9
Parametric Inference

We now turn our attention to parametric models, that is, models of the form

$$\mathfrak{F} = \left\{ f(x;\theta) : \ \theta \in \Theta \right\} \tag{9.1}$$

where the $\Theta \subset \mathbb{R}^k$ is the parameter space and $\theta = (\theta_1, \ldots, \theta_k)$ is the parameter. The problem of inference then reduces to the problem of estimating the parameter θ.

Students learning statistics often ask: how would we ever know that the distribution that generated the data is in some parametric model? This is an excellent question. Indeed, we would rarely have such knowledge which is why nonparametric methods are preferable. Still, studying methods for parametric models is useful for two reasons. First, there are some cases where background knowledge suggests that a parametric model provides a reasonable approximation. For example, counts of traffic accidents are known from prior experience to follow approximately a Poisson model. Second, the inferential concepts for parametric models provide background for understanding certain nonparametric methods.

We begin with a brief discussion about parameters of interest and nuisance parameters in the next section, then we will discuss two methods for estimating θ, the method of moments and the method of maximum likelihood.

L. Wasserman, *All of Statistics*

9.1 Parameter of Interest

Often, we are only interested in some function $T(\theta)$. For example, if $X \sim N(\mu, \sigma^2)$ then the parameter is $\theta = (\mu, \sigma)$. If our goal is to estimate μ then $\mu = T(\theta)$ is called the **parameter of interest** and σ is called a **nuisance parameter**. The parameter of interest might be a complicated function of θ as in the following example.

9.1 Example. Let $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$. The parameter is $\theta = (\mu, \sigma)$ and the parameter space is $\Theta = \{(\mu, \sigma) : \ \mu \in \mathbb{R}, \ \sigma > 0\}$. Suppose that X_i is the outcome of a blood test and suppose we are interested in τ, the fraction of the population whose test score is larger than 1. Let Z denote a standard Normal random variable. Then

$$\begin{aligned} \tau &= \mathbb{P}(X > 1) = 1 - \mathbb{P}(X < 1) = 1 - \mathbb{P}\left(\frac{X-\mu}{\sigma} < \frac{1-\mu}{\sigma}\right) \\ &= 1 - \mathbb{P}\left(Z < \frac{1-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{1-\mu}{\sigma}\right). \end{aligned}$$

The parameter of interest is $\tau = T(\mu, \sigma) = 1 - \Phi((1-\mu)/\sigma)$. ■

9.2 Example. Recall that X has a Gamma(α, β) distribution if

$$f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta}, \quad x > 0$$

where $\alpha, \beta > 0$ and

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$$

is the Gamma function. The parameter is $\theta = (\alpha, \beta)$. The Gamma distribution is sometimes used to model lifetimes of people, animals, and electronic equipment. Suppose we want to estimate the mean lifetime. Then $T(\alpha, \beta) = \mathbb{E}_\theta(X_1) = \alpha\beta$. ■

9.2 The Method of Moments

The first method for generating parametric estimators that we will study is called the method of moments. We will see that these estimators are not optimal but they are often easy to compute. They are are also useful as starting values for other methods that require iterative numerical routines.

Suppose that the parameter $\theta = (\theta_1, \ldots, \theta_k)$ has k components. For $1 \leq j \leq k$, define the j^{th} **moment**

$$\alpha_j \equiv \alpha_j(\theta) = \mathbb{E}_\theta(X^j) = \int x^j dF_\theta(x) \tag{9.2}$$

and the j^{th} **sample moment**

$$\widehat{\alpha}_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j. \tag{9.3}$$

9.3 Definition. *The* **method of moments estimator** $\widehat{\theta}_n$ *is defined to be the value of* θ *such that*

$$\begin{aligned} \alpha_1(\widehat{\theta}_n) &= \widehat{\alpha}_1 \\ \alpha_2(\widehat{\theta}_n) &= \widehat{\alpha}_2 \\ &\vdots \quad \vdots \quad \vdots \\ \alpha_k(\widehat{\theta}_n) &= \widehat{\alpha}_k. \end{aligned} \tag{9.4}$$

Formula (9.4) defines a system of k equations with k unknowns.

9.4 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Then $\alpha_1 = \mathbb{E}_p(X) = p$ and $\widehat{\alpha}_1 = n^{-1}\sum_{i=1}^n X_i$. By equating these we get the estimator

$$\widehat{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \quad \blacksquare$$

9.5 Example. Let $X_1, \ldots, X_n \sim \text{Normal}(\mu, \sigma^2)$. Then, $\alpha_1 = \mathbb{E}_\theta(X_1) = \mu$ and $\alpha_2 = \mathbb{E}_\theta(X_1^2) = \mathbb{V}_\theta(X_1) + (\mathbb{E}_\theta(X_1))^2 = \sigma^2 + \mu^2$. We need to solve the equations[1]

$$\begin{aligned} \widehat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \widehat{\sigma}^2 + \widehat{\mu}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2. \end{aligned}$$

This is a system of 2 equations with 2 unknowns. The solution is

$$\widehat{\mu} = \overline{X}_n$$

[1] Recall that $\mathbb{V}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$. Hence, $\mathbb{E}(X^2) = \mathbb{V}(X) + (\mathbb{E}(X))^2$.

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2. \quad \blacksquare$$

9.6 Theorem. *Let $\widehat{\theta}_n$ denote the method of moments estimator. Under appropriate conditions on the model, the following statements hold:*

1. *The estimate $\widehat{\theta}_n$ exists with probability tending to 1.*
2. *The estimate is consistent: $\widehat{\theta}_n \xrightarrow{\text{P}} \theta$.*
3. *The estimate is asymptotically Normal:*

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow N(0, \Sigma)$$

where

$$\Sigma = g\mathbb{E}_\theta(YY^T)g^T,$$

$Y = (X, X^2, \ldots, X^k)^T$, $g = (g_1, \ldots, g_k)$ *and* $g_j = \partial\alpha_j^{-1}(\theta)/\partial\theta$.

The last statement in the theorem above can be used to find standard errors and confidence intervals. However, there is an easier way: the bootstrap. We defer discussion of this until the end of the chapter.

9.3 Maximum Likelihood

The most common method for estimating parameters in a parametric model is the **maximum likelihood method**. Let $X_1, \ldots, X_n$ be IID with PDF $f(x;\theta)$.

9.7 Definition. *The* **likelihood function** *is defined by*

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta). \tag{9.5}$$

The **log-likelihood function** *is defined by* $\ell_n(\theta) = \log \mathcal{L}_n(\theta)$.

The likelihood function is just the joint density of the data, except that we **treat it is a function of the parameter θ.** Thus, $\mathcal{L}_n : \Theta \to [0, \infty)$. The likelihood function is not a density function: in general, it is **not** true that $\mathcal{L}_n(\theta)$ integrates to 1 (with respect to θ).

9.8 Definition. *The* **maximum likelihood estimator** MLE, *denoted by* $\widehat{\theta}_n$, *is the value of θ that maximizes $\mathcal{L}_n(\theta)$.*

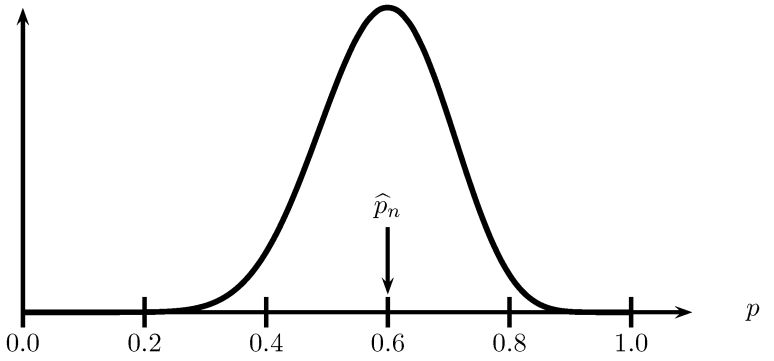

FIGURE 9.1. Likelihood function for Bernoulli with $n = 20$ and $\sum_{i=1}^n X_i = 12$. The MLE is $\widehat{p}_n = 12/20 = 0.6$.

The maximum of $\ell_n(\theta)$ occurs at the same place as the maximum of $\mathcal{L}_n(\theta)$, so maximizing the log-likelihood leads to the same answer as maximizing the likelihood. Often, it is easier to work with the log-likelihood.

9.9 Remark. If we multiply $\mathcal{L}_n(\theta)$ by any positive constant c (not depending on θ) then this will not change the MLE. Hence, we shall often drop constants in the likelihood function.

9.10 Example. Suppose that $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. The probability function is $f(x; p) = p^x(1-p)^{1-x}$ for $x = 0, 1$. The unknown parameter is p. Then,

$$\mathcal{L}_n(p) = \prod_{i=1}^n f(X_i;\, p) = \prod_{i=1}^n p^{X_i}(1-p)^{1-X_i} = p^S(1-p)^{n-S}$$

where $S = \sum_i X_i$. Hence,

$$\ell_n(p) = S \log p + (n - S) \log(1 - p).$$

Take the derivative of $\ell_n(p)$, set it equal to 0 to find that the MLE is $\widehat{p}_n = S/n$. See Figure 9.1. ∎

9.11 Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. The parameter is $\theta = (\mu, \sigma)$ and the likelihood function (ignoring some constants) is:

$$\begin{aligned} \mathcal{L}_n(\mu, \sigma) &= \prod_i \frac{1}{\sigma} \exp\left\{ -\frac{1}{2\sigma^2}(X_i - \mu)^2 \right\} \\ &= \sigma^{-n} \exp\left\{ -\frac{1}{2\sigma^2} \sum_i (X_i - \mu)^2 \right\} \end{aligned}$$

$$= \quad \sigma^{-n} \exp\left\{-\frac{nS^2}{2\sigma^2}\right\} \exp\left\{-\frac{n(\overline{X}-\mu)^2}{2\sigma^2}\right\}$$

where $\overline{X} = n^{-1}\sum_i X_i$ is the sample mean and $S^2 = n^{-1}\sum_i (X_i - \overline{X})^2$. The last equality above follows from the fact that $\sum_i (X_i - \mu)^2 = nS^2 + n(\overline{X}-\mu)^2$ which can be verified by writing $\sum_i (X_i - \mu)^2 = \sum_i (X_i - \overline{X} + \overline{X} - \mu)^2$ and then expanding the square. The log-likelihood is

$$\ell(\mu, \sigma) = -n \log \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\overline{X}-\mu)^2}{2\sigma^2}.$$

Solving the equations

$$\frac{\partial \ell(\mu,\sigma)}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial \ell(\mu,\sigma)}{\partial \sigma} = 0,$$

we conclude that $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma} = S$. It can be verified that these are indeed global maxima of the likelihood. ■

9.12 Example (A Hard Example)**.** Here is an example that many people find confusing. Let $X_1, \ldots, X_n \sim Unif(0, \theta)$. Recall that

$$f(x;\theta) = \begin{cases} 1/\theta & 0 \leq x \leq \theta \\ 0 & \text{otherwise.} \end{cases}$$

Consider a fixed value of θ. Suppose $\theta < X_i$ for some i. Then, $f(X_i;\theta) = 0$ and hence $\mathcal{L}_n(\theta) = \prod_i f(X_i;\theta) = 0$. It follows that $\mathcal{L}_n(\theta) = 0$ if any $X_i > \theta$. Therefore, $\mathcal{L}_n(\theta) = 0$ if $\theta < X_{(n)}$ where $X_{(n)} = \max\{X_1, \ldots, X_n\}$. Now consider any $\theta \geq X_{(n)}$. For every X_i we then have that $f(X_i;\theta) = 1/\theta$ so that $\mathcal{L}_n(\theta) = \prod_i f(X_i;\theta) = \theta^{-n}$. In conclusion,

$$\mathcal{L}_n(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^n & \theta \geq X_{(n)} \\ 0 & \theta < X_{(n)}. \end{cases}$$

See Figure 9.2. Now $\mathcal{L}_n(\theta)$ is strictly decreasing over the interval $[X_{(n)}, \infty)$. Hence, $\widehat{\theta}_n = X_{(n)}$. ■

The maximum likelihood estimators for the multivariate Normal and the multinomial can be found in Theorems 14.5 and 14.3.

9.4 Properties of Maximum Likelihood Estimators

Under certain conditions on the model, the maximum likelihood estimator $\widehat{\theta}_n$ possesses many properties that make it an appealing choice of estimator. The main properties of the MLE are:

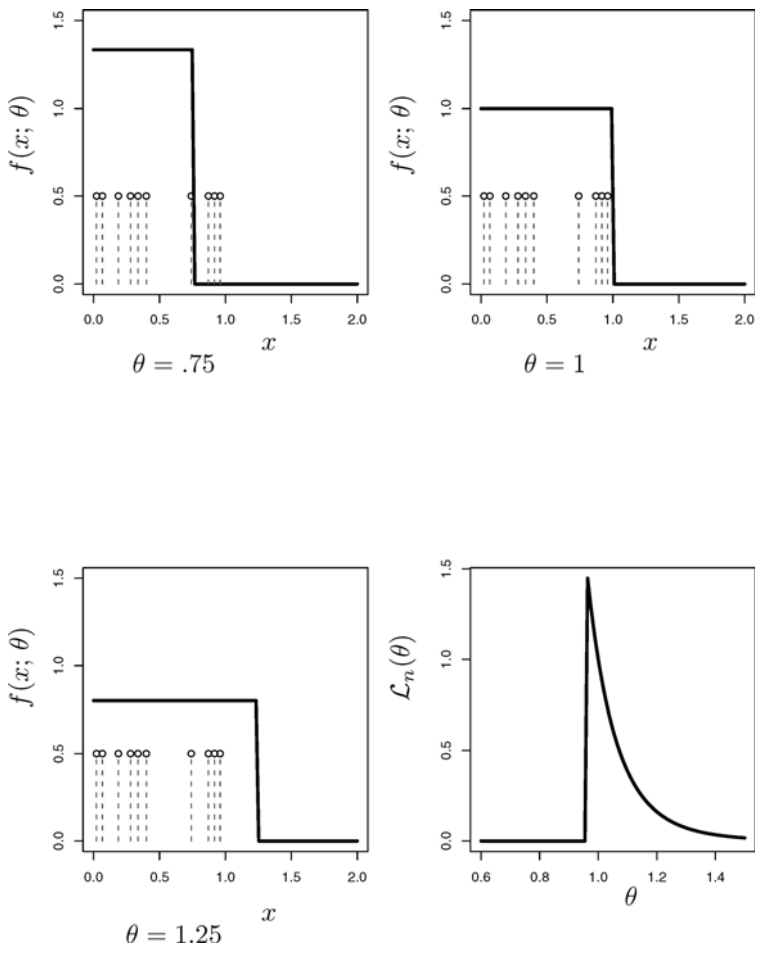

FIGURE 9.2. Likelihood function for Uniform $(0, \theta)$. The vertical lines show the observed data. The first three plots show $f(x;\theta)$ for three different values of θ. When $\theta < X_{(n)} = \max\{X_1, \ldots, X_n\}$, as in the first plot, $f(X_{(n)};\theta) = 0$ and hence $\mathcal{L}_n(\theta) = \prod_{i=1}^n f(X_i;\theta) = 0$. Otherwise $f(X_i;\theta) = 1/\theta$ for each i and hence $\mathcal{L}_n(\theta) = \prod_{i=1}^n f(X_i;\theta) = (1/\theta)^n$. The last plot shows the likelihood function.

1. The MLE is **consistent**: $\widehat{\theta}_n \xrightarrow{\text{P}} \theta_\star$ where $\theta_\star$ denotes the true value of the parameter θ;

2. The MLE is **equivariant**: if $\widehat{\theta}_n$ is the MLE of θ then $g(\widehat{\theta}_n)$ is the MLE of $g(\theta)$;

3. The MLE is **asymptotically Normal**: $(\widehat{\theta} - \theta_\star)/\widehat{\mathsf{se}} \rightsquigarrow N(0,1)$; also, the estimated standard error $\widehat{\mathsf{se}}$ can often be computed analytically;

4. The MLE is **asymptotically optimal** or **efficient**: roughly, this means that among all well-behaved estimators, the MLE has the smallest variance, at least for large samples;

5. The MLE is approximately the Bayes estimator. (This point will be explained later.)

We will spend some time explaining what these properties mean and why they are good things. In sufficiently complicated problems, these properties will no longer hold and the MLE will no longer be a good estimator. For now we focus on the simpler situations where the MLE works well. The properties we discuss only hold if the model satisfies certain **regularity conditions.** These are essentially smoothness conditions on $f(x;\theta)$. **Unless otherwise stated, we shall tacitly assume that these conditions hold.**

9.5 Consistency of Maximum Likelihood Estimators

Consistency means that the MLE converges in probability to the true value. To proceed, we need a definition. If f and g are PDF's, define the **Kullback-Leibler distance** [2] between f and g to be

$$D(f,g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) dx. \tag{9.6}$$

It can be shown that $D(f,g) \geq 0$ and $D(f,f) = 0$. For any $\theta, \psi \in \Theta$ write $D(\theta,\psi)$ to mean $D(f(x;\,\theta), f(x;\,\psi))$.

We will say that the model $\mathfrak{F}$ is **identifiable** if $\theta \neq \psi$ implies that $D(\theta,\psi) > 0$. This means that different values of the parameter correspond to different distributions. We will assume from now on the the model is identifiable.

[2] This is not a distance in the formal sense because $D(f,g)$ is not symmetric.

Let $\theta_\star$ denote the true value of θ. Maximizing $\ell_n(\theta)$ is equivalent to maximizing

$$M_n(\theta) = \frac{1}{n}\sum_i \log \frac{f(X_i;\,\theta)}{f(X_i;\,\theta_\star)}.$$

This follows since $M_n(\theta) = n^{-1}(\ell_n(\theta) - \ell_n(\theta_\star))$ and $\ell_n(\theta_\star)$ is a constant (with respect to θ). By the law of large numbers, $M_n(\theta)$ converges to

$$\begin{aligned}
\mathbb{E}_{\theta_\star}\left(\log \frac{f(X_i;\,\theta)}{f(X_i;\,\theta_\star)}\right) &= \int \log\left(\frac{f(x;\,\theta)}{f(x;\,\theta_\star)}\right) f(x;\,\theta_\star)dx \\
&= -\int \log\left(\frac{f(x;\,\theta_\star)}{f(x;\,\theta)}\right) f(x;\,\theta_\star)dx \\
&= -D(\theta_\star,\theta).
\end{aligned}$$

Hence, $M_n(\theta) \approx -D(\theta_\star,\theta)$ which is maximized at $\theta_\star$ since $-D(\theta_\star,\theta_\star) = 0$ and $-D(\theta_\star,\theta) < 0$ for $\theta \neq \theta_\star$. Therefore, we expect that the maximizer will tend to $\theta_\star$. To prove this formally, we need more than $M_n(\theta) \xrightarrow{\text{P}} -D(\theta_\star,\theta)$. We need this convergence to be uniform over θ. We also have to make sure that the function $D(\theta_\star,\theta)$ is well behaved. Here are the formal details.

9.13 Theorem. *Let $\theta_\star$ denote the true value of θ. Define*

$$M_n(\theta) = \frac{1}{n}\sum_i \log \frac{f(X_i;\,\theta)}{f(X_i;\,\theta_\star)}$$

and $M(\theta) = -D(\theta_\star,\theta)$. Suppose that

$$\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\text{P}} 0 \tag{9.7}$$

and that, for every $\epsilon > 0$,

$$\sup_{\theta:|\theta-\theta_\star|\geq\epsilon} M(\theta) < M(\theta_\star). \tag{9.8}$$

Let $\widehat{\theta}_n$ denote the MLE. *Then $\widehat{\theta}_n \xrightarrow{\text{P}} \theta_\star$.*

The proof is in the appendix.

9.6 Equivariance of the MLE

9.14 Theorem. *Let $\tau = g(\theta)$ be a function of θ. Let $\widehat{\theta}_n$ be the* MLE *of θ. Then $\widehat{\tau}_n = g(\widehat{\theta}_n)$ is the* MLE *of τ.*

PROOF. Let $h = g^{-1}$ denote the inverse of g. Then $\widehat{\theta}_n = h(\widehat{\tau}_n)$. For any τ, $\mathcal{L}(\tau) = \prod_i f(x_i; h(\tau)) = \prod_i f(x_i; \theta) = \mathcal{L}(\theta)$ where $\theta = h(\tau)$. Hence, for any τ, $\mathcal{L}_n(\tau) = \mathcal{L}(\theta) \leq \mathcal{L}(\widehat{\theta}) = \mathcal{L}_n(\widehat{\tau})$. ■

9.15 Example. Let $X_1, \ldots, X_n \sim N(\theta, 1)$. The MLEfor θ is $\widehat{\theta}_n = \overline{X}_n$. Let $\tau = e^\theta$. Then, the MLE for τ is $\widehat{\tau} = e^{\widehat{\theta}} = e^{\overline{X}}$. ■

9.7 Asymptotic Normality

It turns out that the distribution of $\widehat{\theta}_n$ is approximately Normal and we can compute its approximate variance analytically. To explore this, we first need a few definitions.

9.16 Definition. *The* **score function** *is defined to be*

$$s(X; \theta) = \frac{\partial \log f(X; \theta)}{\partial \theta}. \tag{9.9}$$

The **Fisher information** *is defined to be*

$$\begin{aligned} I_n(\theta) &= \mathbb{V}_\theta \left(\sum_{i=1}^n s(X_i; \theta) \right) \\ &= \sum_{i=1}^n \mathbb{V}_\theta \left(s(X_i; \theta) \right). \end{aligned} \tag{9.10}$$

For $n = 1$ we will sometimes write $I(\theta)$ instead of $I_1(\theta)$. It can be shown that $\mathbb{E}_\theta(s(X; \theta)) = 0$. It then follows that $\mathbb{V}_\theta(s(X; \theta)) = \mathbb{E}_\theta(s^2(X; \theta))$. In fact, a further simplification of $I_n(\theta)$ is given in the next result.

9.17 Theorem. $I_n(\theta) = nI(\theta)$. *Also,*

$$\begin{aligned} I(\theta) &= -\mathbb{E}_\theta \left(\frac{\partial^2 \log f(X; \theta)}{\partial \theta^2} \right) \\ &= -\int \left(\frac{\partial^2 \log f(x; \theta)}{\partial \theta^2} \right) f(x;\, \theta) dx. \end{aligned} \tag{9.11}$$

9.18 Theorem (Asymptotic Normality of the MLE). *Let* $\mathsf{se} = \sqrt{\mathbb{V}(\widehat{\theta}_n)}$. *Under appropriate regularity conditions, the following hold:*

1. $\mathsf{se} \approx \sqrt{1/I_n(\theta)}$ *and*

$$\frac{(\widehat{\theta}_n - \theta)}{\mathsf{se}} \rightsquigarrow N(0,1). \tag{9.12}$$

2. *Let* $\widehat{\mathsf{se}} = \sqrt{1/I_n(\widehat{\theta}_n)}$. *Then,*

$$\frac{(\widehat{\theta}_n - \theta)}{\widehat{\mathsf{se}}} \rightsquigarrow N(0,1). \tag{9.13}$$

The proof is in the appendix. The first statement says that $\widehat{\theta}_n \approx N(\theta, \mathsf{se})$ where the approximate standard error of $\widehat{\theta}_n$ is $\mathsf{se} = \sqrt{1/I_n(\theta)}$. The second statement says that this is still true even if we replace the standard error by its estimated standard error $\widehat{\mathsf{se}} = \sqrt{1/I_n(\widehat{\theta}_n)}$.

Informally, the theorem says that the distribution of the MLE can be approximated with $N(\theta, \widehat{\mathsf{se}}^2)$. From this fact we can construct an (asymptotic) confidence interval.

9.19 Theorem. *Let*

$$C_n = \left(\widehat{\theta}_n - z_{\alpha/2}\,\widehat{\mathsf{se}},\ \widehat{\theta}_n + z_{\alpha/2}\,\widehat{\mathsf{se}}\right).$$

Then, $\mathbb{P}_\theta(\theta \in C_n) \to 1 - \alpha$ *as* $n \to \infty$.

PROOF. Let Z denote a standard normal random variable. Then,

$$\begin{aligned}
\mathbb{P}_\theta(\theta \in C_n) &= \mathbb{P}_\theta\left(\widehat{\theta}_n - z_{\alpha/2}\,\widehat{\mathsf{se}} \le \theta \le \widehat{\theta}_n + z_{\alpha/2}\,\widehat{\mathsf{se}}\right) \\
&= \mathbb{P}_\theta\left(-z_{\alpha/2} \le \frac{\widehat{\theta}_n - \theta}{\widehat{\mathsf{se}}} \le z_{\alpha/2}\right) \\
&\to \mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha. \quad \blacksquare
\end{aligned}$$

For $\alpha = .05$, $z_{\alpha/2} = 1.96 \approx 2$, so:

$$\widehat{\theta}_n \pm 2\ \widehat{\mathsf{se}} \tag{9.14}$$

is an approximate 95 percent confidence interval.

When you read an opinion poll in the newspaper, you often see a statement like: the poll is accurate to within one point, 95 percent of the time. They are simply giving a 95 percent confidence interval of the form $\widehat{\theta}_n \pm 2\,\widehat{\mathsf{se}}$.

9.20 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. The MLE is $\widehat{p}_n = \sum_i X_i/n$ and $f(x;p) = p^x(1-p)^{1-x}$, $\log f(x;p) = x \log p + (1-x)\log(1-p)$,

$$s(X;p) = \frac{X}{p} - \frac{1-X}{1-p},$$

and

$$-s'(X;p) = \frac{X}{p^2} + \frac{1-X}{(1-p)^2}.$$

Thus,

$$I(p) = \mathbb{E}_p(-s'(X;p)) = \frac{p}{p^2} + \frac{(1-p)}{(1-p)^2} = \frac{1}{p(1-p)}.$$

Hence,

$$\widehat{\mathsf{se}} = \frac{1}{\sqrt{I_n(\widehat{p}_n)}} = \frac{1}{\sqrt{nI(\widehat{p}_n)}} = \left\{\frac{\widehat{p}(1-\widehat{p})}{n}\right\}^{1/2}.$$

An approximate 95 percent confidence interval is

$$\widehat{p}_n \pm 2\left\{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}\right\}^{1/2}. \quad \blacksquare$$

9.21 Example. Let $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$ where σ^2 is known. The score function is $s(X;\theta) = (X-\theta)/\sigma^2$ and $s'(X;\theta) = -1/\sigma^2$ so that $I_1(\theta) = 1/\sigma^2$. The MLE is $\widehat{\theta}_n = \overline{X}_n$. According to Theorem 9.18, $\overline{X}_n \approx N(\theta, \sigma^2/n)$. In this case, the Normal approximation is actually exact. ■

9.22 Example. Let $X_1, \ldots, X_n \sim \text{Poisson}(\lambda)$. Then $\widehat{\lambda}_n = \overline{X}_n$ and some calculations show that $I_1(\lambda) = 1/\lambda$, so

$$\widehat{\mathsf{se}} = \frac{1}{\sqrt{nI(\widehat{\lambda}_n)}} = \sqrt{\frac{\widehat{\lambda}_n}{n}}.$$

Therefore, an approximate $1-\alpha$ confidence interval for λ is $\widehat{\lambda}_n \pm z_{\alpha/2}\sqrt{\widehat{\lambda}_n/n}$. ■

9.8 Optimality

Suppose that $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$. The MLE is $\widehat{\theta}_n = \overline{X}_n$. Another reasonable estimator of θ is the sample median $\widetilde{\theta}_n$. The MLE satisfies

$$\sqrt{n}(\widehat{\theta}_n - \theta) \rightsquigarrow N(0, \sigma^2).$$

It can be proved that the median satisfies

$$\sqrt{n}(\widetilde{\theta}_n - \theta) \rightsquigarrow N\left(0, \sigma^2 \frac{\pi}{2}\right).$$

This means that the median converges to the right value but has a larger variance than the MLE.

More generally, consider two estimators T_n and U_n and suppose that

$$\sqrt{n}(T_n - \theta) \rightsquigarrow N(0, t^2),$$

and that

$$\sqrt{n}(U_n - \theta) \rightsquigarrow N(0, u^2).$$

We define the asymptotic relative efficiency of U to T by $\text{ARE}(U, T) = t^2/u^2$. In the Normal example, $\text{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n) = 2/\pi = .63$. The interpretation is that if you use the median, you are effectively using only a fraction of the data.

9.23 Theorem. *If $\widehat{\theta}_n$ is the* MLE *and $\widetilde{\theta}_n$ is any other estimator then* [3]

$$\text{ARE}(\widetilde{\theta}_n, \widehat{\theta}_n) \leq 1.$$

Thus, the MLE *has the smallest (asymptotic) variance and we say that the* MLE *is* **efficient** *or* **asymptotically optimal.**

This result is predicated upon the assumed model being correct. If the model is wrong, the MLE may no longer be optimal. We will discuss optimality in more generality when we discuss decision theory in Chapter 12.

9.9 The Delta Method

Let $\tau = g(\theta)$ where g is a smooth function. The maximum likelihood estimator of τ is $\widehat{\tau} = g(\widehat{\theta})$. Now we address the following question: what is the distribution of $\widehat{\tau}$?

9.24 Theorem (The Delta Method). *If $\tau = g(\theta)$ where g is differentiable and $g'(\theta) \neq 0$ then*

$$\frac{(\widehat{\tau}_n - \tau)}{\widehat{\text{se}}(\widehat{\tau})} \rightsquigarrow N(0, 1) \tag{9.15}$$

[3] The result is actually more subtle than this but the details are too complicated to consider here.

where $\widehat{\tau}_n = g(\widehat{\theta}_n)$ *and*

$$\widehat{\mathsf{se}}(\widehat{\tau}_n) = |g'(\widehat{\theta})|\, \widehat{\mathsf{se}}(\widehat{\theta}_n) \tag{9.16}$$

Hence, if

$$C_n = \left(\widehat{\tau}_n - z_{\alpha/2}\, \widehat{\mathsf{se}}(\widehat{\tau}_n),\ \widehat{\tau}_n + z_{\alpha/2}\, \widehat{\mathsf{se}}(\widehat{\tau}_n) \right) \tag{9.17}$$

then $\mathbb{P}_\theta(\tau \in C_n) \to 1 - \alpha$ *as* $n \to \infty$.

9.25 Example. Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ and let $\psi = g(p) = \log(p/(1-p))$. The Fisher information function is $I(p) = 1/(p(1-p))$ so the estimated standard error of the MLE $\widehat{p}_n$ is

$$\widehat{\mathsf{se}} = \sqrt{\frac{\widehat{p}_n(1-\widehat{p}_n)}{n}}.$$

The MLE of ψ is $\widehat{\psi} = \log \widehat{p}/(1-\widehat{p})$. Since, $g'(p) = 1/(p(1-p))$, according to the delta method

$$\widehat{\mathsf{se}}(\widehat{\psi}_n) = |g'(\widehat{p}_n)|\widehat{\mathsf{se}}(\widehat{p}_n) = \frac{1}{\sqrt{n\widehat{p}_n(1-\widehat{p}_n)}}.$$

An approximate 95 percent confidence interval is

$$\widehat{\psi}_n \pm \frac{2}{\sqrt{n\widehat{p}_n(1-\widehat{p}_n)}}. \quad \blacksquare$$

9.26 Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Suppose that μ is known, σ is unknown and that we want to estimate $\psi = \log \sigma$. The log-likelihood is $\ell(\sigma) = -n \log \sigma - \frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2$. Differentiate and set equal to 0 and conclude that

$$\widehat{\sigma}_n = \sqrt{\frac{\sum_i (X_i - \mu)^2}{n}}.$$

To get the standard error we need the Fisher information. First,

$$\log f(X; \sigma) = -\log \sigma - \frac{(X-\mu)^2}{2\sigma^2}$$

with second derivative

$$\frac{1}{\sigma^2} - \frac{3(X-\mu)^2}{\sigma^4},$$

and hence

$$I(\sigma) = -\frac{1}{\sigma^2} + \frac{3\sigma^2}{\sigma^4} = \frac{2}{\sigma^2}.$$

Therefore, $\widehat{\mathsf{se}} = \widehat{\sigma}_n/\sqrt{2n}$. Let $\psi = g(\sigma) = \log\sigma$. Then, $\widehat{\psi}_n = \log\widehat{\sigma}_n$. Since $g' = 1/\sigma$,

$$\widehat{\mathsf{se}}(\widehat{\psi}_n) = \frac{1}{\widehat{\sigma}_n}\frac{\widehat{\sigma}_n}{\sqrt{2n}} = \frac{1}{\sqrt{2n}},$$

and an approximate 95 percent confidence interval is $\widehat{\psi}_n \pm 2/\sqrt{2n}$. ■

9.10 Multiparameter Models

These ideas can directly be extended to models with several parameters. Let $\theta = (\theta_1, \ldots, \theta_k)$ and let $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ be the MLE. Let $\ell_n = \sum_{i=1}^n \log f(X_i; \theta)$,

$$H_{jj} = \frac{\partial^2 \ell_n}{\partial \theta_j^2} \quad \text{and} \quad H_{jk} = \frac{\partial^2 \ell_n}{\partial \theta_j \partial \theta_k}.$$

Define the **Fisher Information Matrix** by

$$I_n(\theta) = -\begin{bmatrix} \mathbb{E}_\theta(H_{11}) & \mathbb{E}_\theta(H_{12}) & \cdots & \mathbb{E}_\theta(H_{1k}) \\ \mathbb{E}_\theta(H_{21}) & \mathbb{E}_\theta(H_{22}) & \cdots & \mathbb{E}_\theta(H_{2k}) \\ \vdots & \vdots & \vdots & \vdots \\ \mathbb{E}_\theta(H_{k1}) & \mathbb{E}_\theta(H_{k2}) & \cdots & \mathbb{E}_\theta(H_{kk}) \end{bmatrix}. \tag{9.18}$$

Let $J_n(\theta) = I_n^{-1}(\theta)$ be the inverse of I_n.

9.27 Theorem. *Under appropriate regularity conditions,*

$$(\widehat{\theta} - \theta) \approx N(0, J_n).$$

Also, if $\widehat{\theta}_j$ is the j^{th} component of $\widehat{\theta}$, then

$$\frac{(\widehat{\theta}_j - \theta_j)}{\widehat{\mathsf{se}}_j} \rightsquigarrow N(0,1) \tag{9.19}$$

where $\widehat{\mathsf{se}}_j^2 = J_n(j,j)$ is the j^{th} diagonal element of J_n. The approximate covariance of $\widehat{\theta}_j$ and $\widehat{\theta}_k$ is $\mathsf{Cov}(\widehat{\theta}_j, \widehat{\theta}_k) \approx J_n(j,k)$.

There is also a multiparameter delta method. Let $\tau = g(\theta_1, \ldots, \theta_k)$ be a function and let

$$\nabla g = \begin{pmatrix} \frac{\partial g}{\partial \theta_1} \\ \vdots \\ \frac{\partial g}{\partial \theta_k} \end{pmatrix}$$

be the gradient of g.

9.28 Theorem (Multiparameter delta method). *Suppose that ∇g evaluated at $\widehat{\theta}$ is not 0. Let $\widehat{\tau} = g(\widehat{\theta})$. Then*

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

where

$$\widehat{\mathsf{se}}(\widehat{\tau}) = \sqrt{(\widehat{\nabla} g)^T \widehat{J}_n (\widehat{\nabla} g)}, \tag{9.20}$$

$\widehat{J}_n = J_n(\widehat{\theta}_n)$ *and* $\widehat{\nabla} g$ *is* ∇g *evaluated at* $\theta = \widehat{\theta}$.

9.29 Example. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Let $\tau = g(\mu, \sigma) = \sigma/\mu$. In Exercise 8 you will show that

$$I_n(\mu, \sigma) = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{bmatrix}.$$

Hence,

$$J_n = I_n^{-1}(\mu, \sigma) = \frac{1}{n} \begin{bmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2} \end{bmatrix}.$$

The gradient of g is

$$\nabla g = \begin{pmatrix} -\frac{\sigma}{\mu^2} \\ \frac{1}{\mu} \end{pmatrix}.$$

Thus,

$$\widehat{\mathsf{se}}(\widehat{\tau}) = \sqrt{(\widehat{\nabla} g)^T \widehat{J}_n (\widehat{\nabla} g)} = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{\widehat{\mu}^4} + \frac{\widehat{\sigma}^2}{2\widehat{\mu}^2}}. \quad \blacksquare$$

9.11 The Parametric Bootstrap

For parametric models, standard errors and confidence intervals may also be estimated using the bootstrap. There is only one change. In the nonparametric bootstrap, we sampled $X_1^*, \ldots, X_n^*$ from the empirical distribution $\widehat{F}_n$. In the parametric bootstrap we sample instead from $f(x; \widehat{\theta}_n)$. Here, $\widehat{\theta}_n$ could be the MLE or the method of moments estimator.

9.30 Example. Consider example 9.29. To get the bootstrap standard error, simulate $X_1, \ldots, X_n^* \sim N(\widehat{\mu}, \widehat{\sigma}^2)$, compute $\widehat{\mu}^* = n^{-1} \sum_i X_i^*$ and $\widehat{\sigma}^{2*} = n^{-1} \sum_i (X_i^* - \widehat{\mu}^*)^2$. Then compute $\widehat{\tau}^* = g(\widehat{\mu}^*, \widehat{\sigma}^*) = \widehat{\sigma}^*/\widehat{\mu}^*$. Repeating this B times yields bootstrap replications

$$\widehat{\tau}_1^*, \ldots, \widehat{\tau}_B^*$$

and the estimated standard error is

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

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

9.12 Checking Assumptions

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

9.13 Appendix

9.13.1 Proofs

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

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

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

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

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

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

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

9.31 Lemma. *The score function satisfies*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Outline of Proof of Theorem 9.24. Write

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

Thus,

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

and hence

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

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

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

or, in other words,

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

where

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

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

9.13.2 Sufficiency

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The partitions induced by these statistics are:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

9.13.3 Exponential Families

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

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

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

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

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

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

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

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

In this case,

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

and

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

■

We can rewrite an exponential family as

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

This is exponential with

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

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

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

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

As before we can write an exponential family as

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

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

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

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

9.13.4 Computing Maximum Likelihood Estimates

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

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

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

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

Solving for $\widehat{\theta}$ gives

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

This suggests the following iterative scheme:

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

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

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

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

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

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

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

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

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

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

Now we describe the EM algorithm.

The EM Algorithm

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

(1) (The E-step): Calculate

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

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

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

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

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

and hence

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

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

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

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

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

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

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

Hence, the complete likelihood is

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

The complete log-likelihood is then

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

And so

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

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

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

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

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

and

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

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

9.14 Exercises

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

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

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

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

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

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

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

   (a) Find the MLE of τ.

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

   (c) Suppose the data are:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

8. Find the Fisher information matrix for Example 9.29.

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

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

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

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

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

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