Appendix B: Statistics

The material in this appendix can be supplemented with the classic text in statistics *The Theory of Point Estimation* (New York: Wiley, 1983) by E. Lehmann.

Contents

1.1 SUFFICIENCY, COMPLETENESS, AND UNBIASED ESTIMATION

In statistics, we often represent our data, in many cases a sample of size n from some population, as a random vector $X = (X_1, \ldots, X_n)$. The model can be written in the form $\{f_\theta(x); \theta \in \Omega\}$, where Ω is the *parameter space* or set of permissible values of the parameter and $f_\theta(x)$ is the probability density function. A *statistic*, $T(X)$, is a function of the data that does not depend on the unknown parameter θ. Although a statistic, $T(X)$, is not a function of θ, its distribution can depend on θ. An estimator is a statistic considered for the purpose of estimating a given parameter. One of our objectives is to find a good estimator of the parameter θ, in some sense of the word "good." How do we ensure that a statistic $T(X)$ is estimating the correct parameter and is not consistently too large or too small, and that as much variability as possible has been removed? The problem of estimating the correct parameter is often dealt with by requiring that the estimator be unbiased.

We will denote an expected value under the assumed parameter value θ by $E_\theta(.)$. Thus, in the continuous case,

$$E_\theta[h(X)] = \int_{-\infty}^{\infty} h(x) f_\theta(x) dx$$

and in the discrete case,

$$E_\theta[h(X)] = \sum_{all\ x} h(x) f_\theta(x)$$

provided the integral/sum converges absolutely. In the discrete case, $f_\theta(x) = P_\theta[X = x]$, the probability function of X under this parameter value θ.

Definition A statistic $T(X)$ is an *unbiased estimator* of θ if $E_\theta[T(X)] = \theta$ for all $\theta \in \Omega$.

For example, suppose that X_i are independent, each with the Poisson distribution with parameter θ, $i = 1, \ldots, n$. Notice that the statistic

$$T = \frac{1}{n(n+1)} \sum_{i=1}^{n} X_i$$

is such that

$$\begin{aligned} E_\theta(T) &= \frac{1}{n} \sum_{i=1}^{n} E_\theta X_i = \frac{1}{n} \sum_{i=1}^{n} \theta \\ &= \theta \end{aligned}$$

and so T is an unbiased estimator of θ. This means that it is centered in the correct place, but it does not mean it is a best estimator in any sense.

In *decision theory*, in order to determine whether a given estimator or statistic $T(X)$ does well for estimating θ, we consider a loss function or distance function between the estimator and the true value. Call this $\delta(\theta, T(X))$. Then this is averaged over all possible values of the data to obtain the risk:

$$\text{Risk} = E_\theta\{\delta(\theta, T(X))\}$$

A good estimator is one with little risk; a bad estimator is one whose risk is high. One particular risk function is called *mean squared error* (MSE) and corresponds to $\delta(\theta, T(X)) = [T(X) - \theta]^2$. The mean squared error has a useful decomposition into two components, the variance of the estimator and the square of its bias:

$$\text{MSE}(\theta, T) = E_\theta\{[T(X) - \theta]^2\} = \text{var}_\theta(T(X)) + [E_\theta T(X) - \theta]^2$$

For example, if X has a normal$(\theta, 1)$ distribution, the mean squared error of $T_1 = X$ is 1 for all θ because the bias $E_\theta\{T(X)\} - \theta$ is zero. On the other hand, the estimator $T_2 = X/2$ has bias $E_\theta T(X) - \theta = \frac{\theta}{2}$ and variance $\frac{1}{4}$, so the mean squared error is $\frac{1}{4}(1 + \theta^2)$. Obviously, T_2 has smaller mean squared error provided that θ is around 0 (more precisely, provided $\theta^2 < 3$), but for θ large, T_1 is preferable. Of these two estimators, only T_1 is unbiased.

In general, in fact, there is usually no one estimator that outperforms all other estimators at all values of the parameter if we use mean squared error as our basis for comparison. In order to achieve an optimal estimator, it is unfortunately necessary to restrict ourselves to a specific class of estimators and select the best within the class. Of course, the best within this class will only be as good as the class itself (best in a class of one is not much of a recommendation), and therefore we must ensure that restricting ourselves to this class is not unduly restrictive. The class of *all* estimators is usually too large to obtain a meaningful solution. One common restriction is to the class of all unbiased estimators.

Definition An estimator $T(X)$ is said to be a *uniformly minimum-variance unbiased estimator* (UMVUE) of the parameter θ if

1. It is an unbiased estimator of θ and
2. Among *all unbiased estimators of* θ it has the smallest mean squared error and therefore the smallest variance.

A *sufficient statistic* is one that, from a certain perspective, contains all the necessary information for making inferences (e.g., estimating the parameter with a point estimator or confidence interval, or conducting a test of a

hypothesized value) about the unknown parameters in a given model. It is important to remember that a statistic is sufficient for inference on a specific parameter. It does not necessarily contain all relevant information in the data for other inferences. For example, if you wished to test whether the family of distributions is an adequate fit to the data (a goodness-of-fit test), the sufficient statistic for the parameter in the model does not contain the relevant information.

Suppose the data is in a vector X and $T = T(X)$ is a sufficient statistic for θ. The intuitive basis for sufficiency is that if the conditional distribution of X given $T(X)$ does not depend on θ, then X provides no additional value in addition to T for estimating θ. The assumption is that random variables carry information on a statistical parameter θ only insofar as their distributions (or conditional distributions) change with the value of the parameter and that since, given $T(X)$, we can randomly generate values for the X without knowledge of the parameter and with the correct distribution, these randomly generated values cannot carry additional information. All of this, of course, assumes that the model is correct and θ is the only unknown. The distribution of X given a sufficient statistic T will often have value for other purposes, such as measuring the variability of the estimator or testing the validity of the model.

Definition A statistic $T(X)$ is *sufficient* for a statistical model $\{f_\theta(x); \theta \in \Omega\}$ if the distribution of the data $(X_1, \ldots, X_n)$ given $T(X) = t$ does not depend on the unknown parameter θ.

The use of a sufficient statistic is formalized in the *the sufficiency principle*, which states that if $T(X)$ is a sufficient statistic for a model $\{f_\theta(x); \theta \in \Omega\}$ and x_1, x_2 are two different possible observations that have identical values of the sufficient statistic,

$$T(x_1) = T(x_2)$$

then whatever inference we would draw from observing x_1 should also be drawn from observing x_2.

Sufficient statistics are **not unique**. For example, if the sample mean $\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ is a sufficient statistic, then any other statistic that allows us to obtain $\overline{X}$ is also sufficient. This will include all one-to-one functions of $\overline{X}$ (these are essentially equivalent) such as $\overline{X}^3$ and all statistics $T(X)$ for which we can write $\overline{X} = g(T)$ for some, possibly many-to-one function g. One result that is normally used to verify whether a given statistic is sufficient is the *factorization criterion for sufficiency:* Suppose $X = (X_1, \ldots, X_n)$ has probability density function $\{f_\theta(x); \theta \in \Omega\}$ and $T(X)$ is a statistic. Then

$T(X)$ is a sufficient statistic for $\{f_\theta(x);\ \theta \in \Omega\}$ if and only if there exist two nonnegative functions $g(.)$ and $h(.)$ such that we can factor the probability density function $f_\theta(x) = g(T(x);\theta)h(x)$ for all x. This factorization into two pieces, one that involves both the statistic T and the unknown parameter θ, and the other that may be a constant or depend on x but does not depend on the unknown parameter, need only hold on a set A of possible values of X that carries the full probability. That is, for some set A with $P_\theta(X \in A) = 1$, for all $\theta \in \Omega$, we require

$$f_\theta(x) = g(T(x);\theta)h(x) \quad \text{for all } x \in A,\ \theta \in \Omega$$

Definition A statistic $T(X)$ is a *minimal sufficient statistic* for $\{f_\theta(x);\ \theta \in \Omega\}$ if it is sufficient and if for **any other sufficient statistic** $U(X)$, there exists a function $g(.)$ such that $T(X) = g(U(X))$.

This definition says in effect that a minimal sufficient statistic can be recovered from any other sufficient statistic. A statistic $T(X)$ implicitly partitions the sample space into events of the form $[T(X) = x]$for varying x, and if $T(X)$ is minimal sufficient, it induces the coarsest possible partition (i.e., the largest possible sets) in the sample space among all sufficient statistics. This partition is called the *minimal sufficient partition*.

The property of *completeness* is useful for determining the uniqueness of estimators and verifying in some cases that a minimal sufficient reduction has been found. It bears no relation to the notion of a complete market in finance, or the mathematical notion of a complete metric space. Let $(X_1, \ldots, X_n)$ denote the observations from a distribution with probability density function $\{f_\theta(x);\ \theta \in \Omega\}$. Suppose $T(X)$ is a statistic and $u(T)$, a function of T, is an unbiased estimator of θ so that $E_\theta[u(T)] = \theta$ for all $\theta \in \Omega$. Under what circumstances is this the only unbiased estimator that is a function of T? To answer this question, suppose $u_1(T)$ and $u_2(T)$ are both unbiased estimators of θ and consider the difference $h(T) = u_1(T) - u_2(T)$. Since $u_1(T)$ and $u_2(T)$ are both unbiased estimators of the parameter θ, we have $E_\theta[h(T)] = 0$ for all $\theta \in \Omega$. Now if the only function $h(T)$ that satisfies $E_\theta[h(T)] = 0$ for all $\theta \in \Omega$ is the zero function $h(t) = 0$, then the two unbiased estimators must be identical. A statistic T with this property is said to be *complete*. Technically, it is not the statistic that is complete, but the family of distributions of T in the model $\{f_\theta(x);\ \theta \in \Omega\}$.

Definition The statistic $T(X)$ is *complete* if

$$E_\theta[h(T(X))] = 0, \quad \text{for all } \theta \in \Omega$$

for any function h implies

$$P_\theta[h(T(X)) = 0] = 1 \quad \text{for all } \theta \in \Omega$$

For example, let $(X_1, \ldots, X_n)$ be a random sample from the normal$(\theta, 1)$ distribution. Consider $T(X) = (X_1, \sum_{i=1}^n X_i)$. Then T is sufficient for $\{f_\theta(x); \theta \in \Omega\}$ but is not complete. It is easy to see that it is not complete, because the function

$$h(T) = X_1 - \frac{1}{n}\sum_{i=1}^n X_i$$

is a function of T that has zero expectation for all values of θ, and yet the function is not identically zero. The fact that the statistic $(X_1, \sum_{i=1}^n X_i)$ is sufficient but not complete is a hint that further reduction is possible, that it is not minimal sufficient. In fact, in this case, as we will show a little later, taking only the second component of T, namely $\sum_{i=1}^n X_i$, provides a minimal sufficient, complete statistic.

Theorem B1 *If $T(X)$ is a complete and sufficient statistic for the model $\{f_\theta(x); \theta \in \Omega\}$, then $T(X)$ is a minimal sufficient statistic for the model.*

The converse to the above theorem is **not true**. Let $(X_1, \ldots, X_n)$ be a random sample from the continuous uniform distribution on the interval $(\theta - 1, \theta + 1)$. This distribution has probability density function

$$f_\theta(x) = \frac{1}{2} \quad \text{for } \theta - 1 < x < \theta + 1$$

Then using the factorization criterion above, the joint probability density function for a sample of n independent observations from this density is

$$\begin{aligned} f_\theta(x_1, \ldots, x_n) &= \frac{1}{2^n} \quad \text{if } \theta - 1 < x_{(1)} < x_{(n)} < \theta + 1, \text{ and zero otherwise,} \\ &= \frac{1}{2^n} I(\theta - 1 < x_{(1)}) I(\theta + 1 > x_{(n)}) \end{aligned}$$

where $I(\theta - 1 < x_{(1)})$ is 1 or 0 as the inequality holds or does not hold, and $x_{(1)}, x_{(n)}$ are the smallest and the largest values in the sample $(x_1, x_2, \ldots, x_n)$. Obviously, $f_\theta(x_1, \ldots, x_n)$ can be written as a function $g(T(x); \theta)$ where $T(X) = (X_{(1)}, X_{(n)})$ and so $T(X)$ is sufficient. Moreover, it is not difficult to show that no further reduction (for example, to $X_{(1)}$ alone) is possible or we can no longer provide such a factorization, so $T(X)$ is minimal sufficient. Nevertheless, if $T(X) = (X_{(1)}, X_{(n)})$ and the function h is defined by

$$h(T) = \frac{X_{(n)} - X_{(1)}}{2} - \frac{n-1}{n+1}$$

(clearly a nonzero function), then $E_\theta[h(T)] = 0$ for all $\theta \in \Omega$ and therefore $T(X)$ is not a complete statistic.

Theorem B2 *For any random variables X and Y,*

$$E_\theta(X) = E_\theta[E_\theta(X|Y)]$$

and

$$\text{var}_\theta(X) = E_\theta[\text{var}_\theta(X|Y)] + \text{var}_\theta[E_\theta(X|Y)]$$

In much of what follows, we wish to be able to estimate a general function of the unknown parameter such as $\tau(\theta)$ instead of the parameter θ itself. We have already seen that if $T(X)$ is a complete statistic, then there is *at most one* function of $T(X)$ that provides an unbiased estimator of any function of a given $\tau(\theta)$. In fact, if we can find such a function, $g(T(X))$, then it automatically has minimum variance among all possible unbiased estimators of $\tau(\theta)$ that are based on the same data.

Theorem B3 *If $T(X)$ is a complete sufficient statistic for the model $\{f_\theta(x); \theta \in \Omega\}$ and $E_\theta[g(T(X))] = \tau(\theta)$, then $g(T(X))$ is the* UMVUE *of $\tau(\theta)$.*

When we have a complete sufficient statistic, and we are able to find an unbiased estimator, even a bad one, of $\tau(\theta)$, then there is a simple recipe for determining the UMVUE of $\tau(\theta)$.

Theorem B4 *If $T(X)$ is a complete sufficient statistic for the model $\{f_\theta(x); \theta \in \Omega\}$ and $U(X)$ is any unbiased estimator of $\tau(\theta)$, then $E(U|T)$ is the* UMVUE *of $\tau(\theta)$.*

Note that we did not subscript the conditional expectation $E(U|T)$ with θ because whenever T is a sufficient statistic, the conditional distribution of $U(X)$ given T does not depend on the underlying value of the parameter θ.

Definition Suppose $X = (X_1, \ldots, X_p)$ has a (joint) probability density function of the form

$$f_\theta(x) = C(\theta)\ \exp\left\{\sum_{j=1}^{k} q_j(\theta) T_j(x)\right\} h(x) \tag{1.1}$$

for functions $q_j(\theta)$, $T_j(x)$, $h(x)$, $C(\theta)$. Then we say that the density is a member of the *exponential family of densities*. We call $(T_1(X), \ldots, T_k(X))$ the *natural sufficient statistic*.

A member of the exponential family can be re-expressed in different ways, and so the natural sufficient statistic is not unique. For example, we may multiply a given T_j by a constant and divide the corresponding q_j by the same constant, resulting in the same probability density function $f_\theta(x)$. Various other conditions need to be applied as well — for example, ensuring that the $T_j(x)$ are all essentially different functions of the data. One of the important properties of the exponential family is its closure under repeated independent sampling. In general, if $X_i, i = 1, \ldots, n$ are independent identically distributed with an exponential family distribution, then their joint distribution $(X_1, \ldots, X_n)$ is also an exponential family distribution.

Theorem B5 *Let $(X_1, \ldots, X_n)$ be a random sample from the distribution with probability density function given by* (1.). *Then $(X_1, \ldots, X_n)$ also has an exponential family form, with joint probability density function*

$$f_\theta(x_1, \ldots, x_n) = C^n(\theta) \exp\left\{\sum_{j=1}^{k} q_j(\theta)\left[\sum_{i=1}^{n} T_j(x_i)\right]\right\} \prod_{i=1}^{n} h(x_i)$$

In other words, C is replaced by C^n and $T_j(x)$ by $\sum_{i=1}^{n} T_j(x_i)$. The natural sufficient statistic is

$$\left(\sum_{i=1}^{n} T_1(X_i), \ldots, \sum_{i=1}^{n} T_k(X_i)\right)$$

It is usual to *reparameterize* equation (2.1) by replacing $q_j(\theta)$ by a new parameter η_j. This results in a more efficient representation, the *canonical form* of the exponential family density:

$$f_\eta(x) = C(\eta) \exp\left\{\sum_{j=1}^{k} \eta_j T_j(x)\right\} h(x)$$

The *natural parameter space* in this form is the set of all values of η for which the above function is integrable, that is,

$$\left\{\eta;\ \int_{-\infty}^{\infty} \exp\left\{\sum_{j=1}^{k} \eta_j T_j(x)\right\} h(x)dx < \infty\right\}$$

We would like this parameter space to be large enough to allow intervals for each of the components of the vector η, and so we will later need to assume that the natural parameter space contains a k-dimensional rectangle.

If the statistic satisfies a linear constraint, for example, $\sum_{j=1}^{k} T_j(X) = 0$ with probability 1, then the number of terms k could be reduced and a more efficient representation of the probability density function is possible. Similarly, if the parameters η_j satisfy a linear relationship, they are not all statistically meaningful because one of the parameters is obtainable from the others. These are all situations that we would handle by reducing the model to a more efficient and nonredundant form. So in the remaining discussion, we will generally assume such a reduction has already been made and that the exponential family representation is minimal in the sense that neither the η_j nor the T_j satisfy any linear constraints.

Definition We will say that X has a *regular* exponential family distribution if it is in canonical form, is of full rank in the sense that neither the T_j nor the η_j satisfy any linear constraints permitting a reduction in the value of k, and the natural parameter space contains a k-dimensional rectangle.

By Theorem B5, if X_i has a regular exponential family distribution, then $X = (X_1, \ldots, X_n)$ also has a regular exponential family distribution.

The main advantage of identifying a distribution as a member of the regular exponential family is that it allows us to quickly identify the minimal sufficient statistic and conclude that it is complete.

Theorem B6 *If X has a regular exponential family distribution, then $(T_1(X), \ldots, T_k(X))$ is a complete sufficient statistic.*

Example Let $(X_1, X_2, \ldots, X_n)$ be independent observations all from the normal (μ, σ^2) distribution. Notice that with the parameter

$$\theta = (\mu, \sigma^2)$$

we can write the probability density function of each X_i as

$$f_\theta(x) = C \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} = C \exp\left\{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2\right\}$$

where $C = C(m, r^2)$ so the natural parameters are $\eta_1 = \frac{\mu}{\sigma^2}$ and $\eta_2 = -\frac{1}{2\sigma^2}$ and the natural sufficient statistic is (X, X^2). For a sample of size n from this density we have the same natural parameters, and by the above theorem, a complete sufficient statistic is $(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2)$. For example, if you wished to find a UMVUE of any function of η_1, η_2, such as the parameter $\eta_1 = \mu/\sigma^2$, we need only find some function of the compete sufficient statistic

that has the correct expected value. For example, in this case, with the sample mean $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$, it is not difficult to show that

$$E\left(\frac{\overline{X}}{S^2}\right) = \frac{n-1}{n-3}\frac{\mu}{\sigma^2}$$

and so, provided $n > 3$,

$$\frac{n-3}{n-1}\frac{\overline{X}}{S^2}$$

is an unbiased estimator of n_1 and a function of the complete sufficient statistic, so it is the desired UMVUE. Suppose one of the parameters, say σ^2, is assumed known. Then the normal distribution is still in the regular exponential family, since it has a representation

$$f_\theta(x) = C(\mu, \sigma)\exp\left\{\frac{\mu}{\sigma^2}x\right\}h(x)$$

with the function h completely known. In this case, for a sample of size n from this distribution, the statistic $\sum_{i=1}^{n} X_i$ is complete sufficient for μ and so any function of it, say $\overline{X}$, that is an unbiased estimator of μ is automatically UMVUE.

The following table gives various members of the regular exponential family and the corresponding complete sufficient statistic.

Member of the Regular Exponential Family		Complete Sufficient Statistic
Poisson(θ)		$\sum_{i=1}^{n} X_i$
Binomial(n, θ)		$\sum_{i=1}^{n} X_i$
Negative binomial(k, θ)		$\sum_{i=1}^{n} X_i$
Geometric(θ)		$\sum_{i=1}^{n} X_i$
Normal(μ, σ^2)	σ^2 known	$\sum_{i=1}^{n} X_i$
	μ known	$\sum_{i=1}^{n} (X_i - \mu)^2$
		$\left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right)$
Gamma(α, β) (includes exponential)	α known	$\sum_{i=1}^{n} X_i$
	β known	$\prod_{i=1}^{n} X_i$
		$\left(\sum_{i=1}^{n} X_i, \prod_{i=1}^{n} X_i\right)$

Differentiating under the Integral

For a regular exponential family, it is possible to differentiate under the integral, that is,

$$\frac{\partial^m}{\partial \eta_i^m} \int C(\eta) \exp\left\{\sum_{j=1}^{k} \eta_j T_j(x)\right\} h(x)dx$$
$$= \int \frac{\partial^m}{\partial \eta_i^m} C(\eta) \exp\left\{\sum_{j=1}^{k} \eta_j T_j(x)\right\} h(x)dx$$

for any $m = 1, 2, \ldots$ and any η in the interior of the natural parameter space.

Let $X = (X_1, \ldots, X_n)$ denote observations from a distribution with probability density function $\{f_\theta(x); \theta \in \Omega\}$, and let $U(X)$ be a statistic. The information on the parameter θ is provided by the sensitivity of the distribution, of a statistic to changes in the parameter. For example, suppose a modest change in the parameter value leads to a large change in the expected value of the distribution, resulting in a large shift in the data. Then the parameter can be estimated fairly precisely. On the other hand, if a statistic U has no sensitivity at all in distribution to the parameter, then it would appear to contain little information for point estimation of this parameter. A statistic of the second kind is called an *ancillary* statistic.

Definition $U(X)$ is an *ancillary statistic* if its distribution does not depend on the unknown parameter θ.

Ancillary statistics are, in a sense, orthogonal or perpendicular to minimal sufficient statistics and are analogous to the residuals in a multiple regression, while the complete sufficient statistics are analogous to the estimators of the regression coefficients. It is well known that the residuals are uncorrelated with the estimators of the regression coefficients (and independent in the case of normal errors). However, the "irrelevance" of the ancillary statistic seems to be limited to the case when it is not part of the minimal (preferably complete) sufficient statistic, as the following example illustrates.

Example Suppose a fair coin is tossed to determine a random variable $N = 1$ with probability 1/2 and $N = 100$ otherwise. We then observe a binomial random variable X with parameters (N, θ). Then the minimal sufficient statistic is (X, N), but N is an ancillary statistic since its distribution does not depend on the unknown parameter θ. Is N completely irrelevant to inference about θ? If you reported to your boss an estimator of θ such as X/N without telling

him or her the value of N, how long would you expect to keep your job? Clearly, any sensible inference about θ should include information about the precision of the estimator, and this inevitably requires knowing the value of N. Although the distribution of N does not depend on the unknown parameter θ so that N is ancillary, it carries important information about precision. The following theorem allows us to use the properties of completeness and ancillarity to prove the independence of two statistics without finding their joint distribution.

Theorem B7 (Basu's Theorem) *Consider X with probability density function $\{f_\theta(x); \theta \in \Omega\}$. Let $T(X)$ be a complete sufficient statistic. Then $T(X)$ is independent of every ancillary statistic $U(X)$.*

Example Assume X_t represents the market price of a given asset such as a portfolio of stocks at time t, and x_0 is the value of the portfolio at the beginning of a given time period (assume that the analysis is conditional on x_0 so that x_0 is fixed and known). The process X_t is assumed to be a Brownian motion and so the distribution of X_t for any fixed time t is normal$(x_0 + \mu t, \sigma^2 t)$ for $0 < t \leq 1$. Suppose that for a period of length 1, we record both the period high $\max_{\{0 \leq t \leq 1\}} X_t$ and the close X_1. Define random variables $M = \max_{\{t \leq 1\}} X_t - x_0$ and $Y = X_1 - x_0$. Then the joint probability density function of (M, Y) can be shown to be

$$f_\theta(m, y) = \frac{2(2m - y)}{\sqrt{2\pi}\sigma^3} \exp\{[2\mu y - \mu^2 - (2m - y)^2]/(2\sigma^2)\}$$
$$-\infty < y < m, \quad m > 0, \quad \theta = (\mu, \sigma^2)$$

It is not hard to show that this is a member of the regular exponential family of distributions with both parameters assumed unknown. If one parameter is known, for example, σ^2, it is again a regular exponential family distribution with $k = 1$. Consequently, if we record independent pairs of observations (M_i, Y_i), $i = 1, \ldots, n$ on the portfolio for a total of n distinct time periods (and if we assume no change in the parameters), then the statistic

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

is a complete sufficient statistic for the drift parameter μ. Since it is also an unbiased estimator of μ, it is the UMVUE of μ. By Basu's theorem it will be independent of any ancillary statistic, i.e. any statistic whose distribution does not depend on the parameter μ. One such statistic is $Z = \sum_i M_i(M_i - Y_i)$, which is therefore independent of $\overline{Y}$.

1.2 MAXIMUM-LIKELIHOOD ESTIMATION

Suppose we have observed n independent discrete random variables all with probability density function

$$P_\theta(X = x) = f_\theta(x)$$

where the scalar parameter θ is unknown. Suppose our observations are $x_1, \ldots, x_n$. Then the probability of the observed data is

$$\prod_{i=1}^{n} P_\theta(X = x_i) = \prod_{i=1}^{n} f_\theta(x_i)$$

When the observations have been substituted, this becomes a function of the parameter only, referred to as the *likelihood function* and denoted $L(\theta)$. Its natural logarithm is usually denoted $\ell(\theta) = \ln(L(\theta))$. In the absence of any other information, it seems logical that we should estimate the parameter θ using a value most compatible with the data. For example, we might choose the value maximizing the likelihood function $L(\theta)$ or equivalently maximizing $\ell(\theta)$. We call such a maximizer the *maximum-likelihood (ML) estimate* provided it exists and satisfies any restrictions placed on the parameter. We denote it by $\hat{\theta}$. Obviously, it is a function of the data, that is, $\hat{\theta} = \hat{\theta}(x)$. The corresponding estimator is $\hat{\theta} = \hat{\theta}(X)$. In practice we are usually satisfied with a *local maximum* of the likelihood function provided that it is reasonable, partly because the global maximization problem is often quite difficult and partly because the global maximum is not always better than a local maximum near a preliminary estimator that is known to be consistent. In the case of a twice differentiable log-likelihood function on an open interval, this local maximum is usually found by solving the equation $S(\theta) = 0$ for a solution $\hat{\theta}$, where $S(\theta) = \ell'(\theta)$ is called the *score function*. The equation $S(\theta) = 0$ is called the *(maximum-) likelihood equation or score equation*. To verify a local maximum we compute the second derivative $\ell''(\hat{\theta})$ and show that it is negative, or alternatively show $I(\hat{\theta}) = -\ell''(\hat{\theta}) > 0$. The function $I(\theta) = -\ell''(\theta)$ is called the *information function*. In a sense to be investigated later, $I(\hat{\theta}) = -\ell''(\hat{\theta})$, the *observed information*, indicates how much information about a parameter is available in a given experiment. The larger the value, the more curved is the log-likelihood function and the easier it is to find the maximum.

Although we view the likelihood, log-likelihood, score, and information functions as functions of θ, they are, of course, also functions of the observed data $x = (x_1, \ldots, x_n)$. When it is important to emphasize the dependence on the data x we will write $L(\theta; x)$, $S(\theta; x)$, and so on. Also, when we wish to determine the sampling properties of these functions as functions of the

random variable $X = (X_1, \ldots, X_n)$ we will write $L(\theta; X)$, $S(\theta; X)$, and so on.

Definition The *Fisher* or *expected information function* is the expected value of the observed information function $J(\theta) = E_\theta[I(\theta; X)]$.

Likelihoods for Continuous Models

Suppose a random variable X has a continuous probability density function $f_\theta(x)$ with parameter θ. We will often observe only the value of X rounded to some degree of precision (say 1 decimal place), in which case the actual observation is a discrete random variable. For example, suppose we observe X correct to one decimal place. Then

$$P(\text{we observe } 1.1) = \int_{1.05}^{1.15} f_\theta(x)dx \approx (0.1)f_\theta(1.1)$$

assuming the function $f_\theta(x)$ is quite smooth over the interval. More generally, if we observe X rounded to the nearest Δ (assumed small), then the likelihood of the observation is approximately Δf_θ (observation). Since the precision Δ of the observation does not depend on the parameter, maximizing the discrete likelihood of the observation is essentially equivalent to maximizing the the probability density function f_θ(observation) over the parameter. This partially justifies the use of the probability density function in the continuous case as the likelihood function.

Similarly, if we observed n independent values $x_1, \ldots, x_n$ of a continuous random variable, we would maximize the likelihood $L(\theta) = \prod_{i=1}^n f_\theta(x_i)$ (or, more commonly, its logarithm) to obtain the maximum-likelihood estimator of θ.

The *relative-likelihood function* $R(\theta)$, defined as $R(\theta) = L(\theta)/L(\hat{\theta})$, is the ratio of the likelihood to its maximum value and takes on values between 0 and 1. It is used to rank possible parameter values according to their plausibility in light of the data. If $R(\theta_1) = 0.1$, say, then θ_1 is rather an implausible parameter value because the data is ten times more likely when $\theta = \hat{\theta}$ than when $\theta = \theta_1$. The set of θ-values for which $R(\theta) \geq p$ is called a $100p\%$ *likelihood region* for θ. When the parameter θ is one-dimensional, and θ_0 is its true value,

$$-2 \log R(\theta_0; X)$$

converges in distribution as the sample size $n \to \infty$ to a chi-squared distribution with 1 degree of freedom. More generally, the numbers of degrees of freedom of the limiting chi-squared distribution is the dimension of the

parameter θ. We can use this to construct a confidence interval for the unknown value of the parameter. For example, if b is chosen to be the 0.95 quantile of the chi-squared(1) distribution ($b = 3.84$), then

$$\begin{aligned}\{\theta : -2\log R(\theta; x) < b\} &= \{\theta : R(\theta; x) > e^{-b/2}\} \\ &\approx \{\theta : R(\theta; x) > 0.15\}\end{aligned}$$

so a 15% likelihood interval is an approximate 95% confidence interval for θ. This seems to indicate that the confidence interval tolerates a considerable difference in the likelihood. The likelihood at a parameter value must differ from the maximum likelihood by a factor of more than 6 before it is excluded by a 95% confidence interval or rejected by a test with level of significance 5%.

Properties of the Score and Information

Consider a continuous model with a family of probability density functions $\{f_\theta(x); \theta \in \Omega\}$. Suppose all of the densities are supported on a common set $\{x : f_\theta(x) > 0\} = A$. Then

$$\int_A f_\theta(x)dx = 1$$

and therefore

$$\int_A \frac{\partial}{\partial\theta} f_\theta(x)dx = \frac{\partial}{\partial\theta}\int_A f_\theta(x)dx = 0$$

provided that the integral can be interchanged with the derivative. Models that permit this interchange, and calculation of the Fisher information, are called *regular* models.

Regular Models

Consider a statistical model $\{f_\theta(x); \theta \in \Omega\}$, $x \in A$ with each density supported by a common set A. Suppose Ω is an open interval in the real line and $f_\theta(x) > 0$ for all $\theta \in \Omega$ and $x \in A$. Suppose in addition

1. $\ln[f_\theta(x)]$ is a continuous, three times differentiable function of θ for all $x \in A$.
2. $\frac{\partial^k}{\partial\theta^k}\int_A f_\theta(x)dx = \int_A \frac{\partial^k}{\partial\theta^k} f_\theta(x)dx,\ k = 1, 2.$
3. $|\frac{\partial^3 \ln f_\theta(x)}{\partial\theta^3}| < M(x)$ for some function $M(x)$ satisfying $\sup_\theta E_\theta[M(X)] < \infty$.

4. $0 < E_\theta\{[S(\theta; X)]^2\} < \infty$.

Then we call this a *regular* family of distributions or a regular model. Similarly, if these conditions hold with X a discrete random variable and the integrals replaced by sums, the family is also called *regular*. Conditions like these permitting the interchange of expected values and derivative are sometimes referred to as the Cramer conditions. In general, they are used to justify passage of a derivative under an integral.

Theorem B8 *If* $X = (X_1, \ldots, X_n)$ *is a random sample from a regular model* $\{f_\theta(x);$
$\theta \epsilon \Omega\}$, *then*

$$E_\theta[S(\theta; X)] = 0$$

and

$$\text{var}_\theta[S(\theta; X)] = E_\theta\{[S(\theta; X)]^2\} = E_\theta[I(\theta; X)] = J(\theta)$$

The Multiparameter Case

The case of several parameters is exactly analogous to the scalar parameter case. Suppose $\theta = (\theta_1, \ldots, \theta_k)'$. In this case the "parameter" can be thought of as a column vector of k scalar parameters. The score function $S(\theta)$ is a k-dimensional column vector whose ith component is the derivative of $\ell(\theta)$ with respect to the ith component of θ, that is,

$$S(\theta) = \left[\frac{\partial}{\partial\theta_1}\ell(\theta), \ldots, \frac{\partial}{\partial\theta_k}\ell(\theta)\right]'$$

The observed information function $I(\theta)$ is a $k \times k$ matrix whose (i, j) element is $-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta)$:

$$I(\theta) = [I_{ij}(\theta)]_{k\times k} = \left[-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta)\right]_{k\times k}, \quad i, j = 1, \ldots, k$$

The Fisher information is a $k \times k$ matrix whose components are componentwise expectations of the information matrix, that is

$$J_{ij}(\theta) = E_\theta[I_{ij}(\theta; X)], \quad i, j = 1, \ldots, k$$

The definition of a regular family of distributions is similarly extended. For a regular family of distributions

$$E_\theta[S(\theta; X)] = (0, \ldots, 0)'$$

and the covariance matrix of the score function $\text{var}_\theta[S(\theta; X)]$ is the Fisher information, that is,

$$J_{ij}(\theta) = E_\theta\left[\frac{\partial}{\partial\theta_i}\ell(\theta)\frac{\partial}{\partial\theta_j}\ell(\theta)\right]$$

Maximum-Likelihood Estimation in the Exponential Family

Suppose X has a regular exponential family distribution of the form

$$f_\eta(x) = C(\eta)\exp\left\{\sum_{j=1}^{k}\eta_j T_j(x)\right\}h(x)$$

Then

$$E_\eta[T_j(X)] = \frac{-\partial \ln C(\eta)}{\partial\eta_j} \quad \text{for } j = 1,\ldots,k$$

and

$$\text{cov}_\eta(T_i(X), T_j(X)) = \frac{-\partial^2 \ln C(\eta)}{\partial\eta_i\delta\eta_j} \quad \text{for } i,j = 1,\ldots,k$$

Then the maximum-likelihood estimator of η based on a random sample $(X_1,\ldots,X_n)$ from $f_\eta(x)$ is the solution to the k equations

$$E_\eta[T_j(X)] = \frac{1}{n}\sum_{i=1}^{n}T_j(x_i) \quad \text{for } j = 1,\ldots,k$$

The maximum-likelihood estimators are obtained by setting the sample moments of the natural sufficient statistic equal to their expected values and solving for the value of n.

Finding Maximum-Likelihood Estimates Using Newton's Method

Suppose the maximum-likelihood estimate $\hat{\theta}$ is determined by the likelihood equation

$$S(\theta) = 0$$

It frequently happens that an analytic solution for $\hat{\theta}$ cannot be obtained. If we begin with an approximate value for the parameter, $\theta^{(0)}$, we may update that value as follows:

$$\theta^{(i+1)} = \theta^{(i)} + \frac{S(\theta^{(i)})}{I(\theta^{(i)})}, \quad i = 0, 1, 2, \ldots$$

and provided that convergence of $\theta^{(i)}$, $i \to \infty$, obtains, it converges to a solution to the score equation above. In the multiparameter case, where $S(\theta)$ is a vector and $J(\theta)$ is a matrix, then Newton's method becomes

$$\theta^{(i+1)} = \theta^{(i)} + I^{-1}(\theta^{(i)})S(\theta^{(i)}), \quad i = 0, 1, 2, \ldots$$

In both of these cases we can replace the information function by the Fisher information for a similar algorithm.

Suppose we consider estimating a parameter $\tau(\theta)$, where θ is a scalar, using an unbiased estimator $T(X)$. Is there any limit to how well an estimator like this can behave? The answer for unbiased estimators is in the affirmative. A lower bound on the variance is given by the information inequality.

Information Inequality

Suppose $T(X)$ is an unbiased estimator of the parameter $\tau(\theta)$ in a *regular* statistical model $\{f_\theta(x); \theta \in \Omega\}$. Then

$$\text{var}_\theta(T) \geq \frac{[\tau'(\theta)]^2}{J(\theta)} \tag{1.2}$$

Equality holds if and only if $f_\theta(x)$ is regular exponential family with natural sufficient statistic $T(X)$.

If equality holds in (1.2), then we call $T(X)$ an *efficient* estimator of $\tau(\theta)$. The number on the right-hand side of (1.2),

$$\frac{[\tau'(\theta)]^2}{J(\theta)}$$

is called the *Cramér-Rao lower bound* (CRLB). We often express the *efficiency* of an unbiased estimator using the ratio of (CRLB) to the variance of the estimator. Large values of the efficiciency (i.e., near 1) indicate that the variance of the estimator is close to the lower bound.

The special case of the information inequality that is of most interest is the unbiased estimation of the parameter θ. The above inequality indicates that *any unbiased estimator* T of θ has variance at least $1/J(\theta)$. The lower bound is achieved only when $f_\theta(x)$ is regular exponential family with natural sufficient statistic T, so even in the exponential family, only certain parameters are such that we can find unbiased estimators that achieve the CRLB, namely those that are expressible as the expected value of the natural sufficient statistics.

The Multiparameter Case

The right-hand side in the information inequality generalizes naturally to the multiple-parameter case in which θ is a vector. For example, if $\theta = (\theta_1, \ldots, \theta_k)'$, then the Fisher information $J(\theta)$ is a $k \times k$ matrix. If $\tau(\theta)$ is any real-valued function of θ, then its derivative is a column vector $\left(\frac{\partial \tau}{\partial \theta_1}, \ldots, \frac{\partial \tau}{\partial \theta_k}\right)'$. Then if $T(X)$ is any unbiased estimator of $\tau(\theta)$ in a regular model,

$$\mathrm{var}_\theta(T) \geq \left(\frac{\partial \tau}{\partial \theta_1}, \ldots, \frac{\partial \tau}{\partial \theta_k}\right) [J(\theta)]^{-1} \left(\frac{\partial \tau}{\partial \theta_1}, \ldots, \frac{\partial \tau}{\partial \theta_k}\right)'$$

for all $\theta \in \Omega$.

Asymptotic Properties of Maximum-Likelihood Estimators

One of the more successful attempts at justifying estimators and demonstrating some form of optimality has been through *large-sample theory* or the asymptotic behavior of estimators as the sample size $n \to \infty$. One of the first properties one requires is consistency of an estimator. This means that the estimator converges to the true value of the parameter as the sample size (and hence the information) approaches infinity.

Definition Consider a sequence of estimators T_n, where the subscript n indicates that the estimator has been obtained from data $(X_1, \ldots, X_n)$ with sample size n. Then the sequence is said to be a *consistent* sequence of estimators of $\tau(\theta)$ if $T_n \to_p \tau(\theta)$ for all $\theta \in \Omega$.

It is worth a reminder at this point that probability density functions are used to produce probabilities and are unique only up to a point. For example, if two probability density functions $f(x)$ and $g(x)$ were such that they produced the same probabilities, or the same cumulative distribution function—for example,

$$\int_{-\infty}^{x} f(z)dz = \int_{-\infty}^{x} g(z)dz$$

for all x—then we would not consider them distinct probability densities, even though $f(x)$ and $g(x)$ may differ at one or more values of x. When we parameterize a given statistical model using θ as the parameter, it is natural to do so in such a way such that *different values of the parameter lead to distinct probability density functions*. This means, for example, that the cumulative distribution functions associated with these densities are distinct. Without

this assumption, made in the following theorem, it would be impossible to accurately estimate the parameter since two different parameters could lead to the same cumulative distribution function and hence exactly the same behavior of the observations.

Theorem B9 *Suppose* $(X_1, \ldots, X_n)$ *is a random sample from a regular statistical model* $\{f_\theta(x); \theta \in \Omega\}$. *Assume the densities corresponding to different values of the parameters are distinct. Let* $S_1(\theta; X_i) = \frac{\partial}{\partial\theta} \ln f_\theta(X_i)$. *Then with probability tending to* 1 *as* $n \to \infty$, *the likelihood equation*

$$\sum_{i=1}^{n} S_1(\theta; X_i) = 0$$

has a root $\hat{\theta}_n$ *such that* $\hat{\theta}_n$ *converges in probability to* θ_0, *the true value of the parameter, as* $n \to \infty$.

The likelihood equation above does not always have a unique root. The consistency of the maximum-likelihood estimator is one indication that it performs reasonably well. However, it provides no reason to prefer it to some other consistent estimator. The following result indicates that maximum-likelihood estimators perform as well as any reasonable estimator can, at least in the limit as $n \to \infty$. Most of the proofs of these asymptotic results can be found in Lehmann (*The Theory of Point Estimation*, Wiley, New York, 1983).

Theorem B10 *Suppose* $(X_1, \ldots, X_n)$ *is a random sample from a regular statistical model* $\{f_\theta(x); \theta \in \Omega\}$. *Suppose* $\hat{\theta}_n$ *is a consistent root of the likelihood equation as in the theorem above. Let* $J_1(\theta) = E_\theta\{\frac{-\partial^2}{\partial\theta^2} \ln f_\theta(X)\}$, *the Fisher information for a sample of size one. Then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_D Y \sim \mathrm{N}\left(0, \frac{1}{J_1(\theta_0)}\right)$$

where θ_0 *is the true value of the parameter.*

This result may also be written as

$$\sqrt{nJ_1(\theta_0)}\,(\hat{\theta}_n - \theta_0) = \sqrt{J(\theta_0)}\,(\hat{\theta}_n - \theta_0) \to_D Z \sim \mathrm{N}(0, 1)$$

This theorem asserts that, at least under the regularity required, the maximum-likelihood estimator is asymptotically unbiased. Moreover, the asymptotic variance of the maximum-likelihood estimator approaches the Cramér-Rao

lower bound for unbiased estimators. This justifies the comparison of the variance of an estimator T_n based on a sample of size n to the value $[nJ_1(\theta_0)]^{-1}$, which is the *asymptotic variance* of the maximum-likelihood estimator and also the Cramér-Rao lower bound.

It also follows that

$$\sqrt{n}[\tau(\hat{\theta}_n) - \tau(\theta_0)] \rightarrow_D W \sim \mathrm{N}\left(0, \frac{[\tau'(\theta_0)]^2}{J_1(\theta_0)}\right)$$

This indicates that the asymptotic variance of any function $\tau(\hat{\theta}_n)$ of the maximum-likelihood estimator also achieves the Cramér-Rao lower bound.

Definition Suppose T_n is asymptotically normal with mean θ_0 and variance σ_T^2/n. The *asymptotic efficiency* of T_n is defined to be $[\sigma_T^2 J_1(\theta_0)]^{-1}$. This is the ratio of the Cramér-Rao lower bound to the variance of T_n and is typically less than 1, with a value close to 1 indicating the asymptotic efficiency is close to that of the maximum-likelihood estimator.

The Multiparameter Case In the case $\theta = (\theta_1, \ldots, \theta_k)'$, the score function is the vector of partial derivatives of the log-likelihood with respect to the components of θ. Therefore, the likelihood equation is k equations in the k unknown parameters. Under similar regularity conditions to the univariate case, the conclusion of Theorem B9 holds in this case; that is, the components of $\hat{\theta}_n$ each converge in probability to the corresponding component of θ_0. Similarly, the asymptotic normality remains valid in this case with little modification. Let $J_1(\theta)$ be the Fisher information matrix for a sample of size 1 and assume it is a nonsingular matrix. Then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_D Y \sim \mathrm{MVN}\,(0, [J_1(\theta_0)]^{-1})$$

where the multivariate normal distribution with k-dimensional mean vector μ and covariance matrix $B(k \times k)$, denoted $MVN(\mu, B)$ has probability density function defined on $\mathcal{R}^k$,

$$f(x) = \frac{1}{(2\pi)^{k/2}|B|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)'B^{-1}(x-\mu)\right\}$$

It also follows that

$$\sqrt{n}[\tau(\hat{\theta}_n) - \tau(\theta_0)] \rightarrow_D W \sim \mathrm{MVN}\,(0, [D(\theta_0)]'[J_1(\theta_0)]^{-1}D(\theta_0))$$

where $D(\theta) = \left(\frac{\partial\tau}{\partial\theta_1}, \ldots, \frac{\partial\tau}{\partial\theta_k}\right)'$. Once again, the asymptotic variance-covariance matrix is identical to the lower bound given by the multiparameter case of the information inequality.

Joint confidence regions can be constructed based on one of the asymptotic results

$$(\hat{\theta}_n - \theta_0)' J(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \to_D W \sim \chi^2(k)$$
$$(\hat{\theta}_n - \theta_0)' I(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \to_D W \sim \chi^2(k)$$

or

$$\Lambda_n(X) = -2 \ln R(\theta_0) \to_D W \sim \chi^2(k)$$

Confidence intervals for a single parameter, say θ_i, can be based on the approximate normality of

$$\{[J^{-1}(\hat{\theta}_n)]_{ii}\}^{-1/2}[(\hat{\theta}_n)_i - (\theta_0)_i]$$

or

$$\{[I^{-1}(\hat{\theta}_n)]_{ii}\}^{-1/2}[(\hat{\theta}_n)_i - (\theta_0)_i]$$

where $(a)_i$ is the ith entry in the vector a and $[A^{-1}]_{ii}$ is the (i, i) entry in the matrix A^{-1}.

Unidentifiability and Singular Information Matrices

Suppose we observe two independent random variables Y_1, Y_2 having normal distributions with the same variance σ^2 and means $\theta_1 + \theta_2$, $\theta_2 + \theta_3$, respectively. In this case, although the means depend on the parameter $\theta = (\theta_1, \theta_2, \theta_3)$, the value of this vector parameter is *unidentifiable* in the sense that, for some pairs of distinct parameter values, the probability density functions of the observations are identical. For example, the parameter $(1, 0, 1)$ leads to exactly the same joint distribution of Y_1, Y_2 as does the parameter $(0, 1, 0)$. In this case, we we might consider only the two parameters $(\phi_1, \phi_2) = (\theta_1 + \theta_2, \theta_2 + \theta_3)$ and anything derivable from this pair estimable, whereas parameters such as θ_2 that cannot be obtained as functions of ϕ_1, ϕ_2 are consequently unidentifiable. The solution to the original identifiability problem is the reparameterization to the new parameter (ϕ_1, ϕ_2) in this case, and in general, unidentifiability usually means one should seek a new, more parsimonious parameterization.

In the above example, we may compute the Fisher information matrix for the parameter $\theta = (\theta_1, \theta_2, \theta_3)$ as follows.

The log likelihood is

$$\ell(\theta) = -\frac{1}{2\sigma^2}\left\{(y_1 - \theta_1 - \theta_2)^2 + (y_2 - \theta_2 - \theta_3)^2\right\} + c$$

and the Fisher information is the covariance matrix of the score vector

$$S(\theta) = \frac{1}{\sigma^2}\begin{pmatrix} y_1 - \theta_1 - \theta_2 \\ y_1 + y_2 - \theta_1 - 2\theta_2 - \theta_3 \\ y_2 - \theta_2 - \theta_3 \end{pmatrix}$$

and this is

$$J(\theta) = \frac{1}{\sigma^2}\begin{pmatrix} 1 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

Notice that $J(\theta)$ is, in this case, singular. If you were to attempt to compute the asymptotic variance of the maximum likleihood estimator of θ by inverting this information matrix, the inversion is impossible. Attempting to invert a singular matrix is like attempting the inverse of 0, one or more components of the inverse can be taken to be infinite, indicating that, asymptotically at least, one of more of the parameters is unidentifiable. When parameters are unidentifiable, the Fisher information matrix is generally singular. However, when $J(\theta)$ is singular for all values of θ, this may or may not mean that parameters are unidentifiable for finite sample sizes, but it does usually mean one should take a careful look at the parameters with a possible view to adopting another parameterization.

UMVUEs and Maximum-Likelihood Estimators: A Comparison

Which of the two main types of estimators should we use? There is no general consensus among statisticians.

1. If we are estimating the expectation of a natural sufficient statistic $T_i(X)$ in a regular exponential family, both maximum-likelihood and unbiasedness considerations lead to the use of T_i as an estimator.
2. When sample sizes are large, UMVUEs and maximum-likelihood estimators are essentially the same. In that case use is governed by ease of computation. Unfortunately how large "large" needs to be is usually unknown. Some studies have been carried out comparing the behavior of UMVUEs and maximum-likelihood estimators for various small fixed sample sizes. The results are, as might be expected, inconclusive.
3. Maximum-likelihood estimators exist "more frequently," and when they do they are usually easier to compute than UMVUEs. This is essentially because of the appealing invariance property of maximum-likelihood estimators.
4. Simple examples are known for which maximum-likelihood estimators behave badly even for large samples. This is more often the case when there are a large number of parameters, some of which, termed "nuisance parameters," are of no direct interest, but complicate the estimation.
5. UMVUEs and maximum-likelihood estimators are not necessarily robust. A small change in the underlying distribution or the data could result in a large change in the estimator.

1.3 OTHER ESTIMATION CRITERIA

Best Linear Unbiased Estimators

The problem of finding best unbiased estimators is considerably simpler if we limit the class in which we search. If we permit any function of the data, then we usually require the heavy machinery of complete sufficiency to produce UMVUEs. However, the situation is much simpler if we suggest some initial random variables and then require that our estimator be a linear combination of these. Suppose, for example we have random variables Y_1, Y_2, Y_3 with $E(Y_1) = \alpha + \theta$, $E(Y_2) = \alpha - \theta$, $E(Y_3) = \theta$, where θ is the parameter of interest and α is another parameter. What linear combinations of the Y_i provide an unbiased estimator of θ, and among these possible linear combinations which one has the smallest possible variance? To answer these questions, we need to know the covariances $\text{cov}(Y_i, Y_j)$ (at least up to some scalar multiple). Suppose $\text{cov}(Y_i, Y_j) = 0$, $i \neq j$, and $\text{var}(Y_j) = \sigma^2$. Let $Y = (Y_1, Y_2, Y_3)'$ and $\beta = (\alpha, \theta)'$. We can write the model in a form reminiscent of linear regression as

$$Y = X\beta + \epsilon$$

where

$$X = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 1 \end{bmatrix}$$

$\epsilon = (\epsilon_1, \epsilon_2, \epsilon_3)'$ and the ϵ_i are uncorrelated random variables with $E(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma^2$. Then the linear combination of the components of Y that has the smallest variance among all unbiased estimators of β is given by the usual regression formula, $\tilde{\beta} = (\tilde{\alpha}, \tilde{\theta})' = (X'X)^{-1}X'Y$, and $\tilde{\theta} = \frac{1}{3}(Y_1 - Y_2 + Y_3)$ provides the best estimator of θ in the sense of smallest variance. In other words, the linear combination of the components of Y that has the smallest variance among all unbiased estimators of $a'\beta$ is $a'\tilde{\beta}$, where $a' = (0, 1)$.

More generally, we wish to consider a number n of possibly dependent random variables Y_i whose expectations may be related to a parameter θ. These may, for example, be individual observations or a number of competing estimators constructed from these observations. We assume $Y = (Y_1, \ldots, Y_n)'$ has expectation given by

$$E(Y) = X\beta$$

where X is some $n \times k$ matrix having rank k and $\beta = (\beta_1, \ldots, \beta_k)'$ is a vector of unknown parameters. As in multiple regression, the matrix X is known and nonrandom. Suppose the covariance matrix of Y is $\sigma^2 B$, with B a known

nonsingular matrix and σ^2 a possibly unknown scalar parameter. We wish to estimate a linear combination of the components of β, say $\theta = a'\beta$, where a is a known k-dimensional column vector. We restrict our attention to unbiased estimators of θ.

Theorem B11: Gauss-Markov Theorem Suppose Y is a random vector with mean and covariance matrix

$$E(Y) = X\beta$$
$$cov(Y_i, Y_j) = \sigma^2 B$$

where matrices X and B are known and the parameters β and σ^2 unknown. Suppose we wish to estimate a linear combination $\theta = a'\beta$ of the components of β. Then among all linear combinations of the components of Y which are unbiased estimators of the parameter θ, the estimator

$$\widetilde{\theta} = a'(X'B^{-1}X)^{-1}X'B^{-1}Y$$

has the smallest variance. Note that this result does not depend on any assumed normality of the components of Y but only on the first and second moment behavior, that is, the mean and the covariances. The special case when B is the identity matrix is the least squares estimator.

Estimating Equations

To find the maximum-likelihood estimator, we usually solve the likelihood equation

$$\sum_{i=1}^{n} S_1(\theta; X_i) = 0 \tag{1.3}$$

Note that the function on the left-hand side is a function of both the observations and the parameter. Such a function is called an *estimating function*. Most sensible estimators, like the maximum-likelihood estimator, can be described easily through an estimating function. For example, if we know $\text{var}_\theta(X_i) = \theta$ for independent identically distributed X_i, then we can use the estimating function

$$\psi(\theta, X) = \sum_{i=1}^{n} (X_i - \overline{X})^2 - (n-1)\theta \tag{1.4}$$

to estimate the parameter θ, without any other knowledge of the distribution, its density, mean, and so on. The estimating function is set equal to 0

and solved for θ. The above estimating function is *an unbiased estimating function* in the sense that

$$E_\theta[\psi(\theta, X)] = 0, \quad \text{for all } \theta \tag{1.5}$$

This allows us to conclude that the function is at least centered appropriately for the estimation of the parameter θ. Now suppose that ψ is an unbiased estimating function corresponding to a large sample. Often it can be written as the sum of independent components, for example,

$$\psi(\theta, X) = \sum_{i=1}^{n} \psi(\theta, X_i) \tag{1.6}$$

Now suppose $\hat{\theta}$ is a root of the estimating equation

$$\psi(\theta, X) = 0$$

Then for θ sufficiently close to $\hat{\theta}$,

$$\psi(\theta, X) = \psi(\theta, X) - \psi(\hat{\theta}, X) \approx (\theta - \hat{\theta}) \frac{\partial}{\partial \theta} \psi(\theta, X) \tag{1.7}$$

Using the Central Limit Theorem, assuming that θ is the true value of the parameter, and provided ψ is a sum as in (1.6), the left-hand side of (1.7) is approximately normal with mean 0 and variance equal to $\text{var}_\theta[\psi(\theta, X)]$. The term $\frac{\partial}{\partial \theta} \psi(\theta, X)$ is also a sum of similar derivatives of the individual ψ_i. If a law of large numbers applies to these terms, then when divided by n this sum will be asymptotically equivalent to $\frac{1}{n} E_\theta[\partial \psi(X, \theta)/\partial \theta]$. It follows that the root $\hat{\theta}$ will have an approximate normal distribution with mean θ and variance

$$\frac{\text{var}_\theta[\psi(\theta, X)]}{\{E_\theta[\partial \psi(\theta, X)/\partial \theta]\}^2}$$

By analogy with the relation between asymptotic variance of the maximum-likelihood estimator and the Fisher information, we call the reciprocal of the above asymptotic variance formula the *Godambe information* of the estimating function. This information measure is

$$J(\psi, \theta) = \frac{\{E_\theta[\partial \psi(\theta, X)/\partial \theta]\}^2}{\text{var}_\theta[\psi(\theta, X)]} \tag{1.8}$$

Godambe (1960) proved the following result.

Theorem B12 *Among all unbiased estimating functions satisfying the usual regularity conditions, an estimating function that maximizes the Godambe information* (1.8) *is of the form* $c(\theta)S(\theta; X)$*, where* $c(\theta)$ *is nonrandom.*

1.4 BAYESIAN METHODS

There are two major schools of thought on the way in which statistical inference is conducted, the *frequentist* school and the *Bayesian* school. Typically, these schools differ slightly on the actual methodology and the conclusions that are reached, and more substantially on the philosophy underlying the treatment of parameters. So far we have considered a parameter as an unknown constant underlying or indexing the probability density function of the data. It is only the data, and statistics derived from the data, that are random.

The Bayesian begins with the assertion that the parameter θ obtains as the realization of some larger random experiment. The parameter is assumed to have been generated according to some distribution, *the prior distribution* π, and the observations then obtained from the corresponding probability density function f_θ interpreted as the conditional probability density of the data given the value of θ. The prior distribution $\pi(\theta)$ quantifies information about θ prior to any further data being gathered. Sometimes $\pi(\theta)$ can be constructed on the basis of past data. For example, if a quality inspection program has been running for some time, the distribution of the number of defectives in past batches can be used as the prior distribution for the number of defectives in a future batch. The prior can also be chosen to incorporate subjective information based on an expert's experience and personal judgment. The purpose is then to adjust this distribution for θ in the light of the data, to result in the *posterior distribution* for the parameter. Any conclusions about the plausible value of the parameter are to be drawn from the posterior distribution. For a frequentist, statements like $P(1 < \theta < 2)$ are meaningless; all randomness lies in the data, and the parameter is an unknown constant. Frequentists are careful to assure students that if an observed 95% confidence interval for the parameter is $1 < \theta < 2$, this does not imply $P(1 < \theta < 2) = 0.95$. However, a Bayesian will happily quote such a probability, usually conditionally on some observations, for example, $P(1 < \theta < 2|X) = 0.95$. In spite of some distance in the philosophies regarding the (random?) nature of statistical parameters, the two paradigms largely agree for large sample sizes because the prior assumptions of the Bayesian tend to be a small contributor to the conclusion.

Posterior Distributions

Suppose the parameter is initially chosen at random according to the prior distribution $\pi(\theta)$ and then, *given the value of the parameter*, the observations

are independent identically distributed, each with conditional probability (density) function $f_\theta(x)$. Then the *posterior distribution of the parameter* is the conditional distribution of θ given the data $x = (x_1, \ldots, x_n)$:

$$\pi(\theta|x) = c\pi(\theta)\prod_{i=1}^{n} f_\theta(x_i) = c\pi(\theta)L(\theta)$$

where $c = 1/\int_{-\infty}^{\infty} \pi(\theta)L(\theta)d\theta$ is independent of θ and $L(\theta)$ is the likelihood function. Since Bayesian inference is based on the posterior distribution, it depends on the data only through the likelihood function.

Example Suppose a coin is tossed n times with probability of heads θ. It is known that the prior probability of heads is not always identically 1/2 but follows a beta $(10, 10)$ distribution. If the n tosses result in x heads, we wish to find the posterior density function for θ. The prior distribution for the parameter θ is the beta(10,10) distribution with probability density function

$$\pi(\theta) = \frac{\Gamma(20)}{\Gamma(10)\Gamma(10)}\theta^9(1-\theta)^9, \quad 0 < \theta < 1$$

The posterior distribution of θ is therefore proportional to

$$\begin{aligned}\pi(\theta)f_\theta(x) &= \frac{\Gamma(20)}{\Gamma(10)\Gamma(10)}\theta^9(1-\theta)^9\binom{n}{x}\theta^x(1-\theta)^{n-x} \\ &= C\theta^{9+x}(1-\theta)^{9+n-x}, \quad 0 < \theta < 1\end{aligned}$$

where the constant C may depend on x but does not depend on θ. Therefore the posterior distribution is also a beta distribution but with parameters $(10 + x, 10 + n - x)$. Notice that the posterior mean is the expected value of this beta distribution and is

$$\frac{10 + x}{10 + n - x}$$

which, for n and x sufficiently large, is reasonably close to the usual estimator x/n.

Conjugate Prior Distributions

If a prior distribution has the property that the posterior distribution is in the same family of distributions as the prior, then the prior is called a *conjugate prior*.

Suppose $(X_1, \ldots, X_n)$ is a random sample from the exponential family

$$f_\theta(x) = C(\theta)\exp[q(\theta)T(x)]h(x)$$

and θ is assumed to have the prior distribution given by

$$\pi(\theta) = k[C(\theta)]^a \exp[bq(\theta)] \tag{1.9}$$

where a and b are parameters and

$$k = \frac{1}{\int_{-\infty}^{\infty}[C(\theta)]^a \exp[bq(\theta)]d\theta}$$

Then the posterior distribution of θ, given the data $x = (x_1, \ldots, x_n)$, is easily seen to be given by

$$\pi(\theta|x) = c[C(\theta)]^{a+n} \exp\left\{q(\theta)\left[b + \sum_{i=1}^{n} T(x_i)\right]\right\}$$

where

$$c = \frac{1}{\int_{-\infty}^{\infty}[C(\theta)]^{a+n} \exp\{q(\theta)[b + \sum_{i=1}^{n} T(x_i)]\}d\theta}$$

Notice that the posterior distribution is in the same family of distributions as (1.9) and thus $\pi(\theta)$ is a conjugate prior. The value of the parameters of the posterior distribution reflect the choice of parameters in the prior.

Example To find the conjugate prior for $\theta = (\alpha, \beta)$ for a random sample $(X_1, \ldots, X_n)$ from the beta(α, β) distribution with probability density function

$$f_\theta(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1, \quad \text{for } \alpha, \beta > 0$$

we begin by writing this in exponential family form,

$$f_\theta(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \exp\{(\alpha-1)\ln x + (\beta-1)\ln(1-x)\}$$

Then the conjugate prior distribution is the joint probability density function $\pi(\alpha, \beta)$ on (α, β), that is proportional to

$$\pi(\alpha, \beta) \propto \{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\}^a \exp\{-b_1(\alpha-1) - b_2(\beta-1)\} \tag{1.10}$$

for parameters a, b_1, b_2. The posterior distribution takes the same form as (1.10) but with the parameters a, b_1, b_2 replaced by $a+n$, $-b_1 + \sum_{i=1}^{n} \ln(X_i)$, $-b_2 + \sum_{i=1}^{n} \ln(1-X_i)$. Bayesians are sometimes criticized for allowing their

subjective opinions (in this case leading to the choice of the prior parameters a, b_1, b_2) influence the resulting inference, but notice that in this case, and more generally, as the sample size n grows, the value of the parameters of the posterior distribution is mostly determined by the components $n, \sum_{i=1}^{n} \ln(X_i), \sum_{i=1}^{n} \ln(1 - X_i)$ above, which grow in n, eventually washing out the influence of the value of the prior parameters.

Noninformative Prior Distributions

The choice of the prior distribution to be the conjugate prior is often motivated by mathematical convenience. However, a Bayesian would also like the prior to accurately represent the preliminary uncertainty about the plausible values of the parameter, and this may not be easily translated into one of the conjugate prior distributions. Noninformative priors are the usual way of representing ignorance about θ, and they are frequently used in practice. It can be argued that they are more objective than a subjectively assessed prior distribution since the latter may contain personal bias as well as background knowledge. Also, in some applications the amount of prior information available is far less than the information contained in the data. In this case there seems little point in worrying about a precise specification of the prior distribution.

In the coin tossing example above, we assumed a beta(10,10) prior distribution for the probability of heads. If were no reason to prefer one value of θ over any other, then a noninformative or "flat" prior disribution for θ that could be used is the uniform(0, 1) distribution—also, as it turns out, a special case of the beta distribution. Ignorance may not be bliss, but for Bayesians it is most often uniformly distributed. For estimating the mean θ of a $\mathrm{N}(\theta, 1)$ distribution the possible values for θ are $(-\infty, \infty)$. If we take the prior distribution to be uniform on $(-\infty, \infty)$, that is,

$$\pi(\theta) = c, \quad -\infty < \theta < \infty$$

then this is not a proper probability density since

$$\int_{-\infty}^{\infty} \pi(\theta) d\theta = c \int_{-\infty}^{\infty} d\theta = \infty \quad \text{if } c > 0$$

Prior densities of this type are called improper priors. In this case we could consider a sequence of prior distributions such as the uniform$(-M, M)$, which approximates this prior as $M \to \infty$. Suppose we call such a prior density function π_M. Then the posterior distribution of the parameter is given by

$$\pi(\theta|x) = c\pi_M(\theta)L(\theta)$$

and it is easy to see that as $M \to \infty$, this approaches a constant multiple of the likelihood function $L(\theta)$. For reasonably large sample size, $L(\theta)$ is often an integrable function of θ and can therefore be normalized to produce a proper posterior distribution, even though the corresponding prior was improper. This Bayesian development provides an alternative interpretation of the likelihood function. We can consider it as proportional to the posterior distribution of the parameter using a *uniform improper prior* on the whole real line. The language is somewhat sloppy here since, as we have seen, the uniform distribution on the whole real line really makes sense only through taking limits for uniform distributions on finite intervals.

In the case of a scale parameter, which must take positive values such as the normal variance, it is usual to express ignorance of the prior distribution of the parameter by assuming that the logarithm of the parameter is uniform on the real line.

One possible difficulty with using nonformative prior distributions is the concern of whether the prior distribution should be uniform for θ itself or some function of θ, such as θ^2 or $\log(\theta)$. The objective when we used a uniform prior for a probability was to add no more information about the parameter around one possible value than around some other, and so it makes sense to use a uniform prior for a parameter that essentially has uniform information attached to it. For this reason, it is common to use a uniform prior for $\tau = h(\theta)$, where $h(\theta)$ is the function of θ whose Fisher information, $J^*(\tau)$, is constant. This idea is due to Jeffreys and leads to a prior distribution that is proportional to $[J(\theta)]^{1/2}$. Such a prior is referred to as a *Jeffreys' prior*. The reparameterization that leads to a Jeffrey's prior can be carried out as follows: Suppose $\{f_\theta(x);\ \theta \in \Omega\}$ is a regular model and $J_1(\theta) = E_\theta \left\{ \frac{-\partial^2}{\partial \theta^2} \log f_\theta(X) \right\}$ is the Fisher information for a single observation. Then if we choose an abitrary value for θ_0 and define the reparameterization

$$\tau(\theta) = \int_{\theta_0}^{\theta} \sqrt{J_1(u)}\, du \tag{1.11}$$

then in this case, the Fisher information for the parameter τ, $J_1^*(\tau)$, equals 1 for all values of τ, and so Jeffry's prior corresponds to using a uniform prior distribution on the values of τ. Since the asymptotic variance of the maximum-likelihood estimator $\hat{\tau}_n$ is equal to $1/n$, which does not depend on τ, (1.11) is often called a *variance-stabilizing transformation*.

Bayes Point Estimators

One method of obtaining a point estimator of θ is to use the posterior distribution and a suitable loss function.

Theorem B13 *The* Bayes estimator of θ for squared error loss *with respect to the prior* $\pi(\theta)$*given data* X *is the mean of the posterior distribution given by*

$$\tilde{\theta} = \tilde{\theta}(X) = \int_{-\infty}^{\infty} \theta\pi(\theta|X)d\theta$$

This estimator minimizes over all functions $\tilde{\theta}(x)$ *of the data*

$$E[(\tilde{\theta} - \theta)^2|X] = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} \left(\tilde{\theta} - \theta\right)^2 f_\theta(x)dx \right\} \pi(\theta)d\theta$$

Example Suppose $(X_1, \ldots, X_n)$ is a random sample from the distribution with probability density function

$$f_\theta(x) = \theta x^{\theta-1} \qquad 0 < x < 1, \qquad \theta > 1$$

Using a conjugate prior for θ, find the Bayes estimator of θ for squared error loss.

We begin by identifying the conjugate prior distribution. Since

$$f_\theta(x) = \theta \exp\{(\theta - 1)\ln x\} \qquad 0 < x < 1, \qquad \theta > 1$$

the conjugate prior density is

$$\pi(\theta) = k\theta^a \exp\{b\theta\}, \quad \theta > 1$$

which is evidently a gamma distribution restricted to the interval $(1, \infty)$, and if the prior is to be proper, the parameters must be chosen such that

$$k^{-1} = \int_1^{\infty} \theta^a \exp\{b\theta\}d\theta < \infty$$

so $b \leq 0$. Then the posterior distribution takes the same form as the prior but with a replaced by $a + n$ and b by $b + \sum_{i=1}^n \ln(X_i)$. The Bayes estimate of θ for squared error loss is the mean of this posterior distribution, or

$$\frac{\int_1^{\infty} \theta^{a+n+1} \exp\{(b + \sum_{i=1}^n \ln(X_i))\theta\}d\theta}{\int_1^{\infty} \theta^{a+n} \exp\{(b + \sum_{i=1}^n \ln(X_i))\theta\}d\theta}$$

Bayesian Interval Estimates

There remains, after many decades, a controversy between Bayesians and frequentists about which approach to estimation is more suitable to the real world. The Bayesian has advantages at least in the ease of interpretation of

the results. For example, a Bayesian can use the posterior distribution given the data $x = (x_1, \ldots, x_n)$ to determine points $c_1 = c_1(x)$, $c_2 = c_2(x)$ such that

$$\int_{c_1}^{c_2} \pi(\theta|x)d\theta = 0.95$$

and then give a *Bayesian confidence interval* (c_1, c_2) for the parameter. If this results in the interval $(2, 5)$, the Bayesian will state that (in a Bayesian model, subject to the validity of the prior) the conditional probability given the data that the parameter falls in the interval $(2, 5)$ is 0.95. No such probability can be ascribed to a confidence interval for frequentists, who see no randomness in the parameter to which this probability statement is supposed to apply. Bayesian confidence regions are also called *credible regions* in order to make clear the distinction between the interpretation of Bayesian confidence regions and frequentist confidence regions.

Suppose $\pi(\theta|x)$ is the posterior distribution of θ given the data $x = (x_1, \ldots, x_n)$ and A is a subset of Ω. If

$$P(\theta \in A|x) = \int_A \pi(\theta|x)d\theta = p$$

then A is called a p credible region for θ. A credible region can be formed in many ways. If (a, b) is an interval such that

$$P(\theta < a|x) = \frac{1-p}{2} = P(\theta > b|x)$$

then (a, b) is called a p equal-tailed credible region. A *highest posterior density* (HPD) credible region is constructed in a manner similar to likelihood regions. The p highest posterior density credible region is given by $\{\theta : \pi(\theta|x) > c\}$, where c is chosen such that

$$p = \int_{\{\theta:\pi(\theta|x)>c\}} \pi(\theta|x)d\theta$$

A highest posterior density credible region is optimal in the sense that it is the shortest p credible interval for a given value of p.

Example Suppose $(X_1, \ldots, X_n)$ is a random sample from the $N(\mu, \sigma^2)$ distribution, where, σ^2 is known and μ has the conjugate prior. Find the $p = 0.95$ HPD credible region for μ. Compare this to a 95% confidence interval for μ.

Suppose the prior distribution for μ is $N(\mu_0, \sigma_0^2)$ so the prior density is given by

$$\pi(\mu) = C_1 \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right\}$$

and the posterior density by

$$\pi(\mu|X) = C_2 \exp\left\{-\frac{(\mu-\mu_0)^2}{2\sigma_0^2} - \sum_{i=1}^{n}\frac{(X_i-\mu)^2}{2\sigma^2}\right\}$$
$$= C_3 \exp\left\{-\frac{(\mu-\widetilde{\mu}_n)^2}{2\widetilde{\sigma}_n^2}\right\}$$

where the constants C_1, C_2 and C_3 depend on X, σ, σ_0 but not on μ and where

$$\widetilde{\mu}_n = w\overline{X} + (1-w)\mu_0$$
$$w = \frac{n\sigma_0^2}{n\sigma_0^2+\sigma^2} \text{ and}$$
$$\widetilde{\sigma}_n^2 = (\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2})^{-1}$$

Then the posterior distribution of μ is $N(\widetilde{\mu}_n, \widetilde{\sigma}_n^2)$. It follows that the 0.95 H.P.D. credible region is of the form

$$\widetilde{\mu}_n \pm 1.96\widetilde{\sigma}_n$$

Notice that as $n \to \infty$, the weight $w \to 1$ and so $\widetilde{\mu}_n$ is asymptotically equivalent to the sample mean $\overline{X}$. Similarly, as $n \to \infty$, $\widetilde{\sigma}_n^2$ is asymptotically equivalent to σ^2/n. This means that for large values of n, the H.P.D. region is close to the region

$$\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

and the latter is the 95% confidence interval for μ based on the normal distribution of the maximum likelihood estimator $\overline{X}$.

Finally, although statisticians argue whether the Bayesian or the frequentist approach is better, there is really no one right way to do statistics. There is something fundamentalist about the Bayesian paradigm, (though the Reverand Bayes was, as far as we know, far from a fundamentalist) in that it places all objects, parameters and data, in much the same context and treats them similarly. It is a coherent philosophy of statistics, and a Bayesian will vigorously argue that there is an inconsistency in regarding some unknowns as random and others as deterministic. There are certainly instances in which a Bayesian approach seems more sensible—particularly, for example, if the parameter is a measurement on a possibly randomly chosen individual (say the expected total annual claim of a client of an insurance company).

1.5 HYPOTHESIS TESTS

Statistical estimation usually concerns the estimation of the value of a parameter when we know little about it except perhaps that it lies in a given parameter space, and when we have no a priori reason to prefer one value of the parameter over another. If, however, we are asked to decide between two possible values of the parameter, the consequences of one choice of the parameter value may be quite different from another choice. For example, if we believe Y_i is normally distributed with mean $\alpha + \beta x_i$ and variance σ^2 for some explanatory variables x_i, then the value $\beta = 0$ means there is no relation between Y_i and x_i. We need neither collect the values of x_i nor build a model around them. Thus the two choices $\beta = 0$ and $\beta = 1$ are quite different in their consequences. This is often the case.

A hypothesis test involves a (usually natural) separation of the parameter space Ω into two disjoint regions, Ω_0 and $\Omega - \Omega_0$. By the difference between the two sets we mean those points in the former (Ω) that are not in the latter (Ω_0). This partition of the parameter space corresponds to testing the *null hypothesis* that the parameter is in Ω_0. We usually write this hypothesis in the form

$$H_0 : \theta \in \Omega_0$$

The null hypothesis is usually the status quo. For example, in a test of a new drug, the null hypothesis would be that the drug had no effect, or no more of an effect than drugs already on the market. The null hypothesis is rejected only if there is reasonably strong evidence against it. The *alternative hypothesis* determines what departures from the null hypothesis are anticipated. In this case, it might be simply

$$H_1 : \theta \in \Omega - \Omega_0$$

Since we do not know the true value of the parameter, we must base our decision on the observed value of X. The *hypothesis test* is conducted by determining a partition of the sample space into two sets, the *critical* or *rejection region* R and its complement $\bar{R}$, which is called the *acceptance region*. We declare that H_0 is rejected (in favor of the alternative) if we observe $x \in R$.

Definition The *power function* of a test with critical region R is the function

$$\beta(\theta) = P_\theta(X \in R)$$

or the probability that the null hypothesis is rejected as a function of the parameter.

It is obviously desirable, in order to minimize the two types of possible errors in our decision, for the power function $\beta(\theta)$ to be small for $\theta \in \Omega_0$ but large otherwise. The probability of rejecting the null hypothesis when it is true (*type I error*) is a particularly important type of error that we attempt to minimize. This probability determines one important measure of the performance of a test, the level of significance.

Definition A test has *level of significance* α if $\beta(\theta) \leq \alpha$ for all $\theta \in \Omega_0$.

The level of significance is simply an upper bound on the probability of a type I error. There is no assurance that the upper bound is tight, that is, that equality is achieved somewhere. The lowest such upper bound is often called the size of the test.

Definition The *size of a test* is equal to $\sup_{\theta \in \Omega_0} \beta(\theta)$.

Uniformly Most Powerful Tests

Tests are often constructed by specifying the size of the test, which in turn determines the probability of the type I error, and then attempting to minimize the probability that the null hypothesis is accepted when it is false (*type II error*). Equivalently, we try and maximize the power function of the test for $\theta \in \Omega - \Omega_0$.

Definition A test with power function $\beta(\theta)$ is a *uniformly most powerful* (UMP) test of size α if, for all other tests of the same size α having power function $\beta^*(\theta)$, we have $\beta(\theta) \geq \beta^*(\theta)$ for all $\theta \in \Omega - \Omega_0$.

The word "uniformly" above refers to the fact that one function dominates another, that is, $\beta(\theta) \geq \beta^*(\theta)$ uniformly for all $\theta \in \Omega - \Omega_0$. When the alternative $\Omega - \Omega_0$ consists of a single point $\{\theta_1\}$, then the construction of a best test is particularly easy. In this case, we may drop the word "uniformly" and refer to a "most powerful test." The construction of a best test, by this definition, is possible under rather special circumstances. First, we often require a *simple null hypothesis*. This is the case when Ω_0 consists of a single point $\{\theta_0\}$, and so we are testing the null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative $H_1 : \theta = \theta_1$.

Lemma B1 (Neyman-Pearson Lemma) *Let X have probability (density) function $f_\theta(x)$, $\theta \in \Omega$. Consider testing a simple null hypothesis $H_0 : \theta = \theta_0$ against a*

simple alternative $H_1 : \theta = \theta_1$. *For a constant* c, *suppose the critical region defined by*

$$R = \left\{x; \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} > c\right\}$$

corresponds to a test of size α. *Then the test with this critical region is a most powerful test of size* α *for testing* $H_0 : \theta = \theta_0$ *against* $H_1 : \theta = \theta_1$.

Proof. Consider another critical region R_1 with the same size. Then

$$P_{\theta_0}(X \in R) = P_{\theta_0}(X \in R_1) = \alpha \quad \text{or} \quad \int_R f_{\theta_0}(x)dx = \int_{R_1} f_{\theta_0}(x)dx$$

Therefore,

$$\int_{R\cap\overline{R}_1} f_{\theta_0}(x)dx + \int_{R\cap R_1} f_{\theta_0}(x)dx = \int_{R\cap R_1} f_{\theta_0}(x)dx + \int_{\overline{R}\cap R_1} f_{\theta_0}(x)dx$$

and so

$$\int_{R\cap\overline{R}_1} f_{\theta_0}(x)dx = \int_{\overline{R}\cap R_1} f_{\theta_0}(x)dx \tag{1.12}$$

For $x \in R \cap \overline{R}_1$,

$$\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} > c \quad \text{or} \quad f_{\theta_1}(x) > cf_{\theta_0}(x)$$

and thus

$$\int_{R\cap\overline{R}_1} f_{\theta_1}(x)dx > c\int_{R\cap\overline{R}_1} f_{\theta_0}(x)dx \tag{1.13}$$

For $x \in \overline{R} \cap R_1$, $f_{\theta_1}(x) \le cf_{\theta_0}(x)$, and thus

$$-\int_{\overline{R}\cap R_1} f_{\theta_1}(x)dx \ge -c\int_{\overline{R}\cap R_1} f_{\theta_0}(x)dx \tag{1.14}$$

Now

$$\begin{aligned}\beta(\theta_1) &= P_{\theta_1}(X \in R) = P_{\theta_1}(X \in R \cap R_1) + P_{\theta_1}(X \in R \cap \overline{R}_1)\\ &= \int_{R\cap R_1} f_{\theta_1}(x)dx + \int_{R\cap\overline{R}_1} f_{\theta_1}(x)dx\end{aligned}$$

and letting B_1 denote the power function of the test with critical region R_1,

$$\begin{aligned}\beta_1(\theta_1) &= P_{\theta_1}(X \in R_1)\\ &= \int_{R\cap R_1} f_{\theta_1}(x)dx + \int_{\overline{R}\cap R_1} f_{\theta_1}(x)dx\end{aligned}$$

Therefore, using (1.12), (1.13), and (1.14) we have

$$\beta(\theta_1) - \beta_1(\theta_1) = \int_{R\cap\overline{R}_1} f_{\theta_1}(x)dx - \int_{\overline{R}\cap R_1} f_{\theta_1}(x)dx$$

$$\geq c\int_{R\cap\overline{R}_1} f_{\theta_0}(x)dx - c\int_{\overline{R}\cap R_1} f_{\theta_0}(x)dx$$

$$= c[\int_{R\cap\overline{R}_1} f_{\theta_0}(x)dx - \int_{\overline{R}\cap R_1} f_{\theta_0}(x)dx] = 0$$

and the test with critical region R is therefore the most powerful.

Example Suppose we anticipate collecting daily returns from the past n days of a stock, $(X_1, \ldots, X_n)$ assumed to be distributed according to a normal$(\mu\Delta, \sigma^2\Delta)$ distribution. Here Δ is the length of a day measured in years, $\Delta \simeq 1/252$, and μ, σ^2 are the annual drift and volatility parameters. We wish to test whether the stock has zero or positive drift, so we wish to test the hypothesis $H_0 : \mu = 0$ against the alternative $H_1 : \mu > 0$ at level of significance α. We want the probability of the incorrect decision when the drift is 20% per year to be small, so let us choose it to be α as well, which means that when $\mu = 0.2$, the power of the test should be at least $1 - \alpha$. How large a sample must be taken to ensure this?

The test itself is easy to express. We reject the null hypothesis if

$$\left(\frac{n}{\Delta}\right)^{\frac{1}{2}} \frac{\overline{X}}{\sigma} > z_\alpha$$

where the value z_α has been chosen so that $P(Z > z_\alpha) = \alpha$ when Z has a standard normal distribution. The power of the test is the probability

$$P\left[\left(\frac{n}{\Delta}\right)^{\frac{1}{2}} \frac{\overline{X}}{\sigma} > z_\alpha\right]$$

when the parameter $\mu_1 = 0.2$, and this is

$$P\left[\left(\frac{n}{\Delta}\right)^{\frac{1}{2}} \frac{\overline{X} - \mu_1\Delta}{\sigma} > z_\alpha - \frac{\mu_1}{\sigma}(n\Delta)^{1/2}\right] = P\left[Z > z_\alpha - \frac{\sqrt{n}\mu_1\Delta^{1/2}}{\sigma}\right]$$

where Z has a standard normal distribution. Since we want the power to be $1 - \alpha$, the value

$$z_\alpha - \frac{\mu_1}{\sigma}(n\Delta)^{1/2}$$

must be chosen to be $-z_\alpha$. Solving for the value of n,

$$z_\alpha - \frac{\mu_1}{\sigma}(n\Delta)^{1/2} = -z_\alpha$$

$$n = \frac{4\sigma^2 z_\alpha^2}{\mu_1^2 \Delta}$$

Now if we try some reasonable values for the parameters, for example, $\sigma^2 = 0.2$, $\Delta = 1/252$, $\mu_1 = 0.2$, $\alpha = 0.05$, then $n \simeq 14{,}000$, which is about 55 years worth of data, far larger a sample than we could hope to collect. This example shows that the typical variabilities in the market are so large, compared with even fairly high rates of return, that it is almost impossible to distinguish between theoretical rates of return of 0% and 20% per annum using a hypothesis test with daily data.

Relationship betweeen Hypothesis Tests and Confidence Intervals

There is a close relationship between hypothesis tests and confidence intervals, as the following example illustrates. Suppose $(X_1, \ldots, X_n)$ is a random sample from the N(θ, 1) distribution and we wish to test the hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. The critical region $R = \{x; |\bar{x} - \theta_0| > 1.96/\sqrt{n}\}$ is a size $\alpha = 0.05$ critical region that has a corresponding acceptance region $\overline{R} = \{x; |\bar{x} - \theta_0| \leq 1.96/\sqrt{n}\}$. Note that the hypothesis $H_0 : \theta = \theta_0$ would not be rejected at the 0.05 level if $|\bar{x} - \theta_0| \leq 1.96/\sqrt{n}$, or equivalently

$$\bar{x} - 1.96/\sqrt{n} < \theta_0 < \bar{x} + 1.96/\sqrt{n}$$

which is a 95% C.I. for θ.

Problem Let $(X_1, \ldots, X_5)$ be a random sample from the gamma$(2, \theta)$ distribution. Show that

$$R = \left\{ x; \sum_{i=1}^{5} x_i < 4.7955\theta_0 \text{ or } \sum_{i=1}^{5} x_i > 17.085\theta_0 \right\}$$

is a size $\alpha = 0.05$ critical region for testing $H_0 : \theta = \theta_0$. Show how this critical region may be used to construct a 95% C.I. for θ.

Likelihood Ratio Tests

Consider a test of the hypothesis $H_0 : \theta \in \Omega_0$ against $H_1 : \theta \in \Omega - \Omega_0$. We have seen that for prescribed $\theta_0 \in \Omega_0$, $\theta_1 \in \Omega - \Omega_0$, the most powerful test of the

simple null hypothesis $H_0: \theta = \theta_0$ against a simple alternative $H_1: \theta = \theta_1$ is based on the likelihood ratio $f_{\theta_1}(x)/f_{\theta_0}(x)$. By the Neyman-Pearson lemma it has critical region

$$R = \left\{x;\ \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} > c\right\}$$

where c is a constant determined by the size of the test. When either the null or the alternative hypothesis is *composite* (i.e., contains more than one point) and there is no uniformly most powerful test, it seems reasonable to use a test with critical region R for some choice of θ_1, θ_0. The *likelihood ratio test* does this with θ_1 replaced by $\hat{\theta}$, the maximum-likelihood estimator over all possible values of the parameter, and θ_0 replaced by the maximum-likelihood estimator of the parameter when it is restricted to Ω_0. Thus, the likelihood ratio test has critical region $R = \{x; \Lambda(x) > c\}$, where

$$\Lambda(x) = \frac{\sup_{\theta\in\Omega} f_\theta(x)}{\sup_{\theta\in\Omega_0} f_\theta(x)} = \frac{\sup_{\theta\in\Omega} L(\theta)}{\sup_{\theta\in\Omega_0} L(\theta)}$$

and c is determined by the size of the test. In general, the distribution of the test statistic $\Lambda(X)$ may be difficult to find. Fortunately, however, the asymptotic distribution is known under fairly general conditions. In a few cases, we can show that the likelihood ratio test is equivalent to the use of a statistic with known distribution. However, in many cases, we need to rely on the asymptotic chi-squared distribution of Theorem B6.

Example Let $(X_1, \ldots, X_n)$ be a random sample from the $N(\mu, \sigma^2)$ distribution, where μ and σ^2 are unknown. Consider a test of

$$H_0: \mu = 0,\ 0 < \sigma^2 < \infty$$

against the alternative

$$H_1: \mu \neq 0,\ 0 < \sigma^2 < \infty$$

We can show that the likelihood ratio test of H_0 against H_1 has critical region $R = \{x; n\bar{x}^2/s^2 > c\}$. Under H_0 the statistic $T = n\overline{X}^2/S^2$ has a $F(1, n-1)$ distribution and we can thus find a size $\alpha = 0.05$ test for $n = 20$.

Theorem B6 *Suppose $(X_1, \ldots, X_n)$ is a random sample from a regular statistical model $\{f_\theta(x); \theta \in \Omega\}$ with Ω an open set in k-dimensional Euclidean space. Consider a subset of Ω defined by $\Omega_0 = \{\theta(\eta); \eta \in$ open subset of q-dimensional Euclidean space$\}$.Then the likelihood ratio statistic defined by*

$$\Lambda_n(X) = \frac{\sup_{\theta\in\Omega} \prod_i^n f_\theta(X_i)}{\sup_{\theta\in\Omega_0} \prod_i^n f_\theta(X_i)} = \frac{\sup_{\theta\in\Omega} L(\theta)}{\sup_{\theta\in\Omega_0} L(\theta)}$$

is such that, under the hypothesis $H_0 : \theta \in \Omega_0$,

$$2 \log \Lambda_n(X) \rightarrow_D W \backsim \chi^2(k - q)$$

Note: The number of degrees of freedom is the difference between the number of parameters that need to be estimated in the general model and the number of parameters left to be estimated under the restrictions imposed by H_0.

Significance Tests and p-Values

A hypothesis test is a rule that allows us to decide whether to accept the null hypothesis H_0 or to reject it in favor of the alternative hypothesis H_1 based on the observed data. In situations in which H_1 is difficult to specify, a test of significance could be used. A (pure) test of significance is a procedure for measuring the strength of the evidence provided by the observed data against H_0. This method usually involves looking at the distribution of a test statisitic or discrepancy measure T under H_0. The *p-value* or *significance level* for the test is the probability, computed under H_0, of observing a T value at least as extreme as the value observed. The smaller the observed p-value, the stronger the evidence against H_0. The difficulty with this approach is finding statistic with "good properties." The likelihood ratio statistic provides a general test statistic that may be used.

Score and Wald Tests

Score Test Score tests can be viewed as a more general class of tests of $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Omega - \{\theta_0\}$, which tend to have considerable power provided that the values of the parameter under the null and the alternative hypotheses are close. If the usual regularity conditions hold, then under $H_0 : \theta = \theta_0$ we have

$$S(\theta_0; X)[J(\theta_0)]^{-1/2} \rightarrow_D Z \backsim \mathrm{N}(0, 1)$$

and thus, the square

$$R(\theta_0; X) = [S(\theta_0; X)]^2 [J(\theta_0)]^{-1} \rightarrow_D W \backsim \chi^2(1)$$

For a vector $\theta = (\theta_1, \ldots, \theta_k)^t$, we have a similar result,

$$R(\theta_0; X) = [S(\theta_0; X)]^t [J(\theta_0)]^{-1} S(\theta_0; X) \rightarrow_D W \backsim \chi^2(k)$$

The test based on $R(\theta_0; X)$ is called a (Rao) score test. It has critical region

$$R = \{x;\ R(\theta_0; x) > c\}$$

where c is determined by the size of the test; that is, c satisfies $P(W > c) = \alpha$, where $W \backsim \chi^2(k)$. The test based on $R(\theta_0; X)$ is asymptotically equivalent to the likelihood ratio test.

Wald Test Suppose that $\hat{\theta}$ is the maximum-likelihood estimator of θ over all $\theta \in \Omega$ and we wish to test $H_0 : \theta = \theta_0$ against $H_1 : \theta \in \Omega - \{\theta_0\}$. If the usual regularity conditions hold, then under $H_0 : \theta = \theta_0$

$$W(\theta_0; X) = (\hat{\theta} - \theta_0)^t J(\theta_0)(\hat{\theta} - \theta_0) \rightarrow_D W \backsim \chi^2(k)$$

A test based on the test statistic $W(\theta_0; X)$ is called a Wald test. It has critical region

$$R = \{x;\ W(\theta_0; x) > c\}$$

where c is determined by the size of the test. Both the score test and the Wald test are asymptotically equivalent to the likelihood ratio test and the intuitive expanation for these equivalences are quite simple. For large values of the sample size n, the maximum likelihood estimator $\widehat{\theta}_n$ is close to the true value of the parameter θ_0 and so the log likelihood can be approximated by the first two terms in the Taylor series expansion of $\ell(\theta) = \log L(\theta)$ about $\widehat{\theta}_n$, and so

$$\begin{aligned} 2 \log \Lambda_n(X) &= 2\{\ell(\widehat{\theta}_n) - \ell(\theta_0)\} \\ &\simeq 2\left\{(\widehat{\theta}_n - \theta_0)' S(\widehat{\theta}_n; X) + \frac{1}{2}(\widehat{\theta}_n - \theta_0)' I(\widehat{\theta}_n)(\widehat{\theta}_n - \theta_0)\right\} \\ &\simeq (\widehat{\theta}_n - \theta_0)' J(\theta_0)(\widehat{\theta}_n - \theta_0) \end{aligned}$$

since

$$S(\widehat{\theta}_n; X) = 0$$

and the observed information $I(\widehat{\theta}_n)$ is asymptotically equivalent to the Fisher information $J(\theta_0)$. This verifies the equivalence of the likelihood ratio and the Wald test. $J(\theta_0)$ may be replaced by $I(\hat{\theta})$ to give an asymptotically equivalent test statistic.