6 Classic Theory of Point Estimation

Point estimation is usually a starting point for more elaborate inference, such as construction of confidence intervals. Centering a confidence interval at a point estimator which has small variability and small bias generally allows us to construct confidence intervals which are shorter. This makes it necessary to know which point estimators have such properties of low bias and low variability.

The answer to these questions depends on the particular model, and the specific parameter of the model to be estimated. But, general structure and principles that apply simultaneously to a broad class of problems have emerged over the last one hundred years of research on statistical inference. These general principles originate from a relatively small number of concepts, which turn out to be connected. These fundamental concepts are

1. the likelihood function;

2. maximum likelihood estimates;

3. reduction of raw data by restricting to sufficient statistics;

4. defining information;

5. relating statistical information to each of the likelihood function, sufficient statistics, maximum likelihood estimates, and construction of point estimators which are either exactly optimal, or optimal asymptotically.

Many of these concepts and associated mathematical theorems are due to Fisher. Very few statistical ideas have ever emerged that match maximum likelihood estimates in its impact, popularity, and conceptual splendor. Of course, maximum likelihood estimates too can falter, and we will see some of those instances as well.

The classic theory of point estimation revolves around these few central ideas. They are presented with examples and the core theorems in this chapter. General references for this chapter are Bickel and Doksum (2006), Lehmann and Casella (1998), Rao (1973), Stuart and Ord (1991), Cox and Hinkley (1979), and DasGupta (2008). Additional specific references are given within the sections.

6.1 Likelihood Function and MLE: Conceptual Discussion

The concept of a likelihood function was invented by R.A. Fisher. It is an intuitively appealing tool for matching an observed piece of data x to the value of the parameter θ that is most consistent with the particular data value x that you have. Here is a small and simple example to explain the idea.

Example 6.1. (Which Model Generated my Data?) Suppose a binary random variable X has one of two distributions P_0 or P_1, but we do not know which one. We want to let the data help us decide. The distributions P_0, P_1 are as in the small table below.

x	P_0	P_1
5	.1	.8
10	.9	.2

Suppose now in the actual sampling experiment, the data value that we got is $x = 5$. We still do not know if this data value $x = 5$ came from P_0 or P_1. We would argue that if indeed the model generating the data was P_0, then the x value should not have been 5; $x = 5$ does not seem like a good match to the model P_0. On the other hand, if the model generating the data was P_1, then we should fully expect that the data value would be $x = 5$; after all, $P_1(X = 5) = .8$, much higher than $P_0(X = 5)$, which is only .1. So, in matching the observed x to the available models, x matches much better with P_1 than with P_0. If we let a parameter $\theta \in \Theta = \{\theta_0, \theta_1\}$ label the two distributions P_0, P_1, then $\theta = \theta_1$ is a better match to $x = 5$ than $\theta = \theta_0$, because $P_{\theta_1}(X = 5) > P_{\theta_0}(X = 5)$. We would call $\theta = \theta_1$ the *maximum likelihood estimate of* θ. Note that if our sampling experiment had produced the data value $x = 10$, our conclusion would have reversed! Then, our maximum likelihood estimate of θ would have been $\theta = \theta_0$.

So, as you can see, Fisher's thesis was to use $P_\theta(X = x)$ as the yardstick for assessing the credibility of each θ for being the true value of the parameter θ. If $P_\theta(X = x)$ is large at some θ, that θ value is consistent with the data that was obtained; on the other hand, if $P_\theta(X = x)$ is small at some θ, that θ value is inconsistent with the data that was obtained. In the discrete case, Fisher suggested maximizing $P_\theta(X = x)$ over all possible values of θ, and use the maxima as an estimate of θ. This is the celebrated maximum likelihood estimate (MLE). It rests on a really simple idea; ask yourself which model among all the models you are considering is most likely to have produced the data that you have?

Of course, in real problems, the sample size would rarely be $n = 1$, and the observable X need not be discrete. If the observable X is a continuous random variable, then $P_\theta(X = x)$ would be zero for any x and any θ. Therefore, we have to be careful about defining maximum likelihood estimates in general.

Definition 6.1. Suppose given a parameter $\theta, X^{(n)} = (X_1, \cdots, X_n)$ have a joint pdf or joint pmf $f(x_1, \cdots, x_n \,|\, \theta), \theta \in \Theta$. The *likelihood function*, which is a function of θ, is defined as

$$l(\theta) = f(x_1, \cdots, x_n \,|\, \theta).$$

Remark: *It is critical that you understand that the observations $X_1, \cdots, X_n$ need not be iid for a likelihood function to be defined. If they are iid, $X_1, \cdots, X_n \overset{iid}{\sim} f(x|\theta)$, then, the*

likelihood function becomes $l(\theta) = \prod_{i=1}^{n} f(x_i \,|\, \theta)$. *But, in general, we just work with the joint density (or pmf), and the likelihood function would still be defined.*

Remark: In working with likelihood functions, multiplicative factors that are pure constants or involve only the data values $x_1, \cdots, x_n$, *but not* θ, may be ignored. That is, if $l(\theta) = c(x_1, \cdots, x_n) l^*(\theta)$, then we may as well use $l^*(\theta)$ as the likelihood function, because the multiplicative factor $c(x_1, \cdots, x_n)$ does not involve the parameter θ.

Definition 6.2. Suppose given a parameter $\theta, X^{(n)} = (X_1, \cdots, X_n)$ have a joint pdf or joint pmf $f(x_1, \cdots, x_n \,|\, \theta), \theta \in \Theta$. Any value $\hat{\theta} = \hat{\theta}(X_1, \cdots, X_n)$ at which the likelihood function $l(\theta) = f(x_1, \cdots, x_n \,|\, \theta)$ is maximized is called a maximum likelihood estimate (MLE) of θ, provided $\hat{\theta} \in \Theta$, and $l(\hat{\theta}) < \infty$.

Remark: It is important to note that an MLE need not exist, or be unique. By its definition, when an MLE exists, it is necessarily an element of the parameter space Θ; i.e., an MLE must be one of the possible values of θ. In many examples, it exists and is unique for any data set $X^{(n)}$. In fact, this is one great virtue of distributions in the Exponential family; for distributions in the Exponential family, unique MLEs exist for sufficiently large sample sizes. Another important technical matter to remember is that in many standard models, it is more convenient to maximize $L(\theta) = \log l(\theta)$; it simplifies the algebra, without affecting the correctness of the final answer. The final answer is not affected because logarithm is a strictly increasing function on $(0, \infty)$.

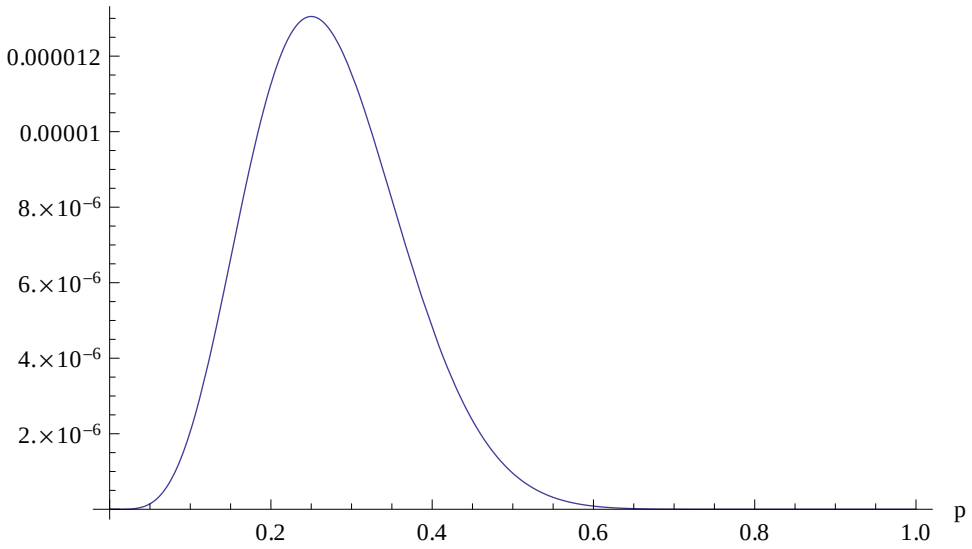
6.1.1 Graphing the Likelihood Function

Plotting the likelihood function is always a good idea. The plot will give you visual information regarding what the observed sample values are saying about the true value of the parameter. It can also be helpful in locating maximum likelihood estimates. The maximum likelihood estimate is mentioned in the examples that follow. But, this is not all that we say about maximum likelihood estimates in this book; in the subsequent sections, maximum likelihood estimates will be treated much more deeply.

Example 6.2. (Likelihood Function in Binomial Case). Let $X_1, \cdots, X_n \overset{iid}{\sim} Ber(p)$, $0 < p < 1$. Then, writing $X = \sum_{i=1}^{n} X_i$ (the total number of successes in these n trials), if the observed value of X is $X = x$, then the likelihood function is

$$l(p) = p^x (1-p)^{(n-x)}, 0 < p < 1.$$

For example, if $n = 20$ and $x = 5$, then $l(p) = p^5 (1-p)^{15}$, and the graph of $l(p)$ is a nice strictly unimodal function of p as in the plot. It is easy to find the point of the global maxima; by elementary calculus, $p^5 (1-p)^{15}$ is maximized at the point $p = \frac{1}{4}$.



Why? Here is a quick verification:

$$\log l(p) = 5 \log p + 15 \log(1-p).$$

The global maxima must either be at a point at which the first derivative of $\log l(p)$ is zero, or it must be attained as a limit as p approaches one of the boundary points $p = 0, 1$. But as $p \to 0, 1, \log l(p) \to -\infty$; so the global maxima cannot be at the boundary points! Hence, it must be at some point within the open interval $0 < p < 1$ where the first derivative $\frac{d}{dp} \log l(p) = 0$.

But,

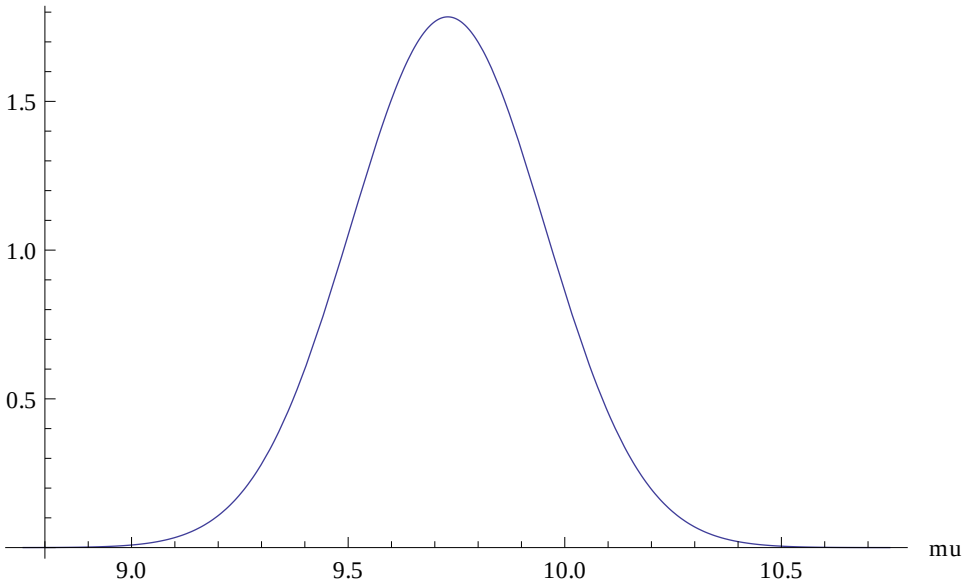
$$\frac{d}{dp} \log l(p) = \frac{5}{p} - \frac{15}{1-p} = \frac{5-20p}{p(1-p)}$$
$$= 0 \Rightarrow p = \frac{5}{20} = \frac{1}{4}.$$

This derivative technique for finding maximum likelihood estimates will often work, but not always! We will discuss this in more generality later.

In this binomial example, the likelihood function is very well behaved; it need not be so well behaved in another problem and we will see examples to that effect.

Example 6.3. (Likelihood Function in Normal Case). Let $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, 1)$, $-\infty < \mu < \infty$. Then, ignoring pure constants (which we can do), the likelihood function is

$$l(\mu) = \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2}} = e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}$$
$$= e^{-\frac{1}{2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\mu-\bar{x})^2\right]}$$



(on using the algebraic identity that for any set of numbers $x_1, \cdots, x_n$ and any fixed real number $a, \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(a - \bar{x})^2$)

$$= e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2} \times e^{-\frac{n}{2}(\mu - \bar{x})^2}.$$

In this expression, the multiplicative term $e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ does not involve μ. So, we may ignore it as well, and take, as our likelihood function the following highly interesting form:

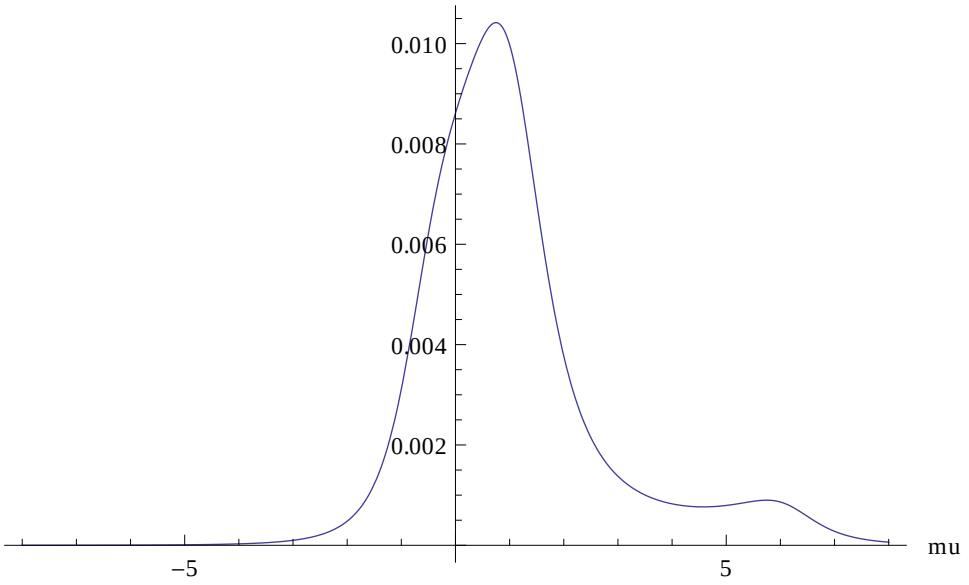
$$l(\mu) = e^{-\frac{n}{2}(\mu - \bar{x})^2}$$

We call it interesting because as a function of μ, this is essentially a *normal density on μ with center at $\bar{x}$ and variance $\frac{1}{n}$. So, if we use the likelihood function as our yardstick for picking the most likely true value of the parameter μ, the likelihood function is going to tell us that the most likely true value of μ is the mean of our sample, $\bar{x}$, a very interesting and neat conclusion, in this specific normal case.*
For the simulated dataset of size $n = 20$ from a normal distribution with mean 10 and variance 1 (thus, in the simulation, the true value of μ is 10):

$$10.78, 9.95, 9.69, 8.39, 10.85, 9.62, 9.61, 9.85, 7.45, 10.37,$$

$$10.81, 10.34, 9.40, 10.78, 9.46, 8.73, 8.17, 10.20, 10.61, 9.43,$$

the sample mean is $\bar{x} = 9.73$, annd the likelihood function is as in our plot below.



Example 6.4. (Likelihood Function May Not Be Unimodal). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} C(\mu, 1)$. Then, ignoring constants, the likelihood function is

$$l(\mu) = \prod_{i=1}^{n} \frac{1}{1 + (\mu - x_i)^2}.$$

No useful further simplification of this form is possible. This example will be revisited; but right now, let us simply mention that this is a famous case where the likelihood function need not be unimodal as a function of μ. Depending on the exact data values, it can have many local maximas and minimas. For the simulated data set $-.61, 1.08, 6.17$ from a standard Cauchy (that is, in the simulation, the true value of μ is zero), the likelihood function is not unimodal, as we can see from its plot below.

6.2 Likelihood Function as a Versatile Tool

The likelihood function is a powerful tool in the statistician's toolbox. The graph of a likelihood function, by itself, is informative. But that is not all. A number of fundamental concepts and theorems are essentially byproducts of the likelihood function; some names are *maximum likelihood estimates, score function, Fisher information, sufficient statistics, and Cramér-Rao inequality.* And, it is remarkable that these are inter-connected. We now introduce these notions and explain their importance one at a time. The mutual connections among them will be revealed as the story unfolds.

6.2.1 Score Function and Likelihood Equation

In a generic problem, let $l(\theta)$ denote the likelihood function and $L(\theta) = \log l(\theta)$; we refer to $L(\theta)$ as *the log likelihood*. In many problems, *but not all*, the point of global maxima of $l(\theta)$, or equivalently the point of global maxima of $L(\theta)$, i.e., a maximum likelihood estimate of θ, can be found by solving the derivative equation $\frac{d}{d\theta}L(\theta) = 0$. If the likelihood function is differentiable, does have a finite global maximum, and if this global maximum is not a boundary point of the parameter space Θ, then calculus tells us that it must be at a point in the *interior* of Θ where the first derivative is zero. If there is exactly one point where the first derivative is zero, then you have found your global maximum. But, in general, there can be multiple points where $\frac{d}{d\theta}L(\theta) = 0$, and then one must do additional work to identify which one among them is the global maximum.
In any case,

$$\frac{d}{d\theta}\log l(\theta) = 0$$

is a good starting point for locating a maximum likelihood estimate, so much so that we have given names to $\frac{d}{d\theta}\log l(\theta)$ and the equation $\frac{d}{d\theta}\log l(\theta) = 0$.

Definition 6.3. Let $l(\theta) = f(x_1, \cdots, x_n \,|\, \theta), \theta \in \Theta \subseteq \mathcal{R}$ denote the likelihood function. Suppose $l(\theta)$ is differentiable on the interior of Θ and strictly positive there. Then, the function

$$U(\theta) = U(\theta, x_1, \cdots, x_n) = \frac{d}{d\theta}\log l(\theta)$$

is called the *score function* for the model f. If the parameter θ is a vector parameter, $\theta \in \Theta \subseteq \mathcal{R}^p$, the score function is defined as

$$U(\theta) = U(\theta, x_1, \cdots, x_n) = \nabla \log l(\theta),$$

where ∇ denotes the gradient vector with respect to θ:

$$\nabla g(\theta) = (\frac{\partial}{\partial\theta_1}g, \cdots, \frac{\partial}{\partial\theta_p}g)'.$$

Definition 6.4. The equation $U(\theta) = \frac{d}{d\theta}\log l(\theta) = 0$ in the scalar parameter case, and the equation $U(\theta) = \nabla \log l(\theta) = 0$ in the vector parameter case is called the *likelihood equation*.
Let us see a few illustrative examples.

Example 6.5. (Score Function in Binomial Case). From Example 7.2, in the binomial case, the score function is

$$U(p) = \frac{d}{dp}\log l(p) = \frac{d}{dp}\left[p^x(1-p)^{n-x}\right]$$

$$= \frac{d}{dp}\Big[x \log p + (n-x)\log(1-p)\Big] = \frac{x}{p} - \frac{n-x}{1-p} = \frac{x-np}{p(1-p)}.$$

As a result, the likelihood equation is

$$x - np = 0,$$

which has a unique root $p = \frac{x}{n}$ inside the interval $(0,1)$ *unless* $x = 0$ *or* n. Indeed, if $x = 0$, directly $l(p) = (1-p)^n$, which is a strictly decreasing function of p on $[0,1]$ and the maximum is at the boundary point $p = 0$. Likewise, if $x = n$, directly $l(p) = p^n$, which is a strictly increasing function of p on $[0,1]$ and the maximum is at the boundary point $p = 1$.

So, putting it all together, if $0 < x < n$, $l(p)$ has its global maximum at the unique root of the likelihood equation , $p = \frac{x}{n}$, while if $x = 0, n$, *the likelihood equation has no roots*, and $l(p)$ does not have a global maximum within the open interval $0 < p < 1$, i.e., *a maximum likelihood estimate of* p *does not exist if* $x = 0, n$.

Example 6.6. (Score Function in Normal Case). From Example 7.3, in the normal case, the score function is

$$U(\mu) = \frac{d}{d\mu}\log l(\mu) = \frac{d}{d\mu}\Big[-\frac{n}{2}(\mu - \bar{x})^2\Big]$$

$$= -\frac{n}{2}\,2(\mu - \bar{x}) = n(\bar{x} - \mu).$$

If μ varies in the entire real line $(-\infty, \infty)$, the likelihood equation

$$\bar{x} - \mu = 0$$

always has a unique root, namely $\mu = \bar{x}$, and this indeed is the unique maximum likelihood estimate of μ. We will see this result as a special case of a general story within the entire Exponential family; this normal distribution result is a basic and important result in classic inference.

6.2.2 Likelihood Equation and MLE in Exponential Family

For ease of understanding, we present in detail only the case of one parameter Exponential family in the canonical form. The multiparameter case is completely analogous, and is stated separately, so as not to confuse the main ideas with the more complex notation of the multiparameter case.

So, suppose that $X_1, \cdots, X_n \overset{iid}{\sim} f(x|\eta) = e^{\eta T(x) - \psi(\eta)} h(x), \eta \in \mathcal{T}$; see Chapter 5 for notation again. We assume below that $\mathcal{T}$ is an open set, so that the interior of $\mathcal{T}$ equals $\mathcal{T}$.

Then, ignoring the multiplicative factor $\prod_{i=1}^n h(x_i)$,

$$l(\eta) = e^{\eta \sum_{i=1}^n T(x_i) - n\psi(\eta)}$$

$$\Rightarrow \log l(\eta) = \eta \sum_{i=1}^{n} T(x_i) - n\psi(\eta)$$

$$\Rightarrow U(\eta) = \frac{d}{d\eta} \log l(\eta) = \sum_{i=1}^{n} T(x_i) - n\psi'(\eta),$$

which, by Corollary 5.1, is a strictly decreasing function of η if the family $f(x|\eta)$ is a nonsingular Exponential family. Therefore, in the nonsingular case, the first derivative of the log likelihood function is strictly decreasing and hence the log likelihood function itself is strictly concave. If the likelihood equation

$$U(\eta) = \sum_{i=1}^{n} T(x_i) - n\psi'(\eta) = 0$$

$$\Leftrightarrow \psi'(\eta) = \frac{1}{n} \sum_{i=1}^{n} T(x_i)$$

has a root, it is the only root of that equation, and by the strict concavity property, that unique root is the unique global maxima of $\log l(\eta)$, and hence also the unique global maxima of $l(\eta)$.

We already know from Example 7.5 that we cannot guarantee that the likelihood equation always has a root. What we can guarantee is the following result; note that it is an all at one time nonsingular Exponential family result.

Theorem 6.1. Let $X_1, X_2, \cdots$ be iid observations from $f(x|\eta)$, a nonsingular canonical one parameter Exponential family, with $\eta \in \mathcal{T}$, an open set in the real line. Then,

(a) For all large n, there is a unique root $\hat{\eta}$ of the likelihood equation

$$\psi'(\eta) = \frac{1}{n} \sum_{i=1}^{n} T(x_i),$$

within the parameter space $\mathcal{T}$, and $\hat{\eta}$ is the unique MLE of η.

(b) For any one-to-one function $g(\eta)$ of η, an unique MLE is $g(\hat{\eta})$.

(c) In particular, the unique MLE of $\psi'(\eta) = E_\eta[T(X_1)]$ is the empirical mean $\bar{T} = \frac{1}{n}\sum_{i=1}^{n} T(X_i)$. In notation, $\hat{E(T)} = \bar{T}$.

We have *almost* proved this theorem. Part (b) of this theorem is known as *invariance of the MLE*. It is in fact always true, not just in the Exponential family. Part (c) follows from part (b), because $\psi'(\eta)$ is a strictly increasing function of η (refer to Corollary 5.1 once again). A complete proof of part (a) involves use of the SLLN (strong law of large numbers), and we choose to not give it here. But since it is a general Exponential family result, a number of interesting conclusions all follow from this theorem. We summarize them below.

Example 6.7. (MLEs in Standard One Parameter Distributions). All of the results in this example follow from Theorem 7.1, by recognizing the corresponding distribution as a one parameter Exponential family, and by identifying in each case the statistic $T(X)$, the natural parameter η, and the function $\psi'(\eta) = E_\eta[T(X)]$. For purposes of recollection, you may want to refer back to Chapter 5 for these facts, which were all derived there.
1) $N(\mu, \sigma^2)$**,** σ^2 **known** In this case, $T(X) = X, \eta = \frac{\mu}{\sigma^2}, \psi(\eta) = \frac{\eta^2\sigma^2}{2}, \psi'(\eta) = \eta\sigma^2 = \mu$, and so, $\bar{T} = \bar{X}$ is the MLE of μ, whatever be σ^2.

2) **Ber**(p) In this case, $T(X) = X, \eta = \log\frac{p}{1-p}, \psi(\eta) = \log(1 + e^\eta), \psi'(\eta) = \frac{e^\eta}{1+e^\eta} = p$, and so, $\bar{T} = \bar{X}$ is the MLE of p. In more familiar terms, if we have n iid observations from a Bernoulli distribution with parameter p, and if X denotes their sum, i.e., the total number of successes, then $\frac{X}{n}$ is the MLE of p, provided $0 < X < n$; no MLE exists in the two boundary cases $X = 0, n$.

3) **Poi**(λ) In this case, $T(X) = X, \eta = \log\lambda, \psi(\eta) = e^\eta, \psi'(\eta) = e^\eta = \lambda$, and so, $\bar{T} = \bar{X}$ is the MLE of λ, provided $\bar{X} > 0$ (which is the same as saying at least one X_i value is greater than zero); no MLE exists if $\bar{X} = 0$.

4) **Exp**(λ) In this case, $T(X) = -X, \eta = \frac{1}{\lambda}, \psi(\eta) = -\log\eta, \psi'(\eta) = -\frac{1}{\eta} = -\lambda$, and so, $\bar{T} = -\bar{X}$ is the MLE of $-\lambda$, which implies that $\bar{X}$ is the MLE of λ.
A few other standard one parameter examples are assigned as chapter exercises. The multiparameter Exponential family case is stated in our next theorem.

Theorem 6.2. Let $X_1, \cdots, X_n$ be iid observations from the density

$$f(x\,|\eta) = e^{\sum_{i=1}^k \eta_i T(x_i) - \psi(\eta)} h(x),$$

where $\eta = (\eta_1, \cdots, \eta_k) \in \mathcal{T} \subseteq \mathcal{R}^k$. Assume that the family is nonsingular and regular. Then,
(a) For all large n, there is a unique root $\hat{\eta}$ of the system of likelihood equations

$$\frac{\partial}{\partial\eta_j}\psi(\eta) = \bar{T}_j = \frac{1}{n}\sum_{i=1}^n T_j(x_i), j = 1, 2, \cdots, k.$$

(b) $\hat{\eta}$ lies within the parameter space $\mathcal{T}$ and is the unique MLE of η.
(c) For any one-to-one function $g(\eta)$ of η, the unique MLE is $g(\hat{\eta})$.
(d) In particular, the unique MLE of $(E_\eta(T_1), E_\eta(T_2), \cdots, E_\eta(T_k))$ is $(\bar{T}_1, \bar{T}_2, \cdots, \bar{T}_k)$.
Caution The likelihood equation may not always have a root within the parameter space $\mathcal{T}$. This is often a problem for small n; in such a case, an MLE of η does not exist. But, usually, the likelihood equation will have a unique root within $\mathcal{T}$, and the Exponential family structure then guarantees you that you are done; that root is your MLE of η.

6.2.3 MLEs Outside the Exponential Family

We have actually already seen an example of a likelihood function for a distribution outside the Exponential family. Example 7.4 with a sample of size $n = 3$ from a location parameter Cauchy density is such an example. We will now revisit the issue of the MLE in that Cauchy example for a general n, and also consider some other illuminating cases outside the Exponential family.

Example 6.8. (MLE of Cauchy Location Parameter). As in Example 7.4, the likelihood function for a general n is

$$l(\mu) = \prod_{i=1}^{n} \frac{1}{1 + (\mu - x_i)^2}$$

$$\Rightarrow \log l(\mu) = -\sum_{i=1}^{n} \log(1 + (\mu - x_i)^2).$$

It is easily seen that $l(\mu) \to 0$ as $\mu \to \pm\infty$; it is also obvious that $l(\mu)$ iis uniformly bounded as a function of μ (can you prove that?). Therefore, $l(\mu)$ has a finite global maximum, and that global maximum is attained somewhere inside the open interval $(-\infty, \infty)$. Hence, any such global maxima is necessarily a point μ at which the likelihood equation holds. What is the likelihood equation in this case? By simple differentiation,

$$\frac{d}{d\mu} \log l(\mu) = -\frac{d}{d\mu} \left[\sum_{i=1}^{n} \log(1 + (\mu - x_i)^2) \right]$$

$$= -2 \sum_{i=1}^{n} \frac{\mu - x_i}{1 + (\mu - x_i)^2}.$$

Each term $\frac{\mu - x_i}{1+(\mu-x_i)^2}$ is a quotient of a polynomial (in μ) of degree one and another polynomial (in μ) of degree two. On summing n such terms, the final answer will be of the following form:

$$\frac{d}{d\mu} \log l(\mu) = \frac{-2P_n(\mu)}{Q_n(\mu)},$$

where $P_n(\mu)$ is a polynomial in μ of degree $2n - 1$, and $Q_n(\mu)$ is a strictly positive polynomial of degree $2n$. The exact description of the coefficients of the polynomial P_n is difficult. Nevertheless, we can now say that

$$\frac{d}{d\mu} \log l(\mu) = 0 \Leftrightarrow P_n(\mu) = 0.$$

Being a polynomial of degree $2n - 1$, $P_n(\mu)$ *can have* up to $2n - 1$ real roots in $(-\infty, \infty)$ (there may be complex roots, depending on the data values $x_1, \cdots, x_n$). An MLE would

be one of the real roots. No further useful description of the MLE is possible for a given general n. It may indeed be shown that a unique MLE always exists; and obviously, it will be exactly one of the real roots of the polynomial $P_n(\mu)$. There is no formula for this unique MLE for general n, and you must be very very careful in picking the right real root that is the MLE. Otherwise, you may end up with a local maxima, *or in the worst case, a local minima!*. This example shows that in an innocuous one parameter problem, computing the MLE can be quite an art. It was proved in Reeds (1985) that for large n, the average number of distinct real roots of $P_n(\mu)$ is about $1+\frac{2}{\pi}$. Because of the extremely heavy tails of Cauchy distributions, sometimes the distribution is suitably truncated when used in a practical problem; Dahiya, Staneski and Chaganty (2001) consider maximum likelihood estimation in truncated Cauchy distributions.

Example 6.9. (MLE of Uniform Scale Parameter). Let $X_1, \cdots, X_n \overset{iid}{\sim} U[0, \theta], \theta > 0$. Then, the likelihood function is

$$l(\theta) = \prod_{i=1}^{n} f(x_i \,|\theta) = \prod_{i=1}^{n} [\frac{1}{\theta} I_{x_i \geq 0, x_i \leq \theta}$$

$$= \frac{1}{\theta^n} \prod_{i=1}^{n} I_{x_i \geq 0, \theta \geq x_i} = \frac{1}{\theta^n} I_{\theta \geq x_{(n)}} I_{x_{(1)} \geq 0},$$

where, as usual, $x_{(1)}$ and $x_{(n)}$ denote the sample minimum and sample maximum respectively.

With probability one under any θ, the sample minimum is positive, and so, $I_{x_{(1)} \geq 0} = 1$ with probability one. Thus, the likelihood function takes the form

$$\begin{aligned} l(\theta) &= 0, \text{ if } \quad \theta < x_{(n)}, \\ &= \tfrac{1}{\theta^n}, \text{ if } \quad \theta \geq x_{(n)}. \end{aligned}$$

Since $\frac{1}{\theta^n}$ is a strictly decreasing function of θ on $[x_{(n)}, \infty)$, we directly have the conclusion that $l(\theta)$ is maximized on $\Theta = (0, \infty)$ at the point $\theta = x_{(n)}$, the sample maximum. This is the unique MLE of θ, and *the likelihood function does not have a zero derivative at* $x_{(n)}$. In fact, at the point $\theta = x_{(n)}$, the likelihood function has a jump discontinuity, and is *NOT* differentiable at that point. *This simple example shows that MLEs need not be roots of a likelihood equation.*

Example 6.10. (Infinitely Many MLEs May Exist). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} U[\mu - \frac{1}{2}, \mu + \frac{1}{2}]$. Then the likelihood function is

$$l(\mu) = I_{\mu - \frac{1}{2} \leq X_{(1)} \leq X_{(n)} \leq \mu + \frac{1}{2}} = I_{X_{(n)} - \frac{1}{2} \leq \mu \leq X_{(1)} + \frac{1}{2}}.$$

Thus, the likelihood function is completely flat, namely equal to the constant value 1 throughout the interval $[X_{(n)} - \frac{1}{2}, X_{(1)} + \frac{1}{2}]$. Hence, *any statistic* $T(X_1, \cdots, X_n)$ *that*

always lies between $X_{(n)} - \frac{1}{2}$ *and* $X_{(1)} + \frac{1}{2}$ *maximizes the likelihood function and is an MLE.* There are, therefore, infinitely many MLEs for this model, e.g.,

$$\frac{X_{(n)} + X_{(1)}}{2};\ .1X_{(n)} + .9X_{(1)} + .4;\ e^{-\bar{X}^2}\left[X_{(n)} - \frac{1}{2}\right] + (1 - e^{-\bar{X}^2})\left[X_{(1)} + \frac{1}{2}\right].$$

Example 6.11. (No MLEs May Exist). This famous example in which an MLE does not exist was given by Jack Kiefer. The example was very cleverly constructed to make the likelihood function unbounded; if the likelihood function can take arbitrarily large values, then there is obviously no finite global maximum, and so there cannot be any MLEs.
Suppose $X_1, \cdots, X_n$ are iid with the density

$$p\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2}} + (1-p)\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2\sigma^2}(x-\mu)^2},$$

where $0 < p < 1$ is known, and $-\infty < \mu < \infty, \sigma > 0$ are considered unknown. Notice that the parameter is two dimensional, $\theta = (\mu, \sigma)$.
For notational ease, consider the case $n = 2$; the same argument will work for any n. The likelihood function is

$$l(\mu, \sigma) = \prod_{i=1}^{2}\left[pe^{-\frac{(x_i-\mu)^2}{2}} + (1-p)\frac{1}{\sigma}e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}\right]$$

If we now take the particular $\theta = (x_1, \sigma)$ (i.e., look at parameter values where μ is taken to be x_1), then, we get,

$$l(x_1, \sigma) = \left[p + \frac{1-p}{\sigma}\right]\left[pe^{-\frac{(x_2-x_1)^2}{2}} + \frac{1-p}{\sigma}e^{-\frac{1}{2\sigma^2}(x_2-x_1)^2}\right]$$

$$\geq p\frac{1-p}{\sigma}e^{-\frac{(x_2-x_1)^2}{2}} \to \infty,$$

as $\sigma \to 0$.
This shows that the likelihood function is unbounded, and hence no MLEs of θ exist. *If we assume that* $\sigma \geq \sigma_0 > 0$*, the problem disappears and an MLE does exist.*

6.2.4 Fisher Information: The Concept

We had remarked before that the graph of a likelihood function gives us information about the parameter θ. In particular, a very spiky likelihood function which falls off rapidly from its peak value at a unique MLE is very informative; it succeeds in efficiently discriminating between the promising values of θ and the unpromising values of θ. On the other hand, a flat likelihood function is uninformative; a lot of values of θ would look almost equally promising if the likelihood function is flat. Assuming that the likelihood function $l(\theta)$ is differentiable, a flat likelihood function should produce a score function $U(\theta) = \frac{d}{d\theta}\log l(\theta)$ of

small magnitude. For example, if $l(\theta)$ was completely flat, i.e., a constant, then $\frac{d}{d\theta}\log l(\theta)$ would be zero! For all distributions in the Exponential family, and in fact more generally, the expectation of $U(\theta)$ is zero. So, a reasonable measure of the magnitude of $U(\theta)$ would be its second moment, $E_\theta[U^2(\theta)]$, which is also $\mathrm{Var}_\theta[U(\theta)]$, if the expectation of $U(\theta)$ is zero.

Fisher proposed that information about θ be defined as $\mathrm{Var}_\theta[U(\theta)]$. This will depend on n and θ, *but not the actual data values*. So, as a measure of information, it is measuring average information, rather than information obtainable *for your specific data set*. There are concepts of such conditional information for the obtained data.

Fisher's clairvoyance was triumphantly observed later, when it turned out that his definition of information function, namely the Fisher information function, shows up as a critical mathematical entity in major theorems of statistics that were simply not there at Fisher's time. *However, the concept of statistical information is elusive and slippery. It is very very difficult to define information in a way that passes every test of intuition. Fisher information does have certain difficulties when subjected to a comprehensive scrutiny. The most delightful and intellectually deep exposition of this is Basu (1975).*

6.2.5 Calculating Fisher Information

For defining the Fisher information function, some conditions on the underlying density (or pmf) must be made. Distributions which do satisfy these conditions are often referred to as *regular*. Any distribution in the Exponential family is regular, according to this convention. Here are those *regularity conditions*. For brevity, we present the density case only; in the pmf case, the integrals are replaced by sums, the only change required.

Regularity Conditions A

$A1$ **The support of** $f(x|\theta)$ **does not depend on** θ**, i.e., the set** $S = \{x : f(x|\theta) > 0\}$ **is the same set for all** θ**.**

$A2$ **For any** x**, the density** $f(x\,|\theta)$ **is differentiable as a function of** θ**.**

$A3$ $\int[\frac{\partial}{\partial\theta}f(x\,|\theta)]dx = 0$ **for all** θ**.**

Remark: Condition $A3$ holds whenever we can interchange a derivative and an integral in the following sense:

$$\frac{d}{d\theta}\int f(x\,|\theta)dx = \int[\frac{\partial}{\partial\theta}f(x\,|\theta)]dx.$$

This is because, if we can do this interchange of the order of differentiation and integration, then, because *any density function always integrates to* 1 *on the whole sample space*, we will have:

$$\int[\frac{\partial}{\partial\theta}f(x\,|\theta)]dx = \frac{d}{d\theta}\int f(x\,|\theta)dx = \frac{d}{d\theta}[1] = 0,$$

which is what $A3$ says.

Suppose then $f(x\,|\theta)$ satisfies Regularity Conditions A, and that $X_1, \cdots, X_n \overset{iid}{\sim} f(x\,|\theta)$. As usual, let $U(\theta)$ be the score function $U(\theta) = \frac{d}{d\theta} \log l(\theta)$. Let $I_n(\theta) = \text{Var}_\theta[U(\theta)]$, assuming that $I_n(\theta) < \infty$. So, $I_n(\theta)$ measures information if your sample size is n. Intuitively, we feel that information should be increasing in the sample size; after all, more data should purchase us more information. We have the following interesting result in the iid case.

(Proposition). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} f(x\,|\theta)$. If Regularity Conditions A hold, then

$$I_n(\theta) = nI_1(\theta),$$

for all θ. That is, information grows linearly with n.

The proof is simple and is left as an exercise. It follows by simply using the familiar fact that variance is additive if we sum independent random variables.

Because of this above proposition, we take the case $n = 1$ as the base, and define information as $I_1(\theta)$. It is a convention to take the case $n = 1$ as the base. We also drop the subscript and just call it $I(\theta)$. Here then is the definition of *Fisher information function in the regular case.*

Definition 6.5. Let $f(x\,|\theta)$ satisfy Regularity Conditions A. Then the Fisher information function corresponding to the model f is defined as

$$I(\theta) = E_\theta[(\frac{d}{d\theta} \log f(X|\theta))^2] = \text{Var}_\theta[\frac{d}{d\theta} \log f(X|\theta)].$$

If we have *an extra regularity condition* that we state below, then $I(\theta)$ also has another equivalent formula,

$$I(\theta) = -E_\theta[\frac{d^2}{d\theta^2} \log f(X|\theta)].$$

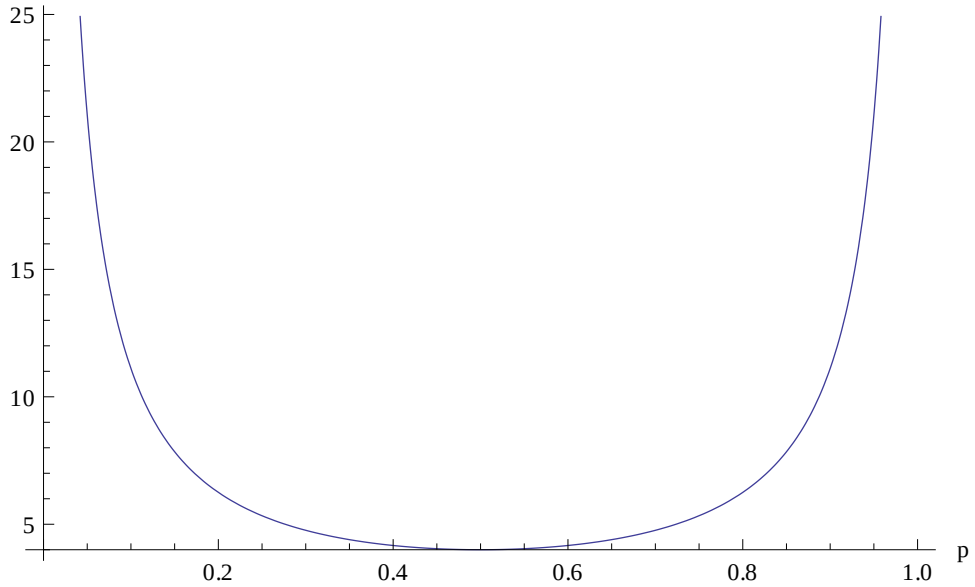
The extra regularity condition we need for this alternative formula of $I(\theta)$ to be valid is this:

Regularity Condition $A4$ For all θ,

$$\int \frac{\partial^2}{\partial\theta^2} f(x|\theta)dx = \frac{d^2}{d\theta^2} \int f(x|\theta)dx = 0.$$

Important Fact Regularity conditions $A1, A2, A3, A4$ all hold for any f (density or pmf) in the one parameter regular Exponential family. They do not hold for densities such as $U[0, \theta], U[\theta, 2\theta], U[\mu - a, \mu + a]$, or a shifted exponential with density $f(x|\mu) = e^{-(x-\mu)}, x \geq \mu$. *Fisher information is never calculated for such densities, because assumption A1 already fails.*

Let us see a few illustrative examples of the calculation of $I(\theta)$.



Example 6.12. (Fisher Information in Bernoulli Case). In the Bernoulli case, $f(x|p) = p^x(1-p)^{1-x}, x = 0, 1$. In other words, Bernoulli is the same as binomial with $n = 1$. Therefore, the score function, from Example 7.5, is $U(p) = \frac{X-p}{p(1-p)}$. Therefore,

$$I(p) = \text{Var}_p[U(p)] = \text{Var}_p\left[\frac{X-p}{p(1-p)}\right] = \frac{\text{Var}_p[X-p]}{[p(1-p)]^2} = \frac{\text{Var}_p[X]}{[p(1-p)]^2}$$

$$= \frac{p(1-p)}{[p(1-p)]^2} = \frac{1}{p(1-p)}.$$

Interestingly, information about the parameter is the smallest if the true value of $p = \frac{1}{2}$; see the plot of $I(p)$.

Example 6.13. (Fisher Information for Normal Mean). In the $N(\mu, \sigma^2)$ case, where σ^2 is considered known, ignoring pure constants,

$$f(x\,|\mu) = e^{-(\mu-x)^2/(2\sigma^2)} \Rightarrow \log f(x\,|\mu) = -\frac{(\mu-x)^2}{2\sigma^2}$$

$$\Rightarrow U(\mu) = \frac{d}{d\mu}\log f(x\,|\mu) = \frac{x-\mu}{\sigma^2}.$$

Therefore,

$$I(\mu) = \text{Var}_\mu\left[\frac{X-\mu}{\sigma^2}\right] = \frac{\text{Var}_\mu[X-\mu]}{\sigma^4} = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}.$$

Notice that $I(\mu)$ is a constant function of μ, and is inversely proportional to the variance of the observation. This is true of general location-parameter densities $f_0(x-\mu)$. In location-parameter families, the Fisher information function, if it is finite, is a constant

function in the location parameter μ, and if we throw in a *known scale parameter* σ, then $I(\mu)$ would be of the form $\frac{c}{\sigma^2}$ for some positive constant c; c will depend only on the formula for the density f_0. For example, in the normal case, $c = 1$. Here is a general theorem that describes this fact about Fisher informations.

Theorem 6.3. (Fisher Information for General Location Parameters). Suppose $f_0(z)$ is a density function on the real line that satisfies the following conditions:
(a) $f_0(z)$ is continuously differentiable at all but a countable number of values of z;
(b) $c(f_0) = \int_{-\infty}^{\infty} \frac{[f_0'(z)]^2}{f_0(z)} dz < \infty$.
Let X be a continuous real valued random variable with the location parameter density $f(x|\mu) = f_0(x - \mu)$. Then, the Fisher information function corresponding to f exists and is the constant function $I(\mu) = c(f_0)$.
The proof follows directly by using the definition of Fisher information as the second moment of the score function, and is omitted.
Remark: Because of assumption (a), the theorem above is applicable to the important double exponential example $f(x|\mu) = \frac{1}{2}e^{-|x-\mu|}$; this corresponds to $f_0(z) = \frac{1}{2}e^{-|z|}$, which is not differentiable at $z = 0$. Condition (a) allows that.

Example 6.14. (Fisher Information is Not Transformation Invariant). Consider the $N(0, \sigma^2)$ density, *treating* σ *as the parameter*. Then, ignoring constants,

$$f(x|\sigma) = \frac{1}{\sigma} e^{-x^2/(2\sigma^2)} \Rightarrow \log f(x|\sigma) = -\log \sigma - \frac{x^2}{2\sigma^2}$$

$$\Rightarrow \frac{d}{d\sigma} \log f(x|\sigma) = -\frac{1}{\sigma} + \frac{x^2}{\sigma^3} \Rightarrow \frac{d^2}{d\sigma^2} \log f(x|\sigma) = \frac{1}{\sigma^2} - \frac{3x^2}{\sigma^4}.$$

Therefore, by our alternative formula for Fisher information under assumption $A4$, which always holds in any Exponential family distribution,

$$I(\sigma) = -E_\sigma[\frac{d^2}{d\sigma^2} \log f(X|\sigma)] = -\frac{1}{\sigma^2} + 3\frac{E_\sigma[X^2]}{\sigma^4}$$

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

Now, instead of σ*, call* σ^2 *your parameter*, and denote it as $\theta; \theta = \sigma^2$. With this new parametrization,

$$f(x|\theta) = \frac{1}{\sqrt{\theta}} e^{-x^2/(2\theta)} \Rightarrow \log f(x|\theta) = -\frac{1}{2} \log \theta - \frac{x^2}{2\theta}$$

$$\Rightarrow \frac{d}{d\theta} \log f(x|\theta) = -\frac{1}{2\theta} + \frac{x^2}{2\theta^2} \Rightarrow \frac{d^2}{d\theta^2} \log f(x|\theta) = \frac{1}{2\theta^2} - \frac{x^2}{\theta^3}.$$

Therefore,

$$I(\theta) = -E_\theta[\frac{d^2}{d\theta^2} \log f(X|\theta)] = -[\frac{1}{2\theta^2} - \frac{\theta}{\theta^3}] = \frac{1}{2\theta^2},$$

which rewritten in terms of σ, is $\frac{1}{2\sigma^4}$. This is different from the Fisher information we got when we treated σ as the parameter. Notice that the information changed; it is not invariant under monotone transformations of the parameter. The final formula for the Fisher information will depend on exactly how you parametrize your family.

To some, this lack of transformation invariance of the Fisher information is troublesome. The discomfort is that information from the experiment depends on exactly how you parametrize your density. You need to be mindful of this lack of transformation invariance.

It would be useful to have a general formula for Fisher information in the entire one parameter Exponential family; this is given in the next theorem.

Theorem 6.4. (General Formula in Exponential Family). Consider a general pdf or pmf $f(x|\theta) = e^{\eta(\theta)T(x)-\psi(\theta)}h(x)$ in the regular one parameter Exponential family. Then, the Fisher information function exists and is given by the general formula

$$I(\theta) = \psi''(\theta) - \frac{\psi'(\theta)\eta''(\theta)}{\eta'(\theta)}.$$

Proof: We have

$$\begin{aligned} U(\theta) = \frac{d}{d\theta}\log f(x|\theta) &= \frac{d}{d\theta}[\eta(\theta)T(x) - \psi(\theta)] \\ &= \eta'(\theta)T(x) - \psi'(\theta). \end{aligned}$$

Since $\text{Var}_\theta[T(X)] = \frac{\psi''(\theta)}{[\eta'(\theta)]^2} - \frac{\psi'(\theta)\eta''(\theta)}{[\eta'(\theta)]^3}$ (refer back to Theorem 5.4), we have

$$\begin{aligned} I(\theta) = \text{Var}_\theta[U(\theta)] &= [\eta'(\theta)]^2\left[\frac{\psi''(\theta)}{[\eta'(\theta)]^2} - \frac{\psi'(\theta)\eta''(\theta)}{[\eta'(\theta)]^3}\right] \\ &= \psi''(\theta) - \frac{\psi'(\theta)\eta''(\theta)}{\eta'(\theta)}. \end{aligned}$$

If the density has already been expressed in its canonical form, so that the parameter is the natural parameter η, then this Fisher information formula simplifies to

$$I(\eta) = \psi''(\eta) - \frac{\psi'(\eta)\times 0}{1} = \psi''(\eta);$$

recheck this with definition 5.6.

Example 6.15. Fisher Information in $N(\theta,\theta)$. The $N(\theta,\theta), \theta > 0$ distribution is used when a user feels that a continuous random variable has equal mean and variance; a discrete analog is Poisson. It turns out that $N(\theta,\theta)$ is in the Exponential family, as can be verfied easily:

$$\frac{1}{\sqrt{\theta}}e^{-(x-\theta)^2/(2\theta)} = e^{-x^2/(2\theta)-\theta/2-\frac{1}{2}\log\theta}e^x,$$

which gives $\eta(\theta) = -\frac{1}{2\theta}, T(x) = x^2, \psi(\theta) = \frac{\theta}{2} + \frac{1}{2}\log\theta$. Plugging into Theorem 7.3 gives

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

6.2.6 Sufficient Statistics and Likelihood : Conceptual Discussion

Related to the likelihood function is another momentous concept invented by Fisher, that of a *sufficient statistic*. Here is the question that Fisher wanted to answer: suppose we have raw data $X_1, \cdots, X_n$ with some joint distribution $P_\theta = P_{\theta,n}$. Can we find a summary statistic $T(X_1, \cdots, X_n)$ which squeezes out all the information that the entire raw data $X_1, \cdots, X_n$ could have furnished? The summary statistic, ideally. should be a low dimensional statistic; for example, the mean of the data, or perhaps the mean of the data and the variance of the data. If such an aspiration did materialize, then we have reduced the n dimensional raw data vector $(X_1, \cdots, X_n)$ to a low dimensional statistic, *without losing any information whatsoever*! We may call such a statistic *sufficient*; i.e., once we know the value of this statistic $T(X_1, \cdots, X_n)$, knowing the individual raw data values is unimportant.

Amazingly, in many parametric models, such lower dimensional sufficient statistics do exist; and the likelihood function will take you there. By inspecting the likelihood function, you will know what a sufficient statistic is, after you have chosen a model for your problem. First, we give the formal definition of a sufficient statistic, due to Fisher.

Definition 6.6. Let $X^{(n)} = (X_1, \cdots, X_n) \sim P_\theta = P_{\theta,n}, \theta \in \Theta$. A (possibly vector valued) statistic $T = T(X_1, \cdots, X_n)$ is called *sufficient* for the model P_θ if for any set B, $P_\theta((X_1, \cdots, X_n) \in B \,|T(X_1, \cdots, X_n) = t)$ is free of the underlying parameter θ.

Remark: The interpretation of the definition is that if we already know what was the value of this distinguished statistic T, then there is no information at all in knowing the values of the specific X_i, because the distribution of $(X_1, \cdots, X_n)$ given $T = t$ will not even involve the parameter θ that we are trying to learn about. Thus, it is *enough to know just T. So, sufficiency is a tool for data reduction.*

You should know a few conceptual facts about sufficiency that are either obvious, or easy to prove:

Simple Facts

1. Sufficient statistics are not unique. If T is sufficient, any one-to-one function of T is also sufficient.

2. If T is sufficient, and you augment it with any other statistic, the augmented statistic will also be sufficient. More precisely, if T is sufficient, then (T, S) is also sufficient for any statistic S.

3. The complete sample, $X^{(n)} = (X_1, \cdots, X_n)$, which is an n-dimensional statistic is always sufficient.

4. Sufficient statistics that are less than n-dimensional need not exist in every problem.

5. If the sample observations $X_1, \cdots, X_n$ happen to be iid continuous random variables, then the order statistics $X_{(1)}, \cdots, X_{(n)}$ are sufficient; in that case, it does not matter what

the model is. However, it is still an n-dimensional sufficient statistic; so, reduction to the order statistics does not reduce dimension.
6. A statistic T which is sufficient for one model, need not be sufficient for another model. In general, sufficient statistics are model dependent.
Certainly defining a sufficient statistic in the way Fisher did makes intuitive sense. A quite different question is how does one find such a sufficient statistic in a given problem? A wild strategy would be to guess what T may be in a given problem, and then somehow verify that it indeed is sufficient by verifying the definition. This is not a practical strategy; guessing T can easily mislead us and waste our time, and actually verifying the definition, even if we made a lucky guess, could be a difficult mathematical exercise.
Fortunately, a famous theorem, first provided by Fisher himself, and later proved rigorously by Neyman, takes the guesswork out of the game. This theorem tells you that if you inspect the likelihood function, you will see what a sufficient statistic for your problem is; this is the factorization theorem.

6.2.7 Finding Sufficient Statistics: Factorization Theorem

Theorem 6.5. For a given model P_θ, a statistic $T(X_1, \cdots, X_n)$ is sufficient if and only if the likelihood function can be factorized in the product form

$$l(\theta) = g(\theta, T(x_1, \cdots, x_n))h(x_1, \cdots, x_n);$$

in other words, in the likelihood function, the only term that includes both θ and the data involves merely the statistic T, and no other statistics besides T.
We will give a proof of the *if* part of the factorization theorem in the discrete case. But, first we must see an illustrative example.

Example 6.16. (Sufficient Statistic for Bernoulli Data). Let $X_1, \cdots, X_n \overset{iid}{\sim} Ber(p)$. Denote the total number of successes by T, $T = \sum_{i=1}^n X_i$. The likelihood function is

$$l(p) = \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} = p^t(1-p)^{n-t}$$

(where $t = \sum_{i=1}^n x_i$)

$$= \left(\frac{p}{1-p}\right)^t (1-p)^n.$$

If we now define $T(X_1, \cdots, X_n) = \sum_{i=1}^n X_i, g(p,t) = \left(\frac{p}{1-p}\right)^t (1-p)^n$, and $h(x_1, \cdots, x_n) \equiv 1$, then we have factorized the likelihood function $l(p)$ in the form

$$l(p) = g(p, T(x_1, \cdots, x_n))h(x_1, \cdots, x_n);$$

so the factorization theorem implies that $T(X_1, \cdots, X_n) = \sum_{i=1}^n X_i$ is a sufficient statistic if the raw data $X_1, \cdots, X_n$ are iid Bernoulli. In plain English, there is no value in knowing which trials resulted in successes once we know how many trials resulted in successes, a statement that makes perfect sense.

Example 6.17. (Sufficient Statistics for General $N(\mu, \sigma^2)$). Let $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where μ, σ^2 are both considered unknown; thus the parameter is two dimensional, $\theta = (\mu, \sigma)$. Ignoring constants, the likelihood function is

$$l(\theta) = \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2}$$

$$= \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}[\sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2]}.$$

(we have used the algebraic identity $\sum_{i=1}^n (x_i - \mu)^2 = \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$)
If we now define the *two dimensional statistic*

$$T(X_1, \cdots, X_n) = (\bar{X}, \sum_{i=1}^n (X_i - \bar{X})^2),$$

and define

$$g(\mu, \sigma, t) = \frac{1}{\sigma^n} e^{-\frac{n}{2\sigma^2}(\bar{x} - \mu)^2} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \bar{x})^2}, h(x_1, \cdots, x_n) \equiv 1,$$

then, again, we have factorized the likelihood function $l(\mu, \sigma)$ in the form

$$l(\mu, \sigma) = g(\mu, \sigma, t) h(x_1, \cdots, x_n),$$

and therefore, the factorization theorem implies that the two dimensional statistic $(\bar{X}, \sum_{i=1}^n (X_i - \bar{X})^2)$ (and hence, also, $\bar{X}, \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2)$) is a sufficient statistic if data are iid from a univariate normal distribution, both μ, σ being unknown.

Example 6.18. (Sufficient Statistic in $U[0, \theta]$ Case). Let $X_1, \cdots, X_n \overset{iid}{\sim} U[0, \theta], \theta > 0$. From Example 7.9, the likelihood function is

$$l(\theta) = \frac{1}{\theta^n} I_{\theta \geq x_{(n)}},$$

where $x_{(n)}$ denotes the sample maximum. If we define $T(X_1, \cdots, X_n) = X_{(n)}, g(\theta, t) = \frac{1}{\theta^n} I_{\theta \geq t}, h(x_1, \cdots, x_n) \equiv 1$, we have factorized the likelihood function $l(\theta)$ as

$$l(\theta) = g(\theta, t) h(x_1, \cdots, x_n),$$

and so, the sample maximum $T(X_1, \cdots, X_n) = X_{(n)}$ is a sufficient statistic if data are iid from $U[0, \theta]$.
We will now give a proof of the factorization theorem in the discrete case.

Proof of Factorization Theorem: We will prove just the *if part*. Thus, suppose the factorization

$$l(\theta) = P_\theta(X_1 = x_1, \cdots, X_n = x_n) = g(\theta, T(x_1, \cdots, x_n))h(x_1, \cdots, x_n)$$

holds for some statistic T; we would like to prove that T is sufficient by verifying the definition of sufficiency. Fix $x_1, \cdots, x_n$, and look at the conditional probability

$$P_\theta(X_1 = x_1, \cdots, X_n = x_n \,|T(X_1, \cdots, X_n) = t) = \frac{P_\theta(X_1 = x_1, \cdots, X_n = x_n, T(X_1, \cdots, X_n) = t)}{P_\theta(T(X_1, \cdots, X_n) = t)}$$

$$= 0, \text{ if } T(x_1, \cdots, x_n) \neq t,$$

because, in that case, the intersection of the two sets $\{X_1 = x_1, \cdots, X_n = x_n\}$ and $\{T(X_1, \cdots, X_n) = t\}$ is the empty set.
Next, if $T(x_1, \cdots, x_n) = t$, then,

$$\frac{P_\theta(X_1 = x_1, \cdots, X_n = x_n, T(X_1, \cdots, X_n) = t)}{P_\theta(T(X_1, \cdots, X_n) = t)} = \frac{P_\theta(X_1 = x_1, \cdots, X_n = x_n)}{P_\theta(T(X_1, \cdots, X_n) = t)}$$

(because the event $\{X_1 = x_1, \cdots, X_n = x_n\}$ implies the event $\{T(X_1, \cdots, X_n) = t\}$ in the case that $T(x_1, \cdots, x_n) = t$)

$$= \frac{g(\theta, t)h(x_1, \cdots, x_n)}{\sum_{z_1, \cdots, z_n : T(z_1, \cdots, z_n) = t} P_\theta(X_1 = z_1, \cdots, X_n = z_n)}$$

$$= \frac{g(\theta, t)h(x_1, \cdots, x_n)}{\sum_{z_1, \cdots, z_n : T(z_1, \cdots, z_n) = t} g(\theta, T(z_1, \cdots, z_n))h(z_1, \cdots, z_n)}$$

$$= \frac{g(\theta, t)h(x_1, \cdots, x_n)}{g(\theta, t)\sum_{z_1, \cdots, z_n : T(z_1, \cdots, z_n) = t} h(z_1, \cdots, z_n)}$$

$$= \frac{h(x_1, \cdots, x_n)}{\sum_{z_1, \cdots, z_n : T(z_1, \cdots, z_n) = t} h(z_1, \cdots, z_n)},$$

i.e., the term that had θ in it, namely $g(\theta, t)$ completely cancels off from the numerator and the denominator, leaving a final answer totally free of θ. Hence, by definition of sufficiency, T is sufficient.

6.2.8 Minimal Sufficient Statistics

We have remarked in Section 7.2.6 that if T is sufficient, then the augmented statistic (T, S) is also sufficient, for any statistic S. From the point of view of maximal data reduction, you would prefer to use T rather than (T, S), because (T, S) adds an extra dimension. To put it another way, the lower dimensional T is a function of (T, S) and so we prefer T over (T, S). A natural question that arises is which sufficient statistic corresponds to the maximum possible data reduction, without causing loss of information. Does such a

sufficient statistic always exist? Do we know how to find it?
Such a most parsimonious sufficient statistic is called *a minimal sufficient statistic*; T is minimal sufficient if for any other sufficient statistic T^*, T is a function of T^*. As regards explicitly identifying a minimal sufficient statistic, the story is very clean for Exponential families. Otherwise, the story is not clean. There are certainly some general theorems that characterize a minimal sufficient statistic in general; but they are not particularly useful in practice. Outside of the Exponential family, anything can happen; for instance, the minimal sufficient statistic may be n-dimensional, i.e., absolutely no dimension reduction is possible without sacrificing information. A famous example of this is the case of $X_1, \cdots, X_n \overset{iid}{\sim} C(\mu, 1)$. It is a one parameter problem; but it may be shown that the minimal sufficient statistic is the vector of order statistics, $(X_{(1)}, \cdots, X_{(n)})$; so no reduction beyond the obvious is possible!

6.2.9 Sufficient Statistics in Exponential Families

In each of the three specific examples of sufficient statistics that we worked out in Section 7.2.5, the dimension of the parameter and the dimension of the sufficient statistic matched. This was particularly revealing in the $N(\mu, \sigma^2)$ example; to be sufficient when both μ, σ are unknown, we must use both the sample mean and the sample variance. Just the sample mean would not suffice, for example; roughly speaking, if you use just the sample mean, you will lose information on σ! Among our three examples, the $U[0, \theta]$ example was a nonregular example. If we consider only regular families, when do we have a sufficient statistic whose dimension exactly matches the dimension of the parameter vector?
It turns out that in the entire Exponential family, *including the multiparameter Exponential family*, we can find a sufficient statistic which has just as many dimensions as has the parameter. A little more precisely, in the k-parameter regular Exponential family, you will be able to find a k-dimensional sufficient statistic (and it will even be minimal sufficient). Here is such a general theorem.

Theorem 6.6. (Minimal Sufficient Statistics in Exponential Families).
(a) Suppose $X_1, \cdots, X_n \overset{iid}{\sim} f(x\,|\theta) = e^{\eta(\theta)T(x) - \psi(\theta)}h(x)$, where $\theta \in \Theta \subseteq \mathcal{R}$. Then, $\sum_{i=1}^n T(X_i)$ (or equivalently, the empirical mean $\bar{T} = \frac{1}{n}\sum_{i=1}^n T(X_i)$) is a one dimensional minimal sufficient statistic.
(b) Suppose $X_1, \cdots, X_n \overset{iid}{\sim}$

$$f(x\,|\theta_1, \cdots, \theta_k) = e^{\sum_{i=1}^k \eta_i(\theta)T_i(x) - \psi(\theta)}h(x),$$

where $\theta = (\theta_1, \cdots, \theta_k) \in \Theta \subseteq \mathcal{R}^k$. Then, the k-dimensional statistic $(\sum_{i=1}^n T_1(X_i), \cdots, \sum_{i=1}^n T_k(X_i))$ (or equivalently, the vector of empirical means $(\bar{T}_1, \cdots, \bar{T}_k)$) is a k-dimensional minimal sufficient statistic.

Remark: 1. *Recall that in the Exponential family, $(\bar{T}_1, \cdots, \bar{T}_k)$ is the maximum likelihood estimate of $(E_\eta(T_1), \cdots, E_\eta(T_k))$. Theorem 7.6 says that maximum likelihood and minimal sufficiency, two apparently different goals, coincide in the entire Exponential family, a beautiful as well as a substantive result.*

Remark: 2. *A deep result in statistical inference is that if we restrict our attention to only regular families, then distributions in the Exponential family are the only ones for which we can have a sufficient statistic whose dimension matches the dimension of the parameter. See, in particular, Barankin and Maitra (1963), and Brown (1964).*

Example 6.19. (Sufficient Statistics in Two Parameter Gamma). We show in this example that the general Gamma distribution is a member of the two parameter Exponential family. To show this, just observe that with $\theta = (\alpha, \lambda) = (\theta_1, \theta_2)$,

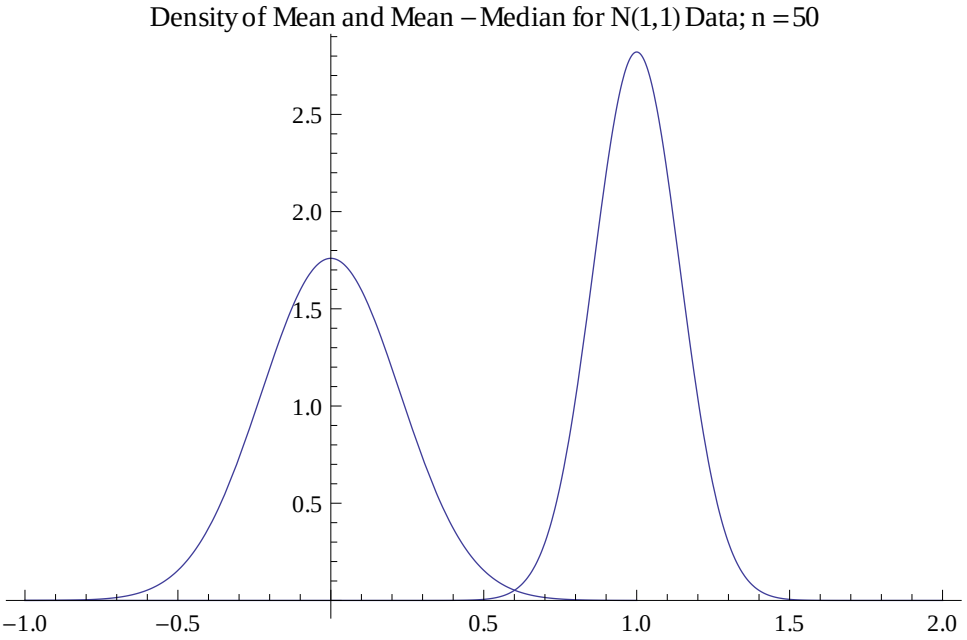
$$f(x\,|\theta) = e^{-\frac{x}{\theta_2} + \theta_1 \log x - \theta_1 \log \theta_2 - \log \Gamma(\theta_1)} \frac{1}{x} I_{x>0}.$$

This is in the two parameter Exponential family with $\eta_1(\theta) = -\frac{1}{\theta_2}, \eta_2(\theta) = \theta_1, T_1(x) = x, T_2(x) = \log x, \psi(\theta) = \theta_1 \log \theta_2 + \log \Gamma(\theta_1)$, and $h(x) = \frac{1}{x} I_{x>0}$. The parameter space in the θ-parametrization is $(0, \infty) \otimes (0, \infty)$.
Hence, by Theorem 7.6, $(\sum_{i=1}^n T_1(X_i), \sum_{i=1}^n T_2(X_i)) = (\sum_{i=1}^n X_i, \sum_{i=1}^n \log X_i)$ is minimal sufficient. Since, $\sum_{i=1}^n \log X_i = \log(\prod_{i=1}^n X_i)$, and logarithm is a one-to-one function, we can also say that $(\sum_{i=1}^n X_i, \prod_{i=1}^n X_i)$ is a two dimensional sufficient statistic for the two dimensional parameter $\theta = (\alpha, \lambda)$.

6.2.10 Are MLEs Always Sufficient?

Unfortunately, the answer is no. Outside of the Exponential family, the maximum likelihood estimate alone in general would not be sufficient. To recover the *ancillary information* that was missed by the maximum likelihood estimate T, one has to couple T with another suitable statistic S so that the augmented statistic (T, S) becomes minimal sufficient. Paradoxically, on its own, S is essentially a worthless statistic in the sense that the distribution of S is free of θ! Such a statistic is called an *ancillary statistic*. But, when the ancillary statistic S joins hands with the MLE T, it provides the missing information and together, (T, S) becomes minimal sufficient! These were basic and also divisive issues in inference in the fifties and the sixties. They did lead to useful understanding. A few references are Basu (1955, 1959, 1964), Lehmann (1981), Buehler (1982), Brown (1990), Ghosh (2002), and Fraser (2004).

Example 6.20. (Example of Ancillary Statistics). We will see some ancillary statistics to have the concept clear in our mind. Suppose $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, 1)$. Let $T(X_1, \cdots, X_n) =$



$\bar{X}$ and $S(X_1, \cdots, X_n) = M_n(X)$, the sample median. Now notice the following; we may write $X_i = \mu + Z_i, i = 1, 2, \cdots, n$, where Z_i are iid $N(0, 1)$. Let $\bar{Z}$ and $M_n(Z)$ denote the mean and the median of these $Z_1, \cdots, Z_n$. Then, we have the representation

$$\bar{X} = \mu + \bar{Z};\ \ M_n(X) = \mu + M_n(Z).$$

Hence, $\bar{X} - M_n(X)$ has the same distribution as $\bar{Z} - M_n(Z)$. But $\bar{Z} - M_n(Z)$ is a function of a set of iid standard normals, and so $\bar{Z} - M_n(Z)$ has some fixed distribution without any parameters in it. Hence, $\bar{X} - M_n(X)$ also has that same fixed distribution without any parameters in it; in other words, in the $N(\mu, 1)$ case $\bar{X} - M_n(X)$ is an ancillary statistic. You may have realized that the argument we gave is valid for any location parameter distribution, not just $N(\mu, 1)$.

We have plotted the density of the sufficient statistic $\bar{X}$ and the ancillary statistic $\bar{X} - M_n(X)$ in the $N(\mu, 1)$ case when the true $\mu = 1$ and $n = 50$. The ancillary statistic peaks at zero, while the sufficient statistic peaks at $\mu = 1$.

Example 6.21. (Another Example of Ancillary Statistics). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} U[0, \theta]$. Let $S(X_1, \cdots, X_n) = \frac{\bar{X}}{X_{(n)}}$. We will show that $S(X_1, \cdots, X_n)$ is an ancillary statistic. The argument parallels the previous example. We may write $X_i = \theta Z_i, i = 1, 2, \cdots, n$, where Z_i are iid $U[0, 1]$. Let $\bar{Z}$ and $Z_{(n)}$ denote the mean and the maximum of these $Z_1, \cdots, Z_n$. Then, we have the representation

$$\bar{X} = \theta\bar{Z};\ \ X_{(n)} = \theta Z_{(n)}.$$

Hence, $\frac{\bar{X}}{X_{(n)}}$ has the same distribution as $\frac{\bar{Z}}{Z_{(n)}}$. But $\frac{\bar{Z}}{Z_{(n)}}$ is a function of a set of iid $U[0, 1]$ variables, and so it has some fixed distribution without any parameters in it. This means

that $\frac{\bar{X}}{X_{(n)}}$ also has that same fixed distribution without any parameters in it; that is, $\frac{\bar{X}}{X_{(n)}}$ is an ancillary statistic. Again, the argument we gave is valid for any scale parameter distribution, not just $U[0, \theta]$.

Example 6.22. (Aberrant Phenomena in Curved Exponential Families). In Exponential families that are *curved*, rather than regular, we have the odd phenomenon that sufficient statistics have dimension *larger* than the dimension of the parameter. Consider the $N(\theta, \theta^2), \theta > 0$ example; this was previously studied in Example 5.14.
In this case, ignoring constants,

$$l(\theta) = \frac{1}{\theta^n} e^{-\frac{\sum_{i=1}^n x_i^2}{2\theta^2} + \frac{\sum_{i=1}^n x_i}{\theta}}.$$

By using the factorization theorem, we get from here that the two-dimensional statistic $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is sufficient, while the parameter θ is one dimensional (there is *only one* free parameter in $N(\theta, \theta^2)$). It is indeed the case that $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is even minimal sufficient.
To complete the example, let us look at the maximum likelihood estimate of θ. For this, we need a little notation. Let

$$U = \sqrt{\sum_{i=1}^n X_i^2},\ S = \frac{\sum_{i=1}^n X_i}{\sqrt{\sum_{i=1}^n X_i^2}}.$$

Then, tt is quite easy to show that the MLE of θ is

$$T = \frac{U[\sqrt{S^2 + 4n} - S]}{2n}.$$

T is one dimensional, and not sufficient. However, the two dimensional statistic (T, S) is a one-to-one function of $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ (verify this easy fact), and hence, (T, S) is minimal sufficient. It may also be easily verified that the distribution of S is a fixed distribution completely devoid of θ; therefore, S is an ancillary statistic, with zero information about θ. But couple it with the insufficient statistic T, and now (T, S) is minimal sufficient.

6.2.11 Basu's Theorem

Remember, however, that in Exponential families the maximum likelihood estimate of the parameter would by itself be minimal sufficient. This property of minimal sufficiency of the MLE also holds in some nonregular cases; e.g., $U[0, \theta], U[-\theta, \theta], U[\theta_1, \theta_2]$, to name a few interesting ones. Now, a minimal sufficient statistic T captures all the information about the parameter θ, while an ancillary statistic S captures none. T has got a lot to do with θ, and S has nothing to do with θ. Intuitively, one would expect the two statistics

to be unrelated. A pretty theorem due to Debabrata Basu (Basu (1955)) says that T and S would actually be independent, under some extra condition. This extra condition need not be verified in Exponential families; it holds. The extra condition also holds in all of the nonregular cases we just mentioned above, $U[0,\theta], U[-\theta,\theta], U[\theta_1,\theta_2]$. You will see for yourself in the next section how useful this theorem is in solving practical problems in a very efficient manner. Here is the theorem. We state it here only for the iid case and Exponential families, although the published theorem is far more general.

Theorem 6.7. (Basu's Theorem). Suppose $X_1,\cdots,X_n$ are iid observations from a multiparameter density or pmf in the Exponential family, $f(x\,|\eta), \eta \in \mathcal{T}$. Assume the following:
(a) The family is regular;
(b) The family is nonsingular;
(c) The parameter space $\mathcal{T}$ contains a ball (however small) of positive radius.
Let $(\bar{T}_1,\cdots,\bar{T}_k)$ be the minimal sufficient statistic and $S(X_1,\cdots,X_n)$ any ancillary statistic. Then, any function $h(\bar{T}_1,\cdots,\bar{T}_k)$ of $(\bar{T}_1,\cdots,\bar{T}_k)$ and $S(X_1,\cdots,X_n)$ are independently distributed under each $\eta \in \mathcal{T}$.
Remark: This theorem also holds if the underlying density is $U[0,\theta], U[-\theta,\theta]$, or $U[\theta_1,\theta_2]$; you have to be careful that the possible values of the parameter(s) contain a ball (an interval if the parameter is a scalar parameter). Thus, Basu's theorem holds in the $U[0,\theta]$ case if $\theta \in (a,b), a < b$; but it would not hold if $\theta = 1, 2$ (only two possible values).

Example 6.23. (Application of Basu's Theorem). Let $X_1,\cdots,X_n \overset{iid}{\sim} N(\mu,1), \mu \in (a,b), -\infty \le a < b \le \infty$. Thus, condition (c) in Basu's theorem holds. We know that $\bar{X}$ is minimal sufficient and $\bar{X} - M_n$ is ancillary in this case, where M_n is the sample median. Therefore, by Basu's theorem, in the iid $N(\mu,1)$ case, $\bar{X}$ and $\bar{X} - M_n$ are independent under any μ.

Example 6.24. (Another (Application of Basu's Theorem). Let $X_1,\cdots,X_n \overset{iid}{\sim} U[0,\theta], \theta \in (a,b), 0 \le a < b \le \infty$. In this case, $X_{(n)}$ is minimal sufficient and $S(X_1,\cdots,X_n) = \frac{\bar{X}}{X_{(n)}}$ is ancillary. By Basu's theorem, $X_{(n)}$ and $S(X_1,\cdots,X_n)$ are independent under any θ. So would be $X_{(n)}$ and $\frac{X_{(1)}}{X_{(n)}}$, or $\frac{\bar{X}-X_{(1)}}{X_{(n)}}$, because these are also ancillary statistics.

6.2.12 Sufficiency in Action: Rao-Blackwellization

The concept of data reduction without losing any information is an appealing idea. But the inquisitive student would ask *what can sufficiency do for me apart from dimension reduction*? It turns out that a sufficient statistic is also sufficient in the literal sense in the unifying framework of decision theory. We had remarked in Chapter 6 that given

two decision rules (procedures) δ_1, δ_2, if one of them always had a smaller risk, we would naturally prefer that procedure. The *Rao-Blackwell theorem*, proved independently by C.R. Rao and David Blackwell (Rao (1945), Blackwell (1947)), provides a concrete benefit of looking at sufficient statistics from a viewpoint of preferring procedures with lower risk. The Rao-Blackwell theorem says that after you have chosen your model, there is no reason to look beyond a minimal sufficient statistic.

Theorem 6.8. (Rao-Blackwell). Consider a general decision problem with a loss function $L(\theta, a)$. Assume that for every fixed $\theta, L(\theta, a)$ is a convex function of the argument a. Let X represent the complete data and T any sufficient statistic. Suppose $\delta_1(X)$ is a general procedure depending on X. Then, there exists an alternative procedure $\delta_2(T)$ depending only on T such that δ_2 has smaller riisk than δ_1:

Such a procedure δ_2 may be chosen to be

$$\delta_2(t) = E_{\theta_0}[\delta_1(X)\,|T = t)];$$

here θ_0 is any arbitrary member of the parameter space Θ, and the choice of θ_0 does not change the final procedure $\delta_2(T)$.

Remark: The better procedure δ_2 is called *the Rao-Blackwellized version of* δ_1. *Note that only because* T *is sufficient,* $\delta_2(T)$ *depends just on* T*, and not on the specific* θ_0 *you chose. If* T *was not sufficient,* $E_\theta[\delta_1(X)\,|T]$ *will depend on* θ!

Proof: The proof follows by using *Jensen's inequality of probability theory* (see Chapter 3). Fix any member $\theta_0 \in \Theta$. Then,

$$R(\theta_0, \delta_2) = E_{\theta_0}[L(\theta_0, \delta_2(T))] = E_{\theta_0}[L(\theta_0, E_{\theta_0}[\delta_1(X)\,|T)]$$

$$\leq E_{\theta_0}[E_{\theta_0}\{L(\theta_0, \delta_1(X))\,|T\}]$$

(by applying Jensen's inequality to the convex function $L(\theta_0, a)$ under the distribution of X given T; understand this step well)

$$= E_{\theta_0}[L(\theta_0, \delta_1(X))]$$

(by using the iterated expectation formula that expectation of a conditional expectation is the unconditional expectation; see Chapter 3, if needed)

$$= R(\theta_0, \delta_1),$$

which shows that at any arbitrary θ_0, the risk of δ_2 is at least as small as the risk of δ_1. Let us see a few illustrative examples.

Example 6.25. (Rao-Blackwellization of Sample Median). Suppose $X_1, \cdots, X_n \sim N(\mu, 1)$, and suppose someone has proposed estimating μ by using the sample median M_n.

But, M_n is not sufficient for the $N(\mu, 1)$ model; however, $\bar{X}$ is (and it is even minimal sufficient). So, the Rao-Blackwell theorem tells us that as long as we use a convex loss function (e.g., squared error loss is certainly convex), we would be able to find a better procedure than the sample median M_n. We also knoow what such a better procedure is; it is,

$$E_{\mu_0}[M_n \,|\bar{X}] \text{ (any } \mu_0) = E_{\mu_0}[M_n - \bar{X} + \bar{X} \,|\bar{X}]$$
$$= E_{\mu_0}[M_n - \bar{X} \,|\bar{X}] + E_{\mu_0}[\bar{X} \,|\bar{X}]$$
$$= E_{\mu_0}[M_n - \bar{X}] + \bar{X}$$

(use the fact that by Basu's theorem, $M_n - \bar{X}$ and $\bar{X}$ are independent; hence, the conditional expectation $E_{\mu_0}[M_n - \bar{X} \,|\bar{X}]$ must be the same as the unconditional expectation $E_{\mu_0}[M_n - \bar{X}]$)

$$= E_{\mu_0}[M_n] - E_{\mu_0}[\bar{X}] + \bar{X} = \mu_0 - \mu_0 + \bar{X} = \bar{X}.$$

So, at the end we have reached a very neat conclusion; the sample mean has a smaller risk than the sample median in the $N(\mu, 1)$ model under any convex loss function! You notice that it was very critical to use Basu's theorem in carrying out this Rao-Blackwellization calculation.

Example 6.26. (Rao-Blackwellization in Poisson). Suppose based on $X_1, \cdots, X_n \overset{iid}{\sim} Poi(\lambda)$, we wish to estimate $P_\lambda(X = 0) = e^{-\lambda}$. A layman's estimate might be the fraction of data values equal to zero:

$$\delta_1(X_1, \cdots, X_n) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i=0}.$$

Rao-Blackwell theorem tells us we can do better by conditioning on $\sum_{i=1}^n X_i$, because $\sum_{i=1}^n X_i$ is sufficient in the iid Poisson case. To calculate this conditional expectation, we have to borrow the probability theory result that if $X_i, 1 \le i \le n$ are iid $Poi(\lambda)$, then any X_i given that $\sum_{i=1}^n X_i = t$ is distributed as $Bin(t, \frac{1}{n})$ (see Chapter 3). Then,

$$E_\lambda[\delta_1(X_1, \cdots, X_n) \,|\, \sum_{i=1}^{n} X_i = t] = \frac{1}{n} \sum_{i=1}^{n} E_\lambda[I_{X_i=0} \,|\, \sum_{i=1}^{n} X_i = t]$$
$$= \frac{1}{n} \sum_{i=1}^{n} P_\lambda[X_i = 0 \,|\, \sum_{i=1}^{n} X_i = t] = \frac{1}{n} \sum_{i=1}^{n} (1 - \frac{1}{n})^t$$
$$= (1 - \frac{1}{n})^t.$$

Thus, in the iid Poisson case, for estimating the probability of the zero value (no events), $(1 - \frac{1}{n})^{\sum_{i=1}^n X_i}$ is better than the layman's estimator $\frac{1}{n} \sum_{i=1}^n I_{X_i=0}$.
Heuristically, this better estimator

$$(1 - \frac{1}{n})^{\sum_{i=1}^n X_i} = (1 - \frac{1}{n})^{n \frac{1}{n} \sum_{i=1}^n X_i} = (1 - \frac{1}{n})^{n\bar{X}}$$

$$= [(1 - \frac{1}{n})^n]^{\bar{X}} \approx [e^{-1}]^{\bar{X}} = e^{-\bar{X}}.$$

Now you can see that the Rao-Blackwellized estimator iis *almost* the same as the more transparent estimator $e^{-\bar{X}}$, which *plugs* $\bar{X}$ for λ in $P_\lambda(X = 0) = e^{-\lambda}$.

Example 6.27. (Rao-Blackwellization in Uniform). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} U[0, \theta]$. Obviously, $E[\bar{X}] = \frac{\theta}{2}$, and so, $E[2\bar{X}] = \theta$. But, $\bar{X}$ is not sufficient in the $U[0, \theta]$ case; a minimal sufficient statistic is the sample maximum $X_{(n)}$. Once again, Rao-Blackwell theorem tells us that we can do better than $2\bar{X}$ and a better estimator would be the conditional expectation $E_{\theta_0}[2\bar{X}\,|X_{(n)}]$. We will choose $\theta_0 = 1$ as a convenient value to use. We will now work oout this conditional expectation $E_{\theta_0=1}[2\bar{X}\,|X_{(n)}]$:

$$E_{\theta_0=1}[2\bar{X}\,|X_{(n)} = t] = 2E_{\theta_0=1}[\bar{X}\,|X_{(n)} = t]$$

$$= 2E_{\theta_0=1}[X_{(n)}\,\frac{\bar{X}}{X_{(n)}}\,|X_{(n)} = t] = 2tE_{\theta_0=1}[\frac{\bar{X}}{X_{(n)}}\,|X_{(n)} = t]$$

$$= 2tE_{\theta_0=1}[\frac{\bar{X}}{X_{(n)}}]$$

(since, by Basu's theorem, $\frac{\bar{X}}{X_{(n)}}$ and $X_{(n)}$ are independent)
Now write

$$E_{\theta_0=1}[\bar{X}] = E_{\theta_0=1}[\frac{\bar{X}}{X_{(n)}}\,X_{(n)}] = E_{\theta_0=1}[\frac{\bar{X}}{X_{(n)}}]E_{\theta_0=1}[X_{(n)}]$$

(because of the same reason, that $\frac{\bar{X}}{X_{(n)}}$ and $X_{(n)}$ are independent)
Hence,

$$E_{\theta_0=1}[\frac{\bar{X}}{X_{(n)}}] = \frac{E_{\theta_0=1}[\bar{X}]}{E_{\theta_0=1}[X_{(n)}]} = \frac{\frac{1}{2}}{\frac{n}{n+1}} = \frac{n+1}{2n}.$$

By plugging this back into $E_{\theta_0=1}[2\bar{X}\,|X_{(n)} = t]$, we get,

$$E_{\theta_0=1}[2\bar{X}\,|X_{(n)} = t] = 2t\,\frac{n+1}{2n} = \frac{n+1}{n}t.$$

So, finally, we have arrived at the conclusion that although $2\bar{X}$ is an unbiased estimate of the uniform endpoint θ, a better estimate under any convex loss function is the Rao-Blackwellized estimate $\frac{n+1}{n}X_{(n)}$. Notice how critical it was to use Basu's theorem to get this result.

6.3 Unbiased Estimation: Conceptual Discussion

We recall the definition of an unbiased estimator.

Definition 6.7. An estimator $\hat{\theta}$ of a parameter θ is called unbiased if for all $\theta \in \Theta$, $E_\theta[|\hat{\theta}|] < \infty$, and $E_\theta[\hat{\theta}] = \theta$.

Unbiasedness is sometimes described as lack of systematic error; recall the discussions in Chapter 6.

Maximum likelihood estimates are widely used, and except in rather rare cases, are very fine estimators in parametric problems with not too many parameters. But, often they are not exactly unbiased. We recall two examples.

Example 6.28. (Biased MLEs). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} U[0, \theta]$. Then, we know that the unique MLE of θ is the sample maximum $X_{(n)}$. Now, it is obvious that $X_{(n)}$, being one of the data values, is certain to be smaller than the true θ; i.e., $P_\theta(X_{(n)} < \theta) = 1$. This means that also on the average, $X_{(n)}$ is smaller than the true θ; $E_\theta[X_{(n)}] < \theta$ for any θ. In fact, $E_\theta[X_{(n)}] = \frac{n}{n+1}\theta$. So, in this example the MLE has a systematic (but small) underestimation error.

Here is another example. Suppose $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where μ, σ are both considered unknown, Then, the MLE of σ^2 is $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$. Its expectation is

$$E[\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2] = E[\frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2] = \frac{n-1}{n}E[\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2]$$

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

Once again, the MLE is biased.

Of course, an MLE is not always biased; it is just that sometimes it can be biased.

At one time in the fifties and the sixties, a very substantial number of statisticians used to prefer estimates which have no bias, i.e., estimates which are unbiased. This led to the development of a very pretty and well structured theory of unbiased estimation, and more importantly, *best unbiased estimation*. The preference for exactly unbiased estimates has all but disappeared in statistical research now. There are a few reasons. First, insisting on exact unbiasedness forces one to eliminate otherwise wonderful estimates in many problems. Second, exact unbiased estimates often buy their zero bias for a higher variance. If we permit a little bias, we can often find estimates which have a smaller MSE than the exactly unbiased estimate. There are too many examples of this. Third, finding best unbiased estimates as soon as we consider distributions outside of the Exponential family is a tremendously difficult mathematical exercise, and not worth it. Still, because of its historical importance, a discussion of best unbiased estimation is wise; we limit ourselves mostly to the Exponential families. Regarding the existence of useful unbiased estimates, there are some very general results; see Liu and Brown (1993).

6.3.1 Sufficiency in Action: Best Unbiased Estimates

First we define a best unbiased estimate. Recall that in general, MSE equals the sum of the variance and the square of the bias. So, for unbiased estimates, MSE is same as variance.

If we minimize the variance, then in the case of unbiased estimates, we automatically minimize the MSE. This explains the following definition of a best unbiased estimate.

Definition 6.8. An unbiased estimator T of a parameter θ (or more generally, a parametric function $g(\theta)$) is called the *best unbiased estimate or uniformly minimum variance unbiased estimate* (UMVUE) of θ if

$$(a)\,\mathrm{Var}_\theta(T) < \infty \text{ for all } \theta;$$

$$(b)\; E_\theta(T) = \theta \text{ for all } \theta;$$

and, for any other unbiased estimate U of θ,

$$(c)\mathrm{Var}_\theta(T) \leq \mathrm{Var}_\theta(U)\text{for all } \theta.$$

You should know the following general facts about best unbiased estimates.

Facts to Know

1. A best unbiased estimate need not exist.
2. If a best unbiased estimate exists, it is unique.
3. In iid cases, if a best unbiased estimate exists, it is a *permutation invariant function* of the sample observations.
4. If a best unbiased estimate exists, it is a function of the minimal sufficient statistic.
5. Unlike maximum likelihood estimates, best unbiased estimates may take values outside the known parameter space. For example, a best unbiased estimate of a positive parameter may assume negative values for some datasets.
6. Outside of Exponential families, and a few simple nonregular problems, best unbiased estimates either do not exist, or are too difficult to find.
7. Inside the Exponential families, there is a neat and often user-friendly description of a best unbiased estimate of general parametric functions. This will be the content of the next theorem, a classic in statistical inference (Lehmann and Scheffe (1950)).

Theorem 6.9. (Lehmann-Scheffe Theorem for Exponential Families).

Let $X_1, \cdots, X_n \overset{iid}{\sim} f(x\,|\eta) = e^{\sum_{i=1}^k \eta_i T_i(x) - \psi(\eta)} h(x), \eta \in \mathcal{T}$, where we assume that
(a) the family is regular;
(b) the family is nonsingular;
(c) $\mathcal{T}$ contains in it a ball (however small) of positive radius.
Let $\theta = g(\eta)$ be any scalar parametric function. Then,
(a) If $\hat{\theta}(\bar{T}_1, \cdots, \bar{T}_k)$ is an unbiased estimate of θ depending only on the minimal sufficient statistic $(\bar{T}_1, \cdots, \bar{T}_k)$, and having a finite variance, $\mathrm{Var}_\eta(\hat{\theta}) < \infty$, it is automatically the best unbiased estimate of θ.

(b) There can be at most one such unbiased estimate of θ that depends only on $(\bar{T}_1, \cdots, \bar{T}_k)$.
(c) Such a best unbiased estimate may be explicitly calculated as follows: start with any arbitrary unbiased estimate $U(X_1, \cdots, X_n)$ of θ, and then find its Rao-Blackwellized version

$$\hat{\theta}(\bar{T}_1, \cdots, \bar{T}_k) = E_{\eta_0}[U(X_1, \cdots, X_n) \,|\bar{T}_1, \cdots, \bar{T}_k];$$

in this, the point η_0 can be chosen to be any convenient point of the parameter space $\mathcal{T}$ (for example, $\eta_0 = 0$, or 1).

Remark: The Lehmann-Scheffe theorem beautifully ties together minimal sufficiency, the technique of Rao-Blackwellization, and the structure that Exponential families enjoy. The Lehmann-Scheffe theorem may be appropriately generalized to distributions not in the Exponential family. However, such generalizations require introduction of other concepts that we have chosen to omit. See Lehmann and Casella (1998) for more general Lehmann-Scheffe theorems.

As usual, let us see some illustrative examples.

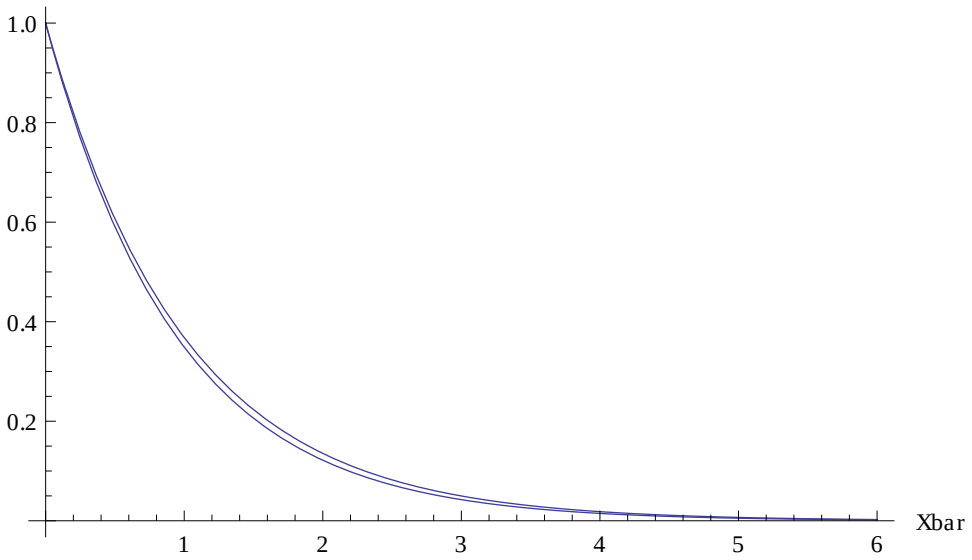
Example 6.29. (UMVUE of Normal Mean and Variance). Let $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, both parameters considered unknown, $-\infty < \infty < \mu, \sigma > 0$. This is a density in the two parameter exponential family, and $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$, or equivallently, $(\bar{X}, s^2)$ is minimal sufficient. Plainly, $E(\bar{X}) = \mu$ and $\bar{X}$ has finite variance. Since $\bar{X}$ is a function of $(\bar{X}, s^2)$, by the Lehmann-Scheffe theorem, it is the unique UMVUE of μ.

Next, equally evidently, $E(s^2) = \sigma^2$ and it is easy to verify that s^2 has finite variance (just use that $\frac{(n-1)s^2}{\sigma^2}$ is a chi-square variable, and chi-squares have finite variance). Since s^2 is a function of $(\bar{X}, s^2)$, by the Lehmann-Scheffe theorem, it is the unique UMVUE of σ^2. To summarize, in the iid univariate normal case, the sample mean and the sample variance are UMVUE of μ and σ^2.

Note one subtle point. It is not necessary to assume that μ, σ are completely unrestricted. If $\mu \in [a, b]$ and $\sigma \in [c, d]$ where $a < b$ and $c < d$, then the "ball with positive radius condition in the Lehmann-Scheffe theorem" still holds. So the UMVUE properties of $\bar{X}, s^2$ continue to hold as long as there is some tiny interval of values contained in the range of μ and the range of σ. If we take this away, then the UMVUE property may be lost. For instance, if we take μ to be an integer, $0, \pm 1, \pm 2, \cdots,$, then all the values of μ are isolated, and not even a tiny interval is contained in the range of μ. And, alas, it may be shown that in this case, there is no UMVUE of μ.

Example 6.30. (UMVUE of Poisson Mean). Let $X_1, \cdots, X_n \overset{iid}{\sim} Poi(\lambda)$. In this case, $\bar{X}$ is minimal sufficient. Obviously, $E(\bar{X}) = \lambda$ and $\text{Var}(\bar{X}) = \frac{\lambda}{n} < \infty$. So, if $a < \lambda < b$ with $a < b$, then, by the Lehmann-Scheffe theorem, $\bar{X}$ is the UMVUE of λ.

Can we think of some other unbiased estimate of λ? Yes; for example, in the Poisson case, $E(s^2)$ is also λ, because mean and variance are equal for any Poisson distribution. So,



remarkably, in the iid Poisson case, the sample variance s^2 is also an unbiased estimate of λ; but $\bar{X}$ is a better unbiased estimate than s^2. In fact, $\bar{X}$ is the best unbiased estimate.

Example 6.31. (Constructing an UMVUE when it is not Obvious). This example illustrates the explicit Rao-Blackwell construction of the UMVUE when the UMVUE is not in plain sight; part (c) of the Lehmann-Schheffe theorem guides us to this Rao-Blackwell construction of the UMVUE.

Let again $X_1, \cdots, X_n \overset{iid}{\sim} Poi(\lambda)$, and consider estimation of $\theta = P_\lambda(X = 0) = e^{-\lambda}$. Easily, we can think of one unbiased estimate of θ (does not matter how silly it might be), namely, $U(X_1, \cdots, X_n) = I_{X_1=0}$. The minimal sufficient statistic is $T = \sum_{i=1}^n X_i$. Part (c) of the Lehmann-Scheffe theorem tells us what the UMVUE of θ will be; it is

$$\hat{\theta} = E[U\,|T] = P(X_1 = 0\,|T) = (1 - \frac{1}{n})^T;$$

(this formula was justified in great detail in Example 7.22) To conclude, for estimating the probability of no events on the basis of iid Poisson data, the unique UMVUE is $(1 - \frac{1}{n})^T$. Contrast this with the MLE of the probability of no events; the MLE is $e^{-\bar{X}}$. For $n = 10$, we have plotted both estimates, and you can see that they are almost the same as functions of $\bar{X}$.

6.3.2 Asymptotic Unbiasedness

We remarked earlier that exact unbiasedness is no longer considered to be important by most researchers. However, large biases are usually dangerous. A large bias will often mean that in many real experiments, the estimate will be unacceptably too far from the

true value of the parameter for moderate or even large sample sizes. Another problem is perception; estimates with large biases are treated with communal suspicion. It is common to demand that the estimate has a vanishing bias. Here is the formal definition.

Definition 6.9. Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \cdots, X_n)$ be a sequence of estimators of a parameter θ. We say that the sequence $\hat{\theta}_n$ is *asymptotically unbiased* if for any given $\theta \in \Theta$,

$$b_n(\theta) = E_\theta[\hat{\theta}_n] - \theta \to 0, \text{ as } n \to \infty.$$

Of course, exactly unbiased estimates are also asymptotically unbiased, because $b_n(\theta) \equiv 0$ for any n and any θ if $\hat{\theta}_n$ is already unbiased for all n. So, asymptotic unbiasedness is relevant for those estimates which are not quite exactly unbiased; e.g., if an MLE in some problem is seen to be biased, we would want to know if it is asymptotically unbiased. Let us see two examples.

Example 6.32. (Asymptotic Unbiasedness of MLE of Variance). Let $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. The MLE of σ^2 is biased (see Example 7.23). Its bias is

$$b_n(\mu, \sigma) = E[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2] - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}$$

$$\to 0, \text{ for any given } \sigma, \mu$$

(the bias does not depend on μ)
So, the MLE of σ^2 is biased, but asymptotically unbiased.

Example 6.33. (Asymptotic Unbiasedness in a Poisson Problem). Let $X_1, \cdots, X_n \overset{iid}{\sim} Poi(\lambda)$, and suppose we want to estimate λ^2. Now, by direct verification,

$$E[\bar{X}^2 - \frac{\bar{X}}{n}] = E[\bar{X}^2] - \frac{1}{n}E[\bar{X}]$$

$$= \text{Var}[\bar{X}] + [E(\bar{X})]^2 - \frac{1}{n}E[\bar{X}] = \frac{\lambda}{n} + \lambda^2 - \frac{\lambda}{n} = \lambda^2.$$

Therefore, by the Lehmann-Scheffe theorem, $\bar{X}^2 - \frac{\bar{X}}{n}$ is the UMVUE of λ^2. But this estimate has a little practical problem; it can take negative values, although λ^2 is strictly positive.
Instead, we may wish to use the *plug-in estimate* $\bar{X}^2$. However, the plug-in estimate is biased:

$$E[\bar{X}^2] = \frac{\lambda}{n} + \lambda^2,$$

and so, the bias is

$$b_n(\lambda) = \frac{\lambda}{n} + \lambda^2 - \lambda^2 = \frac{\lambda}{n}.$$

Note that for any $\lambda, \frac{\lambda}{n} \to 0$ as $n \to \infty$; i.e., the plug-in estimate is never negative, is biased, but asymptotically unbiiased.

Example 6.34. (A Classic Calculation of Fisher). We take a normal example again. Let $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. We know that the sample variance s^2 is the UMVUE of σ^2. But, the sample standard deviation s is not the UMVUE of σ. In fact, s is not even an unbiiased estimate of σ, so obviously cannot be the UMVUE! A simple application of Jensen's inequality shows that $E(s) = E(\sqrt{s^2}) < \sqrt{E(s^2)} = \sigma$.
Let us show that s is asymptotically unbiased. For this, we will need the familiar result

$$V = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}.$$

As a result, V has the density

$$f_V(v) = \frac{e^{-v/2} v^{\frac{n-3}{2}}}{2^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2})}, v > 0.$$

Interestingly, in the normal case, we can calculate $E(s)$ exactly. Indeed,

$$E(s) = \sigma E\left[\sqrt{\frac{V}{n-1}}\right] = \frac{\sigma}{\sqrt{n-1}} \int_0^\infty \sqrt{v} f_V(v) dv$$

$$= \frac{\sigma}{\sqrt{n-1}\, 2^{\frac{n-1}{2}} \Gamma(\frac{n-1}{2})} \int_0^\infty e^{-v/2} v^{\frac{n-2}{2}} dv$$

The integral $\int_0^\infty e^{-v/2} v^{\frac{n-2}{2}} dv$ is a Gamma integral, on making the change of variable $w = \frac{v}{2}$, and this results, on simplification, to the exact formula

$$E(s) = \sigma \frac{\sqrt{2}\Gamma(\frac{n}{2})}{\sqrt{n-1}\Gamma(\frac{n-1}{2})}.$$

We will now need to use *Stirling's approximation* to the Gamma function; for large x,

$$\Gamma(x) \sim e^{-x} x^{x-1/2} \sqrt{2\pi},$$

where the $\sim$ notation means that the quotient of the RHS and the LHS converges to 1 as $x \to \infty$.
If we use Stirling's approximation in the exact formula for $E(s)$ in place of $\Gamma(\frac{n}{2})$, as well as $\Gamma(\frac{n-1}{2})$, then after cancellations, we get the final neat result

$$E(s) = \sigma[1 - \frac{1}{8n} + o(\frac{1}{n})],$$

and hence the bias of s is

$$E(s) - \sigma = -\frac{\sigma}{8n} + o(\frac{1}{n}) \to 0,$$

as $n \to \infty$. Hence, the sample standard deviation s is asymptotically unbiased for the population standard deviation σ in the normal case. This is true for many nonnormal cases also.

6.4 Estimating Functions of Parameters

In practice, we often want to estimate functions of a basic parameter, because the function arises in a natural way. For example, for $Poi(\lambda), e^{-\lambda}$ is the probability of no events, $P(X = 0)$; in Bernoulli, $\frac{p}{1-p}$ represents the odds; in normal, $|\mu|$ represents the magnitude of the mean; again, in $N(\mu, 1), \Phi(a - \mu)$ represents $P(X < a)$, and so on. In some lucky cases, with ingenious calculations, we may be able to work out an UMVUE of the function; for example, $(1 - \frac{1}{n})^{\sum_{i=1}^{n} X_i}$ is the UMVUE of $e^{-\lambda}$ in the Poisson case (refer to Example 7.31).

A much faster strategy is to estimate a function $g(\theta)$ by the estimate $g(\hat{\theta})$, where $\hat{\theta}$ is the MLE of θ. In fact, for any function g, *one-to-one or not*, $g(\hat{\theta})$ is called the MLE of $g(\theta)$. We just plug-in $\hat{\theta}$ in place of θ within the formula of the function $g(\theta)$. Under mild conditions on g and the distribution from which we are sampling, this plug-in estimate works well, or even very well, and it is completely automatic. No new ingenious calculation is needed! You have to be careful about routine use of plug-in if you are simultaneously estimating many functions $g_1(\theta), \cdots, g_p(\theta)$, because the small errors of plug-in for each individual g_i may accumulate to give you a collectively poor plug-in estimate.

6.4.1 Plug-in Estimators

Definition 6.10. Let $X^{(n)} = (X_1, \cdots, X_n) \sim P_\theta = P_{\theta,n}$. Suppose a unique MLE $\hat{\theta} = \hat{\theta}_n(X_1, \cdots, X_n)$ of θ exists. The parametric plug-in estimate of the function $g(\theta)$ is the estimate $g(\hat{\theta})$.

Note that the definition of the parametric plug-in estimate does not assume that the observations are iid; this is because maximum likelihood is a general principle, and is applicable beyond the iid case.

We can also define plug-in estimates without restricting ourselves to parametric models. Such an extension of the plug-in idea is very useful. But, we will define such *nonparametric plug-in estimates* only in the iid case. Let us recall a little background. If F is any CDF on the real line, and we treat F itself as an unknown parameter, we have commented and witnessed before that the empirical CDF F_n is a very effective estimate of F. So, in the nonparametric world, F_n plays the role of the estimate of the basic parameter, similar to the MLE in the parametric world. Now, functions of the parameter will be functions, *rather functionals*, of F (we will see examples), say $h(F)$. Then, the idea is to plug-in F_n for F in $h(F)$. Here is the definition.

Definition 6.11. Let $X_1, \cdots, X_n \overset{iid}{\sim} F$, a CDF on the real line. Let F_n be the empirical CDF of the sample values, and let $h(F)$ be a scalar functional of F. A *nonparametric plug-in estimator* of $h(F)$ is the statistic $h(F_n)$.

Remark: Nonparametric plug-in estimators in this form originated in the work of von

Mises (von Mises (1947)); as such, they are also called *von Mises functionals*. You use nonparametric plug-in estimators when you are not comfortable assuming a parametric model. If you are happy with a parametric model, use the parametric plug-in estimator. In either case, the principle is plug-in.

Let us get the ideas clarified by seeing examples.

Example 6.35. (Parametric Plug-in Estimators). *Suppose* $X_1, \cdots, X_n \overset{iid}{\sim} Exp(\lambda)$*; recall that* λ *is the mean of the distribution. Sometimes, we want to estimate* $\frac{1}{\lambda}$*, because it measures the rate at which events occur. The MLE of* λ *is* $\bar{X}$*; so the parametric plug-in estimator of* $\frac{1}{\lambda}$ *is* $\frac{1}{\bar{X}}$*.*

Suppose $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, 1)$*. We want to know what proportion of the population exceeds a speciific number* a*. Thus, we want to estimate*

$$P(X > a) = 1 - P(X \leq a) = 1 - \Phi(a - \mu) = \Phi(\mu - a).$$

The parametric plug-in estimator would be $\Phi(\bar{X} - a)$*.*

Suppose $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ *and suppose we want to estimate the coefficient of variation* $\frac{\mu}{\sigma}$*. The MLE of* μ *is* $\bar{X}$*; the MLE of* σ *is* $s_0 = \sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}$*. A parametric plug-in estimator would be* $\frac{\bar{X}}{s_0}$*; some people prefer to use* $\frac{\bar{X}}{s}$*, where* $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$*. As estimators, the two are essentially the same unless* n *is very small. But* $\frac{\bar{X}}{s}$ *has some technical advantage when we come to the distribution theory for finite* n*.*

Here is one final example. Suppose $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu_1, 1)$ *and* $Y_1, \cdots, Y_m \overset{iid}{\sim} N(\mu_2, 1)$*. Usually we also assume all* $n + m$ *observations are independent. Suppose we want to estimate* $\max(\mu_1, \mu_2)$*; a parametric plug-in estimator would be* $\max(\bar{X}, \bar{Y})$*.*

Example 6.36. (Nonparametric Plug-in Estimators). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} F$. Let $h(F) = E_F(X)$; the nonparametric plug-in estimator is $h(F_n) = E_{F_n}(X) = \frac{X_1}{n} + \cdots + \frac{X_n}{n} = \bar{X}$, because F_n places probability $\frac{1}{n}$ at each data value.

Suppose $h(F) = P_F(X \leq c)$, where c is some fixed number. The nonparametric plug-in estimator is $h(F_n) = P_{F_n}(X \leq c) = F_n(c)$, the fraction of data values that are c or less.

Suppose $h(F) = F^{-1}(\frac{1}{2})$, i.e., the population median. The nonparametric plug-in estimator is $h(F_n) = F_n^{-1}(\frac{1}{2})$, i.e., the sample median.

Suppose $h(F) = E_F[XI_{F^{-1}(\alpha) \leq X \leq F^{-1}(1-\alpha)}]$; in words, take the population and chop off $100\alpha\%$ of the values from the lower tail and also $100\alpha\%$ of the values from the upper tail (i.e., ignore the *extreme individuals of the population*), and find the mean of the middle $100(1 - 2\alpha)\%$ of the population values. The nonparametric plug-in estimator is

$$h(F_n) = \frac{1}{n(1-2\alpha)} \sum_{i=n\alpha+1}^{n-n\alpha} X_i,$$

which is called a *trimmed sample mean*.

6.4.2 Bias Correction of Plug-in Estimates

Plug-in estimators *usually* have a small bias. Thus, even if an MLE $\hat{\theta}$ was an unbiased estimate of $\theta, g(\hat{\theta})$ would not be an unbiased estimate of $g(\theta)$, except when g is a linear function. In such cases, one sometimes does a *bias correction of the plug-in estimator*. Opinions differ on whether it is necessary to do a bias correction. But the method of doing the bias correction is an interesting and educational process, and you should be familiar with it. We will explain it at a general conceptual level. There are newer *nonparametric bias correction techniques, for example by using Efron's bootstrap. This is treated in a later chapter.*

We start with an illustrative example.

Example 6.37. (Bias Correction of MLE). Suppose based on $X_1, \cdots, X_n \overset{iid}{\sim} Poi(\lambda)$, we wish to estimate $g(\lambda) = e^{-\lambda} = P(X = 0)$. The parametric plug-in estimator is $e^{-\bar{X}}$, which is not unbiased. To find a way to do a bias correction, do a formal Taylor expansion:

$$e^{-\bar{X}} = g(\bar{X}) \approx g(\lambda) + (\bar{X} - \lambda)g'(\lambda) + \frac{(\bar{X} - \lambda)^2}{2}g''(\lambda)$$

$$= e^{-\lambda} - (\bar{X} - \lambda)e^{-\lambda} + \frac{(\bar{X} - \lambda)^2}{2}e^{-\lambda},$$

and formally, taking an expectation on both sides,

$$E[e^{-\bar{X}}] \approx e^{-\lambda} + \frac{E[(\bar{X} - \lambda)^2]}{2}e^{-\lambda} = e^{-\lambda} + \frac{\lambda e^{-\lambda}}{2n},$$

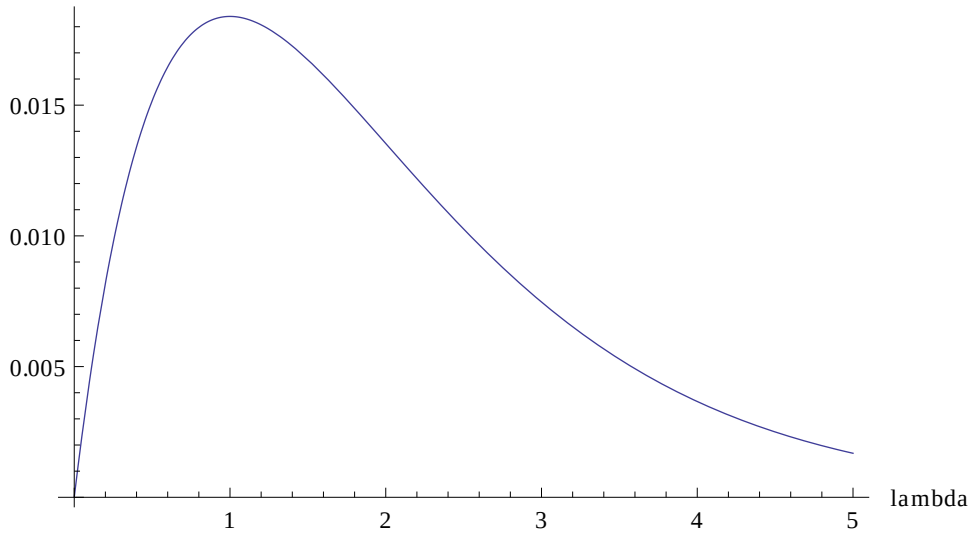
so that the bias

$$E[e^{-\bar{X}}] - e^{-\lambda} \approx \frac{\lambda e^{-\lambda}}{2n}.$$

Plugging in $\bar{X}$ for λ *again*, an estimate of the bias is $\frac{\bar{X}e^{-\bar{X}}}{2n}$, and if we subtract off this estimated bias, we get the bias-corrected estimate

$$e^{-\bar{X}} - \frac{\bar{X}e^{-\bar{X}}}{2n} = e^{-\bar{X}}[1 - \frac{\bar{X}}{2n}].$$

Note that the correction term $\frac{\bar{X}e^{-\bar{X}}}{2n}$ is of a smaller order, namely, because of the n in the denominator, the correction term is of the smaller order of $\frac{1}{n}$; the bias is small in large samples, and so the bias correction term is also a smaller order term in large samples. One other important point is that without a bias correction, the bias of the estimate $e^{-\bar{X}}$ is of the order of $\frac{1}{n}$; *after you do the bias correction, the* $\frac{1}{n}$ *term in the bias cancels, and the bias now becomes of the order of* $\frac{1}{n^2}$. The bias of the estimate $e^{-\bar{X}}$ is plotted here for $n = 10$.



Example 6.38. (Multiplicative Bias Correction). If $X_1, \cdots, X_n$ are iid $N(\mu, \sigma^2)$, then the sample standard deviation s is a biased estimate of σ. In Example 7.34, we worked out the approximation

$$E(s) = \sigma[1 - \frac{1}{8n} + o(\frac{1}{n})].$$

A bias correctioon in this case would lead to the estimate $\frac{s}{1-\frac{1}{8n}} \approx s[1 + \frac{1}{8n}]$. The bias of this estimate would be of the lower order $O(\frac{1}{n^2})$.

Here is a simple enough general theorem that will help you do bias-correction whenever the MLE is a sample mean. In Exponential families, if we do a mean parametrization, then the MLE is $\frac{1}{n}\sum_{i=1}^n T(X_i)$, which is a sample mean (not the mean of the X_i, but still a mean). So, as long as you are careful to parametrize by using $E(T(X))$ as your parameter, the theorem below will be useful in doing a bias correction.

Theorem 6.10. Let $X_1, X_2, \cdots, X_n$ be iid observations with mean μ, variance $\sigma^2 = V(\mu)$, and with a finite fourth moment. Consider any function $g(\mu)$ which has four uniformly bounded derivatives. Then, the bias of the plug-in estimator $g(\bar{X})$ is of the form

$$\frac{g''(\mu)\sigma^2}{2n} + O(\frac{1}{n^2}),$$

and the bias corrected estimator $g(\bar{X}) - \frac{g''(\bar{X})V(\bar{X})}{2n}$ has a bias of the lower order $O(\frac{1}{n^2})$.

Remark: In practice, the theorem is used *even if g does not have four bounded derivatives.* Four bounded derivatives are needed to prove the theorem. But the recipe may work even if the assumption of four bounded derivatives does not hold.

Here is another example.

Example 6.39. (The logarithmic Transformation). Making transformations of data when the underlying distribution is skewed is common practice in statistics. Among the most ubiquitous transformations is the log transformation. For example, if X is Exponential with mean λ, then $Y = \log X$ has mean approximately $\log \lambda$. The MLE of $\log \lambda$ based on n iid observations from $Exp(\lambda)$ is $\log(\bar{X})$; but $\log(\bar{X})$ has some bias as an estimator of $\log \lambda$. It is easy to prove (try it!) that $E[\log \bar{X}] < \log \lambda$. How do we correct the bias? We will use Theorem 7.10 to accomplish the bias correction.
Here, we identify μ with λ, $\sigma^2 = \lambda^2$ with $V(\lambda) = \lambda^2$ and $g(\lambda)$ of course is $\log \lambda$. Thus, $g''(\lambda) = -\frac{1}{\lambda^2}$. So, following the prescription of Theorem 7.10, the bias-corrected estimator of $g(\lambda) = \log \lambda$ is

$$g(\bar{X}) - \frac{g''(\bar{X})V(\bar{X})}{2n} = \log \bar{X} + \frac{\frac{1}{\bar{X}^2}\bar{X}^2}{2n} = \log \bar{X} + \frac{1}{2n},$$

a very simple bias-correction.

6.5 Best Attainable Accuracy

Students often ask the question *what is the best possible estimator in this particular problem I have*? The answer is that there are never any best possible estimators in the class of all estimators. As you would recall from our discussions in Chapter 6, risk functions (e.g., MSE) of reasonable estimators usually cross, and so no one estimator is uniformly the best among all possible estimators. There *may be* best estimators if we suitably restrict our class of estimators. For example, we may restrict our estimators to only those which are linear in the sample observations, $\sum_{i=1}^{n} c_i X_i$, or we may restrict ourselves to estimators which are linear functions of the MLE, $a + b\hat{\theta}$, where $\hat{\theta}$ is the MLE, or, we may decide to restrict ourselves to unbiased estimators, etc. How are we going to know, once we have decided what sort of restricted class of estimators we will consider, how good can the best estimator be? Obviously, we have to define what *good* means. At the outset, we make up our mind that a good estimator is one with a small MSE. Now recall that MSE is always the sum of the variance and the square of the bias. In *regular problems*, the square of the bias of well behaved estimates is a *lower order term* than the variance. More precisely, variance is of the order of $\frac{1}{n}$, but the square of the bias is of the lower order $\frac{1}{n^2}$. So, at least in large samples, for well behaved estimates, variance is the more important term to look at than the square of the bias. A celebrated inequality in statistics places a lower bound on the variance of well behaved estimates in regular estimation problems. The lower bound is the gold standard of variance; it effectively tells you how low can the variance at all be. This is the *Cramér-Rao inequality*, also known as the *information inequality* because it involves the Fisher information function in its statement. It was proved independently by C.R. Rao and Harald Cramér, (Rao (1945), Cramér (1946)). The Cramér-Rao lower

bound (CRLB) is fundamental on its own; but even more, it has inspired numerous further deep and path breaking work in statistics. For the inequality to be true, we need some regularity conditions.

6.5.1 Cramér-Rao Inequality

The purpose of the Cramér-Rao inequality is to place a lower bound on $\mathrm{Var}_\theta[T(X_1, \cdots, X_n)]$ for suitable estimates (statistics) $T(X_1, \cdots, X_n)$. The Cramér-Rao inequality connects Fisher information to the best attainable accuracy in an estimation problem, and essentially says that larger the (Fisher) information, more accurate will be certain estimates, e.g., the MLE, or a best unbiased estimate. Through this direct connection of attainable accuracy to Fisher information, it lends credibility to Fisher information as a well conceived measure of information coming from a statistical experiment.

For simplicity, we will treat only the iid case, although it is possible to state the Cramér-Rao Inequality for more general situations. We will make assumptions on both the underlying family of distributions and the estimate T that we want to look at.

Fortunately, the assumptions on the distribution are the same as what we imposed in order that we could define Fisher information for the family. Thus, we suppose that $f(x\,|\theta)$ is a one parameter family of densities (or pmfs) satisfying the three regularity conditions **A1, A2, A3** of Section 7.2.5, and our observations are iid from $f(x\,|\theta)$.

As regards the estimate $T(X_1, \cdots, X_n)$, we will assume:

Regularity Conditions B

B1 $\mathrm{Var}_\theta[T(X_1, \cdots, X_n)] < \infty$ for all $\theta \in \Theta$.

B2 For all $\theta \in \Theta$, $E_\theta(T)$ is differentiable in θ, and the derivative satisfies

$$\frac{d}{d\theta} \int T(x_1, \cdots, x_n) \prod_{i=1}^{n} f(x_i\,|\theta) dx_1 \cdots dx_n = \int T(x_1, \cdots, x_n)\, [\frac{\partial}{\partial \theta} \prod_{i=1}^{n} f(x_i\,|\theta)] dx_1 \cdots dx_n,$$

i.e., the order of differentiation and integration can be changed. Note that assumption **A3** is the same as assumption **B2** for the special statistic $T \equiv 1$.

Remark: Colloquially, we call an estimate T *well behaved* if **B1**, **B2** hold. On the face of it, we have to verify **B1**, **B2** for the particular T we may have in mind. But this would be a hassle if we have many estimates in mind simultaneously. So, it would be useful to know if we can guarantee **B1**, **B2** for large classes of estimates T at the same time. It turns out that sometimes we can. We may push the burden of nicety of T to the density f by insisting that f be not just nice, *but extra nice*! Here is one such result.

(Proposition) If **B1** holds, then **B2** automatically holds for all $f(x\,|\theta)$ in the regular one parameter Exponential family.

A proof of this requires use of measure theory, and we will omit it. But the proposition says that we can apply the Cramér-Rao lower bound in any regular Exponential family

situation (although those are *NOT* the only ones).
Here is the one parameter Cramér-Rao lower bound; the multiparameter case is delayed till the supplementary section of this chapter.

Theorem 6.11. (Cramér-Rao Inequality). Suppose regularity conditions **A1**, **A2**, **A3**, **B1**, **B2** hold. Let $I(\theta)$ denote the Fisher information function corresponding to the family $f(x\,|\theta)$. Then, for any given n, and for all $\theta \in \Theta$,

$$\mathrm{Var}_\theta[T(X_1, \cdots, X_n)] \geq \frac{[\frac{d}{d\theta} E_\theta(T)]^2}{nI(\theta)}.$$

Proof: The proof is deceptively simple and uses the fact that whenever a correlation between any two random variables is defined, it must be between $+1$ and -1. The task is to cleverly construct the two random variables whose correlation we will look at. Here are the two random variables:

$$U = U(\theta) = \frac{d}{d\theta} \log l(\theta), \;\; V = T(X_1, \cdots, X_n).$$

Note that we are familiiar with U; it is the score function, but considered as a random variable with θ held fixed.
First we compute $\mathrm{Cov}_\theta(U, V)$. To reduce notational clutter, the subscript θ is dropped below:

$$\mathrm{Cov}(U, V) = E(UV) - E(U)E(V) = E(UV)$$

(because, under our regularity assumptions, $E(U) = 0$)

$$= \int [\frac{d}{d\theta} \log l(\theta)]\, T(x_1, \cdots, x_n) \prod_{i=1}^{n} f(x_i\,|\theta) dx_1 \cdots dx_n$$

$$= \int \frac{\frac{d}{d\theta} l(\theta)}{l(\theta)}\, T(x_1, \cdots, x_n) \prod_{i=1}^{n} f(x_i\,|\theta) dx_1 \cdots dx_n$$

$$= \int \frac{\frac{d}{d\theta} [\prod_{i=1}^{n} f(x_i\,|\theta)]}{\prod_{i=1}^{n} f(x_i\,|\theta)}\, T(x_1, \cdots, x_n) \prod_{i=1}^{n} f(x_i\,|\theta) dx_1 \cdots dx_n$$

$$= \int \frac{d}{d\theta} [\prod_{i=1}^{n} f(x_i\,|\theta)\, T(x_1, \cdots, x_n)] dx_1 \cdots dx_n = \frac{d}{d\theta} \int T(x_1, \cdots, x_n) \prod_{i=1}^{n} f(x_i\,|\theta) dx_1 \cdots dx_n$$

(by virtue of regularity condition **B2**)

$$= \frac{d}{d\theta} E_\theta(T).$$

Therefore, the correlation between U and V is:

$$\rho = \rho_{U,V} = \frac{\text{Cov}(U,V)}{\sqrt{\text{Var}(U)\,\text{Var}(V)}} = \frac{\frac{d}{d\theta}\,E_\theta(T)}{\sqrt{(nI(\theta))\,\text{Var}(T)}}.$$

(recall that in the iid case, the variance of the score functioon is $nI(\theta)$)
Now, necessarily, $\rho^2 \leq 1$, which says

$$\frac{[\frac{d}{d\theta}\,E_\theta(T)]^2}{(nI(\theta))\,\text{Var}(T)} \leq 1$$

$$\Rightarrow \text{Var}(T) \geq \frac{[\frac{d}{d\theta}\,E_\theta(T)]^2}{nI(\theta)},$$

as claimed.

Remark: If we inspect the proof, we realize that the Cramér-Rao inequality becomes an exact equality if and only if $\rho^2 = 1$, i.e., U, V are exactly linearly related. This happens in the entire Exponential family provided the particular T is the minimal sufficient statistic $\frac{1}{n}T(X_i)$ of that Exponential family density (consult Chapter 5 for notation again, if needed). We will pick up this matter again in the next section.

Note that there is no assumption anywhere in the Cramér-Rao inequality that T is an unbiased estimate of some specific parameter you have in mind. The Cramér-Rao inequality says T is any well behaved statistic, and its expectation is whatever it is. But what if T *is an unbiased estimate of the parameter* θ? In that case, $E_\theta(T) = \theta$, and so, $\frac{d}{d\theta}\,E_\theta(T) = 1$. This immediately leads to the following productive corollary.

Corollary 6.1. Suppose all the required regularity conditions hold. If T is an unbiased estimator of θ, then

$$(a)\,\forall\,\theta \in \Theta,\ \text{Var}_\theta(T) \geq \frac{1}{nI(\theta)};$$

(b) If a special unbiased estimator T_0 of θ attains the Cramér-Rao lower bound, i.e.,

$$\text{Var}_\theta(T_0) = \frac{1}{nI(\theta)}\ \forall\,\theta \in \Theta,$$

then T_0 is the UMVUE of θ.

Let us see some examples.

Example 6.40. (Cramér-Rao Lower Bound in Binomial). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} Ber(p), 0 < p < 1$, and let $X = \sum_{i=1}^n X_i$, the minimal sufficient statistic. For general a, b, consider estimates (statistics) $T = a + bX$. We have $E_p(T) = a + bnp$, so that $\frac{d}{dp}E_p(T) = bn$. Also, recall that $I(p) = \frac{1}{p(1-p)}$. So, the CRLB (Cramér-Rao Lower Bound) on the variance of T is:

$$\text{Var}_p(T) \geq \frac{(bn)^2}{\frac{n}{p(1-p)}} = nb^2p(1-p).$$

We can certainly calculate $\text{Var}_p(T)$ exactly! Indeed,

$$\text{Var}_p(T) = \text{Var}_p(a + bX) = b^2(np(1-p)) = nb^2p(1-p).$$

So, for all estimates of the form $a + bX$, the CRLB is attained, and hence $a + bX$ is the UMVUE of $E(a + bX) = a + bnp$. In particular, if we take $a = 0, b = \frac{1}{n}$, we get the result that $\frac{X}{n}$ is the UMVUE of p; we may also derive this same result by using the Lehmann-Scheffe theorem. Thus, the Lehmann-Scheffe theorem and the CRLB act as complementary tools in this example.

Example 6.41. (Cramér-Rao Lower Bound in Normal). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where σ is considered known. First consider $T(X_1, \cdots, X_n) = \bar{X}$. It is obviously an unbiased estimate of μ and its exact variance is $\frac{\sigma^2}{n}$. The CRLB on the variance of a general unbiased estimate of μ is

$$\frac{1}{nI(\mu)} = \frac{1}{n\frac{1}{\sigma^2}} = \frac{\sigma^2}{n}.$$

So the particular unbiased estimate $\bar{X}$ attains the CRLB on the variance of an arbitrary unbiased estimate of μ, which means that $\bar{X}$ is the UMVUE of μ; once again, this can also be proved by using the Lehmann-Scheffe theorem.

Now, let us change the problem to estimation of μ^2. We will show that μ^2 does have an UMVUE, but the UMVUE *does not attain the CRLB* on the variance of unbiased estimates of μ^2. In other words, the Cramér-Rao technique works fine for identifying the UMVUE of μ, but it will fail to identify the UMVUE of μ^2, because no unbiased estimate of μ^2, including its UMVUE, hits the CRLB.

Directly, by applying the Lehmann-Scheffe theorem, the UMVUE of μ^2 is $\bar{X}^2 - \frac{\sigma^2}{n}$ (note σ is assumed known). By using standard formulas for moments of a normally distributed random variable,

$$\text{Var}(\bar{X}^2 - \frac{\sigma^2}{n}) = \text{Var}(\bar{X}^2) = E(\bar{X}^4) - [E(\bar{X}^2)]^2$$

$$= \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2}.$$

On the other hand, the CRLB for the variance of an arbitrary unbiased estimates of μ^2 is

$$\frac{[\frac{d}{d\mu}\mu^2]^2}{nI(\mu)} = \frac{4\mu^2}{n\frac{1}{\sigma^2}} = \frac{4\mu^2\sigma^2}{n}.$$

By comparing the exact variance of $\bar{X}^2 - \frac{\sigma^2}{n}$ above with this CRLB, we see that the exact variance is always larger. The exact variance has the extra term $\frac{2\sigma^4}{n^2}$. Thus, although $\bar{X}^2 - \frac{\sigma^2}{n}$ is the UMVUE, as per the Lehmann-Scheffe theorem, it still does not attain the CRLB for this problem.

6.5.2 Cramér-Rao Inequality Attained?

The two examples above give us hints about when the Cramér-Rao inequality is exactly attained. Precisely, here is the question; for which parametric families $f(x\,|\theta)$, and parametric functions $\tau(\theta)$, does there exist an unbiased estimate $V(X_1, \cdots, X_n)$ of $\tau(\theta)$ such that $\text{Var}_\theta(V) = \frac{[\frac{d}{d\theta}\tau(\theta)]^2}{nI(\theta)}$ for all $\theta \in \Theta$?
There is a crisp and somewhat disappointing answer to this.

Theorem 6.12. Let $V(X_1, \cdots, X_n)$ be a statistic with finite variance. Denote $E_\theta(V) = \tau(\theta)$. The equality

$$\text{Var}_\theta(V) = \frac{[\frac{d}{d\theta}\tau(\theta)]^2}{nI(\theta)}$$

holds for all $\theta \in \Theta$ if and only if
(a) $f(x\,|\theta)$ is an one parameter regular exponential family, $f(x\,|\theta) = e^{\eta(\theta)T(x)-\psi(\theta)}h(x)$;
(b) $V(X_1, \cdots, X_n) = a + b\frac{\sum_{i=1}^n T(X_i)}{n}$ for some constants a, b (possibly depending on n);
(c) $\tau(\theta) = a + bE_\theta[T(X)]$ for the same two constants a, b.
Remark: *The theorem says that equality in the Cramér-Rao inequality is attained only for the very narrow problem of estimating the mean of the natural sufficient statistic in regular Exponential families.*
This is the same as saying that equality is attained only when the score function $\frac{d}{d\theta}\log l(\theta)$ is a linear function of V:

$$\frac{d}{d\theta}\log l(\theta) = A(\theta) + B(\theta)V(X_1, \cdots, X_n),$$

for all θ, and $X_1, \cdots, X_n$. In the Exponential families, this happens, as you can easily verify, with $V(X_1, \cdots, X_n) = \frac{\sum_{i=1}^n T(X_i)}{n}$ (or, obviously, any linear function of it, $a + b\frac{\sum_{i=1}^n T(X_i)}{n}$). See Wijsman (1973) for a proof; also see Woodroofe and Simmons (1983) regarding validity almost everywhere of the Cramér-Rao inequality without rrequiring regularity conditions.

6.6 At Instructor's Discretion

6.6.1 Fisher Information Matrix

When there is more than one parameter in the model, Fisher information is defined in essentially the same way as in the one parameter case. Moreover, it has exactly the same connection to the best attainable accuracy as it does in the one parameter case, and this is the content of the multiparameter Cramér-Rao inequality.
First we define Fisher information in the multiparameter case. If the parameter θ is p-dimensional, then Fisher information becomes a $p \times p$ matrix. The (i, j) element of the matrix is defined as

$$I_{ij}(\theta) = -E_\theta[\frac{\partial^2}{\partial\theta_i\,\partial\theta_j}\log f(x\,|\theta)].$$

Of course, in order that this definition makes sense, we assume that the second order partial derivatives all exist, and we also assume that $I_{ij}(\theta)$ itself, which is an expectation, exists. The Fisher information matrix is the $p \times p$ matrix $I(\theta)$ with elements $I_{ij}(\theta), i, j = 1, 2, \cdots, p$. The Fisher information matrix is a nonnegative definite matrix, and in examples, it often is positive definite. In some degenerate cases, it may have rank $< p$. In theoretical work with general densities, we also often impose the assumption that $I(\theta)$ is positive definite, so that we can talk aboout its inverse.
Let us see a few examples.

Example 6.42. (Two Parameter Normal). Consider the general two parameter normal density

$$f(x\,|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}, -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0.$$

On easy calculation, the Fisher information matrix is

$$I(\mu, \sigma) = \begin{pmatrix} \frac{2}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{pmatrix}.$$

We immediately see that the matrix is positive definite. Note that if we change the parametrization to μ, σ^2, the information matrix will change. Note another remarkable property of the information matrix $I(\mu, \sigma)$ above; it is diagonal. When this happens, we refer to the parameters as being *orthogonal*. The diagonal structure in the information matrix is related to the fact that the MLEs of μ, σ are independently distributed in the normal case. If the parameters were not orthogonal to begin with, there are certain mathematical advantages in transforming them to orthogonal parameters. See Reid (2003).

Example 6.43. (Two Parameter Beta). Consider the general two parameter Beta density

$$f(x|\alpha, \beta) = x^{\alpha-1}(1-x)^{\beta-1}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)},$$

$0 < x < 1, \alpha, \beta > 0$.
The log likelihood function is

$$\log f(x|\alpha, \beta) = \log\Gamma(\alpha+\beta) - \log\Gamma(\alpha) - -\log\Gamma(\beta) + (\alpha-1)\log x + (\beta-1)\log(1-x).$$

Let

$$\psi(x) = \frac{d}{dx}\log(\Gamma(x)), \psi'(x) = \frac{d}{dx}\psi(x).$$

Then, on simple calculation, the Fisher information matrix is

$$I(\alpha, \beta) = \begin{pmatrix} \psi'(\alpha) - \psi'(\alpha+\beta) & -\psi'(\alpha+\beta) \\ -\psi'(\alpha+\beta) & \psi'(\beta) - \psi'(\alpha+\beta) \end{pmatrix}.$$

Example 6.44. (Multivariate Normal). This example requires familiarity with derivatives of functions of a matrix with respect to the matrix. You may skip it, if you find it too technical. Consider the p-variate normal distribution $N_p(\mathbf{0}, \Sigma)$, where Σ is assumed positive definite. Then,

$$f(x\,|\Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-x'\Sigma^{-1}x/2}, x \in \mathcal{R}^p.$$

This gives,

$$\log f(x\,|\Sigma) = c - \frac{1}{2}\log|\Sigma| - \frac{x'\Sigma^{-1}x}{2}.$$

On using the matrix theoretic derivative formulas

$$\frac{d}{d\Sigma}\log|\Sigma| = \Sigma^{-1},\ \frac{d}{d\Sigma}(x'\Sigma^{-1}x) = -\Sigma^{-1}xx'\Sigma^{-1},$$

we get

$$\frac{d}{d\Sigma}\log f(x\,|\Sigma) = -\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}xx'\Sigma^{-1},$$

and

$$\frac{d^2}{d\Sigma^2}\log f(x\,|\Sigma) = \frac{1}{2}\Sigma^{-2} - \Sigma^{-3/2}xx'\Sigma^{-3/2}.$$

This gives

$$\begin{aligned} -E[\frac{d^2}{d\Sigma^2}\log f(x\,|\Sigma)] &= -\frac{1}{2}\Sigma^{-2} + \Sigma^{-3/2}\Sigma\Sigma^{-3/2} \\ &= -\frac{1}{2}\Sigma^{-2} + \Sigma^{-2} = \frac{1}{2}\Sigma^{-2}, \end{aligned}$$

which is the Fisher information matrix. Compare this to the Fisher information function we derived in Example 7.14 in the univariate $N(0, \sigma^2)$ case by treating $\theta = \sigma^2$ (rather than σ) as the parameter; the answer we found there was that the information function is equal to $\frac{1}{2\theta^2}$. In the p-dimensional case, the matrix Σ takes the place of θ and you can see that the p-dimensional formula for Fisher information is exactly analogous to the univariate Fisher information formula.

6.6.2 Cramér-Rao Inequality for Many Parameters

The Cramér-Rao inequality for the case of p parameters is very similar to the one parameter Cramér-Rao inequality, except the notation is more involved, and some matrix theory concepts are needed to state it. The conceptually important thing is that its message is the same as that of the one parameter Cramér-Rao inequality; larger the information, better your ability to do accurate estimation.

As in the one parameter case, we need regularity conditions on both $f(x\,|\theta)$ and the estimate (statistic) $T(X_1, \cdots, X_n)$. These regularity conditions parallel the one parameter case conditions. Because θ is now a vector parameter, we have partial derivatives

$\frac{\partial}{\partial\theta_i}E_\theta(T), i = 1, 2, \cdots, p$ to deal with, and exactly as in the one parameter case, every regularity condition holds throughout the multiparameter regular Exponential family and for any statistic $T(X_1, \cdots, X_n)$ which has a finite variance under each θ. The exact regularity conditions may be seen in Lehmann and Casella (1998).

We need one more notation before we can show the multiparameter Cramér-Rao inequality. Given a statistic $T(X_1, \cdots, X_n)$, denote the gradient vector of $E_\theta(T)$ as $\nabla_{\theta,T}$, i.e., $\nabla_{\theta,T}$ is a p-dimensional vector whose ith coordinate is the partial derivative $\frac{\partial}{\partial\theta_i}E_\theta(T)$. We also assume that the $p \times p$ Fisher information matrix is nonsingular for all θ and denote its inverse as $I^{-1}(\theta)$.

Theorem 6.13. (Multiparameter Cramér-Rao Inequality). Let $X_1, \cdots, X_n \overset{iid}{\sim} f(x\,|\theta), \theta \in \Theta \subseteq \mathcal{R}^p, 1 \le p < \infty$. Under regularity conditions, for a scalar statistic $T(X_1, \cdots, X_n)$,

$$\mathrm{Var}_\theta(T) \ge \frac{[\nabla_{\theta,T}]' I^{-1}(\theta)[\nabla_{\theta,T}]}{n},$$

for any given n and for all $\theta \in \Theta$.

6.6.3 Moment Estimates

Moment estimates are special types of plug-in estimates in parametric families, and in some problems the method can produce quick closed form reasonable estimates when more sophisticated estimates, like the MLE, are much harder to find. If all you want is a quick estimate and great accuracy is not needed in some specific problem, you may want to compute a moment estimate.

Suppose $f(x\,|\theta)$ is a p-parameter density or pmf; thus, $\theta = (\theta_1, \cdots, \theta_p)$. Suppoose we have iid observations $X_1, \cdots, X_n$ from f. We choose p linearly independent functions $h_1(X), h_2(X), \cdots, h_p(X)$ which have finite expectations; i.e., for each i, $E_\theta[h_i(X)]$ exists under all θ. Then, we form a system of p simultaneous equations in $\theta_1, \cdots, \theta_p$ by using the plug-in principle. The equations are:

$$E_\theta[h_1(X)] = \frac{1}{n}\sum_{i=1}^n h_1(X_i), \cdots, E_\theta[h_p(X)] = \frac{1}{n}\sum_{i=1}^n h_p(X_i).$$

On solving these p equations simultaneously, we get a unique solution

$$(\hat{\theta}_1(X_1, \cdots, X_n), \cdots, \hat{\theta}_p(X_1, \cdots, X_n))$$

these are called the *moment estimates* of $\theta_1, \cdots, \theta_p$. In practice, the functions $h_1, \cdots, h_p$ are often chosen to be the powers, $h_1(X) = X, h_2(X) = X^2, \cdots, h_p(X) = X^p$. But you could choose them to be some other functions; see Bickel and Doksum (2006) for more on choice of $h_1, \cdots, h_p$.

Let us see a few examples.

Example 6.45. (Two Parameter Gamma). Suppose $X_1, \cdots, X_n \overset{iid}{\sim} G(\alpha, \lambda)$. Then, $E(X) = \alpha\lambda, \text{Var}X = \alpha\lambda^2$. The moment estimates are found by solving the equations

$$\alpha\lambda = \bar{X};\ \alpha\lambda^2 = s^2.$$

Obviously, the unique solutions are

$$\lambda = \frac{s^2}{\bar{X}},\ \alpha = \frac{(\bar{X})^2}{s^2}.$$

Compare the ease of finding these moment estimates to the MLEs of α, λ. This is a two parameter regular Exponential family. Therefore, the unique MLEs are solutions of the likelihood equations, provided the likelihood equations do have solutions (see Section 7.3.2). If we let $P = \prod x_i$, then on some calculation, the likelihood equations are found to be

$$0 = \frac{\partial \log l}{\partial \alpha} = \log P - n \log \lambda - n\Psi(\alpha),$$
$$0 = \frac{\partial \log l}{\partial \lambda} = \frac{\sum x_i}{\lambda^2} - \frac{n\alpha}{\lambda},$$

where $\Psi(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}$. These are transcendental equations in α, λ, and can only be solved by numerical means. As a result, the MLEs have no closed form formula in this case, and it is difficult to study their properties theoretically for given fixed n. On the other hand, for most data sets, once you numerically compute these MLEs, they will be more accurate than the closed form moment estimates.

Example 6.46. (Moment Estimate of Binomial N). We usually know the value of the n parameter in a binomial problem. In some interesting applied problems, the n parameter, as well as the p parameter is unknown. One example is the problem of estimating the number of errors (bugs) in a piece of software. If the errors have a probability p of being detected by a reader, mutually independently, then we have a $Bin(N, p)$ problem with both N, p unknown.

This is an extremely difficult problem. In particular, MLEs may not exist! DasGupta and Rubin (2004) reviews this difficult problem. Two earlier reviews are Olkin, Petkau, and Zidek (1981) and Hall (1994).

Suppose we have iid observations $X_1, \cdots, X_k$ from our $Bin(N, p)$ distribution. Then, moment estimates solve the two equations

$$Np = \bar{X};\ Np(1-p) = s^2.$$

The unique solutions are

$$p = 1 - \frac{s^2}{\bar{X}};\ N = \frac{\bar{X}^2}{\bar{X} - s^2}.$$

It is possible that for some specific data set $\bar{X} \leq s^2$. In that case, the moment estimates would have to be adjusted. It is unlikely that this will happen unless k is too small compared to the true value of N.

(b) If $n = 6$ and the specific sample values are $.5, 1, .2, 2, .8, 1.5$, graph the likelihood function.
(c) Is the likelihood function you plotted uniformly bounded? What is the point of the global maximum?

Exercise 6.3. (Double Exponential Likelihood). Suppose $1, 2.5, 3.5, 5$ are four iid sample values from a location parameter double exponential density $f(x\,|\mu) = \frac{1}{2}e^{-|x-\mu|}, -\infty < \mu < \infty$. Graph the likelihood function and comment on any interesting feature that you find in the plot.

Exercise 6.4. For a general sample of size n from $Exp(\lambda)$, give a proof that the likelihood function is uniformly bounded.

Exercise 6.5. For a general sample of size n from $C(\mu, 1)$, give a proof that the likelihood function is uniformly bounded.

Exercise 6.6. Suppose you have $n = 3$ observations $0 < a < b$ from $C(\mu, 1)$. Characterize the pairs a, b for which the likelihood function is bimodal.

Exercise 6.7. (Poisson Score Function). Suppose $X_1, \cdots, X_n$ are iid from $Poi(\lambda)$. Derive a formula for the score function, and prove that it has expectation zero.

Exercise 6.8. (Unusual Score Function). Suppose $X \sim Ber(p), Y \sim Poi(p)$, and that X, Y are independent. Derive a formula for the score function. Does it have expectation zero?

Exercise 6.9. (Score Function as a Statistic). Suppose $X_1, \cdots, X_n$ are iid from $Exp(\lambda)$. Consider the score function as a statistic with fixed λ. Find the distribution of the score function.

Exercise 6.10. (Score Function as a Statistic). Suppose $X_1, \cdots, X_n$ are iid from a geometric distribution with parameter p.
(a) Derive a formula for the score function.
(b) Consider the score function as a statistic with fixed p. Find the distribution of the score function.

Exercise 6.11. (Uniform Case). Suppose you tried to write a score function in the $U[0, \theta]$ case by formally following the usual definition. Will this score function have a zero expectation? If not, what is the expectation?

Exercise 6.12. (One Parameter Exponential Family). Give a proof that in the regular one parameter Exponential family, the score function has expectation zero.

Exercise 6.13. **($N(\theta, \theta^2)$).** Suppose $X_1, \cdots, X_n$ are iid from $N(\theta, \theta^2), \theta > 0$. Derive a foormula for the score function.
Remark: : $N(\theta, \theta^2)$ is not in the regular one parameter Exponential family.

Exercise 6.14. (Deriving MLEs). For general iid samples of size n, find closed form formulas for the MLE in each of the following cases:
(a) $f(x\,|N) = \frac{1}{N}, x = 1, 2, \cdots, N, N = 1, 2, \cdots$;
(b) $f(x\,|\mu) = e^{-(x-\mu)}, x \geq \mu, -\infty < \mu < \infty$;
(c) $f(x\,|\theta) = \frac{\theta}{x^2}, x \geq \theta, \theta > 0$.

Exercise 6.15. (Deriving MLEs). For general iid samples of size n, find closed form formulas for the MLE in each of the following cases:
(a) $Beta(\alpha, \alpha), \alpha > 0$;
(b) $G(\alpha, \frac{1}{\alpha}), \alpha > 0$;
(c) $N_2(\mu, \mu, 1, 1, \rho_0)$, where the correlation ρ_0 is known.

Exercise 6.16. **($U[-\theta, \theta]$).** Suppose $X_1, \cdots, X_n$ are iid from $U[-\theta, \theta], \theta > 0$. Find a closed form formulas for the MLE of θ.

Exercise 6.17. Give an example of a regular estimation problem where the number of MLEs is always larger than one.

Exercise 6.18. (MLE in a Clinical Trial). Suppose that in a clinical trial, survival times of subjects are iid $Exp(\lambda)$. The trial ran for $T = 2$ years. One subject died after 6 months, and another two survived throughout the period of the study.
(a) Write the likelihood function.
(b) Graph it.
(c) If there is an MLE of λ, find it.

Exercise 6.19. (MLE in Two Parameter Uniform). Suppose $X_1, \cdots, X_n$ are iid from $U[\mu - \sigma, \mu + \sigma]$, both parameters being unknown. Find MLEs of μ, σ.

Exercise 6.20. (MLE in Two Parameter Pareto). Suppose $X_1, \cdots, X_n$ are iid with density $f(x\,|\theta, \alpha) = \alpha\theta^\alpha / x^{\alpha+1}, x \geq \theta, \theta, \alpha > 0$, both parameters being unknown. Find MLEs of θ, α.

Exercise 6.21. (An Unusual MLE Problem). In your physician's office, patients are weighed during a visit. The machine gets stuck and records 225 lbs. if the weight of the patient is more than 225 lbs.
(a) Suppose that the physician keeps only the unstuck data. Write a model for this problem.
(b) Write the likelihood function under your model.
(c) What can you say about MLEs under your model?

Exercise 6.22. (MLE under the Hardy-Weinberg Law). Suppose that the proportions of three genotypes AA, Aa, and aa in a population are $\theta^2, 2\theta(1-\theta), (1-\theta)^2$, where θ is the proportion of genes that appear in the allele form A. In a sample of n people, the observed frequencies of the three genotypes are n_1, n_2, n_3. Find the MLE of θ.

Exercise 6.23. (MLE for Randomized Responses). The problem is to estimate the proportion θ of people in a population who satisfy a certain embarrassing attribute, e.g., have they ever shoplifted. Since people are not going to answer the question truthfully if asked directly, they are first asked to secretly toss a fair die. If they get a six, they answer true or untrue to the question : I have shoplifted. If they do not get a six, they answer true or untrue to the question : I have never shoplifted. If out of a total of n subjects, X give the answer *true*, find the MLE of θ.

Exercise 6.24. (A Difficult MLE Problem). Of Mrs. Smith's three daughters, Ann, Jen, and Julie, one evening Julie came home with a fever, a cough, and a generalized rash. Next morning, the family physician diagnosed it to be measles. Unfortunately, none of the children had been vaccinated.
Measles is a highly contagious disease, and anyone living in the same quarters with an infected person has a risk of contracting the disease for about 7 days from the day of infection of the original patient (i.e., Julie here). Denote by θ the probability that an individual person contracts the disease from an infected person within the 7 day period. If all or none of the rest get infected, the epidemic ends.
However, if the first infected person infects some of the rest, but not all, say only Ann, then the other child, i.e., Jen, runs the risk of contracting it from Ann, etc.
200 three children families in which at least one child was infected were surveyed and the following data were obtained:
One child infected in 90 families; two infected in 70 families; three infected in 40 families.
Numerically compute the maximum likelihood estimate of θ.

Exercise 6.25. (MLE in Multinomial). In a $k+1$ cell multinomial distribution, with parameters $p_1, \cdots, p_k, p_{k+1} = 1 - p_1 - \cdots - p_k$, and n, with n, as usual considered as known, find MLEs of the k cell probabilities $p_1, \cdots, p_k$. Are these MLEs unique?

Exercise 6.26. (Calculating Fisher Information). Find the Fisher information function in the location parameter double exponential case.
Remark: It is a constant.

Exercise 6.27. (Calculating Fisher Information). Find the Fisher information function in the location parameter Cauchy case.

Exercise 6.28. (Calculating Fisher Information). Find the Fisher information function in the scale parameter double exponential case.

Exercise 6.29. (Calculating Fisher Information). Find the Fisher information function in the $N(\theta, \theta)$ and the $N(\theta, \theta^2)$ case. Is one of the calculations more difficult? Why?

Exercise 6.30. (Calculating Fisher Information). Suppose we parametrize a Poisson distribution by using $\theta = \log \lambda$ as the parameter. What would be the Fisher information function?

Exercise 6.31. (Calculating Fisher Information). Suppose we parametrize a Poisson distribution by using $\theta = \sqrt{\lambda}$ as the parameter. What would be the Fisher information function?

Exercise 6.32. Give an example where the Fisher information function is bounded, and another example where the Fisher information function is unbounded.

Exercise 6.33. (Verifying Sufficiency from Definition). Suppose X_1, X_2 are two iid Bernoullis with parameter θ.
(a) Verify by using the definition of a sufficient statistic that $X_1 + X_2$ is sufficient.
(b) Verify by using the definition of a sufficient statistic that X_1 is not sufficient.
(c) Verify by using the definition of a sufficient statistic that $\max(X_1, X_2)$ is not sufficient.

Exercise 6.34. (Verifying Sufficiency from Definition). Suppose X_1, X_2 are two iid normals with mean μ and variance 1. Verify that for any statistic $S(X_1, X_2)$, the conditional exppectation $E_\mu[S(X_1, X_2)\,|X_1 + X_2 = t]$ does not depend on μ. Will this imply that $X_1 + X_2$ is sufficient?

Exercise 6.35. (Factorization Theorem). For each of the following cases, use the factorization theorem to identify a sufficient statistic; assume that you have n iid observations.
(a)$Beta(\alpha, \beta), \alpha, \beta > 0$;
(b) $Beta(\alpha, c\alpha), \alpha > 0$, where $c > 0$ is a fixed constant.

Exercise 6.36. (Normal with Mean and Variance Related). For each of the following cases, use the factorization theorem to identify a sufficient statistic; assume that you have n iid observations.
(a) $N(\theta, \theta), \theta > 0$;
(b) $N(\theta, \theta^2), \theta > 0$;
(c) $N(\theta, \frac{1}{\theta}), \theta > 0$.

Exercise 6.37. (Factorization Theorem). For each of the following cases, use the factorization theorem to identify a sufficient statistic; assume that you have n iid observations.
(a) lognormal$(\mu, \sigma^2), \mu, \sigma$ both unknown;

(b) lognormal$(0, \sigma^2)$;
(c) lognormal$(\mu, 1)$.

Exercise 6.38. (Unusual Problem). Suppose you have n iid observations from the mixture distribution $.5U[0, \theta] + .5U[5\theta, 6\theta]$. What is the minimal sufficient statistic?

Exercise 6.39. (Data Reduction Can Be Difficult). What is the minimal sufficient statistic in the following cases?
(a) location parameter Cauchy;
(b) location parameter double exponential;
(c) location parameter logistic density;
(d) the mixture distribution $.5N(\mu, 1) + .5N(0, \sigma)$, where μ, σ are both unknown.

Exercise 6.40. (Sufficiency in Nonregular Problems). For each of the following cases, use the factorization theorem to identify a sufficient statistic.
(a) $U[a\theta, b\theta], a, b > 0$;
(b) $U[-\theta, \theta]$;
(c) $f(x\,|\mu, \sigma) = \frac{1}{\sigma} e^{-(x-\mu)/\sigma}, x \geq \mu, \mu, \sigma$ both considered unknown.

Exercise 6.41. (Sufficiency in Fisher's Hyperbola). Use the factorization theorem to identify a sufficient statistic for $N(\theta, \frac{c}{\theta}), \theta > 0$, where $c > 0$ is a known constant.

Exercise 6.42. (Sufficiency in Hypernormals). Consider the location parameter density $f(x\,|\mu) = ce^{-(x-\mu)^{2k}}$, where k is a fixed positive integer, and c is an absolute normalizing constant. Find the minimal sufficient statistic.

Exercise 6.43. (Home Range of an Animal). A tiger moves around in a circle of unknown radius r centered at a known point say $(0, 0)$. Paw prints of the tiger have been found at n different points $(x_i, y_i), i = 1, 2, \cdots, n$. Find the MLE of r; state your model clearly.

Exercise 6.44. (Sufficiency in Bivariate Normal). Use the factorization theorem to identify a sufficient statistic for $N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, all five parameters being unknown.

Exercise 6.45. (Sufficiency under Bounds). For the $N(\mu, 1)$ distribution, if we know that $\mu \geq 0$, does it change the minimal sufficient statistic?

Exercise 6.46. (Sufficiency in Clinical Trials). In a clinical trial, survival times of patients are iid $Exp(\lambda)$. The study is conducted for a specified time period $[0, T]$. For patients who die during the study period, we have the exact survival times; for the rest, we only know that they survived past the study period. Find a minimal sufficient statistic.

Exercise 6.47. (Sufficiency and MLE with Covariates Present). Suppose Y_i are independent Poissons with mean λx_i, where x_i are known positive covariates. For instance, x_i may be the size of the ith county in your sample, and Y_i the number of car accidents in the ith county during the last month.
(a) Find a minimal sufficient statistic.
(b) Find the MLE of λ.

Exercise 6.48. (Sufficiency and MLE in Simple Linear Model). Suppose that response variables $Y_i, i = 1, 2, \cdots, n$ are related to fixed covariates $x_i, i = 1, 2, \cdots, n$ through the stochastic model $Y_i = \alpha + \beta x_i + e_i$, where e_i are iid $N(0, \sigma^2)$; σ, α, β are considered unknown.
(a) Write the likelihood function.
(b) Find a sufficient statistic. What theorem enables you to find the minimal sufficient statistic? What is a minimal sufficient statistic?
(c) What are the MLEs of α, β? Are they unbiased?

Exercise 6.49. (Likelihood in Logistic Regression). Suppose that independent binary variables $Y_i, i = 1, 2, \cdots, n$ taking values $0, 1$ are related to fixed covariates $x_i, i = 1, 2, \cdots, n$ through the stochastic model $P(Y_i = 1) = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}$. Write the likelihood function.

Exercise 6.50. (Sufficiency under Truncation). Planetary sizes in extrasolar systems have a gamma density with parameters α, λ, both unknown. Due to limitations of our telescopes, we can detect only those planets whose sizes exceed a given threshold a. We have n observations. Find a minimal sufficient statistic.

Exercise 6.51. (Sufficiency in Binomial N Problem). Suppose $X_1, \cdots, X_k$ are iid from $Bin(N, p)$.
(a) Find a minimal sufficient statistic when N, p are both unknown.
(b) Find a minimal sufficient statistic when only N is unknown.

Exercise 6.52. (Rao-Blackwellization). Suppose $X_1, \cdots, X_n$ are iid $N(\mu, 1)$. Rao-Blackwellize the statistic $\frac{X_{(n)}+X_{(1)}}{2}$.

Exercise 6.53. (Rao-Blackwellization). Suppose $X_1, \cdots, X_n$ are iid $Poi(\lambda)$. Rao-Blackwellize the statistic $I_{X_1>1}$.

Exercise 6.54. (Rao-Blackwellization). Suppose $X_1, \cdots, X_n$ are iid $Exp(\lambda)$. Rao-Blackwellize the statistic $n \min(X_1, \cdots, X_n)$.

Exercise 6.55. (Rao-Blackwellization). Suppose $X_1, \cdots, X_n$ are iid $U[0, \theta]$. Rao-Blackwellize the statistic $(n + 1)X_{(1)}$.

Exercise 6.56. (Ancillarity). Suppose $X_1, \cdots, X_n$ are iid $N(\mu, 1)$. Prove that $X_{(n)} - X_{(1)}$ and $\bar{X}$ are independent.

Exercise 6.57. (Ancillarity). Suppose $X_1, \cdots, X_n$ are iid $N(\mu, 1)$. Find the covariance between $X_{(n)}$ and $\bar{X}$.

Exercise 6.58. (Ancillarity). Suppose $X_1, \cdots, X_n$ are iid $N(\mu, \sigma^2)$. Give a careful proof that

$$E[\frac{X_{(n)} - X_{(1)}}{s}] = \frac{E[X_{(n)} - X_{(1)}]}{E(s)}.$$

Exercise 6.59. (Ancillarity). Suppose $X_1, \cdots, X_n$ are iid $C(\mu, 1)$. List three different ancillary statistics.

Exercise 6.60. (Ancillarity). Suppose $X_1, \cdots, X_n$ are iid $C(0, \sigma)$. List three different ancillary statistics.

Exercise 6.61. (Ancillarity). Suppose $X_1, \cdots, X_n$ are iid $Poi(\lambda)$. Try to find an ancillary statistic.
Remark: But do not try too hard.

Exercise 6.62. (Ancillarity under Measurement Error). Suppose $X_i = \mu + Z_i + e_i$, where $Z_i \sim N(0, 1), e_i \sim U[-1, 1]$, and all variables are mutually independent.
(a) Is $\bar{X}$ ancillary?
(b) Is $\bar{X}$ sufficient?
(c) Is $M_n - \bar{X}$ ancillary? Here, M_n is the median of $X_1, \cdots, X_n$.

Exercise 6.63. (Unbiased Estimation). Suppose $X_1, \cdots, X_n$ are iid $N(\mu, 1)$. Which of the following are unbiased estimates of μ?

$$X_{(n)}; .5X_{(n)} + .5X_{(1)}; 2X_{(n)} - X_{(1)}; .25X_{(n)} + .25X_{(1)} + .25M_n + .25\bar{X},$$

where M_n is the median of $X_1, \cdots, X_n$.

Exercise 6.64. (Unbiased Estimation). Suppose $X \sim Bin(n, p)$. Find an unbiased estimate of p^2.

Exercise 6.65. (Unbiased Estimation). Suppose $X_1, \cdots, X_n$ are iid $Poi(\lambda)$. Find an unbiased estimate of λ^2; of λ^3.

Exercise 6.66. (Unbiased Estimation). Suppose $X_1, \cdots, X_n$ are iid $Poi(\lambda)$. Find an unbiased estimate of $e^{-2\lambda}$.

Exercise 6.67. (Nonexistence of Unbiased Estimates). Suppose $X \sim Bin(n, p)$. Give a proof that $\sqrt{p(1-p)}$ has no unbiased estimates at all.

Exercise 6.68. (Nonexistence of Unbiased Estimates). Suppose $X_1, \cdots, X_n$ are iid $Poi(\lambda)$. Give a proof that $\frac{1}{\lambda}$ has no unbiased estimates at all.

Exercise 6.69. (Nonexistence of Unbiased Estimates). Suppose $X_1, \cdots, X_n$ are iid $N(\mu, 1)$. Show that $|\mu|$ has no unbiased estimates at all.
Remark: : Challenging!

Exercise 6.70. (Normal UMVUE). Suppose $X_1, \cdots, X_n$ are iid $N(\mu, 1)$. Find the UMVUE of $(\mu - 1)^2$.

Exercise 6.71. (Binomial UMVUE). Suppose $X \sim Bin(n, p)$. Find the UMVUE of $p(1-p)$.

Exercise 6.72. (Poisson UMVUE). Suppose $X_1, \cdots, X_n$ are iid $Poi(\lambda)$. Find the UMVUE of $e^{-2\lambda}$.
Remark: A bit challenging!

Exercise 6.73. (Two Parameter Normal UMVUE). Suppose $X_1, \cdots, X_n$ are iid $N(\mu, \sigma^2)$, both parameters unknown. Find the UMVUE of $\mu + \sigma$; of $\frac{\mu}{\sigma}$.

Exercise 6.74. (Uniform UMVUE). Suppose $X_1, \cdots, X_n$ are iid $U[0, \theta]$. Find the UMVUE of θ^2; of θ^k for a general positive integer k.

Exercise 6.75. (Exponential UMVUE). Suppose $X_1, \cdots, X_n$ are iid $Exp(\lambda)$. Find the UMVUE of $P(X > c)$, where c is a fixed positive number.

Exercise 6.76. (Two Parameter Uniform UMVUE). Suppose $X_1, \cdots, X_n$ are iid $U[\mu - \sigma, \mu + \sigma]$. Find the UMVUEs of μ and σ.

Exercise 6.77. (Multinomial UMVUE). Suppose $(X_1, \cdots, X_k)$ has a multinomial distribution with general cell probabilities $p_1, \cdots, p_k$; n is considered known. Find the UMVUE of $p_1 - p_2$; of $p_1 p_2$.

Exercise 6.78. (Uniqueness of UMVUE). Give a proof that if a UMVUE exists, it has to be unique.

Exercise 6.79. (General UMVUE Property). Give a proof that in the iid case, if a UMVUE exists, it has to be a permutation invariant function of the observations.

Exercise 6.80. (General UMVUE Property). Give a proof that if a UMVUE T of some parameter θ exists, and if T^* is another unbiased estimate of θ, then T and $T - T^*$ must be uncorrelated under all θ.

Exercise 6.81. (Bias of MLEs). Suppose $X \sim Bin(n, p)$. Find the MLE of $\max(p, 1-p)$ and prove that it is not unbiased.

Exercise 6.82. (Bias of MLEs). Suppose $X_1, \cdots, X_n$ are iid $N(\mu, 1)$. Find the MLE of $|\mu|$ and prove that it is not unbiased.

Exercise 6.83. (Bias of MLEs). Suppose $X_1, \cdots, X_n$ are iid $Exp(\lambda)$.
(a) Find the MLE of $P(X > 1)$ and calculate its bias.
(b) Show that the MLE is asymptotically unbiased.

Exercise 6.84. (Bias of MLEs). Suppose $X_1, \cdots, X_n$ are iid $N(\mu, \sigma^2)$, both parameters considered unknown.
(a) Find the MLE of $\frac{1}{\sigma}$.
(b) Prove that it is not unbiased.
(c) Show that the MLE is asymptotically unbiased.

Exercise 6.85. (Bias of MLEs). Suppose $X_1, \cdots, X_n$ are iid $U[\theta, 2\theta]$.
(a) Find the MLE of θ.
(b) Prove that it is not unbiased.
(c) Show that the MLE is asymptotically unbiased.

Exercise 6.86. (Cramér-Rao Bound). Suppose $X_1, \cdots, X_n$ are iid $N(\mu, 1)$. Find the Cramér-Rao lower bound on the variance of an arbitrary unbiased estimate of μ^2.

Exercise 6.87. (Cramér-Rao Bound). Suppose $X_1, \cdots, X_n$ are iid $Exp(\lambda)$. Find the Cramér-Rao lower bound on the variance of an arbitrary unbiased estimate of $\lambda; \lambda^2; \frac{1}{\lambda}$.

Exercise 6.88. (Cramér-Rao Bound). Suppose $X_1, \cdots, X_n$ are iid $G(\alpha, 1)$. Find the Cramér-Rao lower bound on the variance of an arbitrary unbiased estimate of α.

Exercise 6.89. (Cramér-Rao Bound). Suppose $X_1, \cdots, X_n$ are iid $N(\theta, \theta)$. Is there any parametric function $\tau(\theta)$ for which the UMVUE attains the Cramér-Rao lower bound? If there is, what is such a function $\tau(\theta)$?

Exercise 6.90. (Cramér-Rao Bound). Suppose $X_1, \cdots, X_n$ are iid $C(\mu, 1)$.
(a) Find the Cramér-Rao lower bound on the variance of an arbitrary unbiased estimate of μ.
(b) Is there an unbiased estimate which attains this bound?

Exercise 6.91. (Cramér-Rao Bound). Suppose $X_1, \cdots, X_n$ are iid from a location parameter double exponential density.
(a) Find the Cramér-Rao lower bound on the variance of an arbitrary unbiased estimate of the location parameter μ.
(b) Is there an unbiased estimate which attains this bound?

Exercise 6.92. (Cramér-Rao Bound). Suppose $X_1, \cdots, X_n$ are iid from $N(\theta, \theta^2), \theta > 0$.

(a) Find the Cramér-Rao lower bound on the variance of an arbitrary unbiased estimate of θ.

(b) Show that the unbiased estimate $\bar{X}$ does not attain this variance lower bound.

(c) Is there an unbiased estimate of θ which attains this bound?

Exercise 6.93. (Moment Estimates in Beta). For the general two parameter Beta density, find moment estimates of the two parameters. Are there simple formulas for MLEs of the two parameters?

Exercise 6.94. (Moment Estimates in Negative Binomial). Consider negative binomial distributions with both parameters r, p unknown, $r = 1, 2, \cdots, 0 < p < 1$.

(a) Find moment estimates of the two parameters.

(b) Is maximum likelihood estimation of r, p straightforward?

Exercise 6.95. (Moment Estimates in Double Exponential). For the general two parameter double Exponential density, find moment estimates of the two parameters.

Exercise 6.96. (Nonparametric Plug-in). For iid observations $X_1, \cdots, X_n$ from a CDF F on the real line, let $h(F) = E_F[|X - F^{-1}(\frac{1}{2})|]$. Write a formula for a nonparametric plug-in estimator of $h(F)$.

Exercise 6.97. (Nonparametric Plug-in). For iid observations $X_1, \cdots, X_n$ from a CDF F on $[0, 1]$, let $h(F) = \sup_{0 \le x \le 1} |F(x) - x|$. Write a formula for a nonparametric plug-in estimator of $h(F)$.

Exercise 6.98. (Cramér-Rao Bound in Multiparameter Case). Suppose θ is a p-dimensional parameter vector and T is an unbiased estimate of θ_i, the ith coordinate of θ. Show that $\text{Var}_\theta(T) \ge I^{ii}(\theta)$, where $I^{ii}(\theta)$ represents the ith diagonal element of $I(\theta)$. Does $I^{ii}(\theta)$ depend only on θ_i?

Exercise 6.99. (Cramér-Rao Bound in Multiparameter Case). Derive the Fisher information matrix for the general two parameter gamma density $G(\alpha, \lambda)$. Are the two parameters orthogonal?

Exercise 6.100. (Cramér-Rao Bound in Multiparameter Case). Derive the Fisher information matrix for the general two parameter $C(\mu, \sigma)$ density. Are the two parameters orthogonal?

Exercise 6.101. (Cramér-Rao Bound in Multiparameter Case). Derive the Fisher information matrix for the general two parameter double exponential density. Are the two parameters orthogonal?

Exercise 6.102. (Binomial Constrained MLE). Let $X \sim Bin(n,p)$ where it is known that $.44 \le p \le .54$. Write a formula for the MLE of p. Is it an unbiased estimate of p?

Exercise 6.103. (Normal Constrained MLE). Let $X_1, \cdots, X_n$ be iid $N(\mu, 1)$ where it is known that $-1 \le \mu \le 1$. Write a formula for the MLE of μ. Is it an unbiased estimate of μ?

Exercise 6.104. (Trick Question). Let $X_1, \cdots, X_n$ be iid $N(\mu, 1)$ where it is known that μ is a rational number. What can you say about the MLE of μ?

Exercise 6.105. (Two Sample Poisson Constrained MLE). Let $X_1, \cdots, X_n$ be iid $Poi(\lambda)$ and let $Y_1, \cdots, Y_m$ be iid $Poi(\mu)$ where it is known that $\lambda \le \mu$. Write formulas for the MLEs of λ and μ. Assume, as usual, that all $n+m$ observations are independent.

Exercise 6.106. (Multivariate Normal Constrained MLE). let $X_1, \cdots, X_n$ be iid $N_p(\mu, I)$ where it is known that the mean vector μ satisfies $\mu'\mu \le 1$. Write a formula for the MLE of μ.

Exercise 6.107. (Basu's Elephants). In the example of *Basu's Elephants*, what would be the circus owner's estimate of the total weight of all the elephants if the selection probabilities, in decreasing order of the weights, were proportional to $\frac{1}{i}, i = 1, 2, \cdots, 50$, and Jumbo was selected?

Exercise 6.108. (Horvitz-Thompson Estimate). Suppose at the first draw, the selection probabilities of the units are $p_i \ge 0, \sum_{i=1}^{N} p_i = 1$. After the first draw, selection of the eligible units is always with an equal probability, i.e., $p_{i,2} = \frac{1}{N-1}, p_{i,3} = \frac{1}{N-2}$, etc. Derive the following formulas:

$$\pi_i - \frac{N-n}{N-1} p_i + \frac{n-1}{N-1};$$

$$\pi_{i,j} = \frac{n-1}{N-1}\left[\frac{N-n}{N-2}(p_i + p_j) + \frac{n-2}{N-2}\right].$$

Exercise 6.109. (EM for Truncated Geometric). Suppose $X_1, \cdots, X_n$ are iid $Geo(p)$, but the value of X_i reported only if it is ≤ 4. Explicitly derive the E-step of the EM algorithm for finding the MLE of p.

Exercise 6.110. (EM in a Genetics Problem). Consider the ABO blood group problem worked out in Example 7.51. For the data values $Y_A = 182, Y_B = 60, Y_{AB} = 17, Y_O = 176$ (McLachlan and Krishnan (2008)), find the first four EM iterates, using the starting values $(p_A, p_B, p_O) = (.264, .093, .643)$.

Exercise 6.111. (EM in a Background plus Signal Model).
(a) Suppose $Y = U + V$, where U, V are independent Poisson variables with mean λ and c respectively, where c is known and λ is unknown. Only Y is observed. Design an EM algotithm for estimating λ, and describe the E-step and the M-step.
(b) Generalize part (a) to the case of independent replications $Y_i = U_i + V_i, 1 \le i \le n, U_i$ with common mean λ and V_i with common mean c.

6.8 References

Barankin, E.W. and Maitra, A. (1963). Generalization of the Fisher-Darmois-Koopman-Pitman theorem on sufficient statistics, Sankhya, A, 25, 217-244.

Basu, D. (1955). On statistics independent of a complete sufficient statistic, Sankhy$\bar{a}$, 15, 377-380.

Basu, D. (1959). The family of ancillary statistics, Sankhya, 21, 247-256.

Basu, D. (1964). Recovery of ancillary information, Sankhya, A, 26, 3-16.

Basu, D. (1971). An essay on the logical foundations of survey sampling, with discussions, in *Foundations of Statistical Inference*, V. P. Godambe and D. A. Sprott eds., Holt, Rinehart, and Winston of Canada, Toronto.

Basu, D. (1975). Statistical information and likelihood, with discussions, Sankhya, Ser. A, 37, 1-71.

Bickel, P. and Doksum,K. (2006). Mathematical Statistics: Basic Ideas and Selected Topics, Prentice Hall, NJ.

Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation, Ann. Math. Statist., 18, 105-110.

Brown, L.D. (1964). Sufficient statistics in the case of independent random variables, Ann. Math. Statist., 35, 1456-1474.

Brown, L. (1990). An ancillarity paradox which appears in multiple linear regression, with discussion, Ann. Statist., 18, 471 -538.

Buehler, R. (1982). Some ancillary statistics and their properties, with discussions, JASA, 77, 581-594.

Chaudhuri, A. and Stenger, H. (2005). Survey Sampling: Theory and Methods, CRC Press, New York.

Cox, D. and Hinkley, D. (1979). Theoretical Statistics, Chapman and Hall, Boca Raton, FL

Cramér, H. (1946). Mathematical Methods of Statistics, Princeton University Press, NJ.

Dahiya, R., Staneski, P. and Chaganty, N.R. (2001). Maximum likelihood estimates of the truncated Cauchy distribution, Comm. Statist., Theory and Methods, 30, 1735-1750.

DasGupta, A. (2008). Asymptotic Theory of Statistics and Probability, Springer, New

York.

DasGupta, A. and Rubin, H. (2004). Estimation of binomial parameters when both n, p are unknown, Jour. Stat. Planning Inf., 130, 391-404.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm, JRSS, Ser. B, 39, 1-38.

Des Raj (1956). Some estimates in sampling with varying probabilities without replacement, JASA, 51, 269-284.

Fraser. D. (2004). Ancillaries and conditional inference, Statist. Sci., 19, 333-369.

Ghosh, M. (2002). Basu's Theorem: A Personalistic Review, Sankhya, A, 64, 509-531.

Godambe, V. and Joshi, V. (1965). Admissibility and Bayes estimation in sampling from finite populations, Ann. Math. Statist., 36, 1708-1722.

Govindarajulu, Z, (1999). Elements of Sampling Theory and Methods, Prentice Hall, Upper Saddle River, NJ.

Hall, P. (1994). On the erratic behavior of estimators of N in the binomial(n,p) distribution, Jour. Amer. Statist. Assoc., 89, 425, 344-352.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe, JASA, 47, 663-685.

Lehmann, E. (1981). An interpretation of completeness and Basu's theorem, JASA, 76, 335-340.

Lehmann, E. L. and Casella, G. (1998). Theory of Point Estimation, Springer, New York.

Lehmann, E. L. and Scheffe, H. (1950). Completeness, similar regions, and unbiased estimation, Sankhya, 10, 305-340.

Liu, R. and Brown, L. (1993). Non-existence of informative unbiased estimators in singular problems, Ann. Statist., 21, 1-13.

Marchand, E. and Strawderman, W. (2004). Estimation in restricted parameter spaces: A Review, LNMS, 45, IMS, Beachwood, OH.

McLachlan, G. and Krishnan, T. (2008). The EM Algorithm and Extensions, Wiley, New York.

Olkin, I., Petkau, A.J. and Zidek, J.V. (1981). A comparison of n estimators for the binomial distribution, Jour. Amer. Statist. Assoc., 76, 375, 637-642.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters, Bull. Calcutta Mmath. Soc., 37, 81-91.

Rao, C. R. (1973), Linear Statistical Inference and Applications, Wiley, New York.

Reeds, J. (1985). The asymptotic distribution of the number of roots of the Cauchy likelihood equation, Ann. Statist., 13, 775-784.

Reid, N. (2003). Asymptotics and the theory of inference, Ann. Statist., 31, 1695-1731.

Shao, P. and Strawderman, W. E. (1996). Improving on the MLE of a positive normal mean, Statist. Sinica, 6, 275-287.