left((\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1-h_{ii}}\right)(\mathbf{X}'\mathbf{y} - y_i\mathbf{x}_i) \\
&= \mathbf{y}'\mathbf{H}\mathbf{y} - 2y_i\hat{y}_i + h_{ii}y_i^2 + \frac{\hat{y}_i^2 - 2h_{ii}y_i\hat{y}_i + h_{ii}^2y_i^2}{1-h_{ii}} \\
&= \mathbf{y}'\mathbf{H}\mathbf{y} + \frac{\hat{y}_i^2 - 2y_i\hat{y}_i + h_{ii}y_i^2}{1-h_{ii}}
\end{aligned}
$$

Subtracting this from (12.56) gives

$$
\begin{aligned}
(n-p-1)\hat{\sigma}^2_{(i)} &= \mathbf{y}'\mathbf{y} - y_i^2 - \mathbf{y}'\mathbf{H}\mathbf{y} - \frac{\hat{y}_i^2 - 2y_i\hat{y}_i + h_{ii}y_i^2}{1-h_{ii}} \\
&= (n-p)\hat{\sigma}^2 - y_i^2 - \frac{\hat{y}_i^2 - 2y_i\hat{y}_i + h_{ii}y_i^2}{1-h_{ii}} \\
&= (n-p)\hat{\sigma}^2 - \frac{\hat{y}_i^2 - 2y_i\hat{y}_i + y_i^2}{1-h_{ii}} \\
&= (n-p)\hat{\sigma}^2 - \frac{\hat{e}_i^2}{1-h_{ii}} \\
&= (n-p)\hat{\sigma}^2 - \hat{\sigma}^2t_i^2
\end{aligned}
$$

and solving for $\hat{\sigma}^2_{(i)}$ gives the assertion of the lemma. □

Linear algebra is really remarkable. Or perhaps it is least squares that is so remarkable. There is something magic, that such a complicated calculation yields such a simple result. Whatever the reason, we now have two simple formulas for identically distributed, standardized residual estimates. Different computer statistical software packages make different choices about which residuals to use. We agree with the R team that the externally studentized residuals are the ones to use.

12.6.2 Quantile-Quantile Plots

How does one check that externally studentized residuals are $t(n-p-1)$ distributed? A widely used method uses a quantile-quantile (Q-Q) plot. Q-Q plots can be applied to any random quantities, not just residuals. So temporarily forget residuals.

Let X_1, X_2, $\ldots$, X_n be data assumed to be i. i. d., and suppose we want to check whether their distribution has a particular distribution with c. d. f. F. A Q-Q plot is a plot of the order statistics $X_{(k)}$ of the data[5] against quantities that are reasonable theoretical positions of these order statistics. We can't be more precise than that, because there are many different proposals about what positions should be used. Two are

$$F^{-1}\left(\frac{k-\frac{1}{2}}{n}\right) \tag{12.57}$$

and

$$F^{-1}\left(\frac{k}{n+1}\right) \tag{12.58}$$

Some more proposals will be discussed below.

We know (Theorem 9 of Chapter 3 in Lindgren) that if F is continuous, then the variables

$$U_i = F(X_i)$$

are i. i. d. $\mathcal{U}(0,1)$, hence of course, the order statistics $U_{(k)}$ are order statistics of a sample of size n from the $\mathcal{U}(0,1)$ distribution, and

$$X_{(k)} = F^{-1}(U_{(k)}).$$

The reason why this is important is that we know the distribution of the $U_{(k)}$

$$X_{(k)} \sim \text{Beta}(k, n-k+1) \tag{12.59}$$

(p. 217 in Lindgren). Hence

$$E\{U_{(k)}\} = \frac{k}{n+1}.$$

[5] Here the $X_{(k)}$ indicate order statistics as in Chapter 7 of these notes, not the leave-one-out quantities of the preceding section. In fact, since the X_i here have nothing to do with regression, the parenthesized subscripts couldn't possibly indicate leave one out.

That is the origin of (12.58). Of course, this doesn't prove that (12.58) is the Right Thing. Far from it. We know that in general

$$g\bigl(E\{X\}\bigr) = E\bigl\{g(X)\bigr\} \tag{12.60}$$

is generally *false*. The only condition we know that makes (12.60) hold is that g be a *linear* function. Now inverse c. d. f.'s are never linear except for the special case of the uniform distribution. Hence

$$E\left\{X_{(k)}\right\} = E\left\{F^{-1}\bigl(U_{(k)}\bigr)\right\} \neq F^{-1}\left(E\bigl\{U_{(k)}\bigr\}\right) = F^{-1}\left(\frac{k}{n+1}\right)$$

Thus, although (12.58) has some theoretical woof associated with it, it does not do exactly the right thing. We can only consider (12.58) *a thing to do* (as opposed to *the* thing to do). Hence the other proposals.

The proposal (12.57) has less theory behind it. It is based on the idea that n in the denominator rather than $n + 1$ is more natural (no theoretical reason for this). Unit spacing between the points also seems natural. Then the requirement that they be placed symmetrically in the interval $(0, 1)$ determines the form $(k - \frac{1}{2})/n$.

Another proposal often seen is so-called *normal scores*. These are $E\bigl\{X_{(k)}\bigr\}$ when the X_i have a standard normal distribution. The are, however, hard to compute. Some statistics packages have them, but not all. R doesn't. Of course, these are only useful when the distribution of interest is the normal distribution. The analogous quantities could be defined for any distribution, but software and tables exist only for the normal.

A proposal that does work for all distributions would be to put the *medians* of the beta distributions (12.59) through the inverse c. d. f., that is, if ζ_k is the median of the Beta$(k, n-k+1)$ distribution use $F^{-1}(\zeta_k)$ as the plotting points. This proposal has the virtue of coming from a correct theoretical argument. The median of $X_{(k)}$ is indeed $F^{-1}(\zeta_k)$, because medians (as opposed to means) do indeed go through quantile transforms.

In practice all of these proposals produce almost the same picture. So nobody worries about the differences, and does what seems simplest. The R function `qqnorm` does a Q-Q plot against the normal distribution and uses the proposal (12.57) for the plotting points. Here's how to do a Q-Q plot against a normal distribution of an arbitrary data vector `x` in R.

```
qqnorm(x)
qqline(x)
```

The first command does the Q-Q plot. The second puts on a line about which the points should cluster. Since we don't know the parameters μ and σ^2 of the population from which `x` was drawn, we don't know in advance which line the points should cluster about (`qqnorm` plots against the *standard* normal. If the data are standard normal, then the points cluster about the line with intercept zero and slope one. If the data are normal (but not standard normal), then the points cluster about the line with intercept μ and slope σ. So if the points cluster

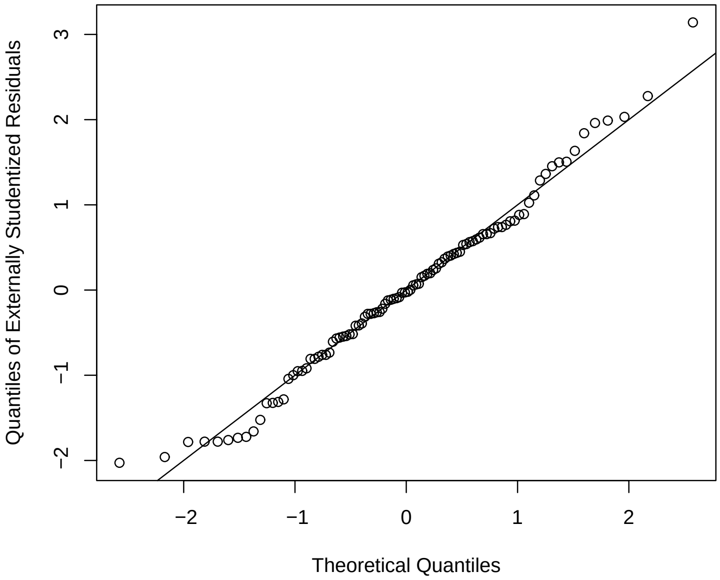


Figure 12.5: A Q-Q Plot.

about *any* line, we conclude they are approximately normally distributed. R picks a reasonable line, one that most of the points cluster about.

Here's how to do a Q-Q plot of the externally studentized residuals in R, assuming we have already fit a linear model and put the result of the `lm` command in a variable `out`

```
qqnorm(rstudent(out))
abline(0, 1)
```

The second command draws a line with intercept zero and slope one. If the residuals are approximately standard normal,[6] then the points in the plot should lie near this line.

Example 12.6.1.
This looks at the residuals from the fifth degree polynomial fit to the data of Example 12.3.2. Figure 12.5 shows the Q-Q plot of the externally studentized residuals.

It's not clear what one is supposed to make of a Q-Q plot. The points never lie *exactly* on the line because of chance variation (the sample is not the population). So how far off the line do they have to be before you should question

[6] Why standard normal when the externally studentized residuals have marginal distribution $t(n-p-1)$? Because they are not independent samples from this distribution. In fact the random parts of the denominators $\hat{\sigma}_{(i)}$ are highly correlated. Thus they look much more like a random sample from a normal than from a t distribution.

the regression assumptions? One sensible recommendation is to "calibrate" your eye by looking at several Q-Q plots for data simulated from the correct distribution (here standard normal).

```
qqnorm(rnorm(n))
abline(0, 1)
```

where `n` is the number of cases in the data. Repeat until you get an idea how much variation you should expect.

But isn't there a hypothesis test we should apply? There are hypothesis tests one can do. The trouble with them is that, when n is large, they tend to find "statistically significant" departures from normality that are fairly minor in their effects. Linear regression is somewhat robust against minor departures from normality of the error distribution. So small departures, "statistically significant" though they may be, can be ignored. What we are looking for here is really obvious and serious nonnormality.

12.7 Model Selection

This section discusses the problem of choosing among many models. Although our examples will be regression problems and some of the discussion and methods will be specific to linear regression, the problem is general. When many models are under consideration, how does one choose the best? Thus we also discuss methods that apply outside the regression context.

There are two main issues.

- Non-nested models
- Many models.

The only methods for model comparison we have studied, the F test for comparison of linear regression models and the likelihood ratio test for comparison of general models, are valid only for comparing two *nested* models. We also want to test *non-nested* models, and for that we need new theory. When more than two models are under consideration, the issue of correction for multiple testing arises. If there are only a handful of models, Bonferroni correction (or some similar procedure) may suffice. When there are many models, a conservative correction like Bonferroni is too conservative. Let's consider how many models might be under consideration in a model selection problem. Consider a regression problem with k predictor variables.

- [**Almost the worst case**] There are 2^k possible submodels formed by choosing a subset of the k predictors to include in the model (because a set with k elements has 2^k subsets).
- [**Actually the worst case**] That doesn't consider all the new predictors that one might "make up" using functions of the old predictors. Thus there are potentially infinitely many models under consideration.

Bonferroni correction for infinitely many tests is undefined. Even for 2^k tests with k large, Bonferroni is completely pointless. It would make nothing statistically significant.

12.7.1 Overfitting

> *Least squares is **good** for model fitting, but **useless** for model selection.*

Why? A bigger model *always* has a smaller residual sum of squares, just because a minimum taken over a larger set is smaller.[7] Thus least squares, taken as a criterion for model selection says "always choose the biggest model." But this is silly. Consider what the principle of choosing the biggest model says about polynomial regression.

Example 12.7.1 (Overfitting in Polynomial Regression).
Consider regression data shown in Figure 12.6. Couldn't ask for nicer data for simple linear regression. The data appear to fit the "simple" model (one non-constant predictor, which we take to be x itself).

But now consider what happens when we try to be a bit more sophisticated. How do we know that the "simple" model is o. k.? Perhaps we should consider some more complicated models. How about trying polynomial regression? But if we consider all possible polynomial models, that's an infinite number of models (polynomials of all orders).

Although an infinite number of models are potentially under consideration, the fact that the data set is finite limits the number of models that actually need to be considered to a finite subset. You may recall from algebra that any set of n points in the plane having n different x values can be interpolated (fit exactly) by a polynomial of degree $n - 1$. A polynomial that fits exactly has residual sum of squares zero (fits perfectly). Can't do better than that by the least squares criterion! Thus all polynomials of degree at least $n - 1$ will give the same fitted values and zero residual sum of squares. Although they may give different predicted values for x values that do not occur in the data, they give the same predicted values at those that do. Hence the polynomials with degree at least $n - 1$ cannot be distinguished by least squares. In fact the polynomials with degree more than $n - 1$ cannot be fit at all, because they

[7] If g is any real-valued function and A and B are two subsets of the domain of g with $A \subset B$ then

$$\inf_{x \in A} g(x) \geq \inf_{y \in B} g(y)$$

simply because every x that occurs on the left hand side also occurs as a y on the right hand side because $A \subset B$. Taking g to be the least squares criterion for a regression model, and A and B the parameter spaces for two nested models gives the result that the larger model always has the smaller residual sum of squares.

Note this is exactly analogous to what happens with maximum likelihood. The larger model always has the larger log likelihood. The reasoning is exactly the same except for the inequality being reversed because of maximizing in maximum likelihood rather than minimizing in least squares.

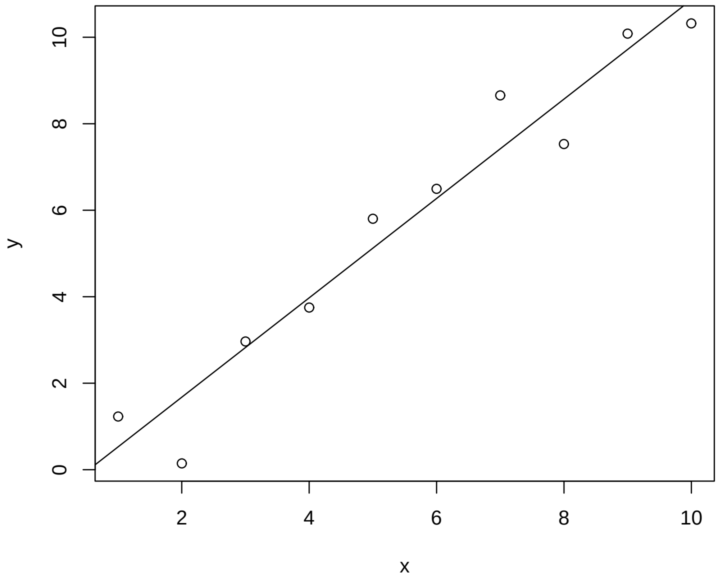


Figure 12.6: Some regression data. With fitted "simple" regression function of the form $y = \hat{\alpha} + \hat{\beta}x$.

have more parameters than there are equations to determine them. $\mathbf{X}'\mathbf{X}$ is a singular matrix and cannot inverted to give unique estimates of the regression coefficients.

This is a general phenomenon, which occurs in all settings, not just with linear regression. Models with more parameters than there are data points are underdetermined. Many different parameter vectors give the same likelihood, or the same empirical moments for the method of moments, or the same for whatever criterion is being used for parameter estimation. Thus in general, even when an infinite number of models are theoretically under consideration, only n models are practically under consideration, where n is the sample size.

Hence the "biggest model" that the least squares criterion selects is the polynomial of degree $n - 1$. What does it look like? Figure 12.7 shows both the best fitting (perfectly fitting!) polynomial of degree $n - 1 = 9$ and the least squares regression line from the other figure.

How well does the biggest model do? It fits the observed data perfectly, but it's hard to believe that it would fit *new* data from the same population as well. The extreme oscillations near the ends of the range of the data are obviously nonsensical, but even the smaller oscillations in the middle seem to be tracking random noise rather than any real features of the population regression function. Of course, we don't actually know what the true population regression function is. It could be either of the two functions graphed in the figure, or it could be some other function. But it's hard to believe, when the linear function fits so

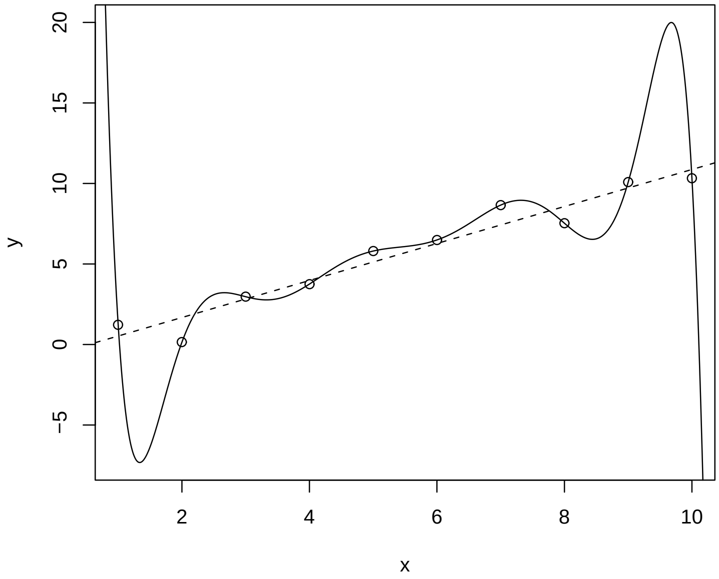


Figure 12.7: Some regression data. With fitted linear regression function (dashed line) and ninth degree polynomial regression function (solid curve).

well, that something as complicated as the ninth degree polynomial is close to the true regression function. We say it "overfits" the data, meaning it's *too* close to the data and not close enough to the true population regression function.

12.7.2 Mean Square Error

So what criterion should we use for model selection if residual sum of squares is no good? One theoretical criterion is mean square error. In the regression setting, it is unclear what quantities mean square error should apply to. Do we use the mean square error of the parameter estimates? We haven't even defined mean square error for vector quantities. That alone suggests we should avoid looking at m. s. e. of regression coefficients. There is also our slogan that regression coefficients are meaningless. Hence we should look at estimates of the regression function.

But here too, there are still issues that need to be clarified. The regression function $h(\mathbf{x}) = E(Y \mid \mathbf{x})$ is a scalar function of $\mathbf{x}$. (Hence we don't need to worry about m. s. e. of a vector quantity). But it is a function of the predictor value $\mathbf{x}$.

$$\operatorname{mse}\{\hat{h}(\mathbf{x})\} = \text{variance} + \text{bias}^2 = \operatorname{var}\{\hat{h}(\mathbf{x})\} + \Bigl(E\{\hat{h}(\mathbf{x})\} - h(\mathbf{x})\Bigr)^2 \qquad (12.61)$$

where $h(\mathbf{x})$ is the true population regression function and $\hat{h}(\mathbf{x})$ is an estimate (we are thinking of least squares regression estimates, but (12.61) applies to any estimate). ("Bias?" did I hear someone say? Aren't linear regression estimates *unbiased*? Yes, the are when the model is *correct*. Here we are considering cases when the model is too small to contain the true regression function.)

To make a criterion that we can minimize to find the best model, we need a single scalar quantity, not a function. There are several things we could do to make a scalar quantity from (12.61). We could *integrate* it over some range of values, obtaining so-called *integrated mean squared error*. A simpler alternative is to sum it over the *design points*, the $\mathbf{x}$ values occurring in the data set under discussion. We'll discuss only the latter.

As we did before, write $\boldsymbol{\mu}$ for the true population means of the responses given the predictors, defined by $\mu_i = h(\mathbf{x}_i)$. Let m index models. The m-th model will have design matrix $\mathbf{X}_m$ and hat matrix $\mathbf{H}_m$. The expected value of the regression predictions $\hat{\mathbf{y}} = \mathbf{H}_m\mathbf{y}$ under the m-th model is

$$E(\hat{\mathbf{y}}) = \mathbf{H}_m\boldsymbol{\mu}.$$

Hence the bias is

$$E(\mathbf{y}) - E(\hat{\mathbf{y}}) = (\mathbf{I} - \mathbf{H}_m)\boldsymbol{\mu}. \tag{12.62}$$

Note that if the model is correct, then $\boldsymbol{\mu}$ is in the range of $\mathbf{H}_m$ and the bias is zero. Models that are incorrect have nonzero bias. (12.62) is, of course, a vector. Its i-th element gives the bias of $\hat{y}_i$. What we decided to study was the sum of the m. s. e.'s at the design points, the "bias" part of which is just the sum of squares of the elements of (12.62), which is the same thing as the squared length of this vector

$$\text{bias}^2 = \|(\mathbf{I} - \mathbf{H}_m)\boldsymbol{\mu}\|^2 = \boldsymbol{\mu}'(\mathbf{I} - \mathbf{H}_m)^2\boldsymbol{\mu} = \boldsymbol{\mu}'(\mathbf{I} - \mathbf{H}_m)\boldsymbol{\mu}. \tag{12.63}$$

The variance is

$$\begin{aligned}
\operatorname{var}(\hat{\mathbf{y}}) &= E\left\{\left\|\hat{\mathbf{y}} - E(\hat{\mathbf{y}})\right\|^2\right\} \\
&= E\left\{\left\|\hat{\mathbf{y}} - \mathbf{H}_m\boldsymbol{\mu}\right\|^2\right\} \\
&= E\left\{\left\|\mathbf{H}_m(\mathbf{y} - \boldsymbol{\mu})\right\|^2\right\} \\
&= E\left\{\mathbf{H}_m(\mathbf{y} - \boldsymbol{\mu})'(\mathbf{y} - \boldsymbol{\mu})\mathbf{H}_m\right\} \\
&= \sigma^2\mathbf{H}_m^2 \\
&= \sigma^2\mathbf{H}_m
\end{aligned}$$

This is, of course, a matrix. What we want though is just the sum of the diagonal elements. The i-th diagonal element is the variance of $\hat{y}_i$, and our decision to focus on the sum of the m. s. e.'s at the design points says we want the sum of these. The sum of the diagonal elements of a square matrix $\mathbf{A}$ is called its *trace*, denoted $\operatorname{tr}(\mathbf{A})$. Thus the "variance" part of the m. s. e. is

$$\text{variance} = \sigma^2 \operatorname{tr}(\mathbf{H}_m). \tag{12.64}$$

And

$$\mathrm{MSE}_m = \sum_{i=1}^{n} \mathrm{mse}(\hat{y}_i) = \sigma^2 \operatorname{tr}(\mathbf{H}_m) + \boldsymbol{\mu}'(\mathbf{I} - \mathbf{H}_m)\boldsymbol{\mu}.$$

12.7.3 The Bias-Variance Trade-Off

Generally, one cannot reduce both bias and variance at the same time. Bigger models have less bias but *more* variance. Smaller models have less variance but *more* bias. This is called the "bias-variance trade-off."

Example 12.7.2.
We again use the data for Example 12.3.2. This example, however, will be purely theoretical. We will look at various polynomial regression models, using the x values, but ignoring the y values. Instead we will do a purely theoretical calculation using the parameter values that were actually used to simulate the y values in the data

$$\mu_i = \sin(x_i) + \sin(2x_i) \tag{12.65a}$$

$$\sigma = 0.2 \tag{12.65b}$$

Since the true population regression curve is a trigonometric rather than a polynomial function, *no* polynomial is unbiased. This is typical of real applications. No model under consideration is exactly correct.

Table 12.1 shows the results of the theoretical calculations for the data in this example. We see that the fifth degree polynomial chosen in Example 12.3.2 is not the best. The ninth degree polynomial is the best. Figure 12.8 shows both the true regression function used to simulate the response values (12.65a) and the sample ninth-degree polynomial regression function. We see that the sample regression function does not estimate the true regression function perfectly (because the sample is not the population), but we now know from the theoretical analysis in this example that no polynomial will fit better. A lower degree polynomial will have less variance. A higher degree will have less bias. But the bias-variance trade-off will be worse for either.

12.7.4 Model Selection Criteria

The theory discussed in the preceding section gives us a framework for discussing model selection. We want the model with the smallest m. s. e., the model which makes the optimal bias-variance trade-off.

Unfortunately, this is useless in practice. Mean square error is a theoretical quantity. It depends on the unknown true regression function and the unknown error variance. Furthermore, there is no obvious way to estimate it. Without knowing which models are good, which is exactly the question we are trying to resolve, we can't get a good estimate of the true regression function (and without that we can't estimate the error variance well either).

Table 12.1: Bias and Variance for Different Polynomial Regression Models. The second column gives the variance (12.64) and the third column gives the "bias^2" (12.63) for polynomial regression models of different degrees. The last column gives their sum (mean squared error). The model with the smallest m. s. e. (degree 7) is the best. The calculations are for the situation with true regression function (12.65a) and true error standard deviation (12.65b).

degree	variance	bias^2	m. s. e.
0	0.04	99.0000	99.0400
1	0.08	33.3847	33.4647
2	0.12	33.3847	33.5047
3	0.16	29.0924	29.2524
4	0.20	29.0924	29.2924
5	0.24	4.8552	5.0952
6	0.28	4.8552	5.1352
7	0.32	0.1633	0.4833
8	0.36	0.1633	0.5233
9	0.40	0.0019	0.4019
10	0.44	0.0019	0.4419
11	0.48	9.8×10^{-6}	0.4800
12	0.52	9.8×10^{-6}	0.5200
13	0.56	2.6×10^{-8}	0.5600
14	0.60	2.6×10^{-8}	0.6000
15	0.64	3.9×10^{-11}	0.6400
16	0.68	3.9×10^{-11}	0.6800
17	0.72	4.9×10^{-14}	0.7200
18	0.76	4.9×10^{-14}	0.7600
19	0.80	1.2×10^{-14}	0.8000
20	0.84	1.4×10^{-14}	0.8400

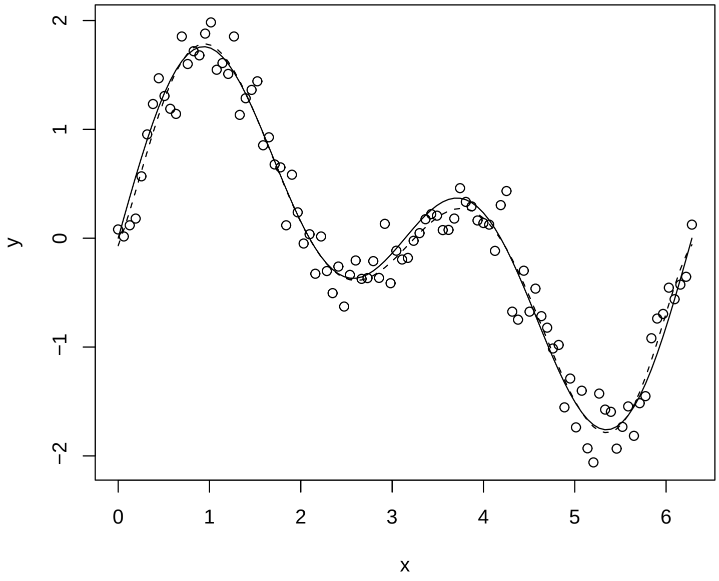


Figure 12.8: Some regression data with true regression function (solid line) and sample regression function (dashed line).

There are several quantities that have been proposed in the literature as estimates of m. s. e. No one is obviously right. We shall not go into the theory justifying them (it is merely heuristic anyway). One fairly natural quantity is

$$\sum_{i=1}^{n} \hat{e}_{(i)}^2 \tag{12.66}$$

Unlike SSResid (the sum of the $\hat{e}_i$), this does not always favor the biggest model. Models that overfit tend to do a bad job of even their leave-one-out predictions. (12.66) is called the *predicted residual sum of squares* (PRESS) or the *cross-validated sum of squares* (CVSS), cross-validation being another term used to describe the leave-one-out idea.

The idea of CVSS or other criteria to be described presently is to pick the model with the smallest value of the criterion. This will not necessarily be the model with the smallest m. s. e., but it is a reasonable estimate of it.

Another criterion somewhat easier to calculate is Mallows' C_p defined by

$$\begin{aligned} C_p &= \frac{\text{SSResid}_p}{\hat{\sigma}^2} + 2p - n \\ &= \frac{\text{SSResid}_p - \text{SSResid}_k}{\hat{\sigma}^2} + p - (k - p) \\ &= (k - p)(F_{k-p,n-k} - 1) + p \end{aligned} \tag{12.67}$$

where SSResid_p is the sum of squares of the residuals for some model with p predictors (including the constant predictor, if present), $\hat{\sigma}^2 = \text{SSResid}_k/(n-k)$ is the estimated error variance for the largest model under consideration, which has k predictors, and $F_{p,k}$ is the F statistic for the F test for comparison of these two models. The F statistic is about one in size if the small model is correct, in which case $C_p \approx p$. This gives us a criterion for finding reasonably fitting models. When many models are under consideration, many of them may have $C_p \approx p$ or smaller. All such models must be considered reasonably good fits. Any might be the correct model or as close to correct as any model under consideration. The quantity estimated by C_p is the mean square error of the model with p predictors, divided by σ^2.

An idea somewhat related to Mallows' C_p, but applicable outside the regression context is the *Akaike information criterion* (AIC).

$$-2 \cdot (\text{log likelihood}) + 2p \tag{12.68}$$

where as in Mallows' C_p, the number of parameters in the model is p. Although (12.67) and (12.68) both have the term $2p$, they are otherwise different. The log likelihood for a linear regression model, with the MLE's plugged in for the parameters is

$$-\frac{n}{2}[1 + \log(\hat{\sigma}^2)]$$

[equation (12) on p. 511 in Lindgren]. Thus for a regression model

$$\text{AIC} = n + n \log(\hat{\sigma}_p^2) + 2p$$

where we have put a subscript p on the estimated error variance to indicate clearly that it is the estimate from the model with p predictors, not the error estimate from the largest model (k predictors used in C_p).

Finally, we add one last criterion, almost the same as AIC, called the *Bayes information criterion* (BIC)

$$-2 \cdot (\text{log likelihood}) + p \log(n) \tag{12.69}$$

In the regression context, this becomes

$$\text{BIC} = n + n \log(\hat{\sigma}_p^2) + p \log(n)$$

Neither AIC nor BIC have a rigorous theoretical justification applicable to a wide variety of models. Both were derived for special classes of models that were easy to analyze and both involve some approximations. Neither can be claimed to be the right thing (nor can anything else). As the "Bayes" in BIC indicates, the BIC criterion is intended to approximate using Bayes tests instead of frequentist tests (although its approximation to true Bayes tests is fairly crude). Note that BIC penalizes models with more parameters more strongly than AIC ($p \log n$ versus $2p$). So BIC always selects a smaller model than AIC.

This gives us four criteria for model selection. There are arguments in favor of each. None of the arguments are completely convincing. All are widely used.

Table 12.2: Model Selection Criteria. Four model selection criteria, CVSS, Mallows' C_p, AIC, and BIC applied to the data of Example 12.3.2.

p	CVSS	C_p	AIC	BIC
1	102.44	2327.01	102.40	105.01
2	40.44	832.80	10.45	15.66
3	41.79	834.29	13.42	21.24
4	36.06	706.84	1.44	11.86
5	37.01	707.20	4.28	17.31
6	10.72	125.33	−124.48	−108.85
7	11.36	127.03	−121.56	−103.32
8	4.52	8.22	−202.20	−181.36
9	4.73	9.75	−199.62	−176.17
10	4.91	11.50	−196.79	−170.74
11	5.22	13.50	−193.67	−165.02
12	5.71	15.49	−190.55	−159.29
13	5.10	14.58	−190.64	−156.77
14	5.74	16.06	−188.07	−151.59
15	5.08	14.62	−188.89	−149.82
16	4.89	16.00	−186.44	−144.76

Example 12.7.3.
We use the data for Example 12.3.2 yet again. Now we fit polynomials of various degrees to the data, and look at our four criteria. The results are shown in Table 12.2. In this example, all four criteria select the same model, the model with $p = 8$ predictors, which is the polynomial of degree 7. This is not the model with the smallest m. s. e. discovered by the theoretical analysis. The criteria do something sensible, but as everywhere else in statistics, there are errors (the sample is not the population).

12.7.5 All Subsets Regression

We know return to the situation in which there are k predictors including the constant predictor and 2^{k-1} models under consideration (the constant predictor is usually included in all models). If k is large and one has no non-statistical reason (e. g., a practical or scientific reason) that cuts down the number of models to be considered, then one must fit them all. Fortunately, there are fast algorithms that allow a huge number of models to be fit or at least quickly checked to see that they are much worse than other models of the same size.

There is a contributed package to R that contains a function `leaps` that does this.

Example 12.7.4 (2^k subsets).
The data set in the URL

`http://www.stat.umn.edu/geyer/5102/ex12.7.4.dat`

consists of multivariate normal data with one response variable y and 20 predictor variables x_1, ..., x_{20}. The predictors are correlated. The distribution from which they were simulated has all correlations equal to one-half. The actual (sample) correlations, of course, are all different because of chance variation.

The true population regression function (the one used to simulate the y values) was

$$y = x_1 + x_2 + x_3 + x_4 + x_5 + e \tag{12.70}$$

with error variance $\sigma^2 = 1.5^2$. If we fit the model

```
foo <- as.matrix(X)
x <- foo[ , -1]
out <- lm(y ~ x)
summary(out)
```

the table of information about the regression coefficients is

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.07879    0.17449  -0.452  0.65286
xx1          1.11887    0.22303   5.017 3.17e-06 ***
xx2          0.79401    0.26519   2.994  0.00367 **
xx3          0.45118    0.25848   1.746  0.08478 .
xx4          0.59879    0.23037   2.599  0.01115 *
xx5          1.09573    0.24277   4.513 2.20e-05 ***
xx6          0.26067    0.24220   1.076  0.28509
xx7         -0.15959    0.21841  -0.731  0.46712
xx8         -0.50182    0.23352  -2.149  0.03470 *
xx9          0.14047    0.22888   0.614  0.54116
xx10         0.37689    0.22831   1.651  0.10275
xx11         0.39805    0.21722   1.832  0.07065 .
xx12         0.38825    0.22396   1.734  0.08689 .
xx13        -0.07910    0.23553  -0.336  0.73788
xx14         0.26716    0.20737   1.288  0.20138
xx15        -0.12016    0.23073  -0.521  0.60398
xx16         0.08592    0.22372   0.384  0.70195
xx17         0.31296    0.22719   1.378  0.17224
xx18        -0.24605    0.23355  -1.054  0.29531
xx19         0.10221    0.21503   0.475  0.63586
xx20        -0.45956    0.23698  -1.939  0.05604 .
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1
```

We've left in the significance codes, bogus though they may be (more on this later), so you can easily spot the regression coefficients that the least squares fit indicates may be important. If we go only with the strongest evidence (two or three stars) we get as "significant" three of the five truly important regression coefficients [recall from (12.70) that the true nonzero regression coefficients are β_1 through β_5]. The other two are missed.

If we use a less stringent standard, say one or more stars, we do pick up another truly nonzero regression coefficient, but we also pick up a false one. Thus we now have both false negatives (we still missed β_3) and false positives (we've picked up β_8).

With the least stringent standard, all the coefficients marked by any of the "significance codes" we now have no false negatives (all five of the truly nonzero regression coefficients are now declared "significant") but we have four false positives.

No matter how you slice it, least squares regression doesn't pick the right model. Of course, this is no surprise. It's just "the sample is not the population." But it does show that the results of such model selection procedures must be treated with skepticism.

Actually, we haven't even started a sensible model selection procedure. Recall the slogan that if you want to know how good a model fits, you have to fit that model. So far we haven't fit any of the models we've discussed. We're fools to think we can pick out the good submodels just by looking at printout for the big model.

There is a function `leaps` in the `leaps` contributed package[8] that fits a huge number of models. By default, it finds the 10 best models of each size (number of regression coefficients) for which there are 10 or more models and finds all the models of other sizes.

It uses the inequality that a bigger model always has a smaller sum of squares to eliminate many models. Suppose we have already found 10 models of size p with SSResid less than 31.2. Suppose there was a model of size $p+1$ that we fit and found its SSResid was 38.6. Finally suppose $\hat{\sigma}^2$ for the big model is 2.05. Now the C_p for the 10 best models of size p already found is

$$C_p = \frac{\text{SSResid}_p}{\hat{\sigma}^2} + 2p - n < \frac{31.2}{2.05} + 2p - n$$

and the C_p for any submodel of size p of the model with SSResid = 38.6 (i. e., models obtained by dropping one predictor from that model) has

$$C_p \geq \frac{38.6}{2.05} + 2p - n$$

This means that no such model can be better than the 10 already found, so they can be rejected even though we haven't bothered to fit them. Considerations of this sort make it possible for `leaps` to pick the 10 best of each size without fitting all or even a sizable fraction of the models of each size. Thus it manages to do in minutes what it couldn't do in a week if it actually had to fit all 2^k models. The reason why `leaps` uses C_p as its criterion rather than one of the

[8] This has to be installed separately. It doesn't come with the "base" package. Like everything else about R, it can be found at `http://cran.r-project.org`.

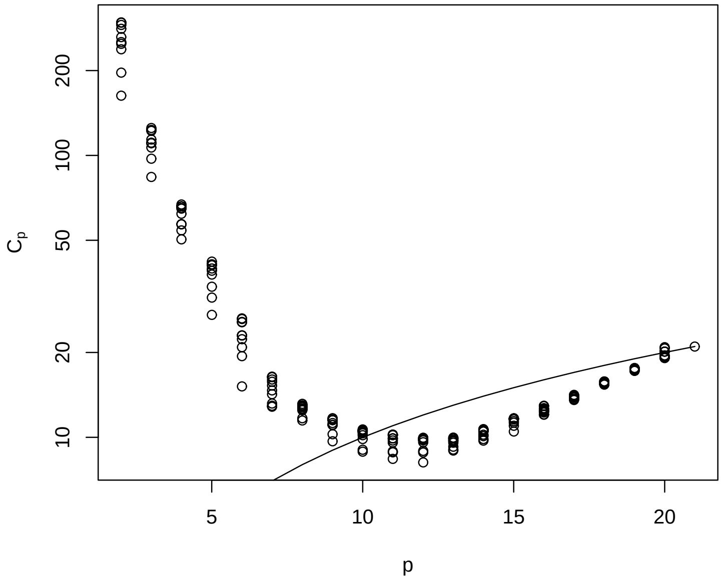


Figure 12.9: C_p plot. Plot of C_p versus p. The dots are the C_p for the 10 best models for each p. The curve is the line $C_p = p$ (curved because of the log scale for C_p).

others is that it is a simple function of SSResid and hence to these inequalities that permit its efficient operation.

We run the leaps function as follows, with the design matrix `x` defined as above,[9]

```
library(leaps)
outs <- leaps(x, y, strictly.compatible=FALSE)
plot(outs$size, outs$Cp, log="y", xlab="p", ylab=expression(C[p]))
lines(outs$size, outs$size)
```

Figure 12.9 shows the plot made by the two plot commands. Every model with $C_p < p$, corresponding to the dots below the line is "good." There are a *huge* number of perfectly acceptable models, because for the larger p there are many more than 10 good models, which are not shown.

The best model according to the C_p criterion is one with $p = 12$, so 11 non-constant predictors, which happen to be x_1 through x_5 (the truly significant predictors) plus x_8, x_{10}, x_{11}, x_{12}, x_{14}, and x_{20}. We can get its regression output as follows.

[9] For some reason, `leaps` doesn't take formula expressions like `lm` does. The reason is probably historical. The equivalent S-plus function doesn't either, because it was written before S had model formulas and hasn't changed. The `strictly.compatible=FALSE` tells R not to be bug-for-bug compatible with S-plus.

```
foo <- x[ , outs$which[outs$Cp == min(outs$Cp)]]
out.best <- lm(y ~ foo)
summary(out.best)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.04925    0.15756  -0.313 0.755316
foox1        1.07411    0.19981   5.376 6.20e-07 ***
foox2        0.88230    0.24483   3.604 0.000519 ***
foox2        0.43560    0.22969   1.896 0.061177 .
foox4        0.72393    0.19912   3.636 0.000466 ***
foox5        1.15609    0.22367   5.169 1.46e-06 ***
foox8       -0.35752    0.21251  -1.682 0.096046 .
foox10       0.43501    0.21885   1.988 0.049957 *
foox11       0.34579    0.20295   1.704 0.091940 .
foox12       0.38479    0.19811   1.942 0.055301 .
foox14       0.28910    0.18838   1.535 0.128455
foox20      -0.49878    0.21736  -2.295 0.024124 *
```

Note that the “stargazing” doesn’t correspond with the notion of the best model by the C_p criterion. One of these coefficients doesn’t even have a dot (so for it $P > 0.10$), and four others only have dots ($0.05 < P < 0.10$). Considering them separately, this would lead us to drop them. But that would be the Wrong Thing (multiple testing without correction). The `leaps` function does as close to the Right Thing as can be done. The only defensible improvement would be to change the criterion, to BIC perhaps, which would choose a smaller “best” model because it penalizes larger models more. However BIC wouldn’t have the nice inequalities that make `leaps` so efficient, which accounts for the use of C_p.

I hope you can see from this analysis that model selection when there are a huge number of models under consideration and no extra-statistical information (scientific, practical, etc.) that can be used to cut down the number is a mug’s game. The best you can do is not very good. The only honest conclusion is that a huge number of models are about equally good, as good as one would expect the correct model to be ($C_p \approx p$).

Thus it is silly to get excited about exactly which model is chosen as the “best” by some model selection procedure (any procedure)! When many models are equally good, the specific features of any one of them can’t be very important.

All of this is related to our slogan about “regression is for prediction, not explanation.” All of the models with $C_p < p$ predict about equally well. So if regression is used for *prediction*, the model selection problem is not serious. Just pick any one of the many good models and use it. For *prediction* it doesn’t matter which good prediction is used. But if regression is used for *explanation*, the model selection problem is insoluble. If you can’t decide which model is “best” and are honest enough to admit that lots of other models are equally good, then how can you claim to have found the predictors which “explain” the response? Of course, if you really understand "correlation is not causation, and regression isn't either," then you know that such "explanations" are bogus anyway, even in the "simple" case (one non-constant predictor) where the model selection problem does not arise.

Problems

12-1. Prove the assertion $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{y}}$ in the proof of Theorem 12.8.

12-2. For the data in Example 12.2.1 give 90% confidence intervals for α and β_1 (the intercept and coefficient for x_1, the coefficient for x_2 was done in Example 12.3.1

12-3. Fill in the details of the proof of Lemma 12.10.

12-4. For the data in Example 12.3.2 in the URL

`http://www.stat.umn.edu/geyer/5102/ex12.3.2.dat`

try fitting some Fourier series models. A Fourier series is sum of sines and cosines of multiples of one fundamental frequency

$$f(x) = a + \sum_{k=1}^{m} b_k \sin(2\pi kx/L) + \sum_{k=1}^{m} c_k \cos(2\pi kx/L)$$

where L is a known constant (the wavelength of the lowest frequency sine and cosine terms) and a, the b_k, and the c_k are adjustable constants. These adjustable constants will be the regression coefficients you fit using linear regression. A Fourier series is always periodic with period L. Since the variable x in this data set does just happen to take values evenly spaced between zero and 2π, and inspection of Figure 12.1 suggests the true regression may be periodic with this period, I recommend using $L = 2\pi$, which gives a regression function of the form

$$E(Y|X) = \alpha + \sum_{k=1}^{m} \beta_k \sin(kX) + \sum_{k=1}^{m} \gamma_k \cos(kX)$$

and we have changed the coefficients to Greek letters to indicate that they are the population regression coefficients, which are unknown constants that we have to estimate.

Use linear regression to find a sample regression function that seems to fit the data better than the polynomial found in Example 12.3.2. (It's up to you to figure out how high m should be and whether all the terms up to order m should be included. You don't have to find the "best" model, whatever that means, just a good model.) Hand in a plot of your fitted sample regression function with the data points also plotted (like Figure 12.2 and the output from the R `summary` command showing the regression fit. (The R functions for sine and cosine are `sin` and `cos`. The R for $\sin(2x)$ is `sin(2 * x)`, because you need the `*` operator for multiplication. You will also need to wrap such terms in the `I()` function, like `I(sin(2 * x))`.)

12-5. The data set in the URL

```
http://www.stat.umn.edu/geyer/5102/prob12-5.dat
```

has three variables `x1` and `x2` (the predictor variables) and `y` (the response variable).

Fit three models to this data set (1) a “linear” model fitting a polynomial of degree one in the two predictor variables, (2) a “quadratic” model fitting a polynomial of degree two, and (3) a “cubic” model fitting a polynomial of degree three. Don’t forget the terms of degree two and three containing products of powers of the two predictor variables.

Print out the ANOVA table for comparing these three models and interpret the P-values in the table. Which model would you say is the best fitting, and why?

12-6. The data set in the URL

```
http://www.stat.umn.edu/geyer/5102/prob12-6.dat
```

has two variables `x` (the predictor variable) and `y` (the response variable). As a glance at a scatter plot of the data done in R by

```
plot(x, y)
```

shows, the relationship between x and y does not appear to be linear. However, it does appear that a so-called *piecewise linear* function with a *knot* at 11 may fit the data well. The means a function having the following three properties.

- It is linear on the interval $x \le 11$.
- It is linear on the interval $x \ge 11$.
- These two linear functions agree at $x = 11$.

Figure out how to fit this model using linear regression. (For some choice of predictor variables, which are functions of x, the regression function of the model is the piecewise linear function described above. Your job is to figure out what predictor variables do this.)

(a) Describe your procedure. What predictors are you using? How many regression coefficients does your procedure have?

(b) Use R to fit the model. Report the parameter estimates (regression coefficients and residual standard error).

The following plot

```
plot(x, y)
lines(x, out$fitted.values)
```

puts a line of predicted values ($\hat{\mathbf{y}}$ in the notation used in the notes and in Lindgren) on the scatter plot. It may help you see when you have got the right thing. You do not have to turn in the plot.

Hint: The `ifelse` function in R defines vectors whose values depend on a condition, for example

```
ifelse(x <= 11, 1, 0)
```

defines the indicator function of the interval $x \leq 11$. (This is *not* one of the predictor variables you need for this problem. It's a hint, but not that much of a hint. The `ifelse` function may be useful, this particular instance is not.)

12-7. For the data in the URL

```
http://www.stat.umn.edu/geyer/5102/ex12.3.2.dat
```

(a) Find the 95% percent prediction interval for an individual with x value 5 using the fifth degree polynomial model fit in Examples 12.3.2 and 12.3.4 (this interval can be read off Figure 12.3, but get the exact numbers from R).

(b) Find the 95% percent confidence interval for the population regression function at the same x value for the same model.

12-8. Prove Theorem 12.13. (Use $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, don't reprove it.)

12-9. The data set

```
http://www.stat.umn.edu/geyer/5102/prob12-9.dat
```

contains three variables, `x`, `y`, and `z`. Each is an i. i. d. sample from some distribution. The three variables are independent of each other (this is not a regression problem). Make Q-Q plots of the variables. One is normal. Which one? Describe the features of the other two plots that make you think the variables plotted are not normal.

It is an interesting comment on the usefulness of Q-Q plots that this problem is essentially undoable at sample size 50 (no differences are apparent in the plots). It's not completely obvious at the sample size 100 used here. Fairly large sample sizes are necessary for Q-Q plots to be useful.

12.8 Bernoulli Regression

As we said in Section 12.5

- Categorical *predictors* are no problem for linear regression. Just use "dummy variables" and proceed normally.

but

- Categorical *responses* do present a problem. Linear regression assumes normally distributed responses. Categorical variables can't be normally distributed.

So now we learn how to deal with at least one kind of categorical response, the simplest, which is Bernoulli.

Suppose the responses are

$$Y_i \sim \text{Ber}(p_i) \tag{12.71}$$

contrast this with the assumptions for linear regression which we write as

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2) \tag{12.72}$$

and

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \tag{12.73}$$

Equations (12.72) and (12.73) express the same assumptions as (12.22) and (12.24). We have just rewritten the "strong" regression assumptions in order to bring out the analogy with what we want to do with Bernoulli regression.

The analogy between (12.71) and (12.72) should be clear. Both assume the data are independent, but not identically distributed. The responses Y_i have distributions in the same family, but not the same parameter values. So all we need to finish the specification of a regression-like model for Bernoulli is an equation that takes the place of (12.73).

12.8.1 A Dumb Idea (Identity Link)

We could use (12.73) with the Bernoulli model, although we have to change the symbol for the parameter from $\boldsymbol{\mu}$ to $\mathbf{p}$

$$\mathbf{p} = \mathbf{X}\boldsymbol{\beta}.$$

This means, for example, in the "simple" linear regression model (with one constant and one non-constant predictor x_i)

$$p_i = \alpha + \beta x_i. \tag{12.74}$$

Before we further explain this, we caution that this is universally recognized to be a dumb idea, so don't get too excited about it.

Now nothing is normal, so least squares, t and F tests, and so forth make no sense. But maximum likelihood, the asymptotics of maximum likelihood estimates, and likelihood ratio tests do make sense.

Hence we write down the log likelihood

$$l(\alpha, \beta) = \sum_{i=1}^{n} \big[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \big]$$

and its derivatives

$$\begin{aligned} \frac{\partial l(\alpha, \beta)}{\partial \alpha} &= \sum_{i=1}^{n} \left[\frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i} \right] \\ \frac{\partial l(\alpha, \beta)}{\partial \beta} &= \sum_{i=1}^{n} \left[\frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i} \right] x_i \end{aligned}$$

and set equal to zero to solve for the MLE's. Fortunately, even for this dumb idea, R knows how to do the problem.

Example 12.8.1 (Bernoulli Regression, Identity Link).
We use the data in

```
http://www.stat.umn.edu/geyer/5102/ex12.8.1.dat
```

which has three variables `x`, `y`, and `z`. For now we will just use the first two. The response `y` is Bernoulli (zero-one-valued). We will do a Bernoulli regression using the model assumptions described above, of `y` on `x`. The following code

```
out <- glm(y ~ x, family=quasi(variance="mu(1-mu)"),
    start=c(0.5, 0))
summary(out, dispersion=1)
```

does the regression and prints out a summary. We have to apologize for the rather esoteric syntax, which results from our choice of introducing Bernoulli regression via this rather dumb example. The printout is

```
Call:
glm(formula = y ~ x, family = quasi(variance = "mu(1-mu)"),
    start = c(0.5, 0))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5443  -1.0371  -0.6811   1.1221   1.8214

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.34750    0.19320  -1.799    0.072 .
x            0.01585    0.00373   4.250 2.14e-05 ***
```

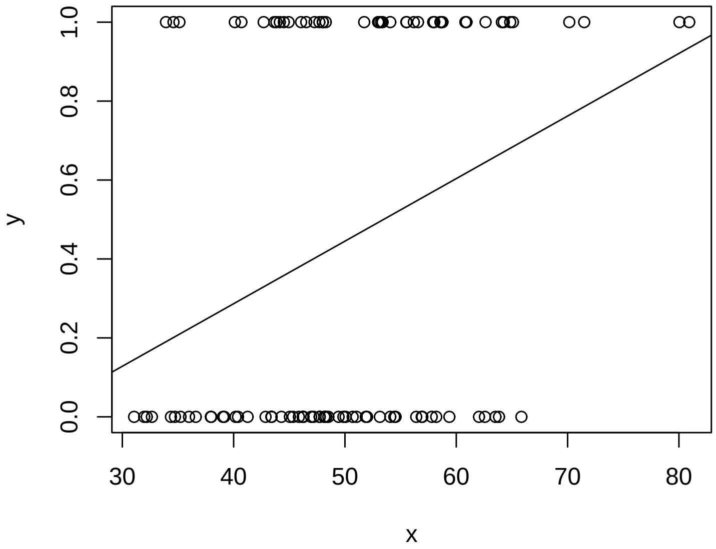


Figure 12.10: Scatter plot and regression line for Example 12.8.1 (Bernoulli regression with an identity link function).

```
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

(Dispersion parameter for quasi family taken to be 1)

    Null deviance: 137.19  on 99  degrees of freedom
Residual deviance: 126.96  on 98  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 3
```

As usual, our main interest is in the table labeled `Coefficients:`, which says the estimated regression coefficients (the MLE's) are $\hat{\alpha} = -0.34750$ and $\hat{\beta} = 0.01585$. This table also gives standard errors, test statistics ("z values") and P-values for the two-tailed test of whether the true value of the coefficient is zero.

The scatter plot with regression line for this regression is somewhat unusual looking. It is produced by the code

```
plot(x, y)
curve(predict(out, data.frame(x=x)), add=TRUE)
```

and is shown in Figure 12.10. The response values are, of course, being Bernoulli, either zero or one, which makes the scatter plot almost impossible to interpret (it is clear that there are more ones for high x values than for low, but it's impossible to see much else, much less to visualize the correct regression line).

That finishes our discussion of the example. So why is it "dumb"? One reason is that nothing keeps the parameters in the required range. The p_i, being probabilities must be between zero and one. The right hand side of (12.74), being a linear function may take any values between $-\infty$ and $+\infty$. For the data set used in the example, it just happened that the MLE's wound up in $(0, 1)$ without constraining them to do so. In general that won't happen. What then? R being semi-sensible will just crash (produce error messages rather than estimates).

There are various ad-hoc ways one could think to patch up this problem. One could, for example, truncate the linear function at zero and one. But that makes a nondifferentiable log likelihood and ruins the asymptotic theory. The only simple solution is to realize that linearity is no longer simple and give up linearity.

12.8.2 Logistic Regression (Logit Link)

What we need is an assumption about the p_i that will always keep them between zero and one. A great deal of thought by many smart people came up with the following general solution to the problem. Replace the assumption (12.73) for linear regression with the following two assumptions

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \tag{12.75}$$

and

$$p_i = h(\eta_i) \tag{12.76}$$

where g is a smooth invertible function that maps $\mathbb{R}$ into $(0, 1)$ so the p_i are always in the required range. We now stop for some important terminology.

- The vector $\boldsymbol{\eta}$ in (12.75) is called the *linear predictor.*

- The function h is called the *inverse link function* and its inverse $g = h^{-1}$ is called the *link function.*

The most widely used (though not the only) link function for Bernoulli regression is the *logit* link defined by

$$g(p) = \operatorname{logit}(p) = \log\left(\frac{p}{1-p}\right) \tag{12.77a}$$

$$h(\eta) = g^{-1}(\eta) = \frac{e^\eta}{e^\eta + 1} \tag{12.77b}$$

The right hand equality in (12.77a) defines the so-called *logit* function, and, of course, the right hand inequality in (12.77b) defines the inverse logit function.

For generality, we will not at first use the explicit form of the link function writing the log likelihood

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \bigl[y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \bigr]$$

where we are implicitly using (12.75) and (12.76) as part of the definition. Then

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i} \right] \frac{\partial p_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j}$$

where the two partial derivatives on the right arise from the chain rule and are explicitly

$$\begin{aligned} \frac{\partial p_i}{\partial \eta_i} &= h'(\eta_i) \\ \frac{\partial \eta_i}{\partial \beta_j} &= x_{ij} \end{aligned}$$

where x_{ij} denotes the i, j element of the design matrix $\mathbf{X}$ (the value of the j-th predictor for the i-th individual). Putting everything together

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \left[\frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i} \right] h'(\eta_i) x_{ij}$$

These equations also do not have a closed form solution, but are easily solved numerically by R

Example 12.8.2 (Bernoulli Regression, Logit Link).
We use the same data in Example 12.8.1. The R commands for logistic regression are

```
out <- glm(y ~ x, family=binomial)
summary(out)
```

Note that the syntax is a lot cleaner for this (logit link) than for the "dumb" way (identity link). The `Coefficients:` table from the printout (the only part we really understand) is

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.56633    1.15871  -3.078  0.00208 **
x            0.06607    0.02257   2.927  0.00342 **
---
Signif. codes:  0  `***'  0.001  `**'  0.01  `*'  0.05  `.'  0.1  ` '  1
```

The regression function for this "logistic regression" is shown in Figure 12.11, which appears later, after we have done another example.

12.8.3 Probit Regression (Probit Link)

Another widely used link function for Bernoulli regression is the *probit* function, which is just another name for the standard normal inverse c. d. f. That is, the link function is $g(p) = \Phi^{-1}(p)$ and the inverse link function is $g^{-1}(\eta) = \Phi(\eta)$. The fact that we do not have closed-form expressions for these functions and must use table look-up or computer programs to evaluate them is no problem. We need computers to solve the likelihood equations anyway.

Example 12.8.3 (Bernoulli Regression, Probit Link).
We use the same data in Example 12.8.1. The R commands for probit regression are

```
out <- glm(y ~ x, family=binomial(link="probit"))
summary(out)
```

The `Coefficients:` table from the printout is

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.20896    0.68646  -3.218  0.00129 **
x            0.04098    0.01340   3.058  0.00223 **
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1
```

Note that there is a huge difference in the regression coefficients for our three examples, but this should be no surprise because the coefficients for the three regressions are not comparable. Because the regressions involve different link functions, the *meaning* of the regression coefficients are not the same. Comparing them is like comparing apples and oranges, as the saying goes. Thus Bernoulli regression in particular and generalized linear models in general give us yet another reason why *regression coefficients are meaningless.* Note that Figure 12.11 shows that the estimated regression functions $E(Y \mid X)$ are almost identical for the logit and probit regressions despite the regression coefficients being wildly different. Even the linear regression function used in our first example is not so different, at least in the middle of the range of the data, from the other two.

> *Regression functions (response predictions) have a direct probabilistic interpretation $E(Y \mid X)$.*
>
> *Regression coefficients don't.*

The regression function $E(Y \mid X)$ for all three of our Bernoulli regression examples, including this one, are shown in Figure 12.11, which was made by the following code, assuming that the results of the `glm` function for the three examples were saved in `out.quasi`, `out.logit`, and `out.probit`, respectively, rather than all three in `out`.

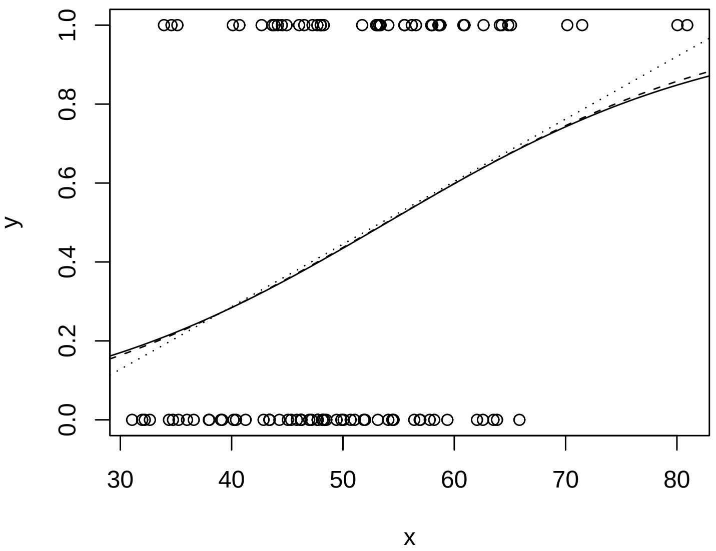


Figure 12.11: Scatter plot and regression functions for Examples 12.8.1, 12.8.2, and 12.8.3. Solid line: regression function for logistic regression (logit link). Dashed line: regression function for probit regression (probit link). Dotted line: regression function for no-name regression (identity link).

```
plot(x, y)
curve(predict(out.logit, data.frame(x=x), type="response"),
   add=TRUE, lty=1)
curve(predict(out.probit, data.frame(x=x), type="response"),
   add=TRUE, lty=2)
curve(predict(out.quasi, data.frame(x=x)), add=TRUE, lty=3)
```

The `type="response"` argument says we want the predicted mean values $g(\boldsymbol{\eta})$, the default being the linear predictor values $\boldsymbol{\eta}$. The reason why this argument is not needed for the last case is because there is no difference with an identity link.

12.9 Generalized Linear Models

A *generalized linear model* (GLM) is a rather general (duh!) form of model that includes ordinary linear regression, logistic and probit regression, and lots more. We keep the regression-like association (12.75) between the regression coefficient vector $\boldsymbol{\beta}$ and the *linear predictor* vector $\boldsymbol{\eta}$ that we used in Bernoulli regression. But now we generalize the probability model greatly. We assume the responses Y_i are independent but not identically distributed with densities of the form

$$f(y \mid \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{\phi / w_i} - c(y, \phi)\right) \tag{12.78}$$

We assume

$$Y_i \sim f(\,\cdot \mid \theta_i, \phi),$$

that is, the *canonical parameter* θ_i is different for each case and is determined (in a way yet to be specified) by the linear predictor η_i but the so-called *dispersion parameter* ϕ is the same for all Y_i. The *weight* w_i is a known positive constant, not a parameter. Also $\phi > 0$ is assumed ($\phi < 0$ would just change the sign of some equations with only trivial effect). The function b is a smooth function but otherwise arbitrary. Given b the function c is determined by the requirement that f integrate to one (like any other probability density).

The log likelihood is thus

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{y_i\theta_i - b(\theta_i)}{\phi / w_i} - c(y_i, \phi)\right) \tag{12.79}$$

Before we proceed to the likelihood equations, let us first look at what the identities derived from differentiating under the integral sign (10.14a) and (10.14b) and their multiparameter analogs (10.44a) and (10.44b) tell us about this model. Note that these identities are exact, not asymptotic, and so can be applied to sample size one and to any parameterization. So let us differentiate one term of (12.79) with respect to its θ parameter

$$\begin{aligned} l(\theta, \phi) &= \frac{y\theta - b(\theta)}{\phi / w} - c(y, \phi) \\ \frac{\partial l(\theta, \phi)}{\partial \theta} &= \frac{y - b'(\theta)}{\phi / w} \\ \frac{\partial^2 l(\theta, \phi)}{\partial \theta^2} &= -\frac{b''(\theta)}{\phi / w} \end{aligned}$$

Applied to this particular situation, the identities from differentiating under the integral sign are

$$\begin{aligned} E_{\theta,\phi}\left\{\frac{\partial l(\theta, \phi)}{\partial \theta}\right\} &= 0 \\ \operatorname{var}_{\theta,\phi}\left\{\frac{\partial l(\theta, \phi)}{\partial \theta}\right\} &= -E_{\theta,\phi}\left\{\frac{\partial^2 l(\theta, \phi)}{\partial \theta^2}\right\} \end{aligned}$$

or

$$\begin{aligned} E_{\theta,\phi}\left\{\frac{Y - b'(\theta)}{\phi / w}\right\} &= 0 \\ \operatorname{var}_{\theta,\phi}\left\{\frac{Y - b'(\theta)}{\phi / w}\right\} &= \frac{b''(\theta)}{\phi / w} \end{aligned}$$

From which we obtain

$$E_{\theta,\phi}(Y) = b'(\theta) \tag{12.80a}$$

$$\operatorname{var}_{\theta,\phi}(Y) = b''(\theta)\frac{\phi}{w} \tag{12.80b}$$

From this we derive the following lemma.

Lemma 12.20. *The function b in* (12.78) *has the following properties*

(i) *b is strictly convex,*

(ii) *b' is strictly increasing,*

(iii) *b'' is strictly positive,*

unless $b''(\theta) = 0$ for all θ and the distribution of Y is concentrated at one point for all parameter values.

Proof. Just by ordinary calculus (iii) implies (ii) implies (i), so we need only prove (iii). Equation (12.80b) and the assumptions $\phi > 0$ and $w > 0$ imply $b''(\theta) \geq 0$. So the only thing left to prove is that if $b''(\theta^*) = 0$ for any one θ^*, then actually $b''(\theta) = 0$ for all θ. By (12.80b) $b''(\theta^*) = 0$ implies $\operatorname{var}_{\theta^*,\phi}(Y) = 0$, so the distribution of Y for the parameter values θ^* and ϕ is concentrated at one point. But now we apply a trick using the distribution at θ^* to calculate for other θ

$$\begin{aligned} f(y \mid \theta, \phi) &= \frac{f(y \mid \theta, \phi)}{f(y \mid \theta^*, \phi)} f(y \mid \theta^*, \phi) \\ &= \exp\left(\frac{y\theta - b(\theta)}{\phi/w_i} - \frac{y\theta^* - b(\theta^*)}{\phi/w_i} \right) f(y \mid \theta^*, \phi) \end{aligned}$$

The exponential term is strictly positive, so the only way the distribution of Y can be concentrated at one point and have variance zero for $\theta = \theta^*$ is if the distribution is concentrated at the same point and hence has variance zero for all other θ. And using (12.80b) again, this would imply $b''(\theta) = 0$ for all θ. □

The "unless" case in the lemma is uninteresting. We never use probability models for data having distributions concentrated at one point (that is, constant random variables). Thus (i), (ii), and (iii) of the lemma hold for any GLM we would actually want to use. The most important of these is (ii) for a reason that will be explained when we return to the general theory after the following example.

Example 12.9.1 (Binomial Regression).
We generalize Bernoulli regression just a bit by allowing more than one one Bernoulli variable to go with each predictor value $\mathbf{x}_i$. Adding those Bernoullis gives a binomial response, that is, we assume

$$Y_i \sim \operatorname{Bin}(m_i, p_i)$$

where m_i is the number of Bernoulli variables with predictor vector $\mathbf{x}_i$. The density for Y_i is

$$f(y_i \mid p_i) = \binom{m_i}{y_i} p_i^{y_i} (1 - p_i)^{1-y_i}$$

we try to match this up with the GLM form. So first we write the density as an exponential

$$\begin{aligned} f(y_i \mid p_i) &= \exp\left[y_i \log(p_i) + (m_i - y_i) \log(1 - p_i) + \log \binom{m_i}{y_i} \right] \\ &= \exp\left[y_i \log\left(\frac{p_i}{1 - p_i} \right) + m_i \log(1 - p_i) + \log \binom{m_i}{y_i} \right] \\ &= \exp\left\{ m_i \bigl[\bar{y}_i \theta_i - b(\theta_i) \bigr] + \log \binom{m_i}{y_i} \right\} \end{aligned}$$

where we have defined

$$\begin{aligned} \bar{y}_i &= y_i / m_i \\ \theta_i &= \operatorname{logit}(p_i) \\ b(\theta_i) &= -\log(1 - p_i) \end{aligned}$$

So we see that

- The *canonical parameter* for the binomial model is $\theta = \operatorname{logit}(p)$. That explains why the logit link is popular.
- The *weight* w_i in the GLM density turns out to be the number of Bernoullis m_i associated with the i-th predictor value. So we see that the weight allows for grouped data like this.
- There is nothing like a dispersion parameter here. For the binomial family the dispersion is known; $\phi = 1$.

Returning to the general GLM model (a doubly redundant redundancy), we first define yet another parameter, the *mean value parameter*

$$\mu_i = E_{\theta_i, \phi}(Y_i) = b'(\theta_i).$$

By (ii) of Lemma 12.20 b' is a strictly increasing function, hence an invertible function. Thus the mapping between the canonical parameter θ and the mean value parameter μ is an invertible change of parameter. Then by definition of "link function" the relation between the mean value parameter μ_i and the linear predictor η_i is given by the link function

$$\eta_i = g(\mu_i).$$

The link function g is required to be a strictly increasing function, hence an invertible change of parameter.

If, as in logistic regression we take the linear predictor to be the canonical parameter, that determines the link function, because $\eta_i = \theta_i$ implies $g^{-1}(\theta) = b'(\theta)$. In general, as is the case in probit regression, the link function g and the function b' that connects the canonical and mean value parameters are unrelated.

It is traditional in GLM theory to make primary use of the mean value parameter and not use the canonical parameter (unless it happens to be the same as the linear predictor). For that reason we want to write the variance as a function of μ rather than θ

$$\operatorname{var}_{\theta_i,\phi}(Y_i) = \frac{\phi}{w} V(\mu_i) \tag{12.81}$$

where

$$V(\mu) = b''(\theta) \qquad \text{when} \qquad \mu = b'(\theta)$$

This definition of the function V makes sense because the function b' is an invertible mapping between mean value and canonical parameters. The function V is called the *variance function* even though it is only proportional to the variance, the complete variance being $\phi V(\mu)/w$.

12.9.1 Parameter Estimation

Now we can write out the log likelihood derivatives

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_{i=1}^n \left(\frac{y_i - b'(\theta_i)}{\phi / w_i} \right) \frac{\partial \theta_i}{\partial \beta_j} \\
&= \sum_{i=1}^n \left(\frac{y_i - \mu_i}{\phi / w_i} \right) \frac{\partial \theta_i}{\partial \beta_j}
\end{aligned}$$

In order to completely eliminate θ_i we need to calculate the partial derivative. First note that

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i)$$

so by the inverse function theorem

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b''(\theta_i)} = \frac{1}{V(\mu_i)}$$

Now we can write

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{1}{V(\mu_i)} h'(\eta_i) x_{ij} \tag{12.82}$$

where $h = g^{-1}$ is the inverse link function. And we finally arrive at the likelihood equations expressed in terms of the mean value parameter and the linear predictor

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^n \left(\frac{y_i - \mu_i}{V(\mu_i)} \right) w_i h'(\eta_i) x_{ij}$$

These are the equations the computer sets equal to zero and solves to find the regression coefficients. Note that the dispersion parameter ϕ appears only multiplicatively. So it cancels when the partial derivatives are set equal to zero. Thus the regression coefficients can be estimated without estimating the dispersion (just as in linear regression).

Also as in linear regression, the dispersion parameter is not estimated by maximum likelihood but by the method of moments. By (12.81)

$$E \left\{ \frac{w_i (Y_i - \mu_i)^2}{V(\mu_i)} \right\} = \frac{w_i}{V(\mu_i)} \operatorname{var}(Y_i) = \phi$$

Thus

$$\frac{1}{n} \sum_{i=1}^n \frac{w_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

would seem to be an approximately unbiased estimate of ϕ. Actually it is not because $\hat{\boldsymbol{\mu}}$ is not $\boldsymbol{\mu}$, and

$$\hat{\phi} = \frac{1}{n - p} \sum_{i=1}^n \frac{w_i (y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

is closer to unbiased where p is the rank of the design matrix $\mathbf{X}$. We won't bother to prove this. The argument is analogous to the reason for $n - p$ in linear regression.

12.9.2 Fisher Information, Tests and Confidence Intervals

The log likelihood second derivatives are

$$\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^n \left(\frac{y_i - b'(\theta_i)}{\phi / w_i} \right) \frac{\partial^2 \theta_i}{\partial \beta_j \partial \beta_k} - \sum_{i=1}^n \left(\frac{b''(\theta_i)}{\phi / w_i} \right) \frac{\partial \theta_i}{\partial \beta_j} \frac{\partial \theta_i}{\partial \beta_k}$$

This is rather a mess, but because of (12.80a) the expectation of the first sum is zero. Thus the j, k term of the expected Fisher information is, using (12.82) and $b'' = V$,

$$\begin{aligned} -E \left\{ \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} \right\} &= \sum_{i=1}^n \left(\frac{b''(\theta_i)}{\phi / w_i} \right) \frac{\partial \theta_i}{\partial \beta_j} \frac{\partial \theta_i}{\partial \beta_k} \\ &= \sum_{i=1}^n \left(\frac{V(\mu_i)}{\phi / w_i} \right) \frac{1}{V(\mu_i)} h'(\eta_i) x_{ij} \frac{1}{V(\mu_i)} h'(\eta_i) x_{ik} \\ &= \frac{1}{\phi} \sum_{i=1}^n \left(\frac{w_i h'(\eta_i)^2}{V(\mu_i)} \right) x_{ij} x_{ik} \end{aligned}$$

We can write this as a matrix equation if we define $\mathbf{D}$ to be the diagonal matrix with i, i element

$$d_{ii} = \frac{1}{\phi} \sum_{i=1}^n \frac{w_i h'(\eta_i)^2}{V(\mu_i)}$$

Then

$$\mathbf{I}(\boldsymbol{\beta}) = \mathbf{X}'\mathbf{D}\mathbf{X}$$

is the expected Fisher information matrix. From this standard errors for the parameter estimates, confidence intervals, test statistics, and so forth can be derived using the usual likelihood theory. Fortunately, we do not have to do all of this by hand. R knows all the formulas and computes them for us.

12.10 Poisson Regression

The Poisson model is also a GLM. We assume responses

$$Y_i \sim \text{Poi}(\mu_i)$$

and connection between the linear predictor and regression coefficients, as always, of the form (12.75). We only need to identify the link and variance functions to get going. It turns out that the canonical link function is the log function (Problem 12-13). The Poisson distribution distribution has the relation

$$\text{var}(Y) = E(Y) = \mu$$

connecting the mean, variance, and mean value parameter. Thus the variance function is $V(\mu) = \mu$, the dispersion parameter is known ($\phi = 1$), and the weight is also unity ($w = 1$).

Example 12.10.1 (Poisson Regression).
The data set

```
http://www.stat.umn.edu/geyer/5102/ex12.10.1.dat
```

simulates the hourly counts from a not necessarily homogeneous Poisson process. The variables are `hour` and `count`, the first counting hours sequentially throughout a 14-day period (running from 1 to $14 \times 24 = 336$) and the second giving the count for that hour.

The idea of the regression is to get a handle on the mean as a function of time if it is not constant. Many time series have a daily cycle. If we pool the counts for the same hour of the day over the 14 days of the series, we see a clear pattern in the histogram (Figure 12.12). In contrast, if we pool the counts for each day of the week, the histogram is fairly even (not shown). Thus it seems to make sense to model the mean function as being periodic with period 24 hours, and the obvious way to do that is to use trigonometric functions. Let us do a bunch of fits

```
w <- hour / 24 * 2 * pi
out1 <- glm(count ~ I(sin(w)) + I(cos(w)), family=poisson)
summary(out1)
out2 <- glm(count ~ I(sin(w)) + I(cos(w)) + I(sin(2 * w))
   + I(cos(2 * w)), family=poisson)
```

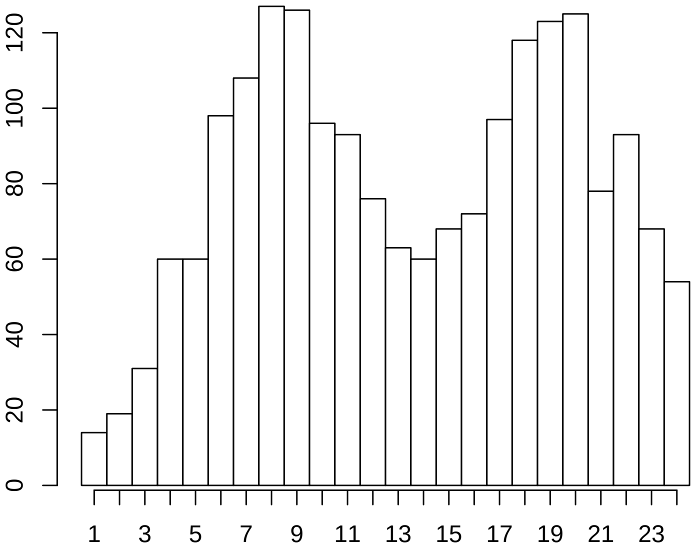


Figure 12.12: Histogram of the total count in each hour of the day for the data for Example 12.10.1.

```
summary(out2)
out3 <- glm(count ~ I(sin(w)) + I(cos(w)) + I(sin(2 * w))
   + I(cos(2 * w)) + I(sin(3 * w)) + I(cos(3 * w)),
   family=poisson)
summary(out3)
```

The `Coefficient:` tables from the printouts are

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.73272    0.02310   75.02  < 2e-16 ***
I(sin(w))   -0.10067    0.03237   -3.11  0.00187 **
I(cos(w))   -0.21360    0.03251   -6.57 5.02e-11 ***

              Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.65917    0.02494  66.521  < 2e-16 ***
I(sin(w))     -0.13916    0.03128  -4.448 8.65e-06 ***
I(cos(w))     -0.28510    0.03661  -7.788 6.82e-15 ***
I(sin(2 * w)) -0.42974    0.03385 -12.696  < 2e-16 ***
I(cos(2 * w)) -0.30846    0.03346  -9.219  < 2e-16 ***

               Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.655430   0.025149  65.826  < 2e-16 ***
I(sin(w))     -0.151196   0.032530  -4.648 3.35e-06 ***
I(cos(w))     -0.301336   0.038244  -7.879 3.29e-15 ***
```

```
I(sin(2 * w)) -0.439789   0.034461 -12.762  < 2e-16 ***
I(cos(2 * w)) -0.312843   0.033919  -9.223  < 2e-16 ***
I(sin(3 * w)) -0.063440   0.033803  -1.877   0.0606 .
I(cos(3 * w))  0.004311   0.033630   0.128   0.8980
```

with the usual "`Signif. codes`". It seems from the pattern of "stars" that maybe it is time to stop. A clearer indication is given by the so-called *analysis of deviance* table, "deviance" being another name for the likelihood ratio test statistic (twice the log likelihood difference between big and small models), which has an asymptotic chi-square distribution by standard likelihood theory.

```
anova(out1, out2, out3, test="Chisq")
```

prints out

```
Analysis of Deviance Table

Model 1: count ~ I(sin(w)) + I(cos(w))
Model 2: count ~ I(sin(w)) + I(cos(w)) + I(sin(2 * w)) + I(cos(2 * w))
Model 3: count ~ I(sin(w)) + I(cos(w)) + I(sin(2 * w)) + I(cos(2 * w)) +
    I(sin(3 * w)) + I(cos(3 * w))
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1       333     651.10
2       331     399.58   2   251.52 2.412e-55
3       329     396.03   2     3.55      0.17
```

The approximate P-value for the likelihood ratio test comparing models 1 and 2 is $P \approx 0$, which clearly indicates that model 1 should be rejected. The approximate P-value for the likelihood ratio test comparing models 2 and 3 is $P = 0.17$, which fairly clearly indicates that model 1 should be accepted and that model 3 is unnecessary. $P = 0.17$ indicates exceedingly weak evidence favoring the larger model. Thus we choose model 2.

The following code

```
hourofday <- (hour - 1) %% 24 + 1
plot(hourofday, count, xlab="hour of the day")
curve(predict(out2, data.frame(w=x/24*2*pi), type="response"),
   add=TRUE)
```

draws the scatter plot and estimated regression function for model 2 (Figure 12.13).

I hope all readers are impressed by how magically statistics works in this example. A glance at Figure 12.13 shows

- Poisson regression is obviously doing more or less the right thing,

- there is no way one could put in a sensible regression function without using theoretical statistics. The situation is just too complicated.

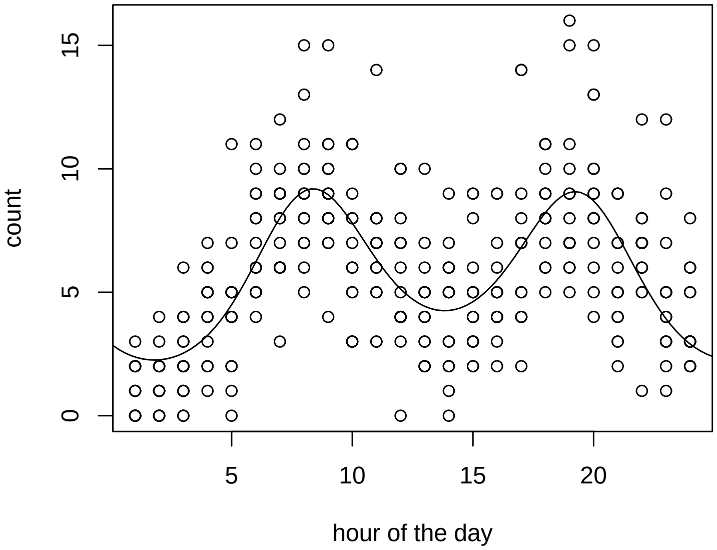


Figure 12.13: Scatter plot and regression curve for Example 12.10.1 (Poisson regression with log link function). The regression function is trigonometric on the scale of the linear predictor with terms up to the frequency 2 per day.

12.11 Overdispersion

So far we have seen only models with unit dispersion parameter ($\phi = 1$). This section gives an example with $\phi \neq 1$ so we can see the point of the dispersion parameter.

The reason $\phi = 1$ for binomial regression is that the mean value parameter $p = \mu$ determines the variance $mp(1 - p) = m\mu(1 - \mu)$. Thus the variance function is

$$V(\mu) = \mu(1 - \mu) \tag{12.83}$$

and the weights are $w_i = m_i$, the sample size for each binomial variable (this was worked out in detail in Example 12.9.1).

But what if the model is wrong? Here is another model. Suppose

$$Y_i \mid W_i \sim \text{Bin}(m_i, W_i)$$

where the W_i are i. i. d. random variables with mean μ and variance τ^2. Then by the usual rules for conditional probability (Axiom CE2 and Theorem 3.7 in Chapter 3 of these notes)

$$E(Y_i) = E\{E(Y_i \mid W_i)\} = E(m_i W_i) = m_i \mu$$

and

$$\begin{aligned}
\operatorname{var}(Y_i) &= E\{\operatorname{var}(Y_i \mid W_i)\} + \operatorname{var}\{E(Y_i \mid W_i)\} \\
&= E\{m_i W_i (1 - W_i)\} + \operatorname{var}\{m_i W_i\} \\
&= m_i \mu - m_i E(W_i^2) + m_i^2 \tau^2 \\
&= m_i \mu - m_i (\tau^2 + \mu^2) + m_i^2 \tau^2 \\
&= m_i \mu (1 - \mu) + m_i (m_i - 1) \tau^2
\end{aligned}$$

This is clearly larger than the formula $m_i \mu (1 - \mu)$ one would have for the binomial model. Since the variance is always larger than one would have under the binomial model.

So we know that if our response variables Y_i are the sum of a random mixture of Bernoullis rather than i. i. d. Bernoullis, we will have overdispersion. But how to model the overdispersion? The GLM model offers a simple solution. Allow for general ϕ so we have, defining $\overline{Y}_i = Y_i / m_i$

$$\begin{aligned}
E(\overline{Y}_i) &= \mu_i \\
\operatorname{var}(\overline{Y}_i) &= \frac{\phi}{m_i} \mu_i (1 - \mu_i) \\
&= \frac{\phi}{m_i} V(\mu_i)
\end{aligned}$$

where V is the usual binomial variance function (12.83).

Example 12.11.1 (Overdispersed Binomial Regression).
The data set

```
http://www.stat.umn.edu/geyer/5102/ex12.11.1.dat
```

contains some data for an overdispersed binomial model. The commands

```
y <- cbind(succ, fail)
out.binom <- glm(y ~ x, family=binomial)
summary(out.binom)
out.quasi <- glm(y ~ x, family=quasibinomial)
summary(out.quasi)
```

fit both the binomial model (logit link and $\phi = 1$) and the "quasi-binomial" (logit link again but ϕ is estimated with the method of moments estimator as explained in the text). Both models have exactly the same maximum likelihood regression coefficients, but because the dispersions differ, the standard errors, z-values, and P-values differ.

The relevant part of the binomial output

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.92155    0.35260  -5.450 5.05e-08 ***
```

```
x            0.07436    0.01227   6.062 1.35e-09 ***
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

(Dispersion parameter for binomial family taken to be 1)
```

and the relevant part of the quasi-binomial output

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.92155    0.41569  -4.623 2.88e-05 ***
x            0.07436    0.01446   5.141 4.97e-06 ***
---
Signif. codes:  0  ‘***’  0.001  ‘**’  0.01  ‘*’  0.05  ‘.’  0.1  ‘ ’  1

(Dispersion parameter for quasibinomial family taken to be 1.38992)
```

Your humble author finds this a bit unsatisfactory. If the data are really overdispersed, then the standard errors and so forth from the latter output are the right ones to use. But since the dispersion was not estimated by maximum likelihood, there is no likelihood ratio test for comparing the two models. Nor could your author find any other test in a brief examination of the literature. Apparently, if one is worried about overdispersion, one should use the model that allows for it. And if not, not. But that's not the way we operate in the rest of statistics. I suppose I need to find out more about overdispersion.

Problems

12-10. Show that (12.77a) and (12.77b) do indeed define a pair of inverse functions.

12-11. Do calculations similar to Example 12.9.1 for the normal problem

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

identifying (a) the canonical parameter θ, the dispersion parameter ϕ, and the weight w_i.

12-12. The data set

`http://www.stat.umn.edu/geyer/5102/ex12.8.1.dat`

contains another predictor vector `z` besides the ones we used in Examples 12.8.1, 12.8.2, and 12.8.3. Perform the logistic regression of `y` on `x` and `z`. Perform a test comparing this new model and the one fit in Example 12.8.2 giving the P-value for the test and the conclusion as to which model the test accepts.

12-13. Do calculations similar to Example 12.9.1 for the Poisson model, showing that the canonical parameter for the Poisson distribution is $\theta = \log(\mu)$.

Appendix A

Greek Letters

Table A.1: Table of Greek Letters (Continued on following page.)

name	capital letter	small letter	pronunciation	sound
alpha	A	α	AL-fah	short a
beta	B	β	BAY-tah	b
gamma	Γ	γ	GAM-ah	g
delta	Δ	δ	DEL-tah	d
epsilon	E	ϵ	EP-si-lon	e
zeta	Z	ζ	ZAY-tah	z
eta	H	η	AY-tah	long a
theta	Θ	θ or ϑ	THAY-thah	soft th (as in thin)
iota	I	ι	EYE-oh-tah	i
kappa	K	κ	KAP-ah	k
lambda	Λ	λ	LAM-dah	l
mu	M	μ	MYOO	m
nu	N	ν	NOO	n
xi	Ξ	ξ	KSEE	x (as in box)
omicron	O	o	OH-mi-kron	o
pi	Π	π	PIE	p
rho	R	ρ	RHOH	rh[1]
sigma	Σ	σ	SIG-mah	s
tau	T	τ	TAOW	t
upsilon	Υ	υ	UP-si-lon	u

[1] The sound of the Greek letter ρ is not used in English. English words, like *rhetoric* and *rhinoceros* that are descended from Greek words beginning with ρ have English pronunciations beginning with an "r" sound rather than "rh" (though the spelling reminds us of the Greek origin).

Table A.2: Table of Greek Letters (Continued.)

	capital	small		
name	letter	letter	pronunciation	sound
phi	Φ	ϕ or φ	FIE	f
chi	X	χ	KIE	guttural ch[2]
psi	Ψ	ψ	PSY	ps (as in stops)[3]
omega	Ω	ω	oh-MEG-ah	o

[2] The sound of the Greek letter χ is not used in English. It is heard in the German *Buch* or Scottish *loch*. English words, like *chemistry* and *chorus* that are descended from Greek words beginning with χ have English pronunciations beginning with a "k" sound rather than "guttural ch" (though the spelling reminds us of the Greek origin).

[3] English words, like *pseudonym* and *psychology* that are descended from Greek words beginning with ψ have English pronunciations beginning with an "s" sound rather than "ps" (though the spelling reminds us of the Greek origin).

Appendix B

Summary of Brand-Name Distributions

B.1 Discrete Distributions

B.1.1 The Discrete Uniform Distribution

The Abbreviation $\mathcal{DU}(S)$.

The Sample Space Any finite set S.

The Density

$$f(x) = \frac{1}{n}, \qquad x \in S,$$

where $n = \operatorname{card}(S)$.

Specialization The case in which the sample space consists of consecutive integers $S = \{m, m+1, \ldots, n\}$ is denoted $\mathcal{DU}(m, n)$.

Moments If $X \sim \mathcal{DU}(1, n)$, then

$$E(X) = \frac{n+1}{2}$$
$$\operatorname{var}(X) = \frac{n^2 - 1}{12}$$

B.1.2 The Binomial Distribution

The Abbreviation $\operatorname{Bin}(n, p)$

The Sample Space The integers 0, $\ldots$, n.

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, \ldots, n.$$

Moments

$$\begin{aligned} E(X) &= np \\ \operatorname{var}(X) &= np(1-p) \end{aligned}$$

Specialization

$$\operatorname{Ber}(p) = \operatorname{Bin}(1, p)$$

B.1.3 The Geometric Distribution, Type II

Note This section has changed. The roles of p and $1 - p$ have been reversed, and the abbreviation Geo(p) is no longer used to refer to this distribution but the distribution defined in Section B.1.8. All of the changes are to match up with Chapter 6 in Lindgren.

The Abbreviation No abbreviation to avoid confusion with the other type defined in Section B.1.8.

Relation Between the Types If $X \sim \operatorname{Geo}(p)$, then $Y = X - 1$ has the distribution defined in this section.

X is the number of *trials* before the first success in an i. i. d. sequence of Ber(p) random variables. Y is the number of *failures* before the first success.

The Sample Space The integers 0, 1, ...

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = p(1-p)^x, \qquad x = 0, 1, \ldots.$$

Moments

$$\begin{aligned} E(X) &= \frac{1}{p} - 1 = \frac{1-p}{p} \\ \operatorname{var}(X) &= \frac{1-p}{p^2} \end{aligned}$$

B.1.4 The Poisson Distribution

The Abbreviation Poi(μ)

The Sample Space The integers 0, 1, ...

The Parameter μ such that $\mu > 0$.

The Density

$$f(x) = \frac{\mu^x}{x!} e^{-\mu}, \qquad x = 0, 1, \ldots.$$

Moments

$$\begin{aligned} E(X) &= \mu \\ \operatorname{var}(X) &= \mu \end{aligned}$$

B.1.5 The Bernoulli Distribution

The Abbreviation Ber(p)

The Sample Space The integers 0 and 1.

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = \begin{cases} p, & x = 1 \\ 1 - p & x = 0 \end{cases}$$

Moments

$$\begin{aligned} E(X) &= p \\ \operatorname{var}(X) &= p(1 - p) \end{aligned}$$

Generalization

$$\operatorname{Ber}(p) = \operatorname{Bin}(1, p)$$

B.1.6 The Negative Binomial Distribution, Type I

The Abbreviation NegBin(k, p)

The Sample Space The integers k, $k + 1$, ...

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \qquad x = k, k+1, \ldots.$$

Moments

$$\begin{aligned} E(X) &= \frac{k}{p} \\ \operatorname{var}(X) &= \frac{k(1-p)}{p^2} \end{aligned}$$

Specialization

$$\operatorname{Geo}(p) = \operatorname{NegBin}(1, p)$$

B.1.7 The Negative Binomial Distribution, Type II

The Abbreviation No abbreviation to avoid confusion with the other type defined in Section B.1.6.

Relation Between the Types If $X \sim \operatorname{NegBin}(k, p)$, then $Y = X - k$ has the distribution defined in this section.

X is the number of *trials* before the k-th success in an i. i. d. sequence of $\operatorname{Ber}(p)$ random variables. Y is the number of *failures* before the k-th success.

The Sample Space The integers 0, 1, ...

The Parameter p such that $0 < p < 1$.

The Density

$$f(y) = \binom{y+k-1}{k-1} p^k (1-p)^{y}, \qquad y = 0, 1, \ldots.$$

Moments

$$\begin{aligned} E(X) &= \frac{k}{p} - k = \frac{k(1-p)}{p} \\ \operatorname{var}(X) &= \frac{k(1-p)}{p^2} \end{aligned}$$

B.1.8 The Geometric Distribution, Type I

The Abbreviation $\operatorname{Geo}(p)$

The Sample Space The integers 1, 2, ...

The Parameter p such that $0 < p < 1$.

The Density

$$f(x) = p(1-p)^{x-1}, \qquad x = 1, 2, \ldots.$$

Moments

$$E(X) = \frac{1}{p}$$
$$\operatorname{var}(X) = \frac{1-p}{p^2}$$

Generalization

$$\operatorname{Geo}(p) = \operatorname{NegBin}(1, p)$$

B.2 Continuous Distributions

B.2.1 The Uniform Distribution

The Abbreviation $\mathcal{U}(S)$.

The Sample Space Any subset S of $\mathbb{R}^d$.

The Density

$$f(x) = \frac{1}{c}, \qquad x \in S,$$

where

$$c = m(S) = \int_S dx$$

is the measure of S (length in $\mathbb{R}^1$, area in $\mathbb{R}^2$, volume in $\mathbb{R}^3$, and so forth).

Specialization The case having $S = (a, b)$ in $\mathbb{R}^1$ and density

$$f(x) = \frac{1}{b-a}, \qquad a < x < b$$

is denoted $\mathcal{U}(a, b)$.

Moments If $X \sim \mathcal{U}(a, b)$, then

$$E(X) = \frac{a+b}{2}$$
$$\operatorname{var}(X) = \frac{(b-a)^2}{12}$$

B.2.2 The Exponential Distribution

The Abbreviation Exp(λ).

The Sample Space The interval $(0, \infty)$ of the real numbers.

The Parameter λ such that $\lambda > 0$.

The Density

$$f(x) = \lambda e^{-\lambda x}, \qquad x > 0.$$

Moments

$$\begin{aligned} E(X) &= \frac{1}{\lambda} \\ \operatorname{var}(X) &= \frac{1}{\lambda^2} \end{aligned}$$

Generalization

$$\operatorname{Exp}(\lambda) = \operatorname{Gam}(1, \lambda)$$

B.2.3 The Gamma Distribution

The Abbreviation Gam(α, λ).

The Sample Space The interval $(0, \infty)$ of the real numbers.

The Parameters α and λ such that $\alpha > 0$ and $\lambda > 0$.

The Density

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} e^{-\lambda x}, \qquad x > 0.$$

where $\Gamma(\alpha)$ is the gamma function (Section B.3.1 below).

Moments

$$\begin{aligned} E(X) &= \frac{\alpha}{\lambda} \\ \operatorname{var}(X) &= \frac{\alpha}{\lambda^2} \end{aligned}$$

Specialization

$$\begin{aligned} \operatorname{Exp}(\lambda) &= \operatorname{Gam}(1, \lambda) \\ \operatorname{chi}^2(k) &= \operatorname{Gam}\left(\tfrac{k}{2}, \tfrac{1}{2}\right) \end{aligned}$$

B.2.4 The Beta Distribution

The Abbreviation Beta(s, t).

The Sample Space The interval $(0, 1)$ of the real numbers.

The Parameters s and t such that $s > 0$ and $t > 0$.

The Density

$$f(x) = \frac{1}{B(s,t)} x^{s-1} (1-x)^{t-1} \qquad 0 < x < 1.$$

where $B(s, t)$ is the *beta function* defined by

$$B(s,t) = \frac{\Gamma(s)\Gamma(t)}{\Gamma(s+t)} \tag{B.1}$$

Moments

$$\begin{aligned} E(X) &= \frac{s}{s+t} \\ \operatorname{var}(X) &= \frac{st}{(s+t)^2(s+t+1)} \end{aligned}$$

B.2.5 The Normal Distribution

The Abbreviation $\mathcal{N}(\mu, \sigma^2)$.

The Sample Space The real line $\mathbb{R}$.

The Parameters μ and σ^2 such that $\sigma^2 > 0$.

The Density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R}.$$

Moments

$$\begin{aligned} E(X) &= \mu \\ \operatorname{var}(X) &= \sigma^2 \\ \mu_4 &= 3\sigma^4 \end{aligned}$$

B.2.6 The Chi-Square Distribution

The Abbreviation chi$^2(k)$.

The Sample Space The interval $(0, \infty)$ of the real numbers.

The Parameter A positive integer k.

The Density

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \qquad x > 0.$$

Moments

$$\begin{aligned} E(X) &= k \\ \operatorname{var}(X) &= 2k \end{aligned}$$

Generalization

$$\operatorname{chi}^2(k) = \operatorname{Gam}\left(\tfrac{k}{2}, \tfrac{1}{2}\right)$$

B.2.7 The Cauchy Distribution

The Abbreviation Cauchy(μ, σ).

The Sample Space The real line $\mathbb{R}$.

The Parameters μ and σ such that $\sigma > 0$.

The Density

$$f(x) = \frac{1}{\pi} \cdot \frac{\sigma}{\sigma^2 + (x - \mu)^2}, \qquad x \in \mathbb{R}.$$

Moments None: $E(|X|) = \infty$.

B.2.8 Student's t Distribution

The Abbreviation $t(\nu)$.

The Sample Space The real line $\mathbb{R}$.

The Parameters ν such that $\nu > 0$, called the "degrees of freedom" of the distribution.

The Density

$$f_\nu(x) = \frac{1}{\sqrt{\nu}} \cdot \frac{1}{B(\frac{\nu}{2}, \frac{1}{2})} \cdot \frac{1}{\left(1 + \frac{x^2}{\nu}\right)^{(\nu+1)/2}}, \qquad -\infty < x < +\infty$$

where $B(s, t)$ is the beta function defined by (B.1).

Moments If $\nu > 1$, then

$$E(X) = 0.$$

Otherwise the mean does not exist. If $\nu > 2$, then

$$\operatorname{var}(X) = \frac{\nu}{\nu - 2}.$$

Otherwise the variance does not exist.

Specialization

$$t(1) = \text{Cauchy}(0, 1)$$

and in a manner of speaking

$$t(\infty) = \mathcal{N}(0, 1)$$

(see Theorem 7.21 of Chapter 7 of these notes).

B.2.9 Snedecor's F Distribution

The Abbreviation $F(\mu, \nu)$.

The Sample Space The interval $(0, \infty)$ of the real numbers.

The Parameters μ and ν such that $\mu > 0$ and $\nu > 0$, called the "numerator degrees of freedom" of the the "denominator degrees of freedom" of the distribution, respectively.

The Density Not derived in these notes.

Moments If $\nu > 2$, then

$$E(X) = \frac{\nu}{\nu - 2}.$$

Otherwise the mean does not exist.

The variance is not derived in these notes.

Relation to the Beta Distribution

$$X \sim F(\mu, \nu)$$

if and only if

$$W \sim \text{Beta}\left(\frac{\mu}{2}, \frac{\nu}{2}\right),$$

where

$$W = \frac{\frac{\mu}{\nu} X}{1 + \frac{\mu}{\nu} X}$$

B.3 Special Functions

B.3.1 The Gamma Function

The Definition

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} \, dx, \qquad \alpha > 0 \tag{B.2}$$

The Recursion Relation

$$\Gamma(\alpha + 1) = \alpha \Gamma(\alpha) \tag{B.3}$$

Known Values

$$\Gamma(1) = 1$$

and hence using the recursion relation

$$\Gamma(n + 1) = n!$$

for any nonnegative integer n.

Also

$$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$$

and hence using the recursion relation

$$\begin{aligned} \Gamma(\tfrac{3}{2}) &= \tfrac{1}{2}\sqrt{\pi} \\ \Gamma(\tfrac{5}{2}) &= \tfrac{3}{2} \cdot \tfrac{1}{2}\sqrt{\pi} \\ \Gamma(\tfrac{7}{2}) &= \tfrac{5}{2} \cdot \tfrac{3}{2} \cdot \tfrac{1}{2}\sqrt{\pi} \end{aligned}$$

and so forth.

B.3.2 The Beta Function

The function $B(s, t)$ defined by (B.1).

B.4 Discrete Multivariate Distributions

B.4.1 The Multinomial Distribution

The Abbreviation $\text{Multi}_k(n, \mathbf{p})$ or $\text{Multi}(n, \mathbf{p})$ if the dimension k is clear from context.

The Sample Space

$$S = \{\, \mathbf{y} \in \mathbb{N}^k : y_1 + \cdots y_k = n \,\}$$

where $\mathbb{N}$ denotes the "natural numbers" 0, 1, 2, ...

The Parameter $\mathbf{p} = (p_1, \ldots, p_k)$ such that $p_i \geq 0$ for all i and $\sum_i p_i = 1$.

The Density

$$f(\mathbf{y}) = \binom{n}{y_1, \ldots, y_k} \prod_{j=1}^{k} p_j^{y_j}, \qquad \mathbf{y} \in S$$

Moments

$$\begin{aligned} E(\mathbf{Y}) &= n\mathbf{p} \\ \operatorname{var}(\mathbf{Y}) &= \mathbf{M} \end{aligned}$$

where $\mathbf{M}$ is the $k \times k$ matrix with elements

$$m_{ij} = \begin{cases} np_i(1 - p_i), & i = j \\ -np_i p_j & i \neq j \end{cases}$$

Specialization The special case $n = 1$ is called the multivariate Bernoulli distribution

$$\operatorname{Ber}_k(\mathbf{p}) = \operatorname{Bin}_k(1, \mathbf{p})$$

but for once we will not spell out the details with a special section for the multivariate Bernoulli. Just take $n = 1$ in this section.

Marginal Distributions Distributions obtained by collapsing categories are again multinomial (Section 5.4.5 in these notes).

In particular, if $\mathbf{Y} \sim \operatorname{Multi}_k(n, \mathbf{p})$, then

$$(Y_1, \ldots, Y_j, Y_{j+1} + \cdots + Y_k) \sim \operatorname{Multi}_{j+1}(n, \mathbf{q}) \tag{B.4}$$

where

$$\begin{aligned} q_i &= p_i, & i \leq j \\ q_{j+1} &= p_{j+1} + \cdots p_k \end{aligned}$$

Because the random vector in (B.4) is degenerate, this equation also gives implicitly the marginal distribution of Y_1, …, Y_j

$$\begin{aligned} &f(y_1, \ldots, y_j) \\ &\quad = \binom{n}{y_1, \ldots, y_j, n - y_1 - \cdots - y_j} p_1^{y_1} \cdots p_j^{y_j} (1 - p_1 - \cdots - p_j)^{n - y_1 - \cdots - y_j} \end{aligned}$$

Univariate Marginal Distributions If $\mathbf{Y} \sim \operatorname{Multi}(n, \mathbf{p})$, then

$$Y_i \sim \operatorname{Bin}(n, p_i).$$

Conditional Distributions If $\mathbf{Y} \sim \text{Multi}_k(n, \mathbf{p})$, then

$$(Y_1, \dots, Y_j) \mid (Y_{j+1}, \dots, Y_k) \sim \text{Multi}_j(n - Y_{j+1} - \dots - Y_k, \mathbf{q}),$$

where

$$q_i = \frac{p_i}{p_1 + \dots + p_j}, \qquad i = 1, \dots, j.$$

B.5 Continuous Multivariate Distributions

B.5.1 The Uniform Distribution

The uniform distribution defined in Section B.2.1 actually made no mention of dimension. If the set S on which the distribution is defined lies in $\mathbb{R}^n$, then this is a multivariate distribution.

Conditional Distributions Every conditional distribution of a multivariate uniform distribution is uniform.

Marginal Distributions No regularity. Depends on the particular distribution. Marginals of the uniform distribution on a rectangle with sides parallel to the coordinate axes are uniform. Marginals of the uniform distribution on a disk or triangle are not uniform.

B.5.2 The Standard Normal Distribution

The distribution of a random vector $\mathbf{Z} = (Z_1, \dots, Z_k)$ with the Z_i i. i. d. standard normal.

Moments

$$\begin{aligned} E(\mathbf{Z}) &= 0 \\ \operatorname{var}(\mathbf{Z}) &= \mathbf{I}, \end{aligned}$$

where $\mathbf{I}$ denotes the $k \times k$ identity matrix.

B.5.3 The Multivariate Normal Distribution

The distribution of a random vector $\mathbf{X} = \mathbf{a} + \mathbf{B}\mathbf{Z}$, where $\mathbf{Z}$ is multivariate standard normal.

Moments

$$\begin{aligned} E(\mathbf{X}) &= \boldsymbol{\mu} = \mathbf{a} \\ \operatorname{var}(\mathbf{X}) &= \mathbf{M} = \mathbf{B}\mathbf{B}' \end{aligned}$$

The Abbreviation $\mathcal{N}_k(\boldsymbol{\mu}, \mathbf{M})$ or $\mathcal{N}(\boldsymbol{\mu}, \mathbf{M})$ if the dimension k is clear from context.

The Sample Space If $\mathbf{M}$ is positive definite, the sample space is $\mathbb{R}^k$.

Otherwise, X is concentrated on the intersection of hyperplanes determined by null eigenvectors of $\mathbf{M}$

$$S = \{\, \mathbf{x} \in \mathbb{R}^k : \mathbf{z}'\mathbf{x} = \mathbf{z}'\boldsymbol{\mu} \text{ whenever } \mathbf{M}\mathbf{z} = 0 \,\}$$

The Parameters The mean vector $\boldsymbol{\mu}$ and variance matrix $\mathbf{M}$.

The Density Only exists if the distribution is nondegenerate ($\mathbf{M}$ is positive definite). Then

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} \det(\mathbf{M})^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})'\mathbf{M}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right), \qquad \mathbf{x} \in \mathbb{R}^k$$

Marginal Distributions All are normal. If

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

is a partitioned random vector with (partitioned) mean vector

$$E(\mathbf{X}) = \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}$$

and (partitioned) variance matrix

$$\operatorname{var}(\mathbf{X}) = \mathbf{M} = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix}$$

and $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{M})$, then

$$\mathbf{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{M}_{11}).$$

Conditional Distributions All are normal. If $\mathbf{X}$ is as in the preceding section and $\mathbf{X}_2$ is nondegenerate, then the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$ is normal with

$$\begin{aligned} E(\mathbf{X}_1 \mid \mathbf{X}_2) &= \boldsymbol{\mu}_1 + \mathbf{M}_{12}\mathbf{M}_{22}^{-1}(\mathbf{X}_2 - \boldsymbol{\mu}_2) \\ \operatorname{var}(\mathbf{X}_1 \mid \mathbf{X}_2) &= \mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21} \end{aligned}$$

If $\mathbf{X}_2$ is degenerate so $\mathbf{M}_{22}$ is not invertible, then the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$ is still normal and the same formulas work if $\mathbf{M}_{22}^{-1}$ is replaced by a generalized inverse.

B.5.4 The Bivariate Normal Distribution

The special case $k = 2$ of the preceeding section.

The Density

$$
\begin{aligned}
f(x, y) = {} & \frac{1}{2 \pi \sigma_X \sigma_Y \sqrt{1 - \rho^2}} \times \\
& \exp\left(- \frac{1}{2(1 - \rho^2)} \left[\frac{(x - \mu_X)^2}{\sigma_X^2} - \frac{2 \rho (x - \mu_X)(y - \mu_Y)}{\sigma_X \sigma_Y} + \frac{(y - \mu_Y)^2}{\sigma_Y^2} \right] \right)
\end{aligned}
$$

Marginal Distributions

$$
Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)
$$

Conditional Distributions The conditional distribution of X given Y is normal with

$$
\begin{aligned}
E(X \mid Y) &= \mu_X + \rho \frac{\sigma_X}{\sigma_Y} (Y - \mu_Y) \\
\operatorname{var}(X \mid Y) &= \sigma_X^2 (1 - \rho^2)
\end{aligned}
$$

where $\rho = \operatorname{cor}(X, Y)$.

Appendix C

Addition Rules for Distributions

"Addition rules" for distributions are rules of the form: if X_1, ..., X_k are independent with some specified distributions, then $X_1 + \cdots + X_k$ has some other specified distribution.

Bernoulli If X_1, ..., X_k are i. i. d. $\text{Ber}(p)$, then

$$X_1 + \cdots + X_k \sim \text{Bin}(k, p). \tag{C.1}$$

- All the Bernoulli distributions must have the *same* success probability p.

Binomial If X_1, ..., X_k are independent with $X_i \sim \text{Bin}(n_i, p)$, then

$$X_1 + \cdots + X_k \sim \text{Bin}(n_1 + \cdots + n_k, p). \tag{C.2}$$

- All the binomial distributions must have the *same* success probability p.
- (C.1) is the special case of (C.2) obtained by setting $n_1 = \cdots = n_k = 1$.

Geometric If X_1, ..., X_k are i. i. d. $\text{Geo}(p)$, then

$$X_1 + \cdots + X_k \sim \text{NegBin}(k, p). \tag{C.3}$$

- All the geometric distributions must have the *same* success probability p.

Negative Binomial If X_1, ..., X_k are independent with $X_i \sim \text{NegBin}(n_i, p)$, then

$$X_1 + \cdots + X_k \sim \text{NegBin}(n_1 + \cdots + n_k, p). \tag{C.4}$$

- All the negative binomial distributions must have the *same* success probability p.
- (C.3) is the special case of (C.4) obtained by setting $n_1 = \cdots = n_k = 1$.

Poisson If X_1, …, X_k are independent with $X_i \sim \text{Poi}(\mu_i)$, then

$$X_1 + \cdots + X_k \sim \text{Poi}(\mu_1 + \cdots + \mu_k). \tag{C.5}$$

Exponential If X_1, …, X_k are i. i. d. $\text{Exp}(\lambda)$, then

$$X_1 + \cdots + X_k \sim \text{Gam}(n, \lambda). \tag{C.6}$$

- All the exponential distributions must have the *same* rate parameter λ.

Gamma If X_1, …, X_k are independent with $X_i \sim \text{Gam}(\alpha_i, \lambda)$, then

$$X_1 + \cdots + X_k \sim \text{Gam}(\alpha_1 + \cdots + \alpha_k, \lambda). \tag{C.7}$$

- All the gamma distributions must have the *same* rate parameter λ.
- (C.6) is the special case of (C.7) obtained by setting $\alpha_1 = \cdots = \alpha_k = 1$.

Chi-Square If X_1, …, X_k are independent with $X_i \sim \text{chi}^2(n_i)$, then

$$X_1 + \cdots + X_k \sim \text{chi}^2(n_1 + \cdots + n_k). \tag{C.8}$$

- (C.8) is the special case of (C.7) obtained by setting

$$\alpha_i = n_i/2 \quad \text{and} \quad \lambda_i = 1/2, \qquad i = 1, \ldots, k.$$

Normal If X_1, …, X_k are independent with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, then

$$X_1 + \cdots + X_k \sim \mathcal{N}(\mu_1 + \cdots + \mu_k, \sigma_1^2 + \cdots + \sigma_k^2). \tag{C.9}$$

Linear Combination of Normals If X_1, …, X_k are independent with $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ and a_1, …, a_k are constants, then

$$\sum_{i=1}^k a_i X_i \sim \mathcal{N}\left(\sum_{i=1}^k a_i \mu_i, \sum_{i=1}^k a_i^2 \sigma_i^2 \right). \tag{C.10}$$

- (C.9) is the special case of (C.10) obtained by setting $a_1 = \cdots = a_k = 1$.

Cauchy If X_1, …, X_k are independent with $X_i \sim \text{Cauchy}(\mu, \sigma)$, then

$$X_1 + \cdots + X_k \sim \text{Cauchy}(n\mu, n\sigma). \tag{C.11}$$

Appendix D

Relations Among Brand Name Distributions

D.1 Special Cases

First there are the special cases, which were also noted in Appendix B.

$$
\begin{aligned}
\operatorname{Ber}(p) &= \operatorname{Bin}(1, p) \\
\operatorname{Geo}(p) &= \operatorname{NegBin}(1, p) \\
\operatorname{Exp}(\lambda) &= \operatorname{Gam}(1, \lambda) \\
\operatorname{chi}^2(k) &= \operatorname{Gam}\left(\tfrac{k}{2}, \tfrac{1}{2}\right)
\end{aligned}
$$

The main point of this appendix are the relationships that involve more theoretical issues.

D.2 Relations Involving Bernoulli Sequences

Suppose X_1, X_2, $\ldots$ are i. i. d. $\operatorname{Ber}(p)$ random variables.

If n is a positive integer and

$$
Y = X_1 + \cdots + X_n
$$

is the number of "successes" in the n Bernoulli trials, then

$$
Y \sim \operatorname{Bin}(n, p).
$$

On the other hand, if y is positive integer and N is the trial at which the y-th success occurs, that is the random number N such that

$$
\begin{aligned}
X_1 + \cdots + X_N &= y \\
X_1 + \cdots + X_k &< y, \qquad k < N,
\end{aligned}
$$

then

$$
N \sim \operatorname{NegBin}(y, p).
$$

D.3 Relations Involving Poisson Processes

In a one-dimensional homogeneous Poisson process with rate parameter λ, the counts are Poisson and the waiting and interarrival times are exponential. Specifically, the number of points (arrivals) in an interval of length t has the $\text{Poi}(\lambda t)$ distribution, and the waiting times and interarrival times are independent and indentically $\text{Exp}(\lambda)$ distributed.

Even more specifically, let X_1, X_2, $\ldots$ be i. i. d. $\text{Exp}(\lambda)$ random variables. Take these to be the waiting and interarrival times of a Poisson process. This means the arrival times themselves are

$$T_k = \sum_{i=1}^{k} X_i$$

Note that

$$0 < T_1 < T_2 < \cdots$$

and

$$X_i = T_i - T_{i-1}, \qquad i > 1$$

so these are the interarrival times and $X_1 = T_1$ is the waiting time until the first arrival.

The characteristic property of the Poisson process, that counts have the Poisson distribution, says the number of points in the interval $(0, t)$, that is, the number of T_i such that $T_i < t$, has the $\text{Poi}(\lambda t)$ distribution.

D.4 Normal, Chi-Square, t, and F

D.4.1 Definition of Chi-Square

If Z_1, Z_2, $\ldots$ are i. i. d. $\mathcal{N}(0, 1)$, then

$$Z_1^2 + \ldots + Z_n^2 \sim \text{chi}^2(n).$$

D.4.2 Definition of t

If Z and Y are independent and

$$\begin{aligned} Z &\sim \mathcal{N}(0, 1) \\ Y &\sim \text{chi}^2(\nu) \end{aligned}$$

then

$$\frac{Z}{\sqrt{Y/\nu}} \sim t(\nu)$$

D.4.3 Definition of F

If X and Y are independent and

$$\begin{aligned} X &\sim \operatorname{chi}^2(\mu) \\ Y &\sim \operatorname{chi}^2(\nu) \end{aligned}$$

then

$$\frac{X/\mu}{Y/\nu} \sim F(\mu, \nu)$$

D.4.4 t as a Special Case of F

If

$$T \sim t(\nu),$$

then

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

Appendix E

Eigenvalues and Eigenvectors

E.1 Orthogonal and Orthonormal Vectors

If $\mathbf{x}$ and $\mathbf{y}$ are vectors of the same dimension, we say they are *orthogonal* if $\mathbf{x}'\mathbf{y} = 0$. Since the transpose of a matrix product is the product of the transposes in reverse order, an equivalent condition is $\mathbf{y}'\mathbf{x} = 0$. Orthogonality is the n-dimensional generalization of perpendicularity. In a sense, it says that two vectors make a right angle.

The *length* or *norm* of a vector $\mathbf{x} = (x_1, \ldots, x_n)$ is defined to be

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}'\mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Squaring both sides gives

$$\|\mathbf{x}\|^2 = \sum_{i=1}^{n} x_i^2,$$

which is one version of the Pythagorean theorem, as it appears in analytic geometry.

Orthogonal vectors give another generalization of the Pythagorean theorem. We say a set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is *orthogonal* if

$$\mathbf{x}_i'\mathbf{x}_j = 0, \qquad i \neq j. \tag{E.1}$$

Then

$$\begin{aligned}
\|\mathbf{x}_1 + \cdots + \mathbf{x}_k\|^2 &= (\mathbf{x}_1 + \cdots + \mathbf{x}_k)'(\mathbf{x}_1 + \cdots + \mathbf{x}_k) \\
&= \sum_{i=1}^k \sum_{j=1}^k \mathbf{x}_i' \mathbf{x}_j \\
&= \sum_{i=1}^k \mathbf{x}_i' \mathbf{x}_i \\
&= \sum_{i=1}^k \|\mathbf{x}_i\|^2
\end{aligned}$$

because, by definition of orthogonality, all terms in the second line with $i \neq j$ are zero.

We say an orthogonal set of vectors is *orthonormal* if

$$\mathbf{x}_i' \mathbf{x}_i = 1. \tag{E.2}$$

That is, a set of vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ is orthonormal if it satisfies both (E.1) and (E.2).

An orthonormal set is automatically linearly independent because if

$$\sum_{i=1}^k c_i \mathbf{x}_i = 0,$$

then

$$0 = \mathbf{x}_j' \left(\sum_{i=1}^k c_i \mathbf{x}_i \right) = c_j \mathbf{x}_j' \mathbf{x}_j = c_j$$

holds for all j. Hence the only linear combination that is zero is the one with all coefficients zero, which is the definition of linear independence.

Being linearly independent, an orthonormal set is always a *basis* for whatever subspace it spans. If we are working in n-dimensional space, and there are n vectors in the orthonormal set, then they make up a basis for the whole space. If there are $k < n$ vectors in the set, then they make up a basis for some proper subspace.

It is always possible to choose an orthogonal basis for any vector space or subspace. One way to do this is the Gram-Schmidt orthogonalization procedure, which converts an arbitrary basis $\mathbf{y}_1$, …, $\mathbf{y}_n$ to an orthonormal basis $\mathbf{x}_1$, …, $\mathbf{x}_n$ as follows. First let

$$\mathbf{x}_1 = \frac{\mathbf{y}_1}{\|\mathbf{y}_1\|}.$$

Then define the $\mathbf{x}_i$ in order. After $\mathbf{x}_1$, …, $\mathbf{x}_{k-1}$ have been defined, let

$$\mathbf{z}_k = \mathbf{y}_k - \sum_{i=1}^{k-1} \mathbf{x}_i \mathbf{x}_i' \mathbf{y}$$

and

$$\mathbf{x}_k = \frac{\mathbf{z}_k}{\|\mathbf{z}_k\|}.$$

It is easily verified that this does produce an orthonormal set, and it is only slightly harder to prove that none of the $\mathbf{x}_i$ are zero because that would imply linear dependence of the $\mathbf{y}_i$.

E.2 Eigenvalues and Eigenvectors

If $\mathbf{A}$ is any matrix, we say that λ is a *right eigenvalue* corresponding to a *right eigenvector* $\mathbf{x}$ if

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

Left eigenvalues and eigenvectors are defined analogously with "left multiplication" $\mathbf{x}'\mathbf{A} = \lambda\mathbf{x}'$, which is equivalent to $\mathbf{A}'\mathbf{x} = \lambda\mathbf{x}$. So the right eigenvalues and eigenvectors of $\mathbf{A}'$ are the left eigenvalues and eigenvectors of $\mathbf{A}$. When $\mathbf{A}$ is symmetric ($\mathbf{A}' = \mathbf{A}$), the "left" and "right" concepts are the same and the adjectives "left" and "right" are unnecessary. Fortunately, this is the most interesting case, and the only one in which we will be interested. From now on we discuss only eigenvalues and eigenvectors of *symmetric* matrices.

There are three important facts about eigenvalues and eigenvectors. Two elementary and one very deep. Here's the first (one of the elementary facts).

Lemma E.1. *Eigenvectors corresponding to distinct eigenvalues are orthogonal.*

This means that if

$$\mathbf{A}\mathbf{x}_i = \lambda_i\mathbf{x}_i \tag{E.3}$$

then

$$\lambda_i \neq \lambda_j \qquad \text{implies} \qquad \mathbf{x}_i'\mathbf{x}_j = 0.$$

Proof. Suppose $\lambda_i \neq \lambda_j$, then at least one of the two is not zero, say λ_j. Then

$$\mathbf{x}_i'\mathbf{x}_j = \frac{\mathbf{x}_i'\mathbf{A}\mathbf{x}_j}{\lambda_j} = \frac{(\mathbf{A}\mathbf{x}_i)'\mathbf{x}_j}{\lambda_j} = \frac{\lambda_i\mathbf{x}_i'\mathbf{x}_j}{\lambda_j} = \frac{\lambda_i}{\lambda_j}\cdot\mathbf{x}_i'\mathbf{x}_j$$

and since $\lambda_i \neq \lambda_j$ the only way this can happen is if $\mathbf{x}_i'\mathbf{x}_j = 0$. □

Here's the second important fact (also elementary).

Lemma E.2. *Every linear combination of eigenvectors corresponding to the same eigenvalue is another eigenvector corresponding to that eigenvalue.*

This means that if

$$\mathbf{A}\mathbf{x}_i = \lambda\mathbf{x}_i$$

then

$$\mathbf{A}\left(\sum_{i=1}^{k} c_i\mathbf{x}_i\right) = \lambda\left(\sum_{i=1}^{k} c_i\mathbf{x}_i\right)$$

Proof. This is just linearity of matrix multiplication. □

The second property means that all the eigenvectors corresponding to one eigenvalue constitute a subspace. If the dimension of that subspace is k, then it is possible to choose an orthonormal basis of k vectors that span the subspace. Since the first property of eigenvalues and eigenvectors says that (E.1) is also satisfied by eigenvectors corresponding to different eigenvalues, all of the eigenvectors chosen this way form an orthonormal set.

Thus our orthonormal set of eigenvectors spans a subspace of dimension m which contains all eigenvectors of the matrix in question. The question then arises whether this set is *complete*, that is, whether it is a basis for the whole space, or in symbols whether $m = n$, where n is the dimension of the whole space ($\mathbf{A}$ is an $n \times n$ matrix and the $\mathbf{x}_i$ are vectors of dimension n). It turns out that the set *is* always complete, and this is the third important fact about eigenvalues and eigenvectors.

Lemma E.3. *Every real symmetric matrix has an orthonormal set of eigenvectors that form a basis for the space.*

In contrast to the first two facts, this is deep, and we shall not say anything about its proof, other than that about half of the typical linear algebra book is given over to building up to the proof of this one fact.

The "third important fact" says that *any* vector can be written as a linear combination of eigenvectors

$$\mathbf{y} = \sum_{i=1}^{n} c_i \mathbf{x}_i$$

and this allows a very simple description of the action of the linear operator described by the matrix

$$\mathbf{A}\mathbf{y} = \sum_{i=1}^{n} c_i \mathbf{A}\mathbf{x}_i = \sum_{i=1}^{n} c_i \lambda_i \mathbf{x}_i \tag{E.4}$$

So this says that *when we use an orthonormal eigenvector basis*, if $\mathbf{y}$ has the representation $(c_1, \ldots, c_n)$, then $\mathbf{A}y$ has the representation $(c_1\lambda_1, \ldots, c_n\lambda_n)$. Let $\mathbf{D}$ be the representation in the orthonormal eigenvector basis of the linear operator represented by $\mathbf{A}$ in the standard basis. Then our analysis above says the i-the element of $\mathbf{D}\mathbf{c}$ is $c_i\lambda_i$, that is,

$$\sum_{j=1}^{n} d_{ij} c_j = \lambda_i c_i.$$

In order for this to hold for all real numbers c_i, it must be that $\mathbf{D}$ is diagonal

$$\begin{aligned} d_{ii} &= \lambda_i \\ d_{ij} &= 0, \qquad i \neq j \end{aligned}$$

In short, using the orthonormal eigenvector basis *diagonalizes* the linear operator represented by the matrix in question.

There is another way to describe this same fact without mentioning bases. Many people find it a simpler description, though its relation to eigenvalues and eigenvectors is hidden in the notation, no longer immediately apparent. Let $\mathbf{O}$ denote the matrix whose columns are the orthonormal eigenvector basis ($\mathbf{x}_1$, …, $\mathbf{x}_n$), that is, if o_{ij} are the elements of $\mathbf{O}$, then

$$\mathbf{x}_i = (o_{1i}, \ldots, o_{ni}).$$

Now (E.1) and (E.2) can be combined as one matrix equation

$$\mathbf{O}'\mathbf{O} = \mathbf{I} \tag{E.5}$$

(where, as usual, $\mathbf{I}$ is the $n \times n$ identity matrix). A matrix $\mathbf{O}$ satisfying this property is said to be *orthogonal*. Another way to read (E.5) is that it says $\mathbf{O}' = \mathbf{O}^{-1}$ (an orthogonal matrix is one whose inverse is its transpose). The fact that inverses are two-sided ($\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ for any invertible matrix $\mathbf{A}$) implies that $\mathbf{O}\mathbf{O}' = \mathbf{I}$ as well.

Furthermore, the eigenvalue-eigenvector equation (E.3) can be written out with explicit subscripts and summations as

$$\sum_{j=1}^{n} a_{ij} o_{jk} = \lambda_k o_{ik} = o_{ik} d_{kk} = \sum_{j=1}^{n} o_{ij} d_{jk}$$

(where $\mathbf{D}$ is the the diagonal matrix with eigenvalues on the diagonal defined above). Going back to matrix notation gives

$$\mathbf{A}\mathbf{O} = \mathbf{O}\mathbf{D} \tag{E.6}$$

The two equations (E.3) and (E.6) may not look much alike, but as we have just seen, they say exactly the same thing in different notation. Using the orthogonality property ($\mathbf{O}' = \mathbf{O}^{-1}$) we can rewrite (E.6) in two different ways.

Theorem E.4 (Spectral Decomposition). *Any real symmetric matrix* $\mathbf{A}$ *can be written*

$$\mathbf{A} = \mathbf{O}\mathbf{D}\mathbf{O}' \tag{E.7}$$

where $\mathbf{D}$ *is diagonal and* $\mathbf{O}$ *is orthogonal.*

Conversely, for any real symmetric matrix $\mathbf{A}$ *there exists an orthogonal matrix* $\mathbf{O}$ *such that*

$$\mathbf{D} = \mathbf{O}'\mathbf{A}\mathbf{O}$$

is diagonal.

(The reason for the name of the theorem is that the set of eigenvalues is sometimes called the *spectrum* of $\mathbf{A}$). The spectral decomposition theorem says nothing about eigenvalues and eigenvectors, but we know from the discussion above that the diagonal elements of $\mathbf{D}$ are the eigenvalues of $\mathbf{A}$, and the columns of $\mathbf{O}$ are the corresponding eigenvectors.

E.3 Positive Definite Matrices

Using the spectral theorem, we can prove several interesting things about positive definite matrices.

Corollary E.5. *A real symmetric matrix* $\mathbf{A}$ *is positive semi-definite if and only if its spectrum is nonnegative. A real symmetric matrix* $\mathbf{A}$ *is positive definite if and only if its spectrum is strictly positive.*

Proof. First suppose that $\mathbf{A}$ is positive semi-definite with spectral decomposition (E.7). Let $\mathbf{e}_i$ denote the vector having elements that are all zero except the i-th, which is one, and define $\mathbf{w} = \mathbf{O}\mathbf{e}_i$, so

$$0 \le \mathbf{w}'\mathbf{A}\mathbf{w} = \mathbf{e}_i'\mathbf{O}'\mathbf{O}\mathbf{D}\mathbf{O}'\mathbf{O}\mathbf{e}_i = \mathbf{e}_i'\mathbf{D}\mathbf{e}_i = d_{ii} \tag{E.8}$$

using $\mathbf{O}'\mathbf{O} = I$. Hence the spectrum is nonnegative.

Conversely, suppose the d_{ii} are nonnegative. Then for any vector $\mathbf{w}$ define $\mathbf{z} = \mathbf{O}'\mathbf{w}$, so

$$\mathbf{w}'\mathbf{A}\mathbf{w} = \mathbf{w}'\mathbf{O}\mathbf{D}\mathbf{O}'\mathbf{w} = \mathbf{z}'\mathbf{D}\mathbf{z} = \sum\nolimits_i d_{ii} z_i^2 \ge 0$$

Hence $\mathbf{A}$ is positive semi-definite.

The assertions about positive definiteness are proved in almost the same way. Suppose that $\mathbf{A}$ is positive definite. Since $\mathbf{e}_i$ is nonzero, $\mathbf{w}$ in (E.8) is also nonzero because $\mathbf{e}_i = \mathbf{O}'\mathbf{w}$ would be zero (and it isn't) if $\mathbf{w}$ were zero. Thus the inequality in (E.8) is actually strict. Hence the spectrum of is strictly positive.

Conversely, suppose the d_{ii} are strictly positive. Then for any nonzero vector $\mathbf{w}$ define $\mathbf{z} = \mathbf{O}'\mathbf{w}$ as before, and again note that $\mathbf{z}$ is nonzero because $\mathbf{w} = \mathbf{O}\mathbf{z}$ and $\mathbf{w}$ is nonzero. Thus $\mathbf{w}'\mathbf{A}\mathbf{w} = \mathbf{z}'\mathbf{D}\mathbf{z} > 0$, and hence $\mathbf{A}$ is positive definite. □

Corollary E.6. *A positive semi-definite matrix is invertible if and only if it is positive definite.*

Proof. It is easily verified that the product of diagonal matrices is diagonal and the diagonal elements of the product are the products of the diagonal elements of the multiplicands. Thus a diagonal matrix $\mathbf{D}$ is invertible if and only if all its diagonal elements d_{ii} are nonzero, in which case $\mathbf{D}^{-1}$ is diagonal with diagonal elements $1/d_{ii}$.

Since $\mathbf{O}$ and $\mathbf{O}'$ in the spectral decomposition (E.7) are invertible, $\mathbf{A}$ is invertible if and only if $\mathbf{D}$ is, hence if and only if its spectrum is nonzero, in which case

$$\mathbf{A}^{-1} = \mathbf{O}\mathbf{D}^{-1}\mathbf{O}'.$$

By the preceding corollary the spectrum of a positive semi-definite matrix is nonnegative, hence nonzero if and only if strictly positive, which (again by the preceding corollary) occurs if and only if the matrix is positive definite. □

Corollary E.7. *Every real symmetric positive semi-definite matrix* $\mathbf{A}$ *has a symmetric square root*

$$\mathbf{A}^{1/2} = \mathbf{O}\mathbf{D}^{1/2}\mathbf{O}' \tag{E.9}$$

where (E.7) *is the spectral decomposition of* $\mathbf{A}$ *and where* $\mathbf{D}^{1/2}$ *is defined to be the diagonal matrix whose diagonal elements are* $\sqrt{d_{ii}}$, *where* d_{ii} *are the diagonal elements of* $\mathbf{D}$.

Moreover, $\mathbf{A}^{1/2}$ *is positive definite if and only if* $\mathbf{A}$ *is positive definite.*

Note that by Corollary E.5 all of the diagonal elements of $\mathbf{D}$ are nonnegative and hence have real square roots.

Proof.

$$\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{O}\mathbf{D}^{1/2}\mathbf{O}'\mathbf{O}\mathbf{D}^{1/2}\mathbf{O}' = \mathbf{O}\mathbf{D}^{1/2}\mathbf{D}^{1/2}\mathbf{O}' = \mathbf{O}\mathbf{D}\mathbf{O}' = \mathbf{A}$$

because $\mathbf{O}'\mathbf{O} = \mathbf{I}$ and $\mathbf{D}^{1/2}\mathbf{D}^{1/2} = \mathbf{D}$.

From Corollary E.5 we know that $\mathbf{A}$ is positive definite if and only if all the d_{ii} are strictly positive. Since (E.9) is the spectral decomposition of $\mathbf{A}^{1/2}$, we see that $\mathbf{A}^{1/2}$ is positive definite if and only if all the $\sqrt{d_{ii}}$ are strictly positive. Clearly $d_{ii} > 0$ if and only if $\sqrt{d_{ii}} > 0$. □

Appendix F

Normal Approximations for Distributions

F.1 Binomial Distribution

The $\mathrm{Bin}(n, p)$ distribution is approximately normal with mean np and variance $np(1-p)$ if n is large.

F.2 Negative Binomial Distribution

The $\mathrm{NegBin}(n, p)$ distribution is approximately normal with mean n/p and variance $n(1-p)/p^2$ if n is large.

F.3 Poisson Distribution

The $\mathrm{Poi}(\mu)$ distribution is approximately normal with mean μ and variance μ if μ is large.

F.4 Gamma Distribution

The $\mathrm{Gam}(\alpha, \lambda)$ distribution is approximately normal with mean α/λ and variance α/λ^2 if α is large.

F.5 Chi-Square Distribution

The $\mathrm{chi}^2(n)$ distribution is approximately normal with mean n and variance $2n$ if n is large.

Appendix G

Maximization of Functions

This appendix contains no statistics. It just reviews some facts from calculus about maximization of functions.

First we distinguish between local and global maxima.[1] A point x is a *global maximum* of a function f if

$$f(x) \geq f(y), \qquad \text{for all } y \text{ in the domain of } f.$$

In words, $f(x)$ is greater than or equal to $f(y)$ for all other y.

Unfortunately, calculus isn't much help in finding global maxima, hence the following definition, which defines something calculus is much more helpful in finding. A point x is a *local maximum* of the function f if

$$f(x) \geq f(y), \qquad \text{for all } y \text{ in some neighborhood of } x.$$

The point is that saying x is a local maximum doesn't say anything at all about whether a global maximum exists or whether x is also a global maximum.

> *Every global maximum is a local maximum, but not all local maxima are global maxima.*

G.1 Functions of One Variable

The connection between calculus and local maxima is quite simple.

Theorem G.1. *Suppose f is a real-valued function of one real variable and is twice differentiable at the point x. A sufficient condition that x be a local maximum of f is*

$$f'(x) = 0 \quad \textit{and} \quad f''(x) < 0. \tag{G.1}$$

[1] An irregular plural following the Latin rather than the English pattern. Singular: *maximum.* Plural: *maxima.*

In words, to find a local maximum, find a point x where the derivative is zero. Then check the second derivative. If $f''(x)$ is negative, then x is a local maximum. If $f''(x)$ is positive, then x is definitely not a local maximum (in fact it's a local minimum). If $f''(x)$ is zero, you are not sure. Consider $f(x) = x^3$ and $g(x) = -x^4$. Both have second derivative zero at $x = 0$, but f is strictly increasing (draw a graph) and hence does not have a maximum (local or global), whereas g does have a local maximum at $x = 0$.

That takes care of local maxima that occur at interior points of the domain of the function being maximized. What about local maxima that occur at boundary points? Here the situation becomes more complicated.

Our first problem is that ordinary derivatives don't exist, but there still may be one-sided derivatives. In the following discussion all the derivatives are one-sided.

Theorem G.2. *Suppose f is a twice differentiable real-valued function defined on a closed interval of the real line. A sufficient condition that a lower boundary point x of the interval be a local maximum of f is*

$$f'(x) < 0. \tag{G.2a}$$

Another sufficient condition is

$$f'(x) = 0 \quad \textit{and} \quad f''(x) < 0. \tag{G.2b}$$

If x is an upper boundary point, the conditions are the same except the inequality in (G.2a) *is reversed: $f'(x) > 0$.*

This theorem is so complicated that you're excused if you want to ignore it and just draw a graph of the function. The main point of the theorem is that a local maximum can occur at a boundary point when the derivative is not zero. Consider the function $f(x) = -x$ defined on the interval $0 \le x < +\infty$. Since f is strictly decreasing, the only global (and local) maximum occurs at $x = 0$ where $f'(x) = -1$. This satisfies the condition (G.2a) of the theorem, but notice the derivative is not zero.

Thus it is *not* enough to just look for points where the first derivative is zero. You also have to check the boundary points, where the more complicated test applies.

Are we done with maximization theory? Not if we are interested in global maxima. Even if you find all the local maxima, it is not necessarily true that a global maximum exists. Consider the function $f(x) = x^3 - x$ graphed in Figure G.1. The first derivative is

$$f'(x) = 3x^2 - 1,$$

which has zeros at $\pm 1/\sqrt{3}$. From the graph it is clear that $-1/\sqrt{3}$ is a local maximum and $+1/\sqrt{3}$ is a local *minimum*. But there is no global maximum since $f(x) \to +\infty$ as $x \to +\infty$.

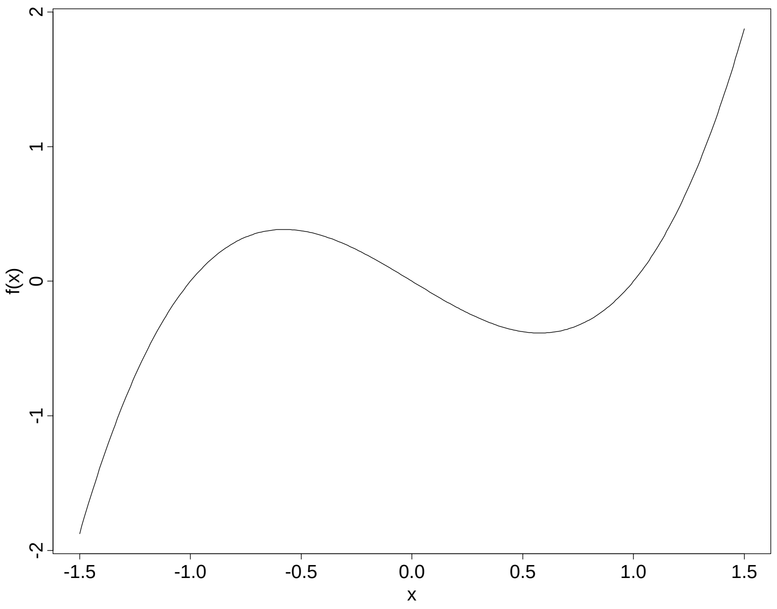


Figure G.1: Graph of $f(x) = x^3 - x$.

If a global maximum exists, then it is also a local maximum. So if you find all local maxima, they must include any global maxima. But the example shows that local maxima can exist when there is no global maximum. Thus calculus can help you find local maxima, but it is no help in telling you which of them are global maxima or whether any of them are global maxima.

G.2 Concave Functions of One Variable

There is one situation in which maximization is much simpler.

Definition G.2.1 (Strictly Concave Functions).
A continuous real-valued function f defined on an interval of the real line and twice differentiable at interior points of the interval is strictly concave *if the inequality $f''(x) < 0$ holds at every interior point x of the interval.*

There is a more general definition of "concave function" that does not require differentiability, but we will not use it.[2]

The concavity property $f''(x) < 0$ is easily recognized from a graph. It says the function curves downward at every point. Figure G.2 is an example.

Strictly concave functions are very special. For them, there is no difference between local and global maxima.

[2] Functions that are twice differentiable are concave but not strictly concave if $f''(x) \leq 0$ at all interior points and $f''(x) = 0$ at some points. But we won't have any use for this concept.

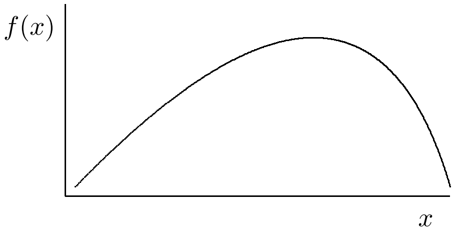


Figure G.2: A Concave Function.

Theorem G.3. *A continuous real-valued strictly concave function f defined on an interval of the real line and twice differentiable at interior points of the interval has at most one local maximum. If it has a local maximum, then that point is also the global maximum.*

Also, f' has at most one zero. If it has a zero, this is the global maximum.

The theorem says the the situation for strictly concave functions is very simple. If you find a local maximum or a zero of the first derivative, then you have found the unique global maximum.

To summarize, there is a very important distinction between general functions and strictly concave ones. The derivative tests are almost the same, but there is a subtle difference. If

$$f'(x) = 0 \text{ and } f''(x) < 0$$

then you know x is a local maximum of f, but you don't know that x is a global maximum or whether there is any global maximum. But if

$$f'(x) = 0 \text{ and } f''(y) < 0 \text{ for all } y$$

then you know that f is strictly concave and hence that x is the unique global maximum. The only difference is whether you just check the sign of the second derivative only at the point x or at all points y in the domain of f.

G.3 Functions of Several Variables

In Chapter 5 (p. 22 of the notes) we learned about the first derivative of a vector-valued function of a vector variable. This derivative was used in the multivariable delta method.

To develop the multivariable analog of the theory of the preceeding sections, we need to develop first and *second* derivatives of a scalar-valued function of a vector variable. (Fortunately, we don't need second derivatives of *vector*-valued functions of a vector variable. They're a mess.)

According to the theory developed in Chapter 5, if $\mathbf{g}$ is a function that maps vectors of dimension n to vectors of dimension m, then its derivative at the point $\mathbf{x}$ is the $m \times n$ matrix $\nabla \mathbf{g}(\mathbf{x})$ having elements

$$g_{ij} = \frac{\partial g_i(\mathbf{x})}{\partial x_j}$$

Here we are interested in the case $m = 1$ (a *scalar*-valued function) so the derivative is a $1 \times n$ matrix (a row vector).

Thus if f is a real-valued ("scalar-valued" means the same thing) function of a vector variable of dimension n, its first derivative is the row vector $\nabla f(\mathbf{x})$ having elements

$$\frac{\partial f(\mathbf{x})}{\partial x_i}, \qquad i = 1, \ldots, n.$$

It is pronounced "del f".

So what might the second derivative be? It is clear from the pattern that it should involve partial derivatives, in this case second derivatives. There are a lot of them. If f is a real-valued function of a vector variable of dimension n, its second derivative is the $n \times n$ matrix $\nabla^2 f(\mathbf{x})$ having elements

$$\frac{\partial f(\mathbf{x})}{\partial x_i \partial x_j}, \qquad i = 1, \ldots, n \text{ and } j = 1, \ldots, n.$$

It is pronounced "del squared f". Note that by the properties of partial derivatives

$$\frac{\partial f(\mathbf{x})}{\partial x_i \partial x_j} = \frac{\partial f(\mathbf{x})}{\partial x_j \partial x_i}$$

the second derivative matrix is a *symmetric* matrix.

Before we can state the multivariate analogue of Theorem G.2, we need to develop one more concept. Recall from Section 5.1.8 of last semester's notes (or look at Section 12.5 in Lindgren) that a symmetric square matrix $\mathbf{A}$ is *positive semidefinite* (Lindgren says *nonnegative definite*) if

$$\mathbf{c}'\mathbf{A}\mathbf{c} \geq 0, \qquad \text{for all vectors } \mathbf{c}, \tag{G.3a}$$

and $\mathbf{A}$ is *positive definite* if the inequality is strict

$$\mathbf{c}'\mathbf{A}\mathbf{c} > 0, \qquad \text{for all nonzero vectors } \mathbf{c}. \tag{G.3b}$$

As a shorthand we write $\mathbf{A} \geq 0$ to indicate (G.3a) and $A > 0$ to indicate (G.3b). No confusion should arise, because these can't have any other meaning (matrices aren't naturally ordered). We also write $\mathbf{A} \leq 0$ and $\mathbf{A} < 0$ to mean that $-\mathbf{A}$ is positive semidefinite or positive definite, respectively. When $\mathbf{A} \leq 0$ we say that it is *negative semidefinite*, and when $\mathbf{A} < 0$ we say that it is *negative definite*.

The place where positive (semi)definiteness arose last semester was that fact that every variance matrix is positive semidefinite (Corollary 5.5 in these notes, Theorem 9 of Chapter 12 in Lindgren) and actually positive definite if the random variable in question is not concentrated on a hyperplane (Corollary 5.6 in last semester's notes).

With these concepts we can now state the multivariate analogue of Theorem G.2.

Theorem G.4. *Suppose f is a real-valued function of a vector variable and is twice differentiable at the point $\mathbf{x}$. A sufficient condition that $\mathbf{x}$ be a local maximum of f is*

$$\nabla f(\mathbf{x}) = 0 \quad \textit{and} \quad \nabla^2 f(\mathbf{x}) < 0, \tag{G.4}$$

Recall from the discussion just before the theorem that the last part of the condition means $\nabla^2 f(\mathbf{x})$ is a *negative definite* matrix.

Unfortunately, the condition that a matrix is negative definite is impossible to check by hand except in a few special cases. However, it is fairly easy to check by computer. Compute the eigenvalues of the matrix (either R or Mathematica can do this) if all the eigenvalues are positive (resp. nonnegative), then the matrix is positive definite (resp. positive semidefinite), and if all the eigenvalues are negative (resp. nonpositive), then the matrix is negative definite (resp. negative semidefinite).

In fact, the first condition of the theorem isn't very easy to handle either except in very special cases. It's hard to find an $\mathbf{x}$ such that $\nabla f(\mathbf{x})$ holds. Recall this is a *vector* equation so what we are really talking about is solving n equations in n unknowns. Since these are generally *nonlinear* equations, there is no general method of finding a solution. In fact, it is much easier if you don't use first derivative information alone. The way to find the maximum of a function is to have the computer go uphill until it can't make any more progress.

Fortunately R has a function that minimizes functions of several variables (and maximizing f is equivalent to minimizing $-f$). So that can be used to solve all such problems.

Example G.3.1.
Consider minimizing the function[3]

$$f(x_1, x_2) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2$$

Here's how we minimize f using R

```
> f <- function(x) 100 * (x[2] - x[1]^2)^2 + (1 - x[1])^2
> out <- nlm(f, c(0,0), hessian=TRUE)
```

The first line defines an R function `f` of one variable `x` which is in this case a vector with two components `x[1]` and `x[2]`. The `nlm` function minimizes the function `f` using the second argument `c(0,0)` as the starting point for its iterative procedure. The starting point also specifies the dimension of the problem. The `x` that `nlm` will pass to `f` will have the same length as this starting point (in this case length 2). The argument `hessian=TRUE` tells `nlm` that we

[3] This function has no relevance to either probability or statistics. It's just a commonly used test case books about optimization. It's called Rosenbrock's function.

want the second derivative matrix too (Hessian matrix is another name for second derivative matrix)[4]

The result out returned by the minimization procedure is a list of components, only the first four of which are interesting (we omit the rest).

```
> out
$minimum
[1] 4.023726e-12

$estimate
[1] 0.999998 0.999996

$gradient
[1] -7.328278e-07  3.605688e-07

$hessian
          [,1]      [,2]
[1,]  802.2368 -400.0192
[2,] -400.0192  200.0000
```

The component estimate is the point $\mathbf{x}$ at which the function is minimized (or at least at which nlm claims it is minimized), and the component minimum is the value $f(\mathbf{x})$ of the function at that point. The component gradient is the first derivative ("gradient" is another name for a derivative vector). Notice that it is as close to zero as computer arithmetic allows. And the component hessian is, as we said above, the second derivative matrix. To check whether the second derivative matrix is positive definite, calculate eigenvalues

```
> eigen(out$hessian)
$values
[1] 1001.8055799    0.4312236

$vectors
           [,1]       [,2]
[1,] -0.8948213 -0.4464245
[2,]  0.4464245 -0.8948213
```

Since both eigenvalues (the elements of the values) component of the result list returned by the eigen function are positive this is a positive definite matrix, from which we conclude that the point found is a local minimum.

Positive definite? Doesn't the theorem say the second derivative should be *negative definite*? It does. This is the difference between maximization

[4] Unlike Mathematica, R doesn't know any calculus, so it calculates derivatives by finite differences

$$\frac{df(x)}{dx} \approx \frac{f(x+h) - f(x)}{h}$$

for small h, and similarly for partial derivatives. For second derivatives, apply this idea twice.

and minimization. Since maximizing f is equivalent to minimizing $-f$ and $\nabla^2 f(\mathbf{x}) = -\nabla^2(-f(\mathbf{x}))$, the condition is

negative definite Hessian at a local *maximum*,
positive definite Hessian at a local *minimum*.

Unfortunately, in statistics we are often interested in maximization, but most optimization theory and optimization software uses minimization, so we're always having to convert between the two.

G.4 Concave Functions of Several Variables

Definition G.4.1 (Convex Sets).
A subset of $\mathbb{R}^n$ is convex *if for any two points in the set the line segment between them lies entirely in the set.*

Definition G.4.2 (Strictly Concave Functions).
A continuous real-valued function f defined on a convex subset of $\mathbb{R}^n$ with a nonempty interior and twice differentiable at interior points of the subset is strictly concave *if the inequality $\nabla^2 f(\mathbf{x}) < 0$ holds at every interior point $\mathbf{x}$.*

As in the single-variable case, strictly concave functions are very special.

Theorem G.5. *A continuous real-valued strictly concave function f defined on a convex subset of $\mathbb{R}^n$ with a nonempty interior and twice differentiable at interior points of the interval has at most one local maximum. If it has a local maximum, then that point is also the global maximum.*

Also, ∇f has at most one zero. If it has a zero, this is the global maximum.

The theorem says the the situation for strictly concave functions is very simple. If you find a local maximum or a zero of the first derivative, then you have found the unique global maximum.

To summarize, there is a very important distinction between general functions and strictly concave ones. The derivative tests are almost the same, but there is a subtle difference. If

$$\nabla f(\mathbf{x}) = 0 \text{ and } \nabla^2 f(\mathbf{x}) < 0$$

then you know x is a local maximum of f, but you don't know that x is a global maximum or whether there is any global maximum. But if

$$\nabla f(\mathbf{x}) = 0 \text{ and } \nabla^2 f(\mathbf{y}) < 0 \text{ for all } \mathbf{y}$$

then you know that f is strictly concave and hence that $\mathbf{x}$ is the unique global maximum. The only difference is whether you just check negative definiteness the second derivative only at the point $\mathbf{x}$ or at all points $\mathbf{y}$ in the domain of f.

Appendix H

Projections and Chi-Squares

H.1 Orthogonal Projections

A matrix $\mathbf{A}$ is said to be an *orthogonal projection* if it is *symmetric* ($\mathbf{A}' = \mathbf{A}$) and *idempotent* ($\mathbf{A}^2 = \mathbf{A}$). The linear transformation represented by the matrix maps onto the subspace range($\mathbf{A}$). We say that $\mathbf{A}$ is the orthogonal projection onto range($\mathbf{A}$). The *rank* of $\mathbf{A}$, denoted rank($\mathbf{A}$) is the dimension of its range.

A typical element of range($\mathbf{A}$) has the form $\mathbf{y} = \mathbf{A}\mathbf{z}$ for an arbitrary vector $\mathbf{z}$. The idempotence property implies

$$\mathbf{A}\mathbf{y} = \mathbf{y}, \qquad \mathbf{y} \in \text{range}(\mathbf{A}),$$

that is, the linear transformation represented by $\mathbf{A}$ behaves like the identity mapping on range($\mathbf{A}$). Any idempotent matrix (symmetric or not) has this property, and all such matrices are called projections.

The reason why the symmetric projections $\mathbf{A}$ are called *orthogonal* projections is because the vector from $\mathbf{y}$ to its projection $\mathbf{A}\mathbf{y}$ is orthogonal to the subspace range($\mathbf{A}$), which means

$$(\mathbf{y} - \mathbf{A}\mathbf{y})'(\mathbf{A}\mathbf{z}) = \mathbf{y}'(\mathbf{I} - \mathbf{A})'\mathbf{A}\mathbf{z} = \mathbf{y}'(\mathbf{I} - \mathbf{A})\mathbf{A}\mathbf{z} = 0, \qquad \text{for all vectors } \mathbf{y} \text{ and } \mathbf{z},$$

which is equivalent to

$$(\mathbf{I} - \mathbf{A})\mathbf{A} = 0. \tag{H.1}$$

But this is just the same thing as the idempotence property.

Since an orthogonal projection is symmetric, it has a spectral decomposition (E.7). Combining the spectral decomposition with the idempotence property gives

$$\mathbf{A} = \mathbf{O}\mathbf{D}\mathbf{O}' = \mathbf{A}^2 = \mathbf{O}\mathbf{D}\mathbf{O}'\mathbf{O}\mathbf{D}\mathbf{O}' = \mathbf{O}\mathbf{D}^2\mathbf{O}'$$

(because $\mathbf{O}'\mathbf{O} = \mathbf{I}$). Multiplying this by $\mathbf{O}$ on the right and $\mathbf{O}'$ on the left gives $\mathbf{D} = \mathbf{D}^2$ (again using $\mathbf{O}'\mathbf{O} = \mathbf{O}\mathbf{O}' = \mathbf{I}$). Since $\mathbf{D}$ is diagonal, so is $\mathbf{D}^2$, and the diagonal elements of $\mathbf{D}^2$ are just the squares of the corresponding diagonal elements of $\mathbf{D}$. Thus the diagonal elements of $\mathbf{D}$ must be idempotent numbers, satisfying $x^2 = x$, and the only such numbers are zero and one.

Hence we have another characterization of orthogonal projections

> *An orthogonal projection is a symmetric matrix having all eigenvalues either zero or one. Its rank is the number of nonzero eigenvalues.*

(The comment about the rank follows from the fact that $\mathbf{D}$ clearly has this rank, since it maps an arbitrary vector to one having this many nonzero components, and the fact that an orthogonal matrix, being invertible maps one subspace to another of the same dimension.)

We say that a pair of orthogonal projections $\mathbf{A}$ and $\mathbf{B}$ are *orthogonal* to each other if $\mathbf{AB} = 0$. Using the fact that the transpose of a product is the product of the transposes in reverse order, we see that for any symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ (projections or not)

$$(\mathbf{AB})' = \mathbf{B}'\mathbf{A}' = \mathbf{BA} \tag{H.2}$$

Thus for orthogonal projections $\mathbf{AB} = 0$ implies $\mathbf{BA} = 0$ and vice versa.

The terminology here may be a bit confusing, because we are using "orthogonal" to mean two slightly different but closely related things. When applied to one matrix, it means symmetric and idempotent. When applied to two matrices, it means the product is zero. The relationship between the two usages is as follows. If $\mathbf{A}$ is an orthogonal projection, then so is $\mathbf{I} - \mathbf{A}$, because

$$(\mathbf{I} - \mathbf{A})^2 = \mathbf{I}^2 - 2\mathbf{IA} + \mathbf{A}^2 = \mathbf{I} - 2\mathbf{A} + \mathbf{A} = \mathbf{I} - \mathbf{A}.$$

Then (H.1) says these two orthogonal projections (in the first usage) are orthogonal to each other (in the second usage).

We say a set $\{\, \mathbf{A}_i : i = 1, \ldots, k \,\}$ of orthogonal projections is *orthogonal* (that is, it is an *orthogonal* set of *orthogonal* projections) if $\mathbf{A}_i\mathbf{A}_j = 0$, when $i \neq j$.

Another useful fact about orthogonal projections is the following.

Lemma H.1. *If orthogonal projections* $\mathbf{A}$ *and* $\mathbf{B}$ *satisfy*

$$\operatorname{range}(\mathbf{A}) \subset \operatorname{range}(\mathbf{B}), \tag{H.3}$$

then

$$\mathbf{A} = \mathbf{AB} = \mathbf{BA}. \tag{H.4}$$

Proof. $\mathbf{A} = \mathbf{BA}$ follows from the fact that $\mathbf{B}$ behaves like the identity map on range($\mathbf{B}$) which includes range($\mathbf{A}$). But this implies that $\mathbf{BA}$ is a symmetric matrix, hence (H.2) implies the other equality in (H.4). □

H.2 Chi-Squares

Theorem H.2. *Suppose* $\mathbf{X} = (X_1, \ldots, X_n)$ *is a multivariate normal random vector with mean vector zero and variance matrix* $\mathbf{M}$ *that is an orthogonal projection having rank* k*, then*

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} X_i^2 \sim \operatorname{chi}^2(k).$$

Proof. Being an orthogonal projection, the variance matrix has a spectral decomposition $\mathbf{M} = \mathbf{ODO}'$ in which the diagonal elements of $\mathbf{D}$ are k ones and $n - k$ zeros. By reordering the indices, we can arrange the first k to be ones.

Define $\mathbf{Y} = \mathbf{O}'\mathbf{X}$. Then

$$\operatorname{var}(Y) = \mathbf{O}'\mathbf{MO} = \mathbf{D}.$$

Thus the components of $\mathbf{Y}$ are uncorrelated (because $\mathbf{D}$ is diagonal) and hence independent ($\mathbf{Y}$ being a linear transformation of a multivariate normal is multivariate normal, and uncorrelated implies independent for multivariate normal). The first k components are standard normal, and the last $n-k$ are concentrated at zero (because their variance is zero). Thus

$$\mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{k} Y_i \sim \operatorname{chi}^2(k).$$

But

$$\mathbf{Y}'\mathbf{Y} = (\mathbf{O}'\mathbf{X})'(\mathbf{O}'\mathbf{X}) = \mathbf{X}'\mathbf{OO}'\mathbf{X} = \mathbf{X}'\mathbf{X}$$

So $\mathbf{X}'\mathbf{X}$ also has this distribution, which is what the theorem asserts. □

Note that by the definition of the length (norm) of a vector

$$\|\mathbf{X}\|^2 = \mathbf{X}'\mathbf{X}$$

so we sometimes call the random variable described by the theorem $\|\mathbf{X}\|^2$.

Theorem H.3. *Suppose* $\mathbf{Z} = (Z_1, \ldots, Z_n)$ *is a multivariate standard normal random vector (that is, the* Z_i *are i. i. d. standard normal) and* $\{ \mathbf{A}_i : i = 1, \ldots, k \}$ *is an orthogonal set of orthogonal projections, then*

$$\mathbf{Y}_i = \mathbf{A}_i \mathbf{Z}, \qquad i = 1, \ldots, k,$$

are independent random variables, and

$$\mathbf{Y}_i'\mathbf{Y}_i \sim \operatorname{chi}^2(\operatorname{rank}(\mathbf{A}_i)).$$

Proof. First note that the $\mathbf{Y}_i$ are jointly multivariate normal (because they are linear transformations of the same multivariate normal random vector $\mathbf{Z}$). Thus by the corollary to Theorem 13 of Chapter 12 in Lindgren they are independent if uncorrelated. Hence we calculate their covariance matrices

$$\begin{aligned} \operatorname{cov}(\mathbf{Y}_i, \mathbf{Y}_j) &= \operatorname{cov}(\mathbf{A}_i \mathbf{Z}, \mathbf{A}_j \mathbf{Z}) \\ &= E\{\mathbf{A}_i \mathbf{Z} (\mathbf{A}_j \mathbf{Z})'\} \\ &= E\{\mathbf{A}_i \mathbf{Z} \mathbf{Z}' \mathbf{A}_j\} \\ &= \mathbf{A}_i E(\mathbf{Z} \mathbf{Z}') \mathbf{A}_j \\ &= \mathbf{A}_i \operatorname{var}(\mathbf{Z}) \mathbf{A}_j \\ &= \mathbf{A}_i \mathbf{A}_j \end{aligned}$$

and this is zero when $i \neq j$ by assumption. That proves the independence assertion.

The chi-square assertion follows from Theorem H.2, because

$$\operatorname{var}(\mathbf{Y}_i) = \operatorname{cov}(\mathbf{Y}_i, \mathbf{Y}_i) = \mathbf{A}_i \mathbf{A}_i = \mathbf{A}_i$$

because $\mathbf{A}_i$ is idempotent. □