Chapter 5. Estimation.

A large class of statistical problems can be formulated as parametric statistical problems: We think of the data, which may be a univariate X or a multivariate $X_1, X_2, \ldots, X_n$, as the outcome of an experiment. Before the experiment is performed, or before the data are observed, their values are uncertain, and they may therefore be considered as random variables. Our goal is to describe the process that produces the data; in our formulation that means describing their probability distribution. What makes the problem a *parametric* statistical problem is that we are willing to adopt, at least provisionally, a parametric model for the class of possible probability distributions. In the univariate case this would mean that we are willing to assume, at least until experience has shown the assumption to be wrong or inadequate, that X has one of a class of distributions $p(x|\theta)$ (in the discrete case) or $f(x|\theta)$ (in the continuous case), where the form of the function describing the distribution is known, and the only remaining question is to determine the value of θ. In some cases θ may have a natural interpretation as a "state of nature" or a "cause", but in general we will simply refer to it as the *parameter* of the family. If the data were multivariate, the model would be of the form $p(x_1, \ldots, x_n|\theta)$ or $f(x_1, \ldots, x_n|\theta)$. We will encounter several examples later where both the data and the parameter are multivariate, but to fix ideas it is useful to start with simpler situations.

Example 5.A (The Survey, continued). In Chapter 4 we considered an example that fits within the present framework (Example 4.C), where the data, X, is a count of the number of Chicago Democrats for the incumbent out of $n = 100$ interviewed, and by the fact of random sampling we are willing to assume

$$\begin{aligned} p(x|\theta) &= b(x; n, \theta) \\ &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad \text{for } x = 0, 1, \ldots, n \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

where θ is a parameter of the distribution $p(x|\theta)$; in this case we can interpret θ as the fraction of all Chicago Democrats for the incumbent. If θ were known, the process that generates the data, as represented by our model $p(x|\theta)$, would be fully specified.

Example 5.B (The Scale, continued). The weighing problem of Example 4.D also fits this framework. The data X (the recorded weight) has, by our assumptions about the known characteristics of the scale, a $N(\theta, \tau^2)$ distribution, with τ^2 known. Then

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}\tau}\, e^{-\frac{(x-\theta)^2}{2\tau^2}} \quad \text{for } -\infty < x < \infty.$$

Here the parameter θ can be interpreted as the true weight; if it were known, the distribution of the data would be fully specified.

It is instructive to contrast our present approach with that of the previous chapter. In both cases, we assume a model for the probability distribution of the data, or rather we assume a class of models, $p(x|\theta)$ or $f(x|\theta)$. There are minor differences in the ways we look at the models. In the previous chapter we stressed these as conditional (given θ) probability distributions, because we were concerned with modelling our uncertainty about θ by a probability distribution and finding the "other" conditional distribution (of θ given X); in this chapter we emphasize considering $p(x|\theta)$ and $f(x|\theta)$ simply as indexed (by θ) families of distributions. But these are superficial differences of stress, not real differences of substance: in both cases this part of the mathematical structure is the same. The real difference is in what we assume about the a priori probabilities of different values of θ. In this chapter, unlike the previous one, we do *not* assume the availability of quantifiable prior information about θ in the form of a probability distribution $f(\theta)$.

In effect, dropping this assumption adds generality to our analysis; whatever results we obtain without assuming an a priori distribution will be accepted by even those scientists who are unable or unwilling to agree upon a particular $f(\theta)$ as representing their prior state of knowledge. But this generality comes with a cost: we can no longer hope to achieve our ideal goal. With limited data (a small survey, a single weighing) we will be uncertain about θ, and our ideal goal is to describe this unavoidable uncertainty, that is, to determine $f(\theta|x)$. But without $f(\theta)$ or its equivalent, this is impossible. Bayes's Theorem tells us $f(\theta|x) \propto f(\theta)f(x|\theta)$, and we do not have one of the required factors.

5.1 Point Estimation.

If our ideal goal, determining $f(\theta|x)$, is beyond reach without $f(\theta)$, we will need to adopt more limited goals. In this chapter we shall begin to explore what can be done without an a priori distribution, where our only assumptions involve the family of distributions of the data, $f(x|\theta)$. One of the simplest inferential problems to state, is, *which* of the distributions $f(x|\theta)$ is the *right* one? Or equivalently, what is our best guess or estimate of the value of θ that, through the distribution $f(x|\theta)$, actually produced the data. A *point estimate* is a function of the data, $\hat{\theta}(X)$ (or in the multivariate case, $\hat{\theta}(X_1,\ldots,X_n)$ that we hope will be close to the actual value of θ. We will write $\hat{\theta}$ for $\hat{\theta}(X)$ or $\hat{\theta}(X_1,\ldots,X_n)$ when there is no confusion as to what is meant. Our problem is to choose a function $\hat{\theta}$ and discuss its accuracy.

Ideally, we want an estimate $\hat{\theta}$ that is likely to be near to θ. What does "likely to be near" mean in this context? The estimate $\hat{\theta} = \hat{\theta}(X)$ will be used after the data is at hand, and so we presumably would want the conditional probability *given* X to be high that $|\hat{\theta}(X) - \theta|$ is small. But we cannot calculate that probability without $f(\theta|x)$, we cannot find $f(\theta|x)$ without $f(\theta)$, and $f(\theta)$ is not available. And so we will fall back upon calculations based on $f(x|\theta)$; we will aim to make the conditional probability *given* θ high that $|\hat{\theta}(X) - \theta|$ is small. That is, we will evaluate the accuracy of the function $\hat{\theta}(X)$ from the point of view of "before the experiment," prior to observing the data. We cannot calculate a posteriori accuracy; we can find the expected accuracy under several alternative hypotheses. In effect, we consider $\hat{\theta}(X)$ as a tool and judge it by its average effectiveness, since we cannot observe its effectiveness in the single instance when we use it. To emphasize that $\hat{\theta}$ is being considered as a function of the data, we will refer to it as an *estimator* (a tool for estimation), whose particular realized values are called *estimates.*

Adopting this "before experiment" or "before data" point of view, we have that for each possible hypothesized or given value of θ, X is a random variable with distribution $f(X|\theta)$. Hence $\hat{\theta}(X)$, a transformation of X, will have a probability distribution, conceivably a different one for every θ. In principle, we can find its distribution, say $f_{\hat{\theta}}(u|\theta)$, and we can check to see if its distribution is concentrated near θ, and if it is centered at θ. To help us discuss accuracy and compare different possible estimators, we will make the following definitions.

[Figure 5.1]

We say $\hat{\theta}$ is *unbiased* if

$$E(\hat{\theta}) = \theta, \quad \text{for all possible values of } \theta. \tag{5.1}$$

The *bias* of $\hat{\theta}$ is given by

$$B(\theta) = E(\hat{\theta}) - \theta. \tag{5.2}$$

The bias $B(\theta)$ represents the average amount by which $\hat{\theta}$ misses θ; positive $B(\theta)$ corresponds to overestimation, negative $B(\theta)$ to underestimation. We will consider two measures of accuracy; the *mean error* of $\hat{\theta}$ is defined to be $E|\hat{\theta} - \theta|$. The *mean squared error* of $\hat{\theta}$ is

$$MSE(\theta) = E(\hat{\theta} - \theta)^2. \tag{5.3}$$

All of these definitions are in terms of the conditional distribution of $\hat{\theta}$ given θ; thus in (5.1) and (5.2), in the continuous case,

$$\begin{aligned} E(\hat{\theta}) &= \int_{-\infty}^{\infty} u f_{\hat{\theta}}(u|\theta)du \\ &= \int_{-\infty}^{\infty} \hat{\theta}(x) f(x|\theta)dx \quad \text{using (2.17),} \end{aligned}$$

and the mean error is

$$E|\hat{\theta} - \theta| = \int_{-\infty}^{\infty} |\hat{\theta}(x) - \theta| f(x|\theta)dx,$$

while the mean squared error (5.3) is given by

$$MSE_{\hat{\theta}} = \int_{-\infty}^{\infty} (\hat{\theta}(x) - \theta)^2 f(x|\theta)dx.$$

Of the two criteria of accuracy, the mean error may seem the more natural, but mean squared error is generally the easier to calculate. In fact, we have

$$\begin{aligned} MSE_{\hat{\theta}}(\theta) &= E(\hat{\theta} - \theta)^2 \\ &= E[(\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta)]^2 \\ &= E[(\hat{\theta} - E(\hat{\theta}))^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) + (E(\hat{\theta}) - \theta)^2] \\ &= E[\hat{\theta} - E(\hat{\theta})]^2 + 2(E(\hat{\theta}) - \theta)E(\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta)^2 \end{aligned}$$

Now the first term is just $\mathrm{Var}(\hat{\theta}|\theta)$, the third term is $(B(\theta))^2$, and the middle term vanishes because $E(\hat{\theta} - E(\hat{\theta})) = E(\hat{\theta}) - E(\hat{\theta}) = 0$, giving us

$$MSE_{\hat{\theta}}(\theta) = \mathrm{Var}(\hat{\theta}|\theta) + (B(\theta))^2, \tag{5.4}$$

or

$$\text{Mean Squared Error } = \text{Variance } + \text{ (Bias)}^2. \tag{5.5}$$

This simple and elegant relationship captures quantitatively a relationship illustrated in Figure 5.2: there can be a tradeoff between variance and bias. The total expected error for a particular

[Figure 5.2]

θ, as measured by $MSE_{\hat{\theta}}(\theta)$, may be considered as due to two sources, the variability of the estimator and its bias, and we may be faced with a choice between a high bias/low variance estimator and a low bias/high variance estimator.

Example 5.A (Continued). Suppose X has Binomial (n, θ) distribution. There are many possible estimators $\hat{\theta}$ that could be used. One obvious choice is the sample fraction,

$$\hat{\theta}_1(X) = \frac{X}{n}.$$

Another is to always guess "1/2," regardless of the data:

$$\hat{\theta}_2(X) = \frac{1}{2}.$$

A third, less obvious choice is

$$\hat{\theta}_3(X) = \frac{X+1}{n+2}.$$

This could be motivated as follows: if we *did* assume that a priori $f(\theta)$ was a Uniform $(0, 1)$ distribution, then given $X = x$, a posteriori θ would have a Beta $(x+1, n-x+1)$ distribution, with $E(\theta|X = x) = \frac{x+1}{n+2}$ (by (4.16)). Now we do not make this assumption, but that does not disqualify us from considering $(X+1)/(n+2)$ as an estimator on its own merits; as (4.16) shows, $\hat{\theta}_3$ is a weighted average of 1/2 and X/n, and hence a compromise between $\hat{\theta}_1$ and $\hat{\theta}_2$:

$$\hat{\theta}_3 = \left(\frac{n}{n+2}\right)\hat{\theta}_1 + \left(\frac{2}{n+2}\right)\hat{\theta}_2.$$

How do these three estimators compare? How well can we expect them to perform? Now

$$\begin{aligned} E(\hat{\theta}_1) &= E(\hat{\theta}_1(X)) \\ &= E\left(\frac{X}{n}\right) \\ &= \frac{E(X)}{n} \\ &= \frac{n\theta}{n} = \theta, \quad \text{using (3.67).} \end{aligned}$$

while $E(\hat{\theta}_2) = \frac{1}{2}$ for all θ, next,

$$\begin{aligned} E(\hat{\theta}_3) &= E\left(\frac{X+1}{n+2}\right) \\ &= \frac{E(X)+1}{n+2} \\ &= \frac{n\theta+1}{n+2}. \end{aligned}$$

Clearly $\hat{\theta}_1$ is unbiased, while $\hat{\theta}_2$ and $\hat{\theta}_3$ are biased: $\hat{\theta}_2$ has bias $B_2(\theta) = 1/2 - \theta$, and $\hat{\theta}_3$ has bias $B_3(\theta) = \frac{n\theta+1}{n+2} - \theta = \frac{1-2\theta}{n+2}$. Both $B_2(\theta)$ and $B_3(\theta)$ are zero only when θ equals 1/2; $B_3(\theta)$ is small if n is large (that is, if the data are extensive), while $B_2(\theta)$ is of course unaffected by n (since $\hat{\theta}_2$ ignores the data).

Because $\hat{\theta}_2$ does not depend upon the data, it is not random and its mean error and mean squared error are easy to calculate:

$$E|\hat{\theta}_2 - \theta| = |\frac{1}{2} - \theta|, \tag{5.6}$$

$$MSE_{\hat{\theta}_2}(\theta) = \left(\frac{1}{2} - \theta\right)^2 = \frac{1}{4} - \theta(1-\theta). \tag{5.7}$$

For $\hat{\theta}_1$, we have

$$\begin{aligned} MSE_{\hat{\theta}_1}(\theta) &= E(\hat{\theta}_1 - \theta)^2 \qquad (5.8) \\ &= E\left(\frac{X}{n} - \theta\right)^2 \\ &= \text{Var}\left(\frac{X}{n}\right) \\ &= \frac{\text{Var}(X)}{n^2} \\ &= \frac{n\theta(1-\theta)}{n^2} \quad \text{using (3.68),} \\ &= \frac{\theta(1-\theta)}{n}. \end{aligned}$$

The mean error of $\hat{\theta}_1$ is extremely difficult to calculate, although it can, with much clever algebraic work, be found to be

$$E|\hat{\theta}_1 - \theta| = 2\binom{n-1}{[n\theta]}\theta^{[n\theta]+1}(1-\theta)^{n-[n\theta]}, \tag{5.9}$$

where by $[n\theta]$ we mean the largest integer that is still no larger than $n\theta$. For large n we have, approximately,

$$E|\hat{\theta}_1 - \theta| \simeq \sqrt{\frac{2}{\pi} \cdot \frac{\theta(1-\theta)}{n}}. \tag{5.10}$$

Next, for $\hat{\theta}_3$, from (5.4),

$$\begin{aligned} MSE_{\hat{\theta}_3}(\theta) &= \mathrm{Var}(\hat{\theta}_3|\theta) + (B_3(\theta))^2 \\ &= \frac{\mathrm{Var}(X)}{(n+2)^2} + \frac{(1-2\theta)^2}{(n+2)^2} \\ &= \frac{n\theta(1-\theta) + (1-2\theta)^2}{(n+2)^2} \end{aligned} \tag{5.11}$$

The remaining measure, $E|\hat{\theta}_3 - \theta|$, is hard to evaluate in other than numerical terms from

$$E|\hat{\theta}_3 - \theta| = \sum_{k=0}^{n} \left| \left(\frac{k+1}{n+2} \right) - \theta \right| b(k; n, \theta). \tag{5.12}$$

How do these estimators compare? From our present "given θ" perspective, we put ourselves in the position of a survey organization, and ask, for what size surveys, and for which types of elections, could we expect $\hat{\theta}_1$ to perform best? $\hat{\theta}_2$? $\hat{\theta}_3$? Figure 5.3 shows their mean errors and mean squared errors for $n = 4$ and $n = 25$. Both of the two measures tell us about the estimators' expected performance under different conditions. For example, if we want an estimator that produces small errors in very close elections, when θ is near 1/2, $\hat{\theta}_2$ does very well; in fact, perfectly when $\theta = 1/2$, even for small surveys. But it does poorly in other situations. If n is as large as 25, the mean errors of both $\hat{\theta}_1$ and $\hat{\theta}_3$ are lower than that of $\hat{\theta}_2$ except for $.46 < \theta < .54$; the mean squared errors are lower except for $.40 < \theta < .60$. The following table shows the interval of values for θ for which the mean and mean squared errors of $\hat{\theta}_2$ are lower than those for $\hat{\theta}_1$.

[Figure 5.3]

Interval where $\hat{\theta}_2$ is better than $\hat{\theta}_1$, when measured by:

$n =$	Mean Error	Mean Squared Error
1	$.29 < \theta < .71$	$.15 < \theta < .85$
4	$.40 < \theta < .60$	$.28 < \theta < .72$
25	$.46 < \theta < .54$	$.40 < \theta < .60$
100	$.48 < \theta < .52$	$.45 < \theta < .55$

[Table 5.1]

We can learn several lessons from these comparisons. First, ignoring the data becomes more costly, the more data there are. If n is at all large, $\hat{\theta}_2$ is never much better than $\hat{\theta}_1$ or $\hat{\theta}_3$ and sometimes much worse. Second, the performance of $\hat{\theta}_1$ improves as n increases and is at its worst when θ is near 1/2. This reflects the fact that Binomial variability is greatest for $\theta = 1/2$; it is hardest to estimate the outcome of close elections (if either all Chicago Democrats or no Chicago Democrats are for the incumbent, the survey with estimator $\hat{\theta}_1$ will give perfect accuracy). And third, while mean error and mean squared error give numerically different answers, they are in qualitative agreement.

Because both of the measures of performance we've discussed convey much the same message, we shall generally adopt mean squared error as our criterion because (5.4) usually makes it simple to calculate and interpret. If an estimator $\hat{\theta}$ is unbiased,

$$MSE_{\hat{\theta}} = \mathrm{Var}(\hat{\theta}|\theta), \tag{5.13}$$

and so the best (in the mean squared error sense) unbiased estimator is the one with the smallest variance, sometimes called *MVUE* (for "Minimum Variance Unbiased Estimator"). We should bear in mind, however, that many of the best estimators we shall consider are actually slightly biased, with bias that decreases with large amounts of data. Still, with unbiased estimators in mind we shall define the *standard error* of an estimator $\hat{\theta}$ to be its standard deviation:

$$\sigma_{\hat{\theta}} = \sqrt{\mathrm{Var}(\hat{\theta}|\theta)}. \tag{5.14}$$

When $\hat{\theta}$ is approximated unbiased, the standard error may reflect the anticipated accuracy of $\hat{\theta}$; if $\hat{\theta}$ is unbiased and approximately normally distributed.

$$P(|\hat{\theta} - \theta| \le \sigma_{\hat{\theta}}|\theta) \simeq 2/3 \tag{5.15}$$

$$P(|\hat{\theta} - \theta| \le 2\sigma_{\hat{\theta}}|\theta) \simeq .95 \tag{5.16}$$

$$P(|\hat{\theta} - \theta| \le 3\sigma_{\hat{\theta}}|\theta) \simeq .998. \tag{5.17}$$

5.2 Maximum Likelihood Estimators.

Where do we get estimators? In the binomial example, one estimator (the sample fraction) seemed like an obvious choice, and another was unreasonable (because it ignored the data) but served as a sort of a baseline. Only in the third case did we motivate the estimator by a statistical argument, and even then the argument was based on an entirely hypothetical assumption. While it is true that any scheme for concocting estimators is permissible, we should expect that the performance of most such ad hoc estimators, as judged by mean squared error, will range from terrible to mediocre. In this section we discuss one principle for finding estimators that generally produces very good results, though there is no absolute guarantee it will yield the best (or even an acceptable) choice. This is the principle of *maximum likelihood.*

We can best motivate this principle by recalling Bayes's theorem:

$$f(\theta|x) \propto f(\theta)f(x|\theta).$$

If we had $f(\theta)$ available, we would be tempted to look for the "most likely" value of θ, namely the value of θ for which $f(\theta|x)$ is largest. Equivalently, we could look for the value of θ for which $f(\theta)f(x|\theta)$ is largest. But $f(\theta)$ is not available; what should we do? If $f(\theta)$ were relatively constant for θ's near the value maximizing $f(\theta)f(x|\theta)$, then it should make little difference whether we look for the value maximizing this product or simply for the value maximizing the second factor, $f(x|\theta)$, and this is what we propose to do.

We shall denote by $L(\theta) = f(x|\theta)$ (or if the data are multivariate by $L(\theta) = f(x_1, x_2, \dots, x_n|\theta)$), viewed as a function of θ, the *likelihood function.* The value of θ, say $\hat{\theta}$, for which $L(\theta)$ achieves its maximum is called a *maximum likelihood estimator* of θ.

Example 5.A (Continued). For the Binomial example we have

$$\begin{aligned} L(\theta) &= p(x|\theta) \\ &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \quad \text{for } 0 \le \theta \le 1 \end{aligned}$$

[Figure 5.4]

Previously, we had been looking at $p(x|\theta)$ as a function of x for a given θ; here we are looking at it as a function of θ for a fixed x. We emphasize this by our notation, $L(\theta)$, which suppresses the argument x. Figure 5.4 shows $L(\theta)$ for the case $n = 100$ and $x = 40$. We are looking for the value of θ for which $L(\theta)$ is maximum. Often this is found from solving

$$\frac{d}{d\theta} L(\theta) = 0$$

for $\hat{\theta}$, and then verifying that for this choice of θ,

$$\frac{d^2}{d\theta^2} L(\theta) < 0,$$

so $\hat{\theta}$ is at least a relative maximum of $L(\theta)$. In this example, the differentiation is cumbersome; we have

$$\begin{aligned}\frac{d}{d\theta}L(\theta) &= \binom{n}{x}\left[\frac{d}{d\theta}\theta^x(1-\theta)^{n-x}\right]\\ &= \binom{n}{x}\left[\theta^x\frac{d}{d\theta}(1-\theta)^{n-x}+(1-\theta)^{n-x}\frac{d}{d\theta}\theta^x\right]\\ &= \binom{n}{x}\left[\theta^x(n-x)(-1)(1-\theta)^{n-x-1}+(1-\theta)^{n-x}\cdot x\cdot\theta^{x-1}\right]\\ &= \binom{n}{x}\theta^{x-1}(1-\theta)^{n-x-1}[x(1-\theta)-(n-x)\theta].\end{aligned}$$

This equals zero when (if $x > 1$ or $n - x > 1$) $\theta = 0$ or $\theta = 1$, or when

$$[x(1-\theta)-(n-x)\theta]=0.$$

From Figure 5.4 it is clear that $\theta = 0$ and $\theta = 1$ generally correspond to minima. We then have

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

or, solving this equation,

$$\hat{\theta}=\frac{x}{n},$$

the sample fraction. We could go on to calculate $\frac{d^2}{d\theta^2}L(\theta)$, but there is a simpler route. It is frequently true that likelihood functions $L(\theta)$ involve products and exponentials, making it easier to analyze $logL(\theta)$ than $L(\theta)$. Fortunately, for our purposes, this can be done with no loss. Because the logarithm is a monotone function, we will reach the same answer asking, "for what θ is $L(\theta)$ maximum?" as asking, "for what θ is $\log L(\theta)$ maximum?" In fact, plotting $L(\theta)$ on a logarithmic scale is the same as plotting $\log L(\theta)$ on a linear scale. The maximum *values* of the two functions will of course differ, but these maxima will necessarily be achieved for the same value of θ, as Figure 5.4 illustrates.

[Figure 5.4]

In our example we have,

$$\log L(\theta)=\log\binom{n}{x}+x\log\theta+(n-x)\log(1-\theta)$$

and

$$\frac{d}{d\theta}\log L(\theta)=0+\frac{x}{\theta}-\frac{(n-x)}{1-\theta}. \tag{5.18}$$

Setting this equal to zero gives

$$\frac{x}{\hat{\theta}}=\frac{n-x}{1-\hat{\theta}},$$

or

$$\frac{\left(\frac{x}{n}\right)}{\hat{\theta}}=\frac{1-\left(\frac{x}{n}\right)}{1-\hat{\theta}},$$

or

$$\hat{\theta}=\frac{x}{n},$$

as before. Next,

$$\frac{d^2}{d\theta^2}\log L(\theta)=-\frac{x}{\theta^2}-\frac{(n-x)}{(1-\theta)^2} \tag{5.19}$$

which is negative for $\theta = \hat{\theta}$ (indeed, for all $0 < \theta < 1$).

5.3 Interpreting the Likelihood Function.

We were led to consider maximum likelihood estimators by the fact that if the a priori density $f(\theta)$ is constant as a function of θ, then the posterior distribution $f(\theta|x)$ is proportional to $L(\theta)$:

$$\begin{aligned} f(\theta|x) &\propto f(\theta)f(x|\theta) \\ &\propto f(x|\theta) \\ &= L(\theta). \end{aligned}$$

That is, if all values of θ are a priori equally likely, $L(\theta)$ is (but for a normalizing constant) the a posteriori density of θ, and the maximum likelihood estimator is the posterior mode. If $f(\theta)$ is only approximately constant (that is, changes little relative to changes in $f(x|\theta)$ as θ varies), this interpretation still holds at least approximately. Thus one primary interpretation of $L(\theta)$ is a Bayesian interpretation.

In the Binomial example, this means that one way of motivating the sample fraction $\hat{\theta}_1(X) = X/n$ is that it is the posterior mode for a Uniform $(0,1)$ a priori distribution, because if $f(\theta) \equiv 1$ for $0 < \theta < 1$, then $f(\theta|x) \propto p(x|\theta)$. Thus in that case, the argument for $\hat{\theta}_1$ is like that for $\hat{\theta}_3(X) = \frac{X+1}{n+2}$; $\hat{\theta}_1$ is the posterior mode, and $\hat{\theta}_3$ is the posterior expectation. This raises a question: why *maximize* $L(\theta)$, why not find the mean of the density

$$\frac{L(\theta)}{\int_{-\infty}^{\infty} L(u)du} \tag{5.20}$$

in general? Both procedures work well in the Binomial case; indeed Figure 5.3 even gives some grounds for preferring $\hat{\theta}_3$. The answer is that when $f(\theta)$ is not constant (but only approximately constant), the posterior mode usually differs little from the maximum likelihood estimator, but the posterior expectation may be quite far from the expectation of (5.20). If $f(\theta)$ is relatively flat for θ near the maximum of $L(\theta)$, then the maxima of $f(\theta|x)$ and $L(\theta)$ occur near the same place, while the expectations of $f(\theta|x)$ and (5.20) may not.

[Figure 5.5]

That is, the maximum likelihood estimator will be an approximate posterior mode for a wide range of plausible priors; the expectation of (5.20) will not generally be so close to the posterior expectation.

While the Bayesian interpretation is one important way of motivating the consideration of $L(\theta)$ and maximum likelihood estimation, it is not the only way. Without any recourse to a prior probability distribution, we still have that $L(\theta) = p(x|\theta)$ or $f(x|\theta)$. If "x" represents the particular data we observe, $L(\theta)$ gives the probability or probability density of our data for the particular θ:

$$L(\theta) = P(\text{observed data} \mid \text{state of nature } \theta).$$

Then we can think of $L(\theta)$ for two values of θ as giving the relative probability (or "likelihood") of actually observing the data we have for these values of θ; if $L(\theta_1)/L(\theta_2) = 2$, we are twice as likely to observe our data values given θ_1 as given θ_2. Looking for the maximum of $L(\theta)$ amounts, in this view, to looking for the value of θ that best explains our data. We are more likely to observe 40 out of 100 Chicago Democrats for the incumbent if the fraction of θ of all Chicago Democrats for the incumbent equals .4 than if $\theta = .3$ or .5. Observing 40 out of 100 remains unlikely in any event, but it is more likely for $\theta = .4$ than for any other value of θ.

5.4 Properties of Maximum Likelihood Estimators.

Whether the motivation of section 5.3 is compelling or not, the principal reasons for actually using maximum likelihood estimators from our present "before experiment" perspective are that they are feasible to find in a wide variety of problems, and they have been found usually to perform well.

Example 5.C Estimating Average Failure Time. Under certain hypotheses (which we will present later), a computer module will last a time X before failure, where X is a continuous random variable with probability density function

$$\begin{aligned} f(x|\theta) &= \frac{1}{\theta}e^{-\frac{x}{\theta}} \quad \text{for } x > 0 \\ &= 0 \quad \text{for } x \le 0. \end{aligned} \tag{5.21}$$

That is, X has an Exponential $(1/\theta)$ distribution. An easy calculation gives

$$\begin{aligned} E(X) &= \int_0^\infty x f(x|\theta)dx \\ &= \theta, \end{aligned}$$

so θ (which must be greater than zero) can be interpreted as the expected time to failure. It can also be easily shown that $\mathrm{Var}(X) = \theta^2$. The statistical problem we consider is where θ is unknown, but data are available: n separate modules have been tested independently and found to fail at times $X_1, X_2, \ldots, X_n$. We wish to estimate θ. Here the data $(X_1, X_2, \ldots, X_n)$ are multivariate, and since they are independent they have density

$$\begin{aligned} f(x_1, x_2, \ldots, x_n|\theta) &= f(x_1|\theta) \cdot f(x_2|\theta) \cdots f(x_n|\theta) \\ &= \frac{1}{\theta}e^{-\frac{x_1}{\theta}} \cdot \frac{1}{\theta}e^{-\frac{x_2}{\theta}} \cdots \frac{1}{\theta}e^{-\frac{x_n}{\theta}} \\ &= \frac{1}{\theta^n}e^{-\frac{(x_1+\cdots+x_n)}{\theta}} \\ &= \frac{1}{\theta^n}e^{-\sum\limits_{i=1}^n x_i/\theta} \quad \text{for all } x_i > 0 \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

The likelihood function is then

$$\begin{aligned} L(\theta) &= \frac{1}{\theta^n}e^{-\sum\limits_{i=1}^n x_i/\theta} \quad \text{for } \theta > 0 \\ &= 0 \quad \text{for } \theta \le 0; \end{aligned} \tag{5.22}$$

it may (in the spirit of Section 5.3) be considered as giving the relative likelihood for different values of θ.

[Figure 5.6]

To find the maximum likelihood estimator we maximize

$$\log L(\theta) = -n\log\theta - \sum_{i=1}^n x_i/\theta.$$

Differentiating,

$$\frac{d}{d\theta}\log L(\theta) = -\frac{n}{\theta} + \frac{\sum\limits_{i=1}^n x_i}{\theta^2}; \tag{5.23}$$

setting this equal to zero gives

$$\frac{n}{\hat{\theta}} = \frac{\Sigma x_i}{\hat{\theta}^2}$$

or

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}, \tag{5.24}$$

the sample arithmetic mean. To check that this gives a maximum, we find

$$\begin{aligned} \frac{d^2}{d\theta^2}\log L(\theta) &= \frac{n}{\theta^2} - 2\frac{\sum\limits_{i=1}^{n} x_i}{\theta^3} \qquad (5.25)\\ &= \frac{n}{\theta^2}\left(1 - 2\frac{\hat{\theta}}{\theta}\right), \end{aligned}$$

which is clearly negative when $\theta = \hat{\theta}$.

The maximum likelihood estimator is thus $\hat{\theta} = \bar{X}$. It is unbiased: from (3.56) we have $E(\hat{\theta}) = E(\bar{X}) = E(X_1) = \theta$. The mean squared error is then

$$\begin{aligned} MSE_{\hat{\theta}} &= \mathrm{Var}(\hat{\theta}|\theta) \qquad (5.26)\\ &= \mathrm{Var}(\bar{X})\\ &= \frac{\mathrm{Var}(X_1)}{n} \quad \text{from (3.57)}\\ &= \frac{\theta^2}{n}. \end{aligned}$$

[Figure 5.7]

The standard error is $\sigma_{\hat{\theta}} = \theta/\sqrt{n}$. We see that both of these increase with θ and decrease as n increases. This means it is harder to pin down with precision the time to failure for long lasting modules, and that expected accuracy increases as the amount of data available increases.

We first encountered the exponential distribution in a slightly different form, as

$$\begin{aligned} f(x|\lambda) &= \lambda e^{-\lambda x} \quad \text{for } x > 0.\\ &= 0 \quad \text{for } x \le 0. \end{aligned}$$

In the present example, where X is the time until failure, we will later see that λ is the expected number of replacements per unit time, a quantity that is of equal interest to that of $\theta = E(X)$. We could now attack the problem of estimating λ by looking for the maximum likelihood estimator, $\hat{\lambda}$, as that value of λ for which

$$\prod_{i=1}^{n} f(x_i|\lambda)$$

is a maximum. But there is a simpler way, for in fact the simple relationship between θ and λ, namely $\lambda = 1/\theta$, must also hold for their maximum likelihood estimators:

$$\hat{\lambda} = \frac{1}{\hat{\theta}} = \frac{1}{\bar{X}}. \tag{5.27}$$

This relation is general and is referred to as the *invariance* property of maximum likelihood estimators: If $\hat{\theta}$ is a maximum likelihood estimate of θ, and $h(\theta)$ is any function of θ, then $h(\hat{\theta})$ is maximum likelihood estimate of $h(\theta)$. In symbols, $\widehat{h(\theta)} = h(\hat{\theta})$. When, as in the above example, h is a one-to-one function, this property follows quite simply. In that case, the likelihood function could as well be considered as a function of $h(\theta)$, and if any value $h(\theta_0)$ produced a higher likelihood than $h(\hat{\theta})$, this would imply $L(\theta_0) > L(\hat{\theta})$, contradicting the assumption that $\hat{\theta}$ maximizes $L(\theta)$. When h is not one-to-one, we simply *define* $\widehat{h(\theta)} = h(\hat{\theta})$.

It is worth noting that, while the invariance property makes it easy to find maximum likelihood estimators in some situations, it does not guarantee that the properties of $\hat{\theta}$ will carry over to $h(\hat{\theta})$, or even that the properties of $h(\hat{\theta})$ will be easy to determine. In the above example, $\hat{\theta} = \bar{X}$ is an unbiased estimator of θ, but $\hat{\lambda} = 1/\bar{X}$ is not an unbiased estimator of λ (in fact, $E(1/\bar{X}) \neq 1/E(\bar{X})$ unless $\text{Var}(\bar{X}) = 0$). Also, the mean squared error of $\hat{\lambda}$ is not easy to evaluate.

Example 5.D. The General Normal Distribution. Suppose our data consist of $X_1, X_2, \ldots, X_n$, where the X_i's are assumed independent, each with a $N(\mu, \sigma^2)$ distribution, where both μ and σ^2 are unknown. For example, we may have n independent weighings of a single object with true weight μ, made by a scale whose error variance σ^2 is unknown. Thus we wish to estimate two parameters, μ and σ^2. The density of a single X_i is

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

The likelihood function is

$$\begin{aligned} L(\mu, \sigma^2) &= \prod_{i=1}^{n} f(x_i|\mu, \sigma^2) \qquad (5.28) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(x_i-\mu)^2}, \end{aligned}$$

and

$$\log L(\mu, \sigma^2) = \log(2\pi)^{-\frac{n}{2}} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

To simplify notation for the derivation, write $\theta = \sigma^2$, so we have

$$\log L(\mu, \theta) = \log(2\pi)^{-\frac{n}{2}} - \frac{n}{2}\log\theta - \frac{1}{2\theta}\sum_{i=1}^{n}(x_i - \mu)^2.$$

We then find

$$\begin{aligned} \frac{d}{d\mu}\log L(\mu, \theta) &= \frac{1}{\theta}\sum_{i=1}^{n}(x_i - \mu) \qquad (5.29) \\ &= \frac{n}{\theta}(\bar{x} - \mu) \\ \frac{d}{d\theta}\log L(\mu, \theta) &= -\frac{n}{2\theta} + \frac{1}{2\theta^2}\sum_{i=1}^{n}(x_i - \mu)^2. \qquad (5.30) \end{aligned}$$

We set both equal to zero and solve the equations simultaneously for $\hat{\mu}$ and $\hat{\theta}$. The first gives

$$\frac{n}{\hat{\theta}}(\bar{x} - \hat{\mu}) = 0$$

or

$$\hat{\mu} = \bar{x}, \qquad (5.31)$$

the *sample mean.* Substituting this in the second gives

$$-\frac{n}{2\hat{\theta}} + \frac{1}{2\hat{\theta}^2} \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$$

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2; \quad \text{that is}$$

$$\hat{\sigma^2} = \frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{5.32}$$

It remains to verify that these give a maximum, and not a minimum or a saddlepoint. We find

$$\frac{d^2}{d\mu^2}\log L(\mu,\theta) = -\frac{n}{\theta} < 0, \quad \text{since } \theta > 0, \tag{5.33}$$

$$\frac{d^2}{d\theta^2}\log L(\mu,\theta) = \frac{n}{2\theta^2} - \frac{1}{\theta^3}\Sigma(x_i - \mu)^2 \tag{5.34}$$

so for $\theta = \hat{\theta}$ and $\mu = \hat{\mu}$,

$$\begin{aligned}\frac{d^2}{d\theta^2}\log L(\mu,\theta) &= \frac{n}{2\hat{\theta}^2} - \frac{n}{\hat{\theta}^3} \cdot \hat{\theta} \\ &= -\frac{n}{2\hat{\theta}^2} < 0.\end{aligned}$$

Also,

$$\begin{aligned}\frac{d^2}{d\theta d\mu}\log L(\mu,\theta) &= \frac{-n}{\theta^2}(\bar{x} - \mu) \\ &= 0 \quad \text{when } \mu = \hat{\mu},\end{aligned} \tag{5.35}$$

since for $\mu = \hat{\mu}$ and $\theta = \hat{\theta}$

$$\left(\frac{d^2}{d\mu^2}\log L(\mu,\theta)\right)\left(\frac{d^2}{d\theta^2}\log L(\mu,\theta)\right) - \left(\frac{d^2}{d\theta d\mu}\log L(\mu,\theta)\right)^2 > 0$$

and

$$\frac{d^2}{d\mu^2}\log L(\mu,\theta) < 0,$$

we have located a maximum.

Let us investigate the properties of $\hat{\mu}$ and $\hat{\sigma^2}$. First,

$$\begin{aligned}E(\hat{\mu}) &= E(\bar{X}) \\ &= \mu \quad \text{from (3.56)},\end{aligned}$$

so $\hat{\mu}$ is an unbiased estimator of μ. Next, we can write

$$\begin{aligned}\hat{\sigma^2} &= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\ &= \frac{1}{n}\left[\sum_{i=1}^{n}X_i^2 - 2\bar{X}\sum_{i=1}^{n}X_i + n\bar{X}^2\right] \\ &= \frac{1}{n}\left[\Sigma X_i^2 - 2\frac{\left(\sum_{i=1}^{n}X_i\right)^2}{n} + \frac{\left(\sum_{i=1}^{n}X_i\right)^2}{n}\right] \\ &= \frac{1}{n}\left[\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}\right].\end{aligned} \tag{5.36}$$

Now for any random variable W, we have from (2.26) that

$$E(W^2) = \mathrm{Var}(W) + (E(W))^2.$$

Then, since $E(X_i) = \mu, \mathrm{Var}(X_i) = \sigma^2$,

$$\begin{aligned} E\left(\sum_{i=1}^{n} X_i^2\right) &= \sum_{i=1}^{n} E(X_i^2) \\ &= \sum_{i=1}^{n} (\sigma^2 + \mu^2) \\ &= n\sigma^2 + n\mu^2. \end{aligned} \tag{5.37}$$

Also,

$$\begin{aligned} E\left(\sum_{i=1}^{n} X_i\right)^2 &= \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) + \left[E\left(\sum_{i=1}^{n} X_i\right)\right]^2 \\ &= n\sigma^2 + (n\mu)^2 \end{aligned} \tag{5.38}$$

from (3.51) and (3.31), so from (5.36) we get

$$\begin{aligned} E(\hat{\sigma^2}) &= \frac{1}{n}\left[n\sigma^2 + n\mu^2 - \frac{(n\sigma^2 + n^2\mu^2)}{n}\right] \\ &= \left(\frac{n-1}{n}\right)\sigma^2. \end{aligned} \tag{5.39}$$

Thus the maximum likelihood estimator of σ^2 is slightly biased, with bias

$$\begin{aligned} B(\sigma^2) &= \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2 \\ &= -\frac{\sigma^2}{n}, \end{aligned} \tag{5.40}$$

which is small if n is large relative to σ^2. The expression (5.39) suggests a way to eliminate the bias: multiply by $\frac{n}{n-1}$. This gives us the commonly used estimator, the *sample variance*

$$\begin{aligned} \mathbf{s}^2 &= \left(\frac{n}{n-1}\right)\hat{\sigma^2} \\ &= \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2. \end{aligned} \tag{5.41}$$

From (5.39), $E(\mathbf{s}^2) = \sigma^2$; $\mathbf{s}^2$ is an unbiased estimator of σ^2, though for large n $\mathbf{s}^2$ and $\hat{\sigma^2}$ differ but little.

We shall generally prefer $\mathbf{s}^2$ to $\hat{\sigma^2}$ for estimating a normal variance σ^2, but the preference is more due to custom and the form in which certain common tables are presented than the fact that $\mathbf{s}^2$ is unbiased. One reason unbiasedness is relatively unimportant here is that we are usually interested in the standard deviation σ, not σ^2. To find the maximum likelihood estimator of σ we could consider the likelihood function (5.28) as a function of σ and go through a full analysis, but it is far simpler to use the invariance property:

$$\hat{\sigma} = \sqrt{\hat{\sigma^2}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Now, it turns out that *both* $\hat{\sigma}$ and $\mathbf{s} = \sqrt{\mathbf{s^2}}$, the *sample standard deviation,* are biased estimators of σ. In fact, it can be shown that

$$E(\mathbf{s}) = b_n\sigma$$

and so

$$E(\hat{\sigma}) = \sqrt{\frac{n-1}{n}}\, b_n\sigma,$$

where

$$b_n = \sqrt{\frac{2}{n-1}} \cdot \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}. \tag{5.42}$$

n	4	10	100
b_n	.921	.973	.997

Table 5.2

For large n, b_n is near 1, so while in principle we could "correct" $\mathbf{s}$ to be unbiased, using $\mathbf{s}/b_n$ to estimate σ, this is practically never done.

We will later find that

$$\text{Var}(\mathbf{s}^2) = \frac{2\sigma^4}{(n-1)}; \tag{5.43}$$

we can exploit that here to compare the mean squared errors of $\hat{\sigma^2}$ and $\mathbf{s}^2$. We have, since $\mathbf{s}^2$ is unbiased,

$$\begin{aligned} MSE_{\mathbf{s}^2}(\sigma^2) &= \text{Var}(\mathbf{s}^2) \\ &= \frac{2\sigma^4}{n-1} \end{aligned}$$

while from (5.40) and (5.4)

$$\begin{aligned} MSE_{\hat{\sigma^2}}(\sigma^2) &= \text{Var}\left(\left(\frac{n-1}{n}\right)\mathbf{s}^2\right) + \left(-\frac{\sigma^2}{n}\right)^2 \\ &= \left(\frac{n-1}{n}\right)^2 \cdot \frac{2\sigma^4}{(n-1)} + \frac{\sigma^4}{n^2} \\ &= \left(\frac{2n-1}{n^2}\right)\sigma^4 \end{aligned}$$

Since

$$\frac{\left(\frac{2n-1}{n^2}\right)}{\left(\frac{2}{n-1}\right)} = 1 - \left(\frac{3n-1}{2n^2}\right) < 1,$$

we see that

$$MSE_{\hat{\sigma^2}}(\sigma^2) < MSE_{\mathbf{s}^2}(\sigma^2)$$

[Figure 5.8]

for all σ^2, although for large n their ratio is nearly 1. Thus despite its bias, $\hat{\sigma^2}$ has a smaller mean squared error than $\mathbf{s}^2$.

5.5 The Distribution of Sums.

The properties of any estimator $\hat{\theta}$ depend upon its distribution, $f_{\hat{\theta}}(x|\theta)$. For some purposes we do not need to know this distribution in detail; the bias and mean squared error can be determined knowing only $E(\hat{\theta})$ and $\text{Var}(\hat{\theta})$. But for more detailed assessments of accuracy we will need to know more.

Since an estimator $\hat{\theta} = \hat{\theta}(X)$ or $\hat{\theta}(X_1, X_2, \ldots, X_n)$ is a transformation of the data, its distribution can be quite complicated, even analytically intractable. In some cases, though, the distribution can be rather simply described.

Example 5.E. The Binomial Estimators. For the estimator $\hat{\theta}_1(X) = \frac{X}{n}$ of the parameter θ of a binomial distribution, we can apply (1.28) directly. If $h(x) = x/n$, $g(y) = ny$ and

$$\begin{aligned} p_{\hat{\theta}_1}(y) &= P(\hat{\theta}_1 = y) \\ &= p_X(ny) \\ &= b(ny; n, \theta). \end{aligned}$$

This is just a rescaled binomial distribution.

[Figure 5.9]

Other such estimators can be handled in equally simple ways directly: the distribution of $\hat{\theta}_3(X) = (X+1)/(n+2)$ is

$$\begin{aligned} p_{\hat{\theta}_2}(y) &= P(\hat{\theta}_2 = y) \\ &= b((n+2)y - 1; n, \theta). \end{aligned}$$

Many of the estimates we have encountered are based on sums of independent random variables; for example, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ appeared as a maximum likelihood estimator both for the exponential mean θ and the normal mean μ. In many cases it is easy to handle the distribution of sums directly, using the following general result: If (X, Y) is a bivariate random variable with density $f(x, y)$, then the density of

$$Z = X + Y$$

is given by

$$f_Z(z) = \int_{-\infty}^{\infty} f(z - y, y)dy. \tag{5.44}$$

If X and Y are independent, $f(x, y) = f_X(x)f_Y(y)$ and we have

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z - y)f_Y(y)dy. \tag{5.45}$$

It is not hard to prove (5.44); since $F_Z(z) = P(Z \leq z) = P(X + Y \leq z)$ is the probability (X, Y) is in the shaded region of Figure 5.10, it is found by integrating $f(x, y)$ over that region:

$$F_Z(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{z-y} f(x, y)dxdy.$$

[Figure 5.10]

Then

$$\begin{aligned} f_Z(z) &= \frac{d}{dz}F_Z(z) \\ &= \int_{-\infty}^{\infty} \frac{d}{dz}\left(\int_{-\infty}^{z-y} f(x, y)dx\right) dy \\ &= \int_{-\infty}^{\infty} f(z - y, y)dy, \end{aligned}$$

and (5.45) follows directly.

Example 5.F. The Sum of Independent Normal Random Variables. One easy and extremely useful application of (5.45) is to show that if X and Y are independent and each is normally distributed, then $Z = X + Y$ has a normal distribution. This is sometimes called the *reproductive property* of the normal distribution. Suppose X has a $N(\mu, \sigma^2)$ distribution and Y has a $N(\theta, \tau^2)$ distribution. Then (5.45) gives us

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(z-y-\mu)^2} \cdot \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{1}{2\tau^2}(y-\theta)^2} dy \qquad (5.46) \\ &= \frac{1}{2\pi\sigma\tau} \int_{-\infty}^{\infty} e^{-\frac{1}{2}[(z-y-\mu)^2/\sigma^2 + (y-\theta)^2/\tau^2]} dy. \end{aligned}$$

Now the part of the exponent in brackets is a quadratic function of y and z, and so it can be written in the form

$$A(z - B)^2 + C(y - Dz)^2 + E. \qquad (5.47)$$

(The algebra involved in evaluating A, B, C, D, and E is straightforward but tedious, and as we will see, unnecessary.) Thus

$$f_Z(z) = \frac{1}{\sqrt{2\pi}\sigma\tau\sqrt{C}} e^{-\frac{E}{2}} \cdot e^{-\frac{A}{2}(z-B)^2} \cdot \int_{-\infty}^{\infty} \sqrt{\frac{C}{2\pi}} e^{-\frac{C}{2}(y-Dz)^2} dy.$$

Now the integral is just the integral of a $N\left(Dz, \frac{1}{C}\right)$ density, so it must equal 1. Thus

$$f_Z(z) \propto e^{-\frac{A}{2}(z-B)^2}.$$

But this is (except for a constant needed to scale the density to integrate to 1) a $N(B, 1/A)$ density. All that is needed is to find $B = E(Z)$ and $1/A = \text{Var}(Z)$. But we know from Chapter 3, (3.38) and (3.40) that

$$\begin{aligned} E(Z) &= E(X) + E(Y) \\ &= \mu + \theta \end{aligned}$$

and

$$\begin{aligned} \text{Var}(Z) &= \text{Var}(X) + \text{Var}(Y) \\ &= \sigma^2 + \tau^2. \end{aligned}$$

Hence Z has a $N(\mu + \theta, \sigma^2 + \tau^2)$ distribution. (Alternatively, $A = 1/(\sigma^2 + \tau^2)$ and $B = \mu + \theta$ can be found by algebra, by equating (5.47) to the exponent in (5.46).) We shall obtain this same result from a different approach later.

It follows by induction that if $X_1, X_2, X_3, \ldots, X_n$ are independent random variables, where X_i has a $N(\mu_i, \sigma_i^2)$ distribution, then

$$\sum_{i=1}^{n} X_i \quad \text{has a } N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right) \text{ distribution.} \qquad (5.48)$$

In particular, if the X_i's are independent, each with a $N(\mu, \sigma^2)$ distribution, then

$$\sum_{i=1}^{n} X_i \quad \text{has a } N(n\mu, n\sigma^2) \text{ distribution,} \qquad (5.49)$$

and

$$\bar{X} \quad \text{has a } N\left(\mu, \frac{\sigma^2}{n}\right) \text{ distribution.} \qquad (5.50)$$

Thus the maximum likelihood estimator of a normal mean has a very simple distribution, and if we specify σ^2 we can use tables of the normal distribution to calculate $P(|\bar{X}-\mu|<c)$ for any c.

Example 5.G. The Chi-Square Distribution. We have already encountered the Chi-square distribution with 1 degree of freedom in Example 1.M: it is the distribution of U^2 where U has a $N(0,1)$ distribution, and we found its density to be

$$\begin{aligned} f_{U^2}(y) &= \frac{1}{\sqrt{2\pi y}}e^{-y/2} \quad \text{for } y>0 \\ &= 0 \quad \text{for } y \le 0. \end{aligned} \tag{1.34}$$

The *Chi-square distribution with n degrees of freedom* is defined to be the probability distribution of

$$\chi^2(n) = U_1^2 + U_2^2 + \cdots + U_n^2, \tag{5.51}$$

where $U_1, U_2, \ldots, U_n$ are independent, each with a $N(0,1)$ distribution. The appropriateness of the term "degrees of freedom" will be clearer later; for now we note that $\chi^2(n)$ involves n terms that are independent and hence varying freely, one from the other. The density of $\chi^2(n)$ is given by

$$\begin{aligned} f_{\chi^2(n)}(x) &= \frac{1}{2^{n/2}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1}e^{-\frac{x}{2}} \quad \text{for } x>0 \qquad (5.52) \\ &= 0 \quad \text{for } x \le 0. \end{aligned}$$

For $n=1$, $\Gamma\left(\frac{n}{2}\right)=\sqrt{\pi}$ and

$$\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}x^{\frac{n}{2}-1} = \frac{1}{\sqrt{2\pi x}},$$

so (5.52) agrees with (1.34).

[Figure 5.11]

We can use (5.45) to verify that the density of $\chi^2(n)$ is as given by (5.52). We proceed by induction. We have already verified (5.52) for $n=1$, in Example 1.M. Assume it holds also for $n=k-1$. Now, let

$$\begin{aligned} X &= U_1^2 + U_2^2 + \cdots + U_{k-1}^2 \\ Y &= U_k^2 \end{aligned}$$

where $U_1, U_2, \ldots, U_k$ are independent, each $N(0,1)$. Then X and Y are independent, and $\chi^2(k)=X+Y$ has a Chi-square distribution with k degrees of freedom, by definition. From the induction hypothesis, X has density given by (5.52) with $n=k-1$:

$$\begin{aligned} f_X(x) &= \frac{1}{2^{\frac{k-1}{2}}\Gamma\left(\frac{k-1}{2}\right)}x^{\frac{k-1}{2}-1}e^{-\frac{x}{2}} \quad \text{for } x>0 \\ &= 0 \quad \text{for } x \le 0, \end{aligned}$$

and Y has density given by (1.34),

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi y}}e^{-y/2} \quad \text{for } y>0 \\ &= 0 \quad \text{for } y \le 0. \end{aligned}$$

Then from (5.45), the density of $\chi^2(k)$ is, for $z>0$,

$$\begin{aligned} f_{\chi^2(k)}(z) &= \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy \\ &= \int_0^z \frac{1}{2^{\frac{k-1}{2}}\Gamma\left(\frac{k-1}{2}\right)}\cdot(z-y)^{\frac{k-1}{2}-1}e^{-\frac{(z-y)}{2}}\cdot\frac{1}{\sqrt{2\pi y}}e^{-\frac{y}{2}}dy \\ &\text{(since } f_X(z-y)=0 \text{ for } y \ge z\text{)} \\ &= \frac{1}{2^{\frac{k-1}{2}}\Gamma\left(\frac{k-1}{2}\right)\sqrt{2\pi}}e^{-\frac{z}{2}}\int_0^z (z-y)^{\frac{k-3}{2}}y^{-\frac{1}{2}}dy. \end{aligned}$$

Now from a change of variables $u = y/z$ (so $zdu = dy$, $(z-y)^{\frac{k-3}{2}} = z^{\frac{k-3}{2}}(1-u)^{\frac{k-3}{2}}$, $y^{-\frac{1}{2}} = z^{-\frac{1}{2}}u^{-\frac{1}{2}}$) we get

$$\begin{aligned}\int_0^z (z-y)^{\frac{k-3}{2}} y^{-\frac{1}{2}} dy &= z^{\frac{k-3}{2}} \cdot z^{-\frac{1}{2}} \cdot z \int_0^1 (1-u)^{\frac{k-3}{2}} u^{-\frac{1}{2}} du \\ &= z^{\frac{k}{2}-1} \cdot \int_0^1 (1-u)^{\frac{k-3}{2}} u^{-\frac{1}{2}} du.\end{aligned}$$

But this last integral does not depend upon z, so we have established that

$$\begin{aligned}f_{\chi^2(k)}(z) &= Cz^{\frac{k}{2}-1}e^{-\frac{z}{2}} \quad \text{for } z > 0 \\ &= 0 \quad \text{for } z \le 0.\end{aligned}$$

This agrees with (5.52) for $n = k$ except for the constant; since the only role the constant plays is as a scaling factor, to guarantee that $\int_0^\infty f(z)dz = 1$, we must have

$$C = \frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}. \tag{5.53}$$

Hence (5.52) holds for $n = k$. By induction, it gives the density of $\chi^2(n)$ for any n. An alternative way to verify (5.53) is by recognizing the integral as a Beta function;

$$\begin{aligned}\int_0^1 (1-u)^{\frac{k-3}{2}} u^{-\frac{1}{2}} du &= B\left(\frac{1}{2}, \frac{k-1}{2}\right) \\ &= \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{k-1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \\ &= \sqrt{\pi}\, \frac{\Gamma\left(\frac{k-1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)}.\end{aligned}$$

The principle applications of the Chi-square distribution will be explored later. For example, we will show that if $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$ (as in Example 5.D), and $\mathbf{s}^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (X_i - \bar{X})^2$, then $(n-1)\mathbf{s}^2/\sigma^2$ has a $\chi^2(n-1)$ distribution. Since $(n-1)\mathbf{s}^2/\sigma^2 = n\hat{\sigma^2}/\sigma^2$, this latter quantity has the same distribution. Thus the Chi-square will appear as the distribution of a multiple of the sample variance of a normally distributed sample.

One important property of the Chi-square distribution is obvious from the definition (5.51); if $\chi^2(k)$ and $\chi^2(m)$ are independent random variables, with Chi-square distributions with k and m degrees of freedom ("d.f.") respectively, then their sum has a $\chi^2(k+m)$ distribution. Symbolically,

$$\chi^2(k+m) = \chi^2(k) + \chi^2(m), \tag{5.54}$$

keeping in mind that the random variables on the right must be independent. This is clear from the definition: If

$$\chi^2(k) = U_1^2 + \cdots + U_k^2$$

and

$$\chi^2(m) = U_{k+1}^2 + \cdots + U_{k+m}^2$$

then

$$\chi^2(k) + \chi^2(m) = U_1^2 + \cdots + U_{k+m}^2.$$

The definition of the Chi-square distribution also makes it easy to calculate its expectation and variance. Since the U_i are independent $N(0,1)$, $E(U_i^2)=1$ for all i and

$$\begin{aligned} E(\chi^2(n)) &= E\left(\sum_{i=1}^n U_i^2\right) \\ &= \sum_{i=1}^n E(U_i^2) \\ &= n. \end{aligned} \tag{5.55}$$

Also,

$$\begin{aligned} \mathrm{Var}(\chi^2(n)) &= \mathrm{Var}\left(\sum_{i=1}^n U_i^2\right) \\ &= \sum_{i=1}^n \mathrm{Var}(U_i^2) \\ &= n\cdot \mathrm{Var}(U_1^2). \end{aligned}$$

Now $Y=U_1^2$ has density $f_Y(y)$ given by (1.34), so

$$\begin{aligned} E(Y^2) &= \int_0^\infty y^2\cdot\frac{1}{\sqrt{2\pi y}}e^{-\frac{y}{2}}\,dy \\ &= \frac{1}{\sqrt{2\pi}}\int_0^\infty y^{\frac{3}{2}}e^{-\frac{y}{2}}\,dy \\ &= \frac{2^{\frac{5}{2}}\Gamma\left(\frac{5}{2}\right)}{\sqrt{2\pi}}\int_0^\infty \frac{1}{2^{\frac{5}{2}}\Gamma\left(\frac{5}{2}\right)}y^{\frac{5}{2}-1}e^{-\frac{y}{2}}\,dy. \end{aligned}$$

The latter integral is the integral of the density (5.52) for $n=5$, and so it must equal 1. Then

$$\begin{aligned} E(Y^2) &= \frac{2^{\frac{5}{2}}\Gamma\left(\frac{5}{2}\right)}{\sqrt{2\pi}} \\ &= 4\cdot\frac{\Gamma\left(\frac{5}{2}\right)}{\sqrt{\pi}} \\ &= 4\cdot\frac{\frac{3}{2}\cdot\frac{1}{2}\cdot\Gamma\left(\frac{1}{2}\right)}{\sqrt{\pi}} \\ &= 3 \quad \text{from (2.8).} \end{aligned}$$

Thus $E(Y^2)=3$ and $\mathrm{Var}(U_1^2)=3-1=2$, so

$$\mathrm{Var}(\chi^2(n)) = 2n. \tag{5.56}$$

Example 5.H. The Exponential and Gamma Distributions. For the special case $n=2$, the Chi-square distribution is an exponential distribution: for $n=2$, $2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)=2$ and (5.52) gives

$$\begin{aligned} f_{\chi^2(2)}(x) &= \frac{1}{2}e^{-\frac{x}{2}} \quad \text{for } x>0 \\ &= 0 \quad \text{for } x\le 0, \end{aligned} \tag{5.57}$$

an Exponential density with $\theta=2$. This fact can be used to determine the distribution of the maximum likelihood estimator of θ. In Example 5.C we found that if $X_1,\ldots,X_n$ are independent Exponential (θ), then

$$\hat{\theta} = \bar{X}.$$

Now if X_i has density

$$\begin{aligned} f(x|\theta) &= \frac{1}{\theta} e^{-\frac{x}{\theta}} \quad \text{for } x > 0 \\ &= 0 \quad \text{for } x \leq 0, \end{aligned}$$

then $Y_i = \frac{2}{\theta} X_i$ has (from (1.35) with $a = \frac{2}{\theta}, b = 0$) the $\chi^2(2)$ density given by (5.57). Now $\sum\limits_{i=1}^{n} Y_i = \frac{2}{\theta} \sum\limits_{i=1}^{n} X_i$, and since the X_i's are independent, it follows from additive property of the Chi-square distribution (5.54) that $\frac{2}{\theta} \sum\limits_{i=1}^{n} X_i$ has a Chi-square distribution with $\sum\limits_{i=1}^{n} 2 = 2n$ degrees of freedom. That is, $\frac{2n}{\theta} \bar{X} = \frac{2n}{\theta} \hat{\theta}$ has the density $f_{\chi^2(2n)}(y)$. We could then find the density of $\hat{\theta}$ from (1.35) with $a = \frac{\theta}{2n}$ and $b = 0$; it is $\frac{2n}{\theta} f_{\chi^2(2n)} \left(\frac{2nx}{\theta} \right)$

[Figure 5.12]

The Chi-square distributions are the most important special cases of a more general class of distributions called the *Gamma distributions,* $G(\alpha, \beta)$. They depend upon two parameters, α and β, and they are defined in terms of their densities, by

$$\begin{aligned} f(x|\alpha, \beta) &= \frac{x^{\alpha-1} e^{-\frac{x}{\beta}}}{\beta^{\alpha} \Gamma(\alpha)} \quad \text{for } x > 0 \\ &= 0 \quad \text{for } x \leq 0. \end{aligned} \tag{5.58}$$

They are similar to the Chi-square densities in appearance; indeed, the $G\left(\frac{n}{2}, 2\right)$ distribution is a $\chi^2(n)$ distribution, as can be verified by comparing (5.52) and (5.58).

It is easy to see, from (1.35) with $b = 0$, that if Z has a $\chi^2(n)$ distribution with density $f_{\chi^2(n)}(z)$, then $Y = aZ$ has density

$$\frac{1}{a} f_{\chi^2(n)} \left(\frac{y}{a} \right) = \frac{1}{(2a)^{\frac{n}{2}} \Gamma\left(\frac{n}{2}\right)} y^{\frac{n}{2}-1} e^{-\frac{y}{2a}} \quad \text{for } y > 0,$$

a $G\left(\frac{n}{2}, 2a\right)$ distribution. Thus we could describe the distribution of $\hat{\theta}$ as a $G\left(n, \frac{\theta}{n}\right)$ distribution.

If Y has a $G(\alpha, \beta)$ distribution, then (see Problems)

$$E(Y) = \alpha\beta, \tag{5.59}$$

$$\text{Var}(Y) = \alpha\beta^2. \tag{5.60}$$

5.6 The Approximate Distribution of Estimators.

In the previous section we saw how in some cases (the Normal, the Exponential) the probability distribution of a maximum likelihood estimator based upon a sum of random variables could be determined explicitly. Such cases are rare, however, and more frequently we must fall back upon approximations. Fortunately, there is available an elegant result in probability theory that will provide the basis for a broad spectrum of applications. It has come to be called the *Central Limit Theorem,* where "central" should be understood in the sense "fundamental." In its simplest form it states: If $X_1, X_2, \ldots, X_n$ are independent, each with the same distribution with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$, then if n is large,

$$\sum_{i=1}^{n} X_i \quad \text{is approximately } N(n\mu, n\sigma^2), \tag{5.61}$$

$$\bar{X} \quad \text{is approximately } N\left(\mu, \frac{\sigma^2}{n}\right). \tag{5.62}$$

By these statements we mean that for any constants $a < b$,

$$P\left(a < \sum_{i=1}^{n} X_i < b\right) \simeq \Phi\left(\frac{b - n\mu}{\sqrt{n}\sigma}\right) - \Phi\left(\frac{a - n\mu}{\sqrt{n}\sigma}\right). \tag{5.63}$$

Furthermore, the errors in these approximations all become smaller the larger n is, vanishing in the limit as $n \to \infty$. In some respects these results should not seem surprising. If the X_i are Normally distributed, then we saw in Section 5.5 that (5.61) and (5.62) hold exactly; in that case the word "approximately" can be struck out and we have a stronger result. Also, even if the X_i are not Normally distributed, we saw in Chapter 3 that

$$E(\bar{X}) = \mu \tag{3.56}$$

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}, \tag{3.57}$$

which is in agreement with (5.62); a similar agreement holds for (5.61). What is remarkable about the Central Limit Theorem is not the expectation or variance of the approximating distribution, but its form: not only are sums of independent normally distributed random variables normally distributed, but sums of independent random variables with almost *any* distribution are approximately normal. Even more general forms of this theorem are true; subject to some restrictions that guarantee that no one X_i or small set of X_i's dominate the sum $\sum_{i=1}^{n} X_i$, the X_i's do not need to all have the same distribution, and they need only be "approximately" independent.

The Central Limit Theorem's main strength is that it gives us an approximation to the distribution of a sum or average in the standardized form

$$W = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

A weaker corollary applies to unstandardized averages, and is called the *weak law of large numbers,* or simply the *law of large numbers.* It states that as n increases, $\bar{X}$ approaches μ *in probability* (written $\bar{X} \xrightarrow{P} \mu$), which is defined to mean that for any $\epsilon > 0$ (however small), an n can be found so that

$$P(|\bar{X} - \mu| < \epsilon) > 1 - \epsilon. \tag{5.66}$$

Thus the probability that $\bar{X}$ is near μ can be made arbitrarily high by increasing the sample size n. This follows immediately from (5.63):

$$\begin{aligned} P(|\bar{X} - \mu| < \epsilon) &= P(\mu - \epsilon < \bar{X} < \mu + \epsilon) \\ &\simeq \Phi\left(\frac{\epsilon}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{-\epsilon}{\sigma/\sqrt{n}}\right) \\ &= \Phi\left(\sqrt{n}\frac{\epsilon}{\sigma}\right) - \Phi\left(-\sqrt{n}\frac{\epsilon}{\sigma}\right) \\ &= \int_{-\sqrt{n}\frac{\epsilon}{\sigma}}^{\sqrt{n}\frac{\epsilon}{\sigma}} \phi(x)dx. \end{aligned}$$

By choosing n large, this integral can be made as close to

$$\int_{-\infty}^{\infty} \phi(x)dx = 1$$

as desired.

The Law of Large Numbers tells us $\bar{X}$ is likely to be close to μ for large n; the Central Limit Theorem describes its approximate distribution in more detail.

We have already seen graphical illustration of the Central Limit Theorem in one case. The Chi-square distribution with n degrees of freedom, $\chi^2(n)$, is the distribution of a sum $\sum\limits_{i=1}^{n} X_i$, where each $X_i = U_i^2$ has a $\chi^2(1)$ distribution. The Central Limit Theorem says that if n is large, $\sum\limits_{i=1}^{n} X_i$ is approximately $N(n, 2n)$, and Figure 5.11 illustrates this. Even though the $\chi^2(1)$ distribution is extremely non-Normal (see Figure 5.11 (a)), we can see that by the time $n = 30$, the distribution $\chi^2(n)$ is fairly close to the symmetrical Normal shape, nearly centered about 30 (its actual expectation), with standard deviation $\sqrt{60} = 7.75$.

How close an approximation we can expect from (5.61)-(5.65) depends upon the distribution of X_i, upon n, and upon a and b. A rough rule of thumb suggests the approximation is usually adequate for statistical purposes if $n \geq 30$, but if the distribution of X_i is at all similar to the normal (as is frequently the case), the approximation can be excellent for n as low as 5 or 10.

We will not prove the Central Limit Theorem; its proof requires techniques from probability theory which we have not introduced. The following heuristic argument helps make it plausible. Suppose that n is very large, in fact $n = 2m$ where m is large. Suppose $X_1, \ldots, X_n$ are independent; we can write $\sum\limits_{i=1}^{n} X_i$ as a sum of 2 blocks, each consisting of m random variables:

$$\begin{aligned}\sum_{i=1}^{n} X_i = X_1 + \cdots + X_m \\ + X_{m+1} + \cdots + X_n\end{aligned}$$

Write $Y_1 = X_1 + \cdots + X_m$, $Y_2 = X_{m+1} + \cdots + X_n$, getting

$$\sum_{i=1}^{n} X_i = Y_1 + Y_2.$$

Now if there is a general approximating distribution, so that each sum of X's has the same distribution when standardized, then

$$\frac{Y_1 - \mu_{Y_1}}{\sigma_{Y_1}}, \frac{Y_2 - \mu_{Y_2}}{\sigma_{Y_2}}, \frac{(Y_1 + Y_2) - \mu_{Y_1 + \mu_2}}{\sigma_{Y_1 + Y_2}}$$

must have approximately that same distribution. That distribution must share the reproductive property (Example 5.F) of the normal distribution. Now the normal distribution is not the only distribution where Y_1, Y_2 are independent and the distributions of Y_1, Y_2, and $Y_1 + Y_2$ differ only by a linear change of scale (the others are called the *stable distributions),* but it is the only one with a finite variance; hence it is the only possibility for a general-purpose approximating distribution.

The Central Limit Theorem gives an explanation why approximately normal distributions are common in applications, since many measurements are themselves made up of an aggregate of a number of roughly independent components. The weight of a sack of grain is the aggregate of many small weights; the yield of an apple tree is the total yield of its several branches.

The Central Limit Theorem gives an approximation to the distribution of many estimators; for example, $\bar{X}$ may always be considered as an estimator of $E(X_i)$, even though it is the maximum likelihood estimator for only a few distributions of X_i (eg. for $N(\mu, \sigma^2)$ and Exponential (θ)). But what of cases such as that of Example 5.C, where the maximum likelihood of λ, namely $\hat{\lambda} = 1/\bar{X}$, is not equal to a sum or average, and the theorem does not apply? It turns out that if the estimator being considered is a maximum likelihood estimator, it will quite generally be approximately normally distributed. A full statement of conditions under which this is true is beyond the scope of this book, but the following rather loose statement captures the main idea:

Fisher's Approximation.

If the data consist of independent $X_1, X_2, \ldots, X_n$, each with distribution $f(x|\theta)$, and the maximum likelihood estimator $\hat{\theta}$ is found by solving $\frac{d}{d\theta}L(\theta) = 0$ or $\frac{d}{d\theta}\log L(\theta) = 0$, then for large n, $\hat{\theta}$ has approximately a $N\left(\theta, \frac{\tau^2(\theta)}{n}\right)$ distribution, where

$$\frac{1}{\tau^2(\theta)} = E\left(\frac{d}{d\theta}\log f(X_1|\theta)\right)^2 = -E\left[\frac{d^2}{d\theta^2}\log f(X_1|\theta)\right], \tag{5.67}$$

provided $0 < \tau^2(\theta) < \infty$.

This approximation and its counterpart for estimating several parameters at once are extremely useful. Its main use is to describe the distribution of $\hat{\theta}$, namely $f_{\hat{\theta}}(x|\theta)$, but the form of the approximation tells us several other things as well. First, while maximum likelihood estimators are frequently biased, this approximation implies that this need not concern us if n is large: the approximating $N\left(\theta, \frac{\tau^2(\theta)}{n}\right)$ distribution has expectation θ. Furthermore, it tells us that the Law of Large Numbers applies to $\hat{\theta}$ just as it did to $\bar{X}$: for any $c > 0$,

$$P(|\hat{\theta} - \theta| < c) \simeq \Phi\left(\frac{\sqrt{n}c}{\tau(\theta)}\right) - \Phi\left(\frac{-\sqrt{n}c}{\tau(\theta)}\right), \tag{5.68}$$

so if $\frac{\sqrt{n}c}{\tau(\theta)}$ is large, $P(|\hat{\theta} - \theta| < c) \simeq 1$, even if c is very small. That is, $\hat{\theta} \xrightarrow{P} \theta$; with high probability (increasing with n), $\hat{\theta}$ will be as close to θ as desired. Increasing the amount of data increases the accuracy. This property is sometimes referred to as *consistency.* Finally, if $\hat{\theta}$ actually had the approximating distribution $N(\theta, \tau^2(\theta)/n)$, its mean squared error would be $\tau^2(\theta)/n$, and we may use this quantity as a measure of its accuracy. In fact, it can be shown under additional restrictions that no other estimator will have an approximating distribution with a smaller MSE, so that for many problems involving large samples, the maximum likelihood estimator will do as well as is possible.

Example 5.I. Consider the general normal distribution of Example 5.D, but suppose σ^2 is known and need not be estimated. Then X_i is $N(\mu, \sigma^2)$ and the log likelihood function is as before, $\log L(\mu) = \log(2\pi)^{-\frac{n}{2}} - \frac{n}{2}\log\sigma^2 - \frac{n}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$, with

$$\frac{d}{d\mu}\log L(\mu) = \frac{n}{\sigma^2}(\bar{X} - \mu),$$

$$\frac{d^2}{d\mu^2}\log L(\mu) = -\frac{n}{\sigma^2}.$$

The maximum likelihood estimator can, as before, be found from setting $\frac{d}{d\mu}\log L(\mu) = 0$ to be $\hat{\mu} = \bar{X}$. In this case we know $\hat{\mu}$ has *exactly* a $N(\mu, \sigma^2/n)$ distribution. To see what Fisher's Theorem tells us, we consider

$$\begin{aligned}\log f(X_1|\mu) &= \log\left(\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(X_1-\mu)^2}{2\sigma^2}}\right)\\ &= -\log(\sqrt{2\pi}\sigma) - \frac{(X_1 - \mu)^2}{2\sigma^2}.\end{aligned}$$

Now,

$$\frac{d}{d\mu}\log f(X_1|\mu) = \frac{(X_1 - \mu)}{\sigma^2},$$

$$\frac{d^2}{d\mu^2}\log f(X_1|\mu) = -\frac{1}{\sigma^2}.$$

The relationship (5.67) gives us two alternative ways of calculating τ^2. First,

$$\begin{aligned} E\left(\frac{d}{d\mu}\log f(X_1|\mu)\right)^2 &= E\left[\frac{(X_1-\mu)^2}{\sigma^4}\right] \\ &= \frac{E(X_1-\mu)^2}{\sigma^4} \\ &= \frac{\sigma^2}{\sigma^4} \\ &= \frac{1}{\sigma^2}; \end{aligned}$$

second,

$$\begin{aligned} -E\left(\frac{d^2}{d\mu^2}\log f(X_1|\mu)\right) &= -E\left(-\frac{1}{\sigma^2}\right) \\ &= \frac{1}{\sigma^2}. \end{aligned}$$

In either case, we find $\tau^2 = \sigma^2$, and Fisher's Approximation gives as an approximate result what we know in this case to be exactly true. In general, we can choose whichever of the expressions in (5.67) is easiest to evaluate, since when $\log f(x|\theta)$ is twice differentiable with respect to θ and the expectations are well-defined, they will be equal.

Example 5.J. For the Exponential $(1/\lambda)$ case of Example 5.C, where the X_i are independent with density

$$\begin{aligned} f(x|\lambda) &= \lambda e^{-\lambda x} \quad \text{for } x > 0 \\ &= 0 \quad \text{for } x \le 0, \end{aligned}$$

we found that $\hat{\lambda} = 1/\bar{X}$, by using the invariance property. We could have found the same result by solving $\frac{d}{d\lambda}\log L(\lambda) = 0$; Fisher's Approximation applies here. We have

$$\begin{aligned} \log f(X_1|\lambda) &= \log\lambda - \lambda X_1, \\ \frac{d}{d\lambda}\log f(X_1|\lambda) &= \frac{1}{\lambda} - X_1, \\ \frac{d^2}{d\lambda^2}\log f(X_1|\lambda) &= -\frac{1}{\lambda^2}. \end{aligned}$$

Then

$$\begin{aligned} -E\left(\frac{d^2}{d\lambda^2}\log f(X_1|\lambda)\right) &= -E\left(-\frac{1}{\lambda^2}\right) \\ &= \frac{1}{\lambda^2}, \end{aligned}$$

and

$$\tau^2 = \lambda^2.$$

We conclude that $\hat{\lambda}$ has approximately a $N\left(\lambda, \frac{\lambda^2}{n}\right)$ distribution, if n is large. Figure 5.13 shows the exact distribution of $\hat{\lambda}$ (found by applying the methods of Section 1.8 to the distribution of $\bar{X}$ found in Example 5.H) and the approximating distribution, for $n = 10, 20,$ and 40.

[Figure 5.13]

Example 5.K. Fisher's Approximation requires that the maximum likelihood estimator be found by setting $\frac{d}{d\theta}L(\theta) = 0$ or $\frac{d}{d\theta}\log L(\theta) = 0$. Here is an example where that is not the case, and the approximation fails. Suppose $X_1, X_2, \ldots, X_n$ are independent, each with the Uniform $(0,\theta)$ density

$$\begin{aligned} f(x|\theta) &= \frac{1}{\theta} \quad \text{for } 0 \le x \le \theta \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Then

$$\begin{aligned} L(\theta) &= f(x_1|\theta)\cdot f(x_2|\theta)\cdots f(x_n|\theta) \\ &= \frac{1}{\theta^n} \quad \text{for } \theta \geq \text{maximum of the } x_i = \max(x_i) \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

[Figure 5.14]

It is clear from Figure 5.14a that the largest value of $L(\theta)$ is at $\hat{\theta} = \max(x_i)$, so this is the maximum likelihood estimator. Yet this cannot be found by differentiation; indeed $\frac{d}{d\theta}L(\theta) \neq 0$ for *any* θ for which $L(\theta) > 0$. (Similarly, $\frac{d}{d\theta}\log L(\theta) \neq 0$ for *for any* such θ.) Fisher's Approximation does not apply, and in fact the distribution of $\hat{\theta}$ is very far from a Normal distribution for all n: For $0 < y < \theta$,

$$\begin{aligned} F_{\hat{\theta}}(y) &= P(\hat{\theta} \leq y) \\ &= P(\max(X_i) \leq y) \\ &= P(X_i \leq y, X_2 \leq y, \ldots, X_n \leq y) \\ &= \prod_{i=1}^{n} P(X_i \leq y) \\ &= [P(X_1 \leq y)]^n \\ &= \left(\frac{y}{\theta}\right)^n, \end{aligned}$$

and so

$$\begin{aligned} f_{\hat{\theta}}(y|\theta) &= \frac{ny^{n-1}}{\theta^n} \quad \text{for } 0 < y < \theta \\ &= 0 \quad \text{otherwise,} \end{aligned}$$

shown in Figure 5.14b for $n = 10$.

5.7 Finding Maximum Likelihood Estimators.

In the examples considered so far, it has been relatively easy to find the maximum likelihood estimator $\hat{\theta}$, usually by solving the equation

$$\frac{d}{d\theta}\log L(\theta) = 0 \tag{5.69}$$

algebraically. Outside of textbook examples it is common, particularly in multiparameter problems, for this approach to be difficult, and it becomes necessary to solve (5.69) numerically once the data are in hand. Fortunately the standard Newton-Raphson method usually works well here and permits us to evaluate the estimate without an explicit algebraic formula.

Let

$$g(\theta) = \frac{d}{d\theta}\log L(\theta), \tag{5.70}$$

and

$$g'(\theta) = \frac{d^2}{d\theta^2}\log L(\theta). \tag{5.71}$$

We wish to find a root, $\hat{\theta}$, of $g(\theta) = 0$. If $\hat{\theta}$ is near θ (and we both hope and expect that it is), we can apply the mean value theorem to get

$$g(\hat{\theta}) - g(\theta) \simeq (\hat{\theta} - \theta)g'(\theta). \tag{5.72}$$

Now $\hat{\theta}$ is a root of $g(\theta) = 0$; that is, $g(\hat{\theta}) = 0$, and so

$$-g(\theta) \simeq (\hat{\theta} - \theta)g'(\theta),$$

or

$$\hat{\theta} - \theta \simeq \frac{g(\theta)}{g'(\theta)}. \tag{5.73}$$

or

$$\hat{\theta} \simeq \theta - \frac{g(\theta)}{g'(\theta)}. \tag{5.74}$$

This gives $\hat{\theta}$ approximately in terms of θ, which is not known, but it suggests the following iterative scheme: Start with a good initial guess at θ, say $\hat{\theta}_0$. Then for $n = 1, 2, \ldots,$ compute

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \frac{g(\hat{\theta}_n)}{g'(\hat{\theta}_n)}, \tag{5.75}$$

until the estimate changes but little, until the series converges. If the procedure is going to work well, which depends upon both the initial guess $\hat{\theta}_0$ and the nature of $\log L(\theta)$, convergence is usually rapid (2 to 6 iterations). If $\log L(\theta)$ is a quadratic function of θ, then $g(\theta)$ is linear, (5.70) is an exact equation, and (5.73) gives the maximum likelihood estimate exactly for $n = 1$, regardless of $\hat{\theta}_0$. Figure 5.15 illustrates how the iteration works, geometrically. As a practical matter, it is a good idea to compute $L(\theta)$ or $\log L(\theta)$ for $\hat{\theta}$ and several other values, to check that a maximum has been achieved.

[Figure 5.15]

The steps leading to this iteration can be used as a basis of a proof of Fisher's Approximation. The idea of the proof is this: Let

$$\begin{aligned} Z_i(\theta) &= \frac{d}{d\theta} \log f(X_i|\theta) \\ &= \frac{\frac{d}{d\theta} f(X_i|\theta)}{f(X_i|\theta)}. \end{aligned} \tag{5.76}$$

Then since

$$\begin{aligned} \log L(\theta) &= \log \prod_{i=1}^{n} f(X_i|\theta) \\ &= \sum_{i=1}^{n} \log f(X_i|\theta), \end{aligned}$$

we see that

$$g(\theta) = \sum_{i=1}^{n} Z_i(\theta)$$

is a sum of independent random variables. Now $E(Z_i(\theta)) = 0$, since

$$\begin{aligned} E(Z_i(\theta)) &= \int_{-\infty}^{\infty} \left[\frac{\frac{d}{d\theta} f(x|\theta)}{f(x|\theta)} \right] f(x|\theta) dx \quad \text{by (2.17)} \\ &= \int_{-\infty}^{\infty} \frac{d}{d\theta} f(x|\theta) dx \\ &= \frac{d}{d\theta} \int_{-\infty}^{\infty} f(x|\theta) dx \\ &= \frac{d}{d\theta} \cdot 1 \quad \text{(since } f \text{ is a density)} \\ &= 0. \end{aligned}$$

Also,

$$\begin{aligned}\mathrm{Var}(Z_i(\theta)) &= E\left(\frac{d}{d\theta}\log f(X_i|\theta)\right)^2\\ &= \frac{1}{\tau^2(\theta)} \quad \text{from (5.67).}\end{aligned}$$

Then the Central Limit Theorem tells us that $\frac{g(\theta)}{\sqrt{n}}$ has approximately a $N\left(0, \frac{1}{\tau^2(\theta)}\right)$ distribution. We also have that

$$g'(\theta) = \sum_{i=1}^{n}\frac{d}{d\theta}Z_i(\theta) = \sum_{i=1}^{n}\frac{d^2}{d\theta^2}\log f(X_i|\theta)$$

is a sum of independent random variables, so the Law of Large Numbers tells us that

$$\begin{aligned}\frac{1}{n}g'(\theta)\xrightarrow{P}E\left(\frac{d}{d\theta}Z_i(\theta)\right) &= E\left(\frac{d^2}{d\theta^2}\log f(X_i|\theta)\right)\\ &= -\frac{1}{\tau^2(\theta)}\end{aligned}$$

from (5.67). But then

$$\begin{aligned}\sqrt{n}\left(\frac{-g(\theta)}{g'(\theta)}\right) &= \frac{g(\theta)/\sqrt{n})}{(g'(\theta)/n)}\\ &\simeq \frac{g(\theta)/\sqrt{n}}{1/\tau^2(\theta)}\\ &= \tau^2(\theta)\cdot\frac{g(\theta)}{\sqrt{n}}\end{aligned}$$

has approximately a $N\left(0, [\tau^2(\theta)]^2\cdot\frac{1}{\tau^2(\theta)}\right)$ or $N(0,\tau^2(\theta))$ distribution, and so $-g(\theta)/g'(\theta)$ has approximately a $N(0,\tau^2(\theta)/n)$ distribution. But with (5.73), this is Fisher's Approximation. This does not constitute a rigorous proof, but it can serve as the basis of one.

There are other useful versions of Fisher's Approximation that follow from the same line of reasoning. For example, the data $X_1, X_2, \ldots, X_n$ may be assumed independent but not necessarily with the same density. Suppose X_i has density $f_i(x \mid \theta)$, and

$$\sigma_i^2(\theta) = E\left(\frac{d}{d\theta}\log f_i(X_i \mid \theta)\right)^2 = -E\left[\frac{d^2}{d\theta^2}\log f_i(X_i \mid \theta)\right].$$

Then $\sqrt{n}(\hat{\theta}-\theta)$ would be expected to be approximately normal with expectation 0 and variance

$$\frac{1}{\sum\limits_{i=1}^{n}\sigma_i^2(\theta)}. \tag{5.77}$$

(If the f_i are identical, $\sigma_i^2(\theta) = 1/\tau^2(\theta)$, and this variance $= \tau^2(\theta)/n$, agreeing with the earlier approximation.)

The most general form of Fisher's Approximation is that for the multiparameter case; that is, where $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ is itself multidimensional. In that case $\hat{\theta}$ arises from maximizing $L(\theta)$ or $\log L(\theta)$ over the k-dimensional set of possible θ, and when the maximum is a "smooth" maximum, as in the 1-dimensional case we would expect $\sqrt{n}(\hat{\theta}-\theta)$ to have approximately a "k-dimensional" normal distribution with vector expectation 0 and covariance matrix $[I(\theta)]^{-1}$, where the $(i,j)^{\text{th}}$ entry of $I(\theta)$ is a_{ij}, with

$$a_{ij} = E\left(\frac{\partial\log L(\theta)}{\partial\theta_i}\cdot\frac{\partial\log L(\theta)}{\partial\theta_j}\right) = -E\left(\frac{\partial^2\log L(\theta)}{\partial\theta_i\partial\theta_j}\right). \tag{5.78}$$

If the X_i are independent, each with density f, then

$$a_{ij} = nE\left(\frac{\partial \log f(X \mid \theta)}{\partial \theta_i} \cdot \frac{\partial \log f(X \mid \theta)}{\partial \theta_j}\right) = -nE\left(\frac{\partial^2 \log f(X \mid \theta)}{\partial \theta_i \partial \theta_j}\right) \tag{5.79}$$

For more than one parameter, the corresponding Newton-Raphson algorithm would proceed like this: Let $L(\boldsymbol{\theta}) = L(\theta_1, \theta_2, \ldots, \theta_k)$ be the likelihood function where $\theta_1, \theta_2, \ldots, \theta_n$ are the parameters we wish to estimate. Define the vector

$$\mathbf{g}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{d}{d\theta_1} \log L(\boldsymbol{\theta}) \\ \frac{d}{d\theta_2} \log L(\boldsymbol{\theta}) \\ \cdot \\ \cdot \\ \cdot \\ \frac{d}{d\theta_k} \log L(\boldsymbol{\theta}) \end{pmatrix}. \tag{5.80}$$

We wish to find the root $\hat{\boldsymbol{\theta}}$ to $\mathbf{g}(\boldsymbol{\theta}) = 0$; that is, to find $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k$ so that all of the derivatives in $\boldsymbol{g}(\boldsymbol{\theta})$ are zero simultaneously. Let

$$\mathbf{G}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{d^2}{d\theta_1^2} \log L(\boldsymbol{\theta}) & \frac{d^2}{d\theta_1 d\theta_2} \log L(\theta) & \cdots & \frac{d^2}{d\theta_1 d\theta_k} \log L(\theta) \\ \frac{d^2}{d\theta_1 d\theta_2} \log L(\theta) & \frac{d^2}{d\theta_2^2} \log L(\theta) & \cdots & \cdot \\ \vdots & & & \vdots \\ \frac{d^2}{d\theta_1 d\theta_k} \log L(\theta) & & \cdots & \frac{d^2}{d\theta_k^2} \log L(\theta) \end{pmatrix} \tag{5.81}$$

Then the mean value theorem tells us

$$\mathbf{g}(\hat{\boldsymbol{\theta}}) - \mathbf{g}(\boldsymbol{\theta}) \simeq \mathbf{G}(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

and since $\mathbf{g}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$,

$$-\mathbf{g}(\boldsymbol{\theta}) \simeq \mathbf{G}(\theta)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$$

or

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = -\mathbf{G}^{-1}(\boldsymbol{\theta})\mathbf{g}(\boldsymbol{\theta}) \tag{5.82}$$

or

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} - \mathbf{G}^{-1}(\boldsymbol{\theta})\mathbf{g}(\boldsymbol{\theta}), \tag{5.83}$$

and the iteration, starting from an initial guess $\widehat{\boldsymbol{\theta}_n}$, is

$$\hat{\boldsymbol{\theta}}_{n+1} = \hat{\boldsymbol{\theta}}_n - \mathbf{G}^{-1}(\hat{\boldsymbol{\theta}}_n)\mathbf{g}(\hat{\boldsymbol{\theta}}_n). \tag{5.84}$$

Chapter 5 Figures

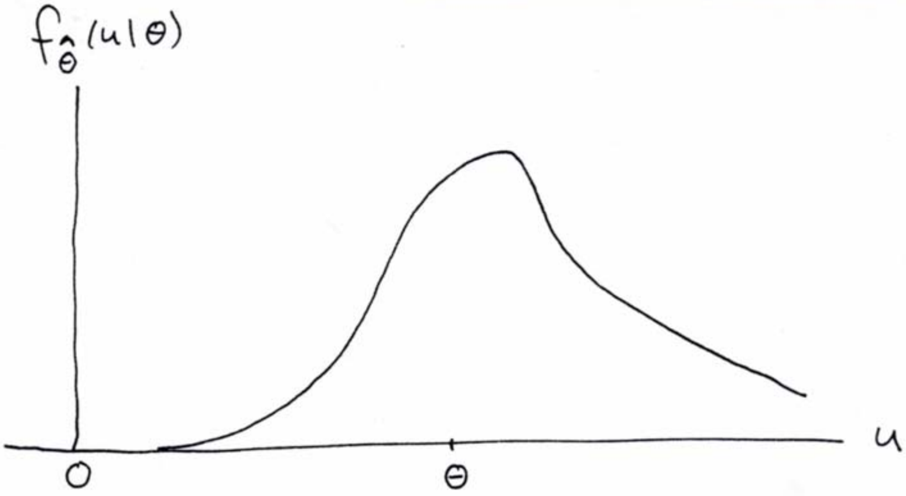
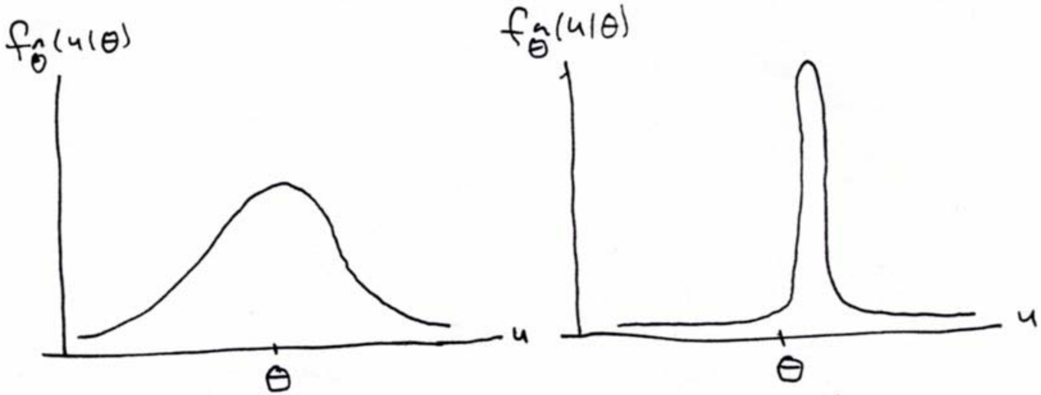


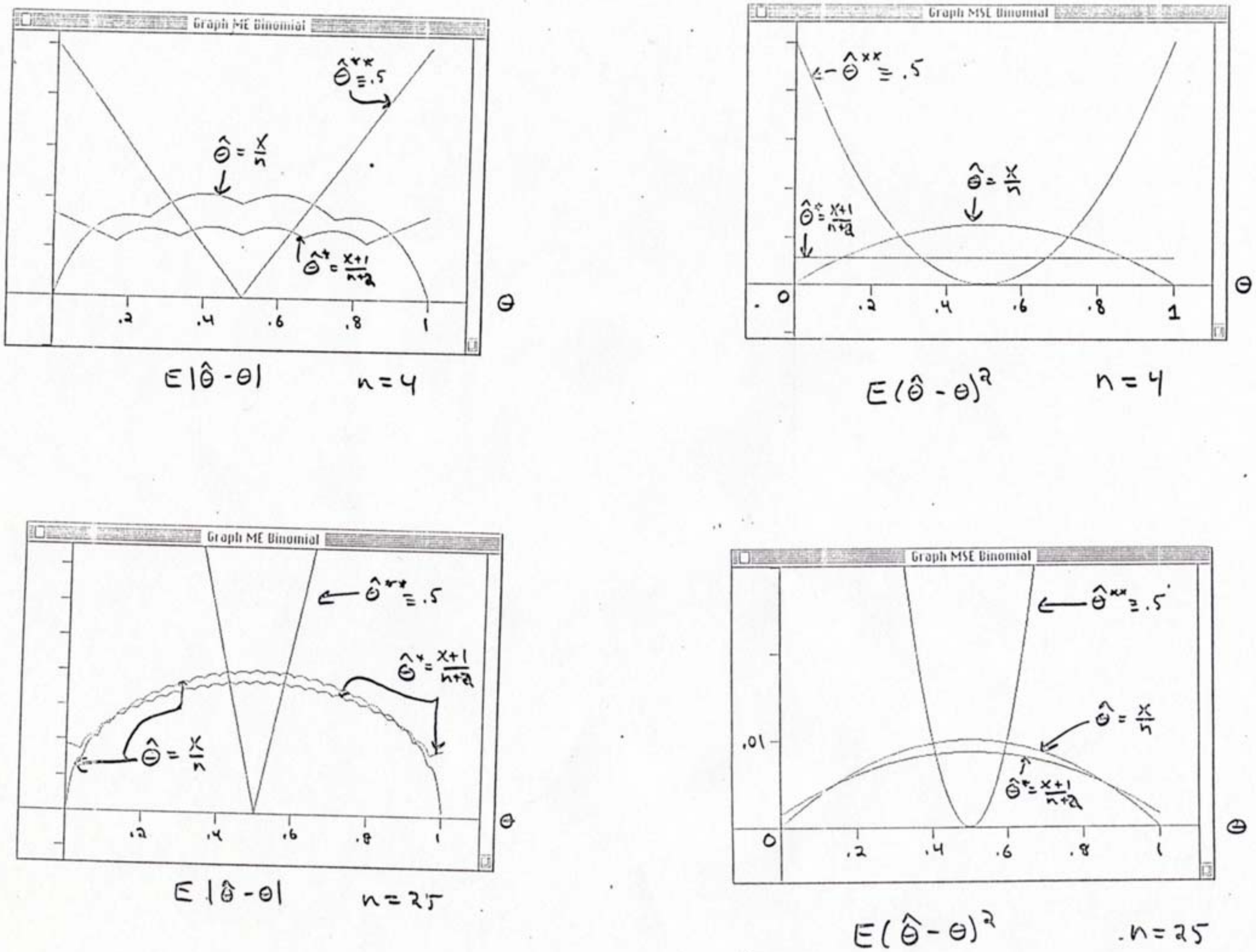
Figure 5.1 The distribution of $\hat{\Theta}$ for a particular value of Θ



no bias
medium variance

some bias
low variance

Figure 5.2



$E|\hat{\theta} - \theta|$ $n = 4$

$E(\hat{\theta} - \theta)^2$ $n = 4$

$E|\hat{\theta} - \theta|$ $n = 25$

$E(\hat{\theta} - \theta)^2$ $n = 25$

Figure 5.3

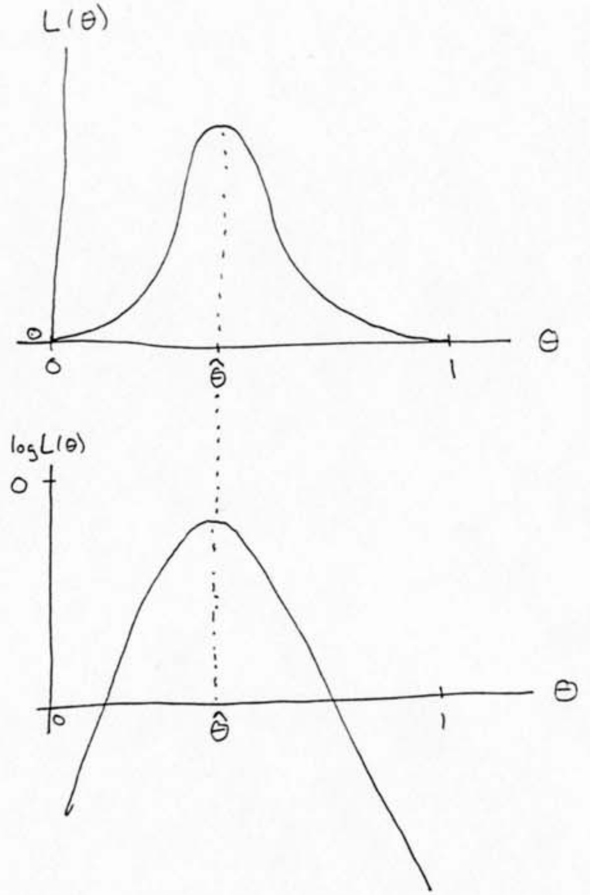


Figure 5.4 The Binomial Likelihood and loglikelihood functions, for n = 40

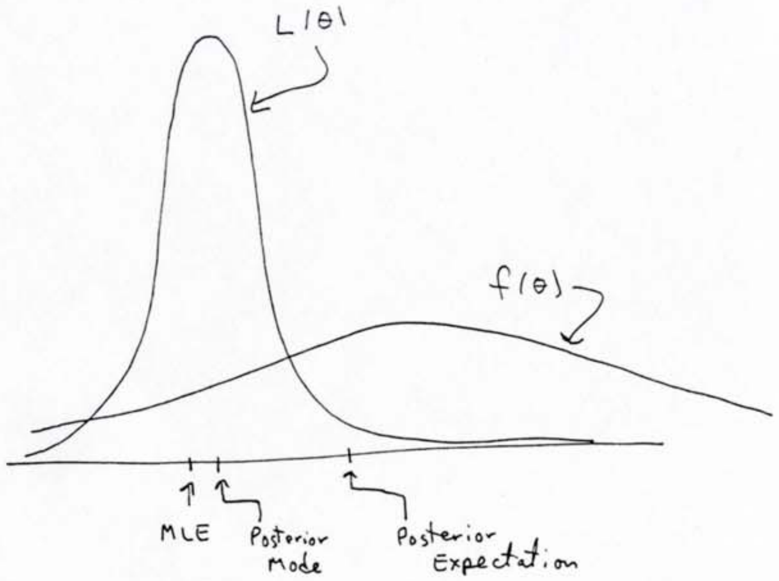


Figure 5.5

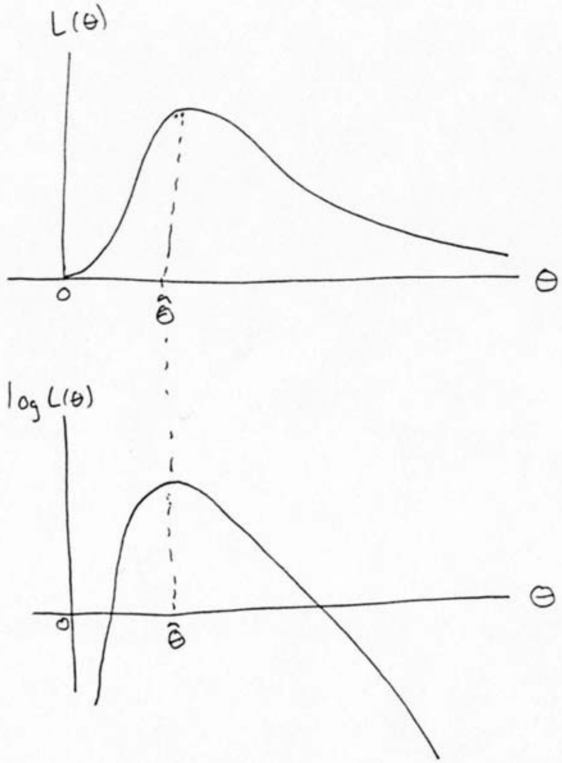


Figure 5.6 The likelihood and loglikelihood functions for the expected failure time θ

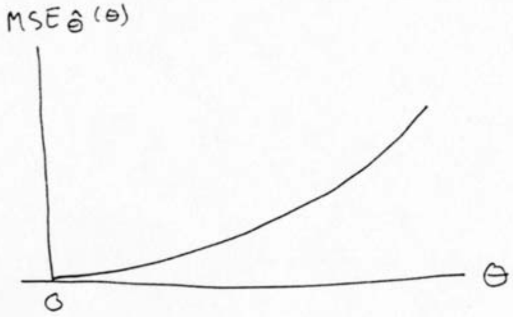


Figure 5.7

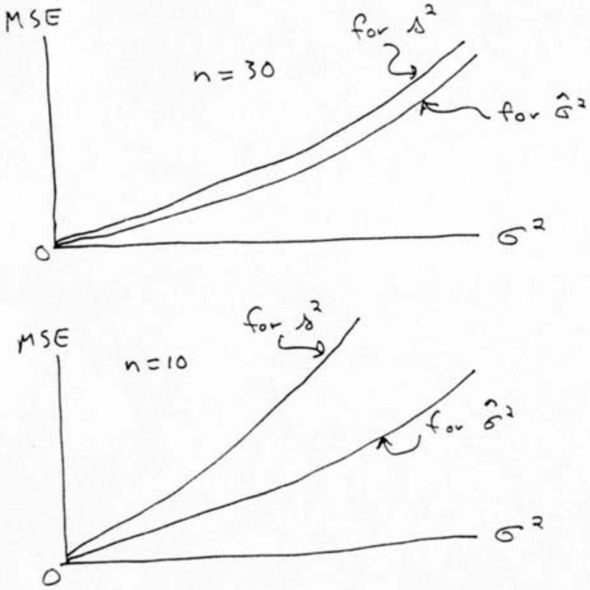


Figure 5.8

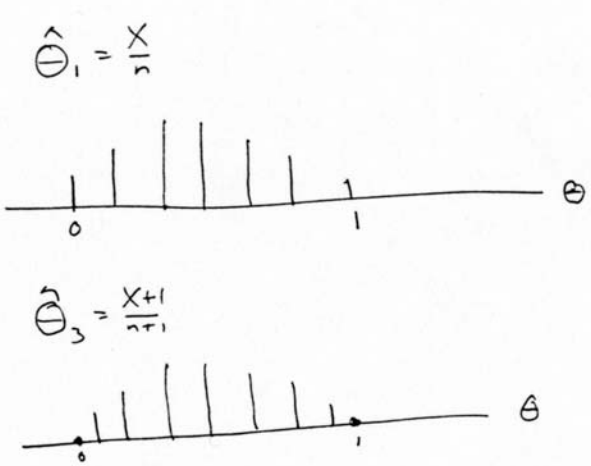


Figure 5.9 Distributions of $\hat{\Theta}_1$, $\hat{\Theta}_3$ $n=6$ $\Theta = .4$

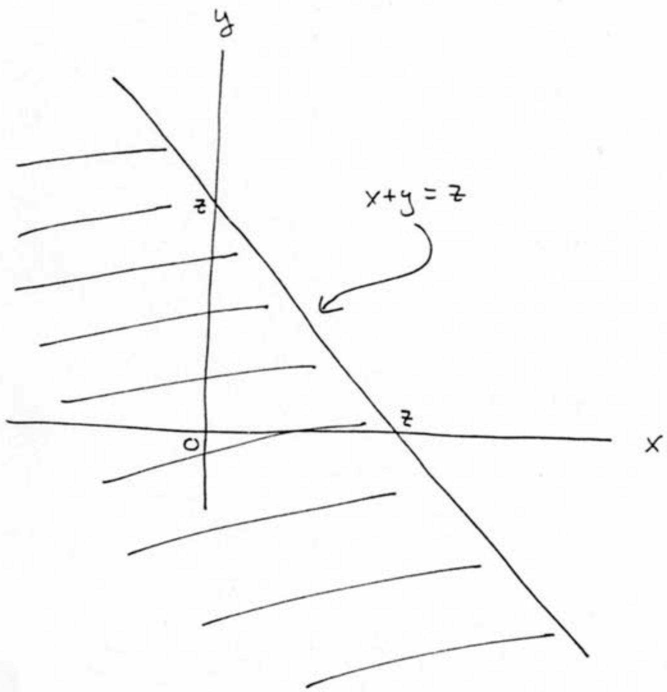


Figure 5.10

The region in the x-y plane where $x+y \leq z$

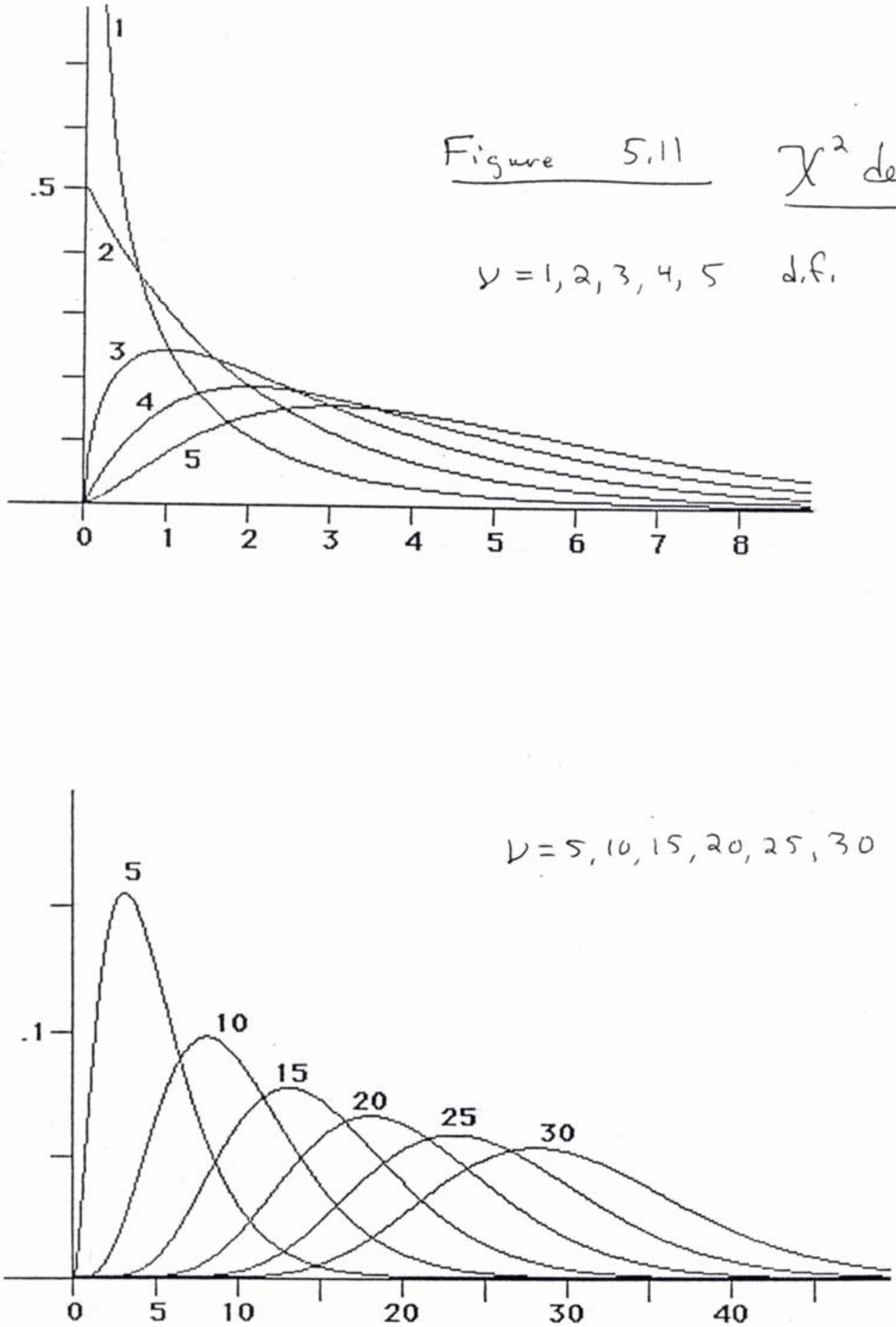


Figure 5.11 χ^2 densities

$\nu = 1, 2, 3, 4, 5$ d.f.

$\nu = 5, 10, 15, 20, 25, 30$ d.f.

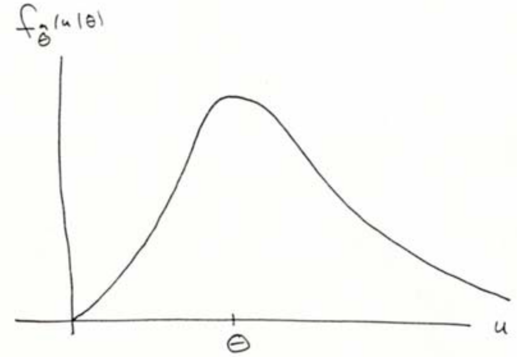


Figure 5.12 The density of $\hat{\Theta}$ for a sample of size 5 from an exponential distribution.

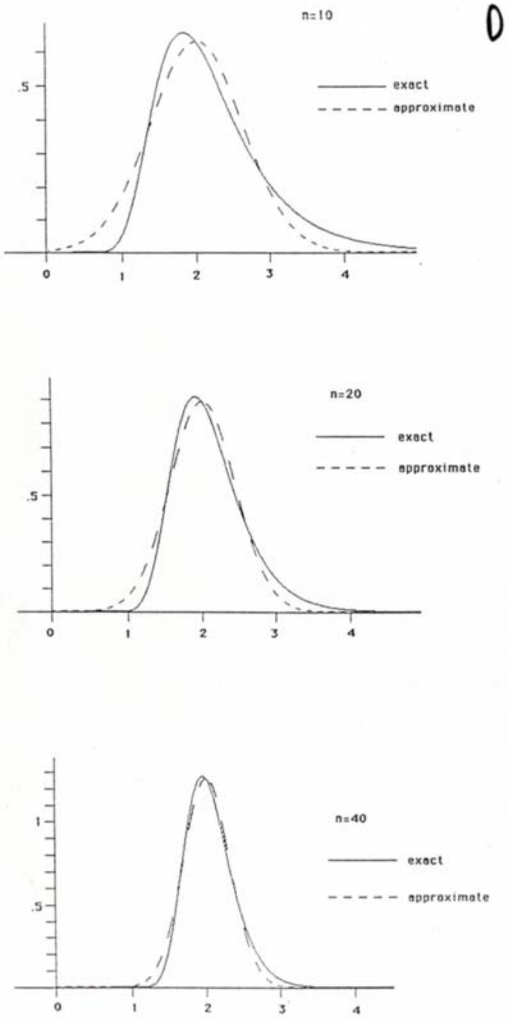


Figure 5.13 Distributions for $\hat{\lambda} = \frac{1}{\bar{X}}$

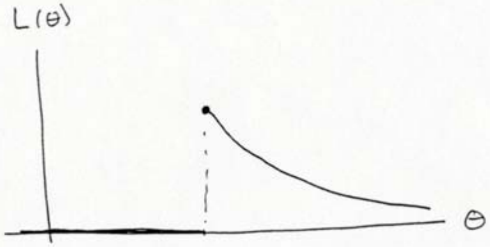


Figure 5.14a

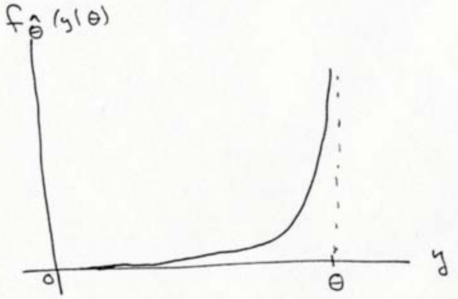
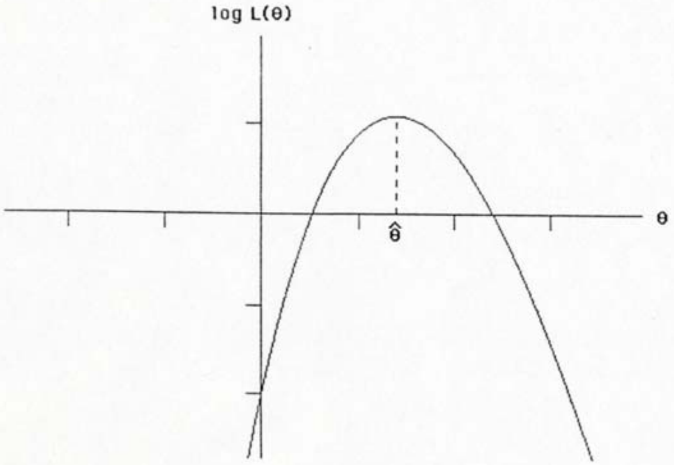


Figure 5.14b



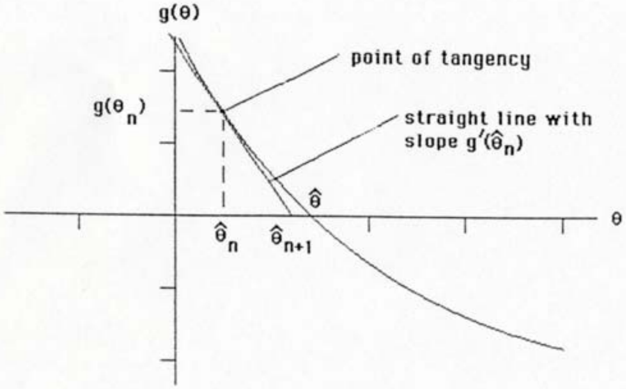


Figure 5.15