Chapter 6. ESTIMATION

1. Desirable Properties of Estimators

Consider data **x** that comes from a DGP which has a density $f(\mathbf{x},\theta_o)$. In most initial applications, we will think of **x** as a simple random sample of size n, $\mathbf{x} = (x_1,...,x_n)$ drawn from a population in which x has a density $f(x,\theta_o)$, so that the DGP density is $f(\mathbf{x},\theta) = f(x_1,\theta_o)\cdot...\cdot f(x_n,\theta_o)$. However, the notation $f(\mathbf{x},\theta_o)$ can also cover more complicated DGP, such as time-series data sets. Suppose that θ_o is unknown, but one knows that this DGP is contained in a family with densities $f(\mathbf{x},\theta)$ indexed by θ. Let $\mathcal{X}$ denote the domain of **x**, and Θ denote the domain of θ. In the case of a simple random sample where an observation x is a point in a space $\mathbb{X}$, one has $\mathcal{X} = \mathbb{X}^n$. The statistical inference task is to estimate θ_o. In Chapter 5, we saw that an estimator $T(\mathbf{x})$ of θ_o was desirable from a Bayesian point of view if $T(\cdot)$ minimized the expected cost of mistakes. For typical cost functions where the larger the mistake, the larger the cost, Bayes estimators will try to get "close" to the true parameter value. That is, the Bayes procedure will seek estimators whose probability densities are concentrated tightly around the true θ_o. Classical statistical procedures lack the expected cost criterion for choosing estimators, but also seek estimators whose probability densities are concentrated around the true θ_o.

Listed below are some of the properties that are deemed desirable for classical estimators. Classical statistics often proceeds by developing some candidate estimators, and then using some of these properties to choose among the candidates. It is often not possible to achieve all of these properties at the same time, and sometimes they can even be incompatible. Some of the properties are defined relative to a *class* of candidate estimators, a set of possible $T(\cdot)$ that we will denote by $\mathcal{T}$. The density of an estimator T(·) will be denoted $\psi(t,\theta_o)$, or when it is necessary to index the estimator, $\psi_T(t,\theta_o)$.

Sufficiency. Suppose there is a one-to-one transformation from the data **x** into variables (**y**,**z**).[1] Then the DGP density f(**x**,θ) can be described in terms of the density of (**y**,**z**), which we might denote g(**y**,**z**,θ) and write as the product of the marginal density of **y** and the conditional density of **z** given **y**, $g(\mathbf{y},\mathbf{z},\theta) = g_1(\mathbf{y},\theta)\cdot g_2(\mathbf{z}|\mathbf{y},\theta)$.[2] Note that in general both the marginal and the conditional densities depend on θ. The variables **y** are said to be *sufficient* for θ if the conditional distribution of **z** given **y** is independent of θ; i.e., $g_2(\mathbf{z}|\mathbf{y},\theta) = g_2(\mathbf{z}|\mathbf{y})$. In this case, all of the information in the sample about θ is summarized in **y**, and once you know **y**, knowing **z** tells you nothing more about θ. (One way to convince yourself of this is to form the posterior density of θ, given **y** and **z**, for any prior. You will find that this posterior density, which is a complete description of what you believe about θ, does not depend on **z** when **y** is sufficient.) Sufficiency of **y** is equivalent to the factorization $g(\mathbf{y},\mathbf{z},\theta) = g_1(\mathbf{y},\theta)\cdot g_2(\mathbf{z}|\mathbf{y})$ of the density into one term depending only on **y** and θ and a second term depending only on **z** and **y**. This characterization is useful for identifying sufficient statistics.

An implication of sufficiency is that there is no reason to consider estimators T(**x**) that depend on **x** except through the sufficient statistics. Then, the

---

[1] This is a known transformation, so it cannot depend on unknown θ.

[2] The relationship of the density f(**x**,θ) and the density g(**y**,**z**,θ) comes from the rules for transforming random variables; see Chapter 3.6. Let **x** = **x**(**y**,**z**) denote the inverse of the one-to-one transformation from **x** to **y** and **z**, and let J denote the *Jacobian* of this mapping; i.e., the determinant of the array of derivatives of **x**(**y**,**z**) with respect to its arguments, signed so that it is positive. Then g(**y**,**z**,θ) = f(**x**(**y**,**z**))·J. The Jacobian J does not depend on θ, so g(**y**,**z**,θ) factors into a term depending only on **y** and θ and a term independent of θ if and only if f(**x**(**y**,**z**)) factors in the same way.

search for a good estimator can be restricted to estimators T(**y**) that depend only on sufficient statistics **y**. This observation is most useful when **y** is of low dimension. In some problems, only the full sample **x** is a sufficient statistic, and you obtain no useful restriction. In others there may be many different transformations of **x** into (**y**,**z**) for which **y** is sufficient. Then, among the alternative sufficient statistics, you will want to choose a **y** that is a *minimal sufficient statistic*. This will be the case if there is no further one-to-one transformation of **y** into variables (**u**,**v**) such that **u** is sufficient for θ and of lower dimension than **y**.

An example shows how sufficiency works. Suppose one has a simple random sample $\mathbf{x} = (x_1,...,x_n)$ from an exponential distribution with an unknown scale parameter λ. The DGP density is the product of univariate exponential densities,

$$f(\mathbf{x},\lambda) = (\lambda \cdot e^{-\lambda x_1}) \cdot ... \cdot (\lambda \cdot e^{-\lambda x_n}) = \lambda^n e^{-\lambda(x_1+...+x_n)} .$$

Make the one-to-one transformation $y = x_1 + ... + x_n$, $z_1 = x_1,..., z_{n-1} = x_{n-1}$, and note that the inverse transformation implies $x_n = y - z_1 - ... - z_{n-1}$. Substitute the inverse transformation into f to obtain $f(\mathbf{x}(y,\mathbf{z})) = \lambda^n \cdot e^{-\lambda y}$. then, f factors into a gamma density $\lambda^n y^{n-1} \cdot e^{-\lambda y}/(n-1)!$ that involves only λ and y, and not z, and a term $(n-1)!/y^{n-1}$ which does not depend on λ and is a trivial function of **z**. Then, y is a sufficient statistic for λ, and one need consider only estimators for λ that are functions of the univariate statistic $y = x_1 + ... + x_n$. In this case, y is a minimal sufficient statistic.

In this exponential example, there are other sufficient statistics that are not minimal. For example, any **y** whose components can be transformed to recover the sum of the x's is sufficient. Knowing only that one can restrict the search for an estimator to functions of such a **y** is not as useful as knowing that one only needs to look at functions of a sum of the x's.

Ancillarity. As in the discussion of sufficiency, suppose there is a one-to-one transformation from the data **x** into variables (**y**,**z**). Then the DGP density can be written the product of the marginal density of **y** and the conditional density of **z** given **y**, $g_1(\mathbf{y},\theta)\cdot g_2(\mathbf{z}|\mathbf{y},\theta)$. Both g_1 and g_2 depend in general on θ. The data **y** are *ancillary* to θ if g_1 does not depend on θ. In this case, all the information about θ that is contained in the data is contained in the conditional distribution of **z** given **y**. This implies that the search for an estimator for θ can be confined to ones derived from the conditional density of **z** given **y**. Ancillarity provides useful restrictions when $g_2(\mathbf{z}|\mathbf{y},\theta)$ depends only on a low-dimensional part of **y**, or when this density is independent of unknown nuisance parameters that enter the marginal density of **y**.

An example where ancillarity is useful arises in data $\mathbf{x} = (x_1,...,x_n)$ where the x_i are independent observations from an exponential density $\lambda\cdot e^{-\lambda x_i}$ and the sample size n is random with a Poisson density $\gamma^{n-1}\cdot e^{-\gamma}/(n-1)!$ for n = 1,2,.... The DGP density is then $\lambda^n\cdot e^{-\lambda(x_1+...+x_n)}\cdot\gamma^{n-1}\cdot e^{-\gamma}/(n-1)!$. This density factors into the density $\lambda^n y^{n-1}\cdot e^{-\lambda y}$, with $y = x_1 +...+ x_n$, that is now the conditional density of y given n, times a term that is a function of n, y, and γ, but not of λ. Then, the principle of ancillarity says that to make inferences on λ, one should condition on n and not be concerned with the nuisance parameter γ that enters the marginal density of n.

Admissibility. An estimator T(·) from a class of estimators $\mathcal{T}$ is *admissible relative to* $\mathcal{T}$ if there is no second estimator T′(·) in $\mathcal{T}$ with the property that for all θ_o, $\mathbf{E}_{\mathbf{x}|\theta_o}(T'(\mathbf{x}) - \theta_o)^2 \le \mathbf{E}_{\mathbf{x}|\theta_o}(T(\mathbf{x}) - \theta_o)^2$, with inequality strict for at least one θ_o. This is the same as the definition of admissibility in statistical decision theory, but with the cost of a mistake defined as *mean squared error* (MSE), the square of the difference between the estimate and the true value of θ. An inadmissible estimator is undesirable because there is an identified alternative estimator that is more closely clustered around the true parameter value. A limitation of admissibility is that there will often be many admissible estimators, and this criterion does not choose between them.

Unbiasedness. An estimator $T(\cdot)$ is *unbiased* for θ_o if $\mathbf{E}_{\mathbf{x}|\theta_o}T(\mathbf{x}) \equiv \theta_o$ for all θ_o; i.e., $\theta \equiv \int_{-\infty}^{+\infty} T(\mathbf{x})f(\mathbf{x}|\theta)d\mathbf{x}$. An estimator with this property is "centered" around the true parameter value, and will not systematically be too high or too low.

Efficiency. An estimator $T(\cdot)$ is *efficient relative to* an estimator $T'(\cdot)$ if for all θ_o, $\mathbf{E}_{\mathbf{x}|\theta_o}(T(\mathbf{x}) - \theta_o)^2 \le \mathbf{E}_{\mathbf{x}|\theta_o}(T'(\mathbf{x}) - \theta_o)^2$. The estimator $T(\cdot)$ is efficient relative to a class of estimators $\mathcal{T}$ if it is efficient relative to $T'(\cdot)$ for all $T'(\cdot)$ in $\mathcal{T}$. An efficient estimator provides estimates that are most closely clustered around the true value of θ, by the squared distance measure, among all the estimators in $\mathcal{T}$. The limitation of efficiency is that for many problems and classes of estimators $\mathcal{T}$, there will be no efficient estimator, in that one cannot satisfy the required inequality uniformly for all θ_o. The following theorem establishes an efficiency result for estimators that are functions of sufficient statistics:

Blackwell Theorem. If $T'(\cdot)$ is any estimator of θ_o from data $\mathbf{x}$, and $\mathbf{y}$ is a sufficient statistic, then there exists an estimator $T(\cdot)$ that is a function solely of the sufficient statistic and that is efficient relative to $T'(\cdot)$. If $T'(\cdot)$ is unbiased, then so is $T(\cdot)$. If an unbiased estimator $T(\cdot)$ is uncorrelated with every unbiased estimator of zero, then T(·) has a smaller variance than any other unbiased estimator.

Proof: Make a one-to-one transformation of the data $\mathbf{x}$ into $(\mathbf{y},\mathbf{z})$, where $\mathbf{y}$ is the sufficient statistic, and let $g_1(\mathbf{y},\theta)\cdot g_2(\mathbf{z}|\mathbf{y})$ denote the DGP density. Define $T(\mathbf{y}) = \mathbf{E}_{\mathbf{z}|\mathbf{y}}T'(\mathbf{y},\mathbf{z})$. Write $T'(\mathbf{y},\mathbf{z}) - \theta_o = T'(\mathbf{y},\mathbf{z}) - T(\mathbf{y}) + T(\mathbf{y}) - \theta_o$. Then

$$\begin{aligned}\mathbf{E}(T'(\mathbf{y},\mathbf{z}) - \theta_o)^2 &= \mathbf{E}(T'(\mathbf{y},\mathbf{z}) - T(\mathbf{y}))^2 + \mathbf{E}(T(\mathbf{y}) - \theta_o)^2 \\ &\quad + 2\cdot\mathbf{E}(T(\mathbf{y}) - \theta_o)\cdot(T'(\mathbf{y},\mathbf{z}) - T(\mathbf{y})) \ .\end{aligned}$$

But the last term satisfies

$$2\cdot\mathbf{E}(T(\mathbf{y}) - \theta_o)\cdot(T'(\mathbf{y},\mathbf{z}) - T(\mathbf{y})) = 2\cdot\mathbf{E}_{\mathbf{y}}(T(\mathbf{y}) - \theta_o)\cdot\mathbf{E}_{\mathbf{z}|\mathbf{y}}(T'(\mathbf{y},\mathbf{z}) - T(\mathbf{y})) = 0 \ .$$

Therefore, $\mathbf{E}(T'(\mathbf{y},\mathbf{z}) - \theta_o)^2 \ge \mathbf{E}(T(\mathbf{y}) - \theta_o)^2$.

If $T'(\mathbf{y},\mathbf{z})$ is unbiased, then $\mathbf{E}T(\mathbf{y}) = \mathbf{E}_{\mathbf{y}}\mathbf{E}_{\mathbf{z}|\mathbf{y}}T'(\mathbf{y},\mathbf{z}) = \theta_o$, and T(·) is also unbiased. Finally, if T(·) is uncorrelated with any unbiased estimator of zero, and T′(·) is any other unbiased estimator, this implies $\mathbf{E}T(\mathbf{x})\cdot(T'(\mathbf{x}) - T(\mathbf{x})) = 0$. Therefore,

$$\begin{aligned}\mathbf{E}(T'(\mathbf{x}) - \theta_o)^2 &= \mathbf{E}(T'(\mathbf{x}) - T(\mathbf{x}) + T(\mathbf{x}) - \theta_o)^2 \\ &= \mathbf{E}(T'(\mathbf{x}) - T(\mathbf{x}))^2 + \mathbf{E}(T(\mathbf{x}) - \theta_o)^2 + 2\cdot\mathbf{E}T(\mathbf{x})\cdot(T'(\mathbf{x}) - T(\mathbf{x})) \\ &= \mathbf{E}(T'(\mathbf{x}) - T(\mathbf{x}))^2 + \mathbf{E}(T(\mathbf{x}) - \theta_o)^2 > \mathbf{E}(T(\mathbf{x}) - \theta_o)^2 \ .\end{aligned}$$

Thus, T′ has a larger variance than T. □

If $\mathcal{T}$ is a class of unbiased estimators, so that $\mathbf{E}_{\mathbf{x}|\theta_o}T'(\mathbf{x}) \equiv \theta_o$ for every estimator T′(·) in this class, then the efficiency criterion is the variance of the estimator, and an efficient estimator is a *minimum variance unbiased estimator* (MVUE).

There are many problems for which no MVUE estimator exists. We next give a lower bound on the variance of an unbiased estimator. If a candidate satisfies this bound, then we can be sure that it is MVUE. However, the converse is not true: There may be a MVUE, its variance may still be larger than this lower bound; i.e., the lower bound may be unobtainable.

Cramer-Rao Bound. Suppose a simple random sample $\mathbf{x} = (x_1,...,x_N)$ with $f(x,\theta_o)$ the density of an observation x. Assume that log $f(x,\theta_o)$ is twice continuously differentiable in θ, and that this function and its derivatives are bounded in magnitude by a function that is independent of θ and has a finite integral in x. Suppose an estimator $T(\mathbf{x})$ has $\mathbf{E}_{\mathbf{x}|\theta}T(\mathbf{x}) \equiv \theta + \mu(\theta)$, and that the bias $\mu(\theta)$ is differentiable. Then, the variance of $T(\mathbf{x})$ satisfies

$$\mathbf{V}_{\mathbf{x}|\theta}(T(\mathbf{x})) \geq (1 + \mu'(\theta))^2/n\cdot\mathbf{E}_{\mathbf{x}|\theta}\,[\nabla_\theta \log f(x,\theta_o)]^2 .$$

If the estimator is unbiased, so $\mu(\theta) \equiv 0$, this bound is

$$\mathbf{V}_{\mathbf{x}|\theta}(T(\mathbf{X})) \geq 1/n\cdot\mathbf{E}_{\mathbf{x}|\theta}[\nabla_\theta \log f(x,\theta_o)]^2 .$$

The expression $\mathbf{E}_{\mathbf{x}|\theta}[\nabla_\theta \log f(x,\theta_o)]^2$ is termed the *Fisher information* contained in an observation; then, the Cramer-Rao bound states that the variance of an unbiased estimator is at least as large as the reciprocal of the Fisher information in the sample. To demonstrate this result, let $L(\mathbf{x},\theta) = \sum_{i=1}^{n} \log f(x_i,\theta)$, so that the DGP density is $f(\mathbf{x},\theta) = e^{L(\mathbf{x},\theta)}$. By construction,

$$1 \equiv \int_{-\infty}^{+\infty} e^{L(\mathbf{x},\theta)}d\mathbf{x} \qquad \text{and} \qquad \theta + \mu(\theta) \equiv \int_{-\infty}^{+\infty} T(\mathbf{x})\cdot e^{L(\mathbf{x},\theta)}d\mathbf{x}.$$

Differentiate each integral with respect to θ to get

$$0 \equiv \int_{-\infty}^{+\infty} \nabla_\theta L(\mathbf{x},\theta)\cdot e^{L(\mathbf{x},\theta)}d\mathbf{x} \qquad \text{and} \qquad 1 + \mu'(\theta) \equiv \int_{-\infty}^{+\infty} T(\mathbf{x})\cdot\nabla_\theta L(\mathbf{x},\theta)\cdot e^{L(\mathbf{x},\theta)}d\mathbf{x} .$$

Combine these to get an expression for the covariance of T and $\nabla_\theta L$,

$$1 + \mu'(\theta) \equiv \int_{-\infty}^{+\infty} [T(\mathbf{x}) - \theta]\cdot\nabla_\theta L(\mathbf{x},\theta)\cdot e^{L(\mathbf{x},\theta)}d\mathbf{x} \ .$$

Now, any covariance has the property that its square is no greater than the product of the variances of its terms. This is called the *Cauchy-Schwartz inequality*. In this case, the inequality can be written

$$\left(1 + \mu'(\theta)\right)^2 = \left[\int_{-\infty}^{+\infty} [T(\mathbf{x}) - \theta]\cdot\nabla_\theta L(\mathbf{x},\theta)\cdot e^{L(\mathbf{x},\theta)}d\mathbf{x}\right]^2 \le \mathbf{V}_{\mathbf{x}|\theta}(T(\mathbf{x}))\cdot\mathbf{E}_{\mathbf{x}|\theta}[\nabla_\theta L(\mathbf{x},\theta)]^2 \ .$$

Dividing both sides by the Fisher information in the sample, which is simply the variance of the sample score, $\mathbf{E}_{\mathbf{x}|\theta}\,[\nabla_\theta L(\mathbf{x},\theta)]$, gives the Cramer-Rao bound. □

Invariance. In some conditions, one would expect that a change in a problem should not alter an estimate of a parameter, or should alter it in a specific way. Generically, these are called invariance properties of an estimator. For example, when estimating a parameter from data obtained by a simple random sample, the estimate should not depend on the indexing of the observations in the sample; i.e., $T(x_1,...,x_n)$ should be *invariant under permutations of the observations*.

Sometimes a parameter enters a DGP in such a way that there is a simple relationship between shifts in the parameter and the shifts one would expect to observe in the data. For example, suppose the density of an observation is of the form $f(x_i|\theta) \equiv h(x_i-\theta)$; in this case, θ is called a location parameter. If the true value of θ shifts up by an amount Δ, one would expect observations on average to shift up by an amount Δ. If $T(x_1,...,x_n)$ is an estimator of θ_o in this problem, a reasonable property to impose on $T(\cdot)$ is that $T(x_1+\Delta,...,x_n+\Delta) = T(x_1,...,x_n) + \Delta$.

In this case, $T(\cdot)$ is *invariant with respect to location*. For this problem, one can restrict attention to estimators with this invariance property.

Another example is *invariance with respect to scale*. Suppose the density of an observation has the form $f(x_i|\theta) \equiv \theta \cdot h(\theta x_i)$. Then θ is called a scale parameter. If θ is reduced by a proportion λ, one would expect observations on average to be scaled up by λ. The corresponding invariance property on an estimator $T(\cdot)$ is that $T(\lambda \cdot x_1,...,\lambda \cdot x_n) = T(x_1,...,x_n)/\lambda$.

To illustrate the use of invariance conditions, consider the example of a simple random sample $\mathbf{x} = (x_1,...,x_n)$ from an exponential distribution with an unknown scale parameter λ, with the DGP density $f(\mathbf{x},\lambda) = \lambda^n e^{-\lambda(x_1+...+x_n)}$. Then $y = x_1 + ... + x_n$ is sufficient and we need consider only estimators $T(y)$. Invariance with respect to scale implies $T(y) = T(1)/y$, and the scale-invariant estimator of λ must be inversely proportional to y.

The next group of properties refer to the limiting behavior of estimators in a sequence of larger and larger samples, and are sometimes called *asymptotic properties*. The rationale for employing these properties is that when one is working with a large sample, then properties that hold in the limit will also hold, approximately, for this sample. The reason for considering such properties at all, rather than concentrating on the sample you actually have, is that one can use these approximate properties to choose among estimators in situations where the exact finite sample property cannot be imposed or is analytically intractable to work out.

Application of asymptotic properties raises several conceptual and technical issues. The first question is what it would mean to increase sample size indefinitely, and whether various methods that might be used to define this limit correspond to approximations that are likely to be relevant to a specific problem. There is no ambiguity when one is drawing simple random samples from an infinite population. However, if one samples from a finite population, a finite sequence of samples of increasing size will terminate in a complete census of the population. While one could imagine sampling with replacement and drawing samples that are larger than the population, it is not obvious why estimators that have some reasonable properties in this limit are necessarily appropriate for the finite population. Put another way, it is not obvious that this limit provides a good approximation to the finite sample. The issue of the appropriate asymptotic limit is particularly acute for time series. One can imagine extending observations indefinitely through time. This may provide approximations that are appropriate in some situations for some purposes, but not for others. For example, if one is trying to estimate the timing of a particular event, a local feature of the time series, it is questionable that extending the time series indefinitely into the past and future leads to a good approximation to the statistical properties of the estimator of the time of an event. Other ways of thinking of increasing sample sizes for time series, such as sampling from more and more "parallel" universes, or sampling at shorter and shorter intervals, have their own idiosyncrasies that make them questionable as useful approximations. The second major issue is how the sequence of estimators associated with various sample sizes is defined. A conceptualization introduced in Chapter 5 defines an estimator to be a functional of the empirical CDF of the data, $T(F_n)$. Then, it is natural to think of $T(F(\cdot,\theta_o))$ as the limit of this sequence of estimators, and the Glivenko-Cantelli theorem stated in Chapter 5.1 establishes an approximation property that the estimator $T(F_n)$ converges almost surely to $T(F(\cdot,\theta_o))$, as long as $T(\cdot)$ satisfies a continuity property at $F(\cdot,\theta_o)$. It is particularly important to avoid reliance on asymptotic arguments when is is clear that the asymptotic approximation is irrelevant to the behavior of the estimator in the range of sample sizes actually encountered. Consider an estimation procedure which says "Ignore the data and estimate θ_o to be zero in all samples of size less than 10 billion, and for larger samples employ some computationally complex but statistically sound estimator". This procedure may technically have good asymptotic properties, but this approximation obviously tells you nothing about the behavior of the estimator in economic sample sizes of a few thousand observations.

Consistency. A sequence of estimators $T_n(\mathbf{x}) = T_n(x_1,...,x_n)$ for samples of size n are *consistent* for θ_o if the probability that they are more than a distance $\varepsilon > 0$ from θ_o goes to zero as n increases; i.e., $\lim_{n\to\infty} P(|T_n(x_1,...,x_n) - \theta_o| > \varepsilon) = 0$. In the terminology of Chapter 4, this is *weak convergence* or *convergence in probability,* written $T_n(x_1,...,x_n) \xrightarrow{p} \theta_o$. One can also talk about *strong consistency*, which holds when $\lim_{n\to\infty} P(\sup_{n'\geq n} |T_{n'}(x_1,...,x_{n'}) - \theta_o| > \varepsilon) = 0$, and corresponds to almost sure convergence, $T_n(x_1,...,x_n) \xrightarrow{as} \theta_o$.

Asymptotic Normality. A sequence of estimators $T_n(\cdot)$ for samples of size n are *consistent asymptotically normal* (CAN) for θ_o if there exists a sequence r_n of scaling constants such that $r_n \to +\infty$ and $r_n\cdot(T_n(\mathbf{x}_n) - \theta_o)$ converges in distribution to a normally distributed random variable with some mean $\mu = \mu(\theta_o)$ and variance $\sigma^2 = \sigma(\theta_o)^2$.[3] The mean μ is termed the *asymptotic bias*, and σ^2 is termed the

[3] If $\Psi_n(t)$ is the CDF of $T_n(\mathbf{x}_n)$, then the CDF of $Q_n = r_n\cdot(T_n(\mathbf{x}_n) - \theta_o)$ is $\Psi_n(\theta_o + q/r_n)$. From Chapter 4, $r_n(T_n(\mathbf{x}_n) - \theta_o) \xrightarrow{d} Z$ with $Z \sim N(\mu,\sigma^2)$ if for each q, the CDF of Q_n satisfies $\lim_{n\to\infty} |\Psi_n(\theta_o + q/r_n) - \Phi((t-\mu)/\sigma)| = 0$. This is the conventional definition of convergence in distribution, with the continuity of the

asymptotic variance. If $\mu = 0$, the estimator is said to be *asymptotically unbiased*. Often, when a sequence of estimators is said to be asymptotically normal, asymptotic unbiasedness is taken to be part of the definition unless stated explicitly to the contrary. The scaling term r_n can be taken to be $\sqrt{n}$ in almost all finite-parameter problems, and unless it is stated otherwise, you can assume that this is the scaling that is being used. When it is important to make this distinction clear, one can speak of *Root-N consistent asymptotically normal* (RCAN) sequences of estimators.

Convergence in distribution to a normal is a condition that holds pointwise for each θ_o. One could strengthen the property by requiring that this convergence be uniform in θ_o; i.e., by requiring for each $\varepsilon > 0$ and q that there be a sample size $n(\varepsilon,q)$ beyond which $\sup_{\theta_o} |\Psi(\theta_o + q/r_n) - \Phi((q-\mu(\theta_o))/\sigma(\theta_o))| < \varepsilon$. If this form of convergence holds, and in addition $\mu(\theta)$ and $\sigma(\theta)^2$ are continuous functions of θ, then the estimator is said to be *consistent uniformly asymptotically normal* (CUAN).

Asymptotic Efficiency. Consider a family $\mathcal{T}$ of sequences of estimators $T_n(\cdot)$ that are CUAN for a parameter θ_o and have asymptotic bias $\mu(\theta) \equiv 0$. An estimator $T^*(\cdot)$ is *asymptotically efficient* relative to class $\mathcal{T}$ if its asymptotic variance is no larger than that of any other member of the family.

Asymptotic sufficiency. In some problems, sufficiency does not provide a useful reduction of dimension in finite samples, but a weaker "asymptotic" form of sufficiency will provide useful restrictions. This could arise if the DGP density can be written $g_1(\mathbf{y},\theta)\cdot g_2(\mathbf{z}|\mathbf{y},\theta)$ for a low-dimensional statistic $\mathbf{y}$, but both g_1 and g_2 depend on θ so $\mathbf{y}$ is not sufficient. However, $g_2(\mathbf{z}|\mathbf{y},\theta)$ may converge in

---

normal CDF Φ permitting us to state the condition without excepting jump points in the limit distribution.

distribution to a density that does not depend on θ. Then, there is a large sample rationale for concentrating on estimators that depend only on $\mathbf{y}$.

2. General Estimation Criteria

It is useful to have some general methods of generating estimators that as a consequence of their construction will have some desirable statistical properties. Such estimators may prove adequate in themselves, or may form a starting point for refinements that improve statistical properties. We introduce several such methods:

Analogy Estimators. Suppose one is interested in a feature of a target population that can be described as a functional of its CDF $F(\cdot)$, such as its mean, median, or variance, and write this feature as $\theta_o = \mu(F)$. An analogy estimator exploits the similarity of a population and of a simple random sample drawn from this population, and forms the estimator $T(\mathbf{x}) = \mu(F_n)$, where μ is the functional that produces the target population feature and F_n is the empirical distribution function. For example, a sample mean will be an analogy estimator for a population mean.

Moment Estimators. Population moments will depend on the parameter index in the underlying DGP. This is true for ordinary moments such as means, variances, and covariances, as well as more complicated moments involving data transformations, such as quantiles. Let $m(x)$ denote a function of an observation and $\mathbf{E}_{x|\theta_o}m(x) = \gamma(\theta_o)$ denote the population moment formed by taking the expectation of $m(x)$. In a sample $\mathbf{x} = (x_1,...,x_n)$, the idea of a moments estimator is to form a sample moment $\frac{1}{n}\sum_{i=1}^{n} m(x_i) \equiv \mathbf{E}_n m(x)$, and then to use the analogy of the population and sample moments to form the approximation $\mathbf{E}_n m(x) \approx \mathbf{E}_{x|\theta_o} = \gamma(\theta_o)$.[4] The moment estimator $T(\mathbf{x})$ solves

[4] The sample average of a function m(x) of an observation can also be interpreted as

$\mathbf{E}_n m(x) = \gamma(T(\mathbf{x}))$. When the number of moment conditions equals the number of parameters, an exact solution is normally obtainable, and $T(\mathbf{x})$ is termed a classical method of moments estimator. When the number of moment conditions exceeds the number of parameters, it is not possible in general to find $T(\mathbf{x})$ that sets them all to zero at once. In this case, one may form a number of linear combinations of the moments equal to the number of parameters to be estimated, and find $T(\mathbf{x})$ that sets these linear combinations to zero. The linear combinations in turn may be derived starting from some metric that provides a measure of the distance of the moments from zero, with $T(\mathbf{x})$ interpreted as a minimand of this metric. This is called *generalized method of moments* estimation.

Maximum likelihood estimators. Consider the DGP density $f(\mathbf{x},\theta)$ for a given sample as a function of θ. The maximum likelihood estimator of the unknown true value θ_o is the function $\theta = T(\mathbf{x})$ that maximizes $f(\mathbf{x},\theta)$. The intuition behind this estimator is that if we guess a value for θ that is far away from the true θ_o, then the probability law for this θ would be very unlikely to produce the data that are actually observed, whereas if we guess a value for θ that is near the true θ_o, then the probability law for this θ would be likely to produce the observed data. Then, the $T(\mathbf{x})$ which maximized this likelihood, as measured by the probability law itself, should be close to the true θ. The maximum likelihood estimator plays a central role in classical statistics, and can be motivated solely in terms of its desirable classical statistical properties in large samples.

---

its expectation with respect to the empirical distribution of the sample; we use the notation $\mathbf{E}_n m(x)$ to denote this empirical expectation.

When the data are a sample of n independent observations, each with density $f(x,\theta)$, then the likelihood of the sample is $f(\mathbf{x},\theta) = \prod_{i=1}^{n} f(x_i,\theta)$. It is often convenient to work with the logarithm of the density, $l(x,\theta) \equiv \text{Log } f(x,\theta)$. Then, the *Log Likelihood* of the sample is $L(\mathbf{x},\theta) \equiv \text{Log } f(\mathbf{x},\theta) = \sum_{n=1}^{N} l(x_i,\theta)$. The *maximum likelihood estimator* is the function $t = T(\mathbf{x})$ of the data that when substituted for θ maximizes $f(\mathbf{x},\theta)$, or equivalently $L(\mathbf{x},\theta)$.

The gradient of the log likelihood of an observation with respect to θ is denoted $s(\mathbf{x},\theta) \equiv \nabla_\theta l(x,\theta)$, and termed the *score*. The maximum likelihood estimator is a zero of the sample expectation of the score, $\mathbf{E}_n s(\mathbf{x},T(\mathbf{x}))$. Then, the maximum likelihood estimator is a special case of a moments estimator.

Maximum likelihood estimators will under quite general regularity conditions be consistent and asymptotically normal. Under uniformity conditions that rule out some odd non-uniform "super-efficient" alternatives, they are also asymptotically efficient. They often have good finite-sample properties, or can be easily modified so that they do. However, their finite-sample properties have to be determined on a case-by-case basis.

3. Estimation in Normally Distributed Populations

Consider a simple random sample $\mathbf{x} = (x_1,...,x_n)$ from a population in which observations are normally distributed with mean μ and variance σ^2. Then, the density of an observation is $\phi((x-\mu)/\sigma)/\sigma$, where $\phi(v) = (2\pi)^{-1/2}e^{-v^2/2}$, and the log likelihood of the sample is $L(\mathbf{x},\mu,\sigma^2) = -\frac{n}{2}\cdot\text{Log}(2\pi) - \frac{n}{2}\cdot\text{Log } \sigma^2 - \frac{1}{2}\cdot\sum_{i=1}^{n}(x_i-\mu)^2/\sigma^2$. We will estimate the parameters μ and σ^2 using the maximum likelihood method, and establish some of the statistical properties of these estimators.

The first-order-conditions for maximizing $L(\mathbf{x},\mu,\sigma^2)$ in μ and σ^2 are

$$0 = \sum_{i=1}^{n} (x_i-\mu)/\sigma^2 \quad \Longrightarrow \quad \hat{\mu} = \bar{x} \equiv \frac{1}{n} \sum_{i=1}^{n} x_i \ ,$$

$$0 = -n/2\sigma^2 + \sum_{i=1}^{n} (x_i-\mu)^2/2\sigma^4 \quad \Longrightarrow \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i-\bar{x})^2 \ .$$

The maximum likelihood estimator of μ is then the sample mean, and the maximum likelihood estimator of σ^2 is the sample variance. Define $s^2 = \hat{\sigma}^2{\cdot}n/(n-1)$, the sample variance with a sample size correction. The following results summarize the properties of these estimators:

(1) $(\bar{x},s^2)$ are joint minimal sufficient statistics for (μ,σ^2)**.**

(2) $\bar{x}$ is an unbiased estimator for μ, and s^2 an unbiased estimator for σ^2**.**

(3) $\bar{x}$ is a Minimum Variance Unbiased Estimator (MVUE) for μ**;** s^2 is MVUE for σ^2**.**

(4) $\bar{x}$ is Normally distributed with mean μ and variance σ^2/n.

(5) $(n-1)s^2/\sigma^2$ has a Chi-square distribution with n-1 degrees of freedom.

(6) $\bar{x}$ and s^2 are statistically independent.

(7) $\sqrt{n}(\bar{x} - \mu)/s$ has a Student's-T distribution with n-1 degrees of freedom.

(8) $n{\cdot}(\bar{x} - \mu)^2/s^2$ has an F-distribution with 1 and n-1 degrees of freedom.

The following paragraphs comment on these properties and prove them.

Consider the sufficiency property (1). Factor the log likelihood function as

$$L(\mathbf{x},\mu,\sigma^2) = - \frac{n}{2}{\cdot}\text{Log}(2\pi) - \frac{n}{2}{\cdot}\text{Log}\ \sigma^2 - \frac{1}{2}{\cdot} \sum_{i=1}^{n} (x_i-\bar{x} + \bar{x}-\mu)^2/\sigma^2$$

$$= -\tfrac{n}{2}\cdot\text{Log}(2\pi) - \tfrac{n}{2}\cdot\text{Log }\sigma^2 - \tfrac{1}{2}\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2/\sigma^2 - \tfrac{1}{2}\cdot\sum_{i=1}^{n}(\bar{x}-\mu)^2/\sigma^2$$

$$= -\tfrac{n}{2}\cdot\text{Log}(2\pi) - \tfrac{n}{2}\cdot\text{Log }\sigma^2 - \tfrac{1}{2}\cdot\frac{(n-1)s^2}{\sigma^2} - \tfrac{n}{2}\cdot\frac{(\bar{x}-\mu)^2}{\sigma^2} .$$

This implies that $\bar{x}$ and s^2 are jointly sufficient for μ and σ^2. Because the dimension of $(\bar{x},s^2)$ is the same as the dimension of (μ,σ^2), they are obviously minimal sufficient statistics.

The expectation of $\bar{x}$ is $\mathbf{E}\bar{x} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}x_i = \mu$, since the expectation of each observation is μ. Hence $\bar{x}$ is unbiased. To establish the expectation of s^2, first form the $n\times n$ matrix

$$M = \mathbf{I} - \mathbf{1}\mathbf{1}'/n = \begin{bmatrix} 1\text{-}1/n & -1/n & \ldots & -1/n & -1/n \\ -1/n & 1\text{-}1/n & \ldots & -1/n & -1/n \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ -1/n & -1/n & \ldots & 1\text{-}1/n & -1/n \\ -1/n & -1/n & \ldots & -1/n & 1\text{-}1/n \end{bmatrix},$$

where $\mathbf{I}$ is the identity matrix and $\mathbf{1}$ is a $n\times 1$ vector of ones. This matrix is idempotent, with $M^2 = M$, and its trace satisfies

$$\text{tr}(M) = \text{tr}(\mathbf{I}) - \text{tr}(\mathbf{1}\mathbf{1}'/n) = n - \text{tr}(\mathbf{1}'\mathbf{1}/n) = n - 1 .$$

Let $Z' = (x_1-\mu,\ldots,x_n-\mu)$ denote the vector of deviations of observations from the population mean. Then, $Z'M = (x_1 - \bar{x},\ldots,x_n - \bar{x})$ and $s^2 = Z'M\cdot MZ/(n-1) = Z'MZ/(n-1)$. Therefore, since with independent observations one has $\mathbf{E}ZZ' = \sigma^2\mathbf{I}$, one obtains

$$\mathbf{E}s^2 = \mathbf{E}(Z'MZ)/(n-1) = \mathbf{E}\ \text{tr}(Z'MZ)/(n-1) = \mathbf{E}\ \text{tr}(MZZ')/(n-1)$$

$$= \text{tr}(M\cdot\mathbf{E}(ZZ'))/(n-1) = \sigma^2\cdot\text{tr}(M)/(n-1) = \sigma^2 .$$

Hence, s^2 is unbiased.

The MVUE property of $\bar{x}$ and s^2 is most easily proved by application of the Blackwell theorem. We already know that these estimators are unbiased. Any other unbiased estimator of μ then has the property that the difference of this estimator and $\bar{x}$, which we will denote by $h(\mathbf{x})$, must satisfy $\mathbf{E}h(\mathbf{x}) \equiv 0$. Alternately, $h(\mathbf{x})$ could be the difference of s^2 and any other unbiased estimator of σ^2. Then,

$$0 \equiv \mathbf{E}h(\mathbf{x}) \equiv \int_{-\infty}^{+\infty} h(\mathbf{x})e^{L(\mathbf{x}|\mu,\sigma^2)}d\mathbf{x},$$

implying

$$0 \equiv \int_{-\infty}^{+\infty} h(\mathbf{x})e^{-(\mathbf{x}-\mu)'(\mathbf{x}-\mu)/2\sigma^2}d\mathbf{x} \ .$$

Differentiate with respect to μ to get

$$0 \equiv \int_{-\infty}^{+\infty} h(\mathbf{x})(\mathbf{x}-\mu)e^{-(\mathbf{x}-\mu)'(\mathbf{x}-\mu)/2\sigma^2}d\mathbf{x} \equiv \int_{-\infty}^{+\infty} h(\mathbf{x})\mathbf{x}e^{-(\mathbf{x}-\mu)'(\mathbf{x}-\mu)/2\sigma^2}d\mathbf{x}$$

Hence, $\mathbf{E}h(\mathbf{x})\cdot x_i \equiv 0$ for each component of $\mathbf{x}$. Averaging then implies $\mathbf{E}h(\mathbf{x})\cdot\bar{x} \equiv 0$. Differentiate again with respect to μ to get

$$0 \equiv \int_{-\infty}^{+\infty} h(\mathbf{x})\mathbf{x}(\mathbf{x}-\mu)'e^{-(\mathbf{x}-\mu)'(\mathbf{x}-\mu)/2\sigma^2}d\mathbf{x} \equiv \int_{-\infty}^{+\infty} h(\mathbf{x})(\mathbf{x}-\mu)(\mathbf{x}-\mu)'e^{-(\mathbf{x}-\mu)'(\mathbf{x}-\mu)/2\sigma^2}d\mathbf{x} \ .$$

Pre-multiply and post-multiply this expectation by the idempotent matrix M to conclude that $\mathbf{E}h(\mathbf{x})M(\mathbf{x}-\mu)(\mathbf{x}-\mu)'M \equiv 0$. Taking the trace of this expression and dividing by (n-1) yields $\mathbf{E}h(\mathbf{x})s^2 \equiv 0$. Then, the estimators $\bar{x}$ and s^2 are uncorrelated with any unbiased estimator of zero. The Blackwell theorem then establishes that they are minimum variance among all unbiased estimators.

Next consider the distribution of $\bar{x}$. We use the fact that linear transformations of multivariate normal random vectors are again multivariate normal: If $Z \sim N(\mu,\Omega)$ and $W = CZ$, then $W \sim N(C\mu,C\Omega C')$. This result holds even if Z and W are of different dimensions, or C is of less than full rank. (If the rank of $C\Omega C'$ is less than full, then the random variable has all its density concentrated on a subspace.) Now $\bar{x} = C\mathbf{x}$, where $C = (1/n,...,1/n)$ and $\mathbf{x} = (x_1,...,x_n)'$, and $\mathbf{x}$ is multivariate normal with mean $\mathbf{1}\cdot\mu$ and covariance matrix $\sigma^2 I$, where $\mathbf{1}$ is a $n\times 1$ vector of ones and I is the $n\times n$ identity matrix. Therefore, $\bar{x} \sim N(\mu C\mathbf{1},\sigma^2 CC') = N(\mu,\sigma^2/n)$.

Next consider the distribution of s^2. We will need the following fact about statistical distributions: <u>The sum of the squares of K independent standard normal random variables has a Chi-Square distribution with K degrees of freedom</u>. (To prove this, first show that it holds for K = 1 by finding the density of the square of a standard normal random variable and noting that it coincides with the density of χ_1^2. Then use the rules for moment generating functions to see that the sum of independent χ_1^2 random variables is χ_K^2.) We also need the matrix result that <u>any idempotent matrix M of dimension n and rank r can be written as M = WW′, where W is n×r and column-orthonormal</u> (i.e., $W'W = I_r$). (To prove this, write M in terms of its singular value decomposition, and apply the conditions $M = M'$ and $M\cdot M = M$.) Consider $M = I - \mathbf{1}\mathbf{1}'/n$ which has rank $n-1 = tr(M)$, and the linear transformation

$$\begin{bmatrix} \bar{x} - \mu \\ u \end{bmatrix} = \begin{bmatrix} (1/n)\mathbf{1}' \\ M \end{bmatrix}(\mathbf{x} - \mathbf{1}\cdot\mu) \equiv C(\mathbf{x} - \mathbf{1}\cdot\mu) \ .$$

The result of this transformation is then multivariate normal,

$$\begin{bmatrix} \bar{x} - \mu \\ u \end{bmatrix} \sim N(0,\sigma^2 CC') \ .$$

But $CC' = \begin{bmatrix} (1/n) & 0 \\ 0 & M \end{bmatrix}$, so that $\bar{x} - \mu$ and u are uncorrelated, hence (for joint normals) independent. Then $\bar{x}$ is independent of any function of u, and specifically of $s^2 = u'u/(n-1)$. The distribution of s^2 is obtained by noting that $u'u = \varepsilon'M\varepsilon = \varepsilon'WW'\varepsilon$, where $\varepsilon = (\mathbf{x} - \mathbf{1}\cdot\mu)$. But $z = W'\varepsilon \sim N(0,\sigma^2 I_{n-1})$, by the matrix result above for idempotent matrices. Hence, $(n-1)s^2/\sigma^2 = u'u/\sigma^2 = z'z/\sigma^2$ is the sum of squares of n-1 independent standard normal random variates, so that it is distributed χ^2_{n-1}.

The results that $\sqrt{n}(\bar{x} - \mu)/s$ has a Student's-T distribution with n-1 degrees of freedom, and that $n\cdot(\bar{x} - \mu)^2/s^2$ has an F-distribution with 1 and n-1 degrees of freedom follow from properties of distributions related to the normal, Chapter 3.9.

4. Large Sample Properties of Maximum Likelihood Estimates

This section provides a brief and informal introduction to the statistical properties of maximum likelihood estimators and similar estimation methods in large samples. Consider a simple random sample $\mathbf{x} = (x_1,...,x_n)$ from a population in which the density of an observation is $f(x,\theta_o)$. The DGP density or likelihood of the sample is then $f(\mathbf{x},\theta) = f(x_1,\theta)\cdot...\cdot f(x_n,\theta)$, with θ_o the true value of θ. The log likelihood of an observation is $l(x,\theta) = \log f(x,\theta_o)$, and the log likelihood of the sample is $L_n(\mathbf{x},\theta) = \frac{1}{n}\sum_{i=1}^{n} l(x_n,\theta)$.[5] The maximum likelihood estimator $T_n(\mathbf{x})$ is a value of θ which maximizes $L_n(\mathbf{x},\theta)$. The first-order condition for this maximum is that the *sample score*,

[5] For the purposes of this section, it will be convenient to scale the sample likelihood by 1/n so that it is an average of the scores of the individual observations. Obviously one can go from this definition to a definition of the sample log likelihood without scaling simply by multiplying by n.

$$\nabla_\theta L_n(\mathbf{x},\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta l(x_i,\theta) \;,$$

equal zero at $\theta = T_n(\mathbf{x})$. The second order condition is that the *sample hessian*

$$\nabla_{\theta\theta} L_n(\mathbf{x},\theta) = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta\theta} l(x_i,\theta) \;,$$

be negative at $\theta = T(\mathbf{x})$.[6]

Under very mild regularity conditions, the expectation of the score of an observation is zero at the true parameter vector. Start from the identity $\int_{-\infty}^{+\infty} e^{l(x,\theta)}\cdot dx \equiv 1$ and differentiate with respect to θ under the integral sign to obtain the condition $\int_{-\infty}^{+\infty} \nabla_\theta l(x,\theta)\cdot e^{l(x,\theta)}\cdot dx \equiv 0$. (Regularity conditions are needed to assure that one can indeed differentiate under the integral.) Then, at the true parameter θ_o, it must be true that $\mathbf{E}_{\mathbf{x}|\theta_o}\nabla_\theta l(x,\theta_o) = 0$. This gives a *population score* condition that $\mathbf{E}_{\mathbf{x}|\theta_o}\nabla_\theta l(x,\theta) = 0$ when $\theta = \theta_o$. Another regularity condition requires that $\mathbf{E}_{\mathbf{x}|\theta_o}\nabla_\theta l(x,\theta) = 0$ only if $\theta = \theta_o$; this has the interpretation of an identification condition. The maximum likelihood estimator can be interpreted as an analogy estimator that chooses $T_n(\mathbf{x})$ to satisfy a sample condition (that the sample score be zero) that is analogous to the population score condition (that the population score of an observation is zero at the true parameter value). One could sharpen the statement of this analogy by writing the population score as an explicit

[6] When the parameter θ is more than one-dimensional, the second-order condition is that the sample hessian is a negative definite matrix.

function of the population DGP, $\mu(\theta,F(\cdot,\theta_o)) \equiv \mathbf{E}_{\mathbf{x}|\theta_o}\nabla_\theta l(x,\theta)$, and writing the sample score as $\mu(\theta,F_n) \equiv \mathbf{E}_n\nabla_\theta l(x,\theta)$. The mapping $\mu(\theta,\cdot)$ is linear in its second argument, and this is enough to assure that it is continuous (in an appropriate sense) in this argument. Then one has almost sure convergence of $\mu(\theta,F_n)$ to $\mu(\theta,F(\cdot,\theta_o))$ for each θ, from the Glivenko-Cantelli theorem. A few additional regularity conditions are enough to ensure that this convergence is uniform in θ, and that a solution $T_n(\mathbf{x})$ that sets the sample score to zero converges almost surely to the value θ_o that sets the population score to zero.

The basic large sample properties of maximum likelihood estimators are that under suitable regularity conditions, T_n converges in probability to the true parameter vector θ_o, and $\sqrt{n}(T_n - \theta_o)$ converges in distribution to a normal random variable with mean zero and a variance which achieves the Cramer-Rao bound for an unbiased estimator. These results imply that in large samples, T_n will become a more and more precise estimate of the true parameter. In large samples, the convergence in distribution to a Normal permits one to use the properties of a Normal population to construct hypothesis tests and confidence bounds, and get good approximations for significance levels and power. The achievement of the Cramer-Rao lower bound on variance indicates that in large samples there are no alternative estimators which are uniformly more precise, so MLE is the "best" one can do.

We next list a series of regularity conditions under which the results stated above can be shown to hold. Only the single parameter case will be presented. However, the conditions and results have direct generalizations to the multiple parameter case. This list is chosen so the conditions are easy to interpret and to check in applications. Note that these are conditions on the population DGP, not on a specific sample. Hence, "checking" means verifying that your model of the DGP and your assumptions on distributions of random variables are logically consistent with the regularity conditions, rather than carrying out an empirical verification using the data. There are alternative forms for the regularity conditions, as well as weaker conditions, which give the same or similar results. These conditions are quite generic and will be satisfied in many economic applications. However, it is a serious mistake to assume without checking that the DGP you assume for your problem is consistent with these conditions. While in many cases the mantra "I assume the appropriate regularity conditions" will work out, you can be acutely embarrassed if your DGP happens to be one of the exceptions that is logically inconsistent with the regularity conditions, particularly if it results in estimators that fail to have desirable statistical properties.

A.1 There is a single parameter θ which is permitted to vary in a closed bounded set Θ. The true value θ_o is in the interior of Θ.

A.2 The sample observations are realizations of independently identically distributed random variables $x_1,...,x_n,...$ with a common density $f(x,\theta_o)$.

A.3 The density $f(x,\theta)$ is continuous in θ, and three times continuously differentiable in θ, for each x, and is "well behaved" (e.g., measurable or piecewise continuous or continuous) in x for each θ.

A.4 There exists a bound $\beta(x)$ on the density and its derivatives, uniform in θ, satisfying $|l(x,\theta)| \le \beta(x)$, $(\nabla_\theta l(x,\theta))^2 \le \beta(x)$, $|\nabla_{\theta\theta} l(x,\theta)| \le \beta(x)$, $|\nabla_{\theta\theta\theta} l(x,\theta)| \le \beta(x)$, and $\int_{-\infty}^{+\infty} \beta(x)^2 f(x|\theta_o)dx < +\infty$.

A.5 The function $\lambda(\theta) = \mathbf{E}_{\mathbf{x}|\theta_o}\, l(x,\theta)$ has $\lambda(\theta) < \lambda(\theta_o)$ and $\nabla_\theta\lambda(\theta) \ne 0$ for $\theta \ne \theta_o$ and $J = -\nabla_{\theta\theta}\lambda(\theta_o) > 0$.

The expression J in A.5 is termed the *Fisher information* in an observation. The first two assumptions mostly set the problem. The restriction of the parameter to a closed bounded set guarantees that a MLE exists, and can be relaxed by adding conditions elsewhere. Requiring θ_o interior to Θ guarantees that the first-order condition $\mathbf{E}_n\nabla_\theta l(x,T_n(\cdot)) = 0$ for a maximum holds for large n, rather than an inequality condition for a maximum at a boundary. This really matters because MLE at boundaries can have different asymptotic distributions and rates of convergence than the standard $\sqrt{n}$ convergence to the normal. The continuity conditions A.3 are satisfied for most economic problems, and in some weak form are critical to the asymptotic distribution results. Condition A.4 gives bounds that permit exchange of the order of differentiation and integration in forming expectations with respect to the population density. Condition A.5 is an identification requirement which implies there cannot be a parameter vector other than θ_o that on average always explains the data as well as θ_o.

The next result establishes that under these regularity conditions, a MLE is consistent and asymptotically normal (CAN):

Theorem: If A.1-A.5 hold, then a MLE T_n satisfies

(1) T_n is consistent for θ_o.

(2) T_n is asymptotically normal: $\sqrt{n}(T_n(\mathbf{x}) - \theta_o) \xrightarrow{d} Z_o \sim N(0,J^{-1})$, with J equal to the Fisher information in an observation, $J = \mathbf{E}_{x|\theta_o}\nabla_\theta l(x,\theta_o)^2$.

(3) $\mathbf{E}_n[\nabla_\theta l(x,T_n)]^2 \xrightarrow{p} J$ and $-\mathbf{E}_n\nabla_{\theta\theta}l(x,T_n) \xrightarrow{p} J$.

(4) Suppose T_n' is any sequence of estimators that solve equations of the form $\mathbf{E}_n g(x,\theta) = 0$, where g is twice continually differentiable, with $\mathbf{E}_{x|\theta_o}g(x,\theta) = 0$ if and only if $\theta = \theta_o$; $\mathbf{E}\nabla_\theta g(x,\theta_o) \neq 0$; bounds $|g(x,\theta)| \leq \beta(x)$, $|\nabla_\theta g(y,\theta)^2| \leq$

$\beta(x)$, $|\nabla_{\theta\theta}g(x,\theta)| \le \beta(x)$, and with $\mathbf{E}\beta(x)^2 < +\infty$; and $R = -\mathbf{E}\nabla_\theta g(y,\theta_o) \neq 0$. Let $S = \mathbf{E}g(x,\theta_o)^2$. Then $T_n' \xrightarrow{p} \theta_o$ and $\sqrt{n}(T_n' - \theta^*) \xrightarrow{d} Z_1 \sim N(0,V)$, where $V = R^{-1}SR'^{-1}$. Further, $V \ge J^{-1}$, so that the MLE T_n is efficient relative to T_n', and $\mathrm{cov}(Z_0, Z_1 - Z_0) = 0$.

Result (2) in this theorem implies that to a good approximation in large samples, the expression $\sqrt{n}(T_n-\theta_o)$ has a normal distribution with a variance which is the inverse of the Fisher information, J^{-1}. Result 3 gives two ways of estimating the asymptotic variance J^{-1} consistently since J^{-1} is a continuous function of J for $J \neq 0$. Result (4) establishes that MLE is efficient relative to a broad class of estimators called *M-estimators*.

An intuitive demonstration of the Theorem will be given rather than formal proofs. Consider first the consistency result. The reasoning is as follows. Consider the expected likelihood of an observation,

$$\lambda(\theta) \equiv \mathbf{E}_{x|\theta_o} l(x,\theta) = \int_{-\infty}^{+\infty} l(x,\theta)f(x,\theta_o)dx.$$

We will argue that $\lambda(\theta)$ has a unique maximum at θ_o. Then we will argue that any function which is uniformly very close to $\lambda(\theta)$ must have its maximum near θ_o. Finally, we argue by applying a uniform law of large numbers that the likelihood function is with probability approaching one uniformly very close to λ for n sufficiently large. Together, these results will imply that with probability approaching one, T_n is close to θ_o for n large.

Assumption A.4 ensures that $\lambda(\theta)$ is continuous, and that one can reverse the order of differentiation and integration to obtain continuous derivatives

$$\nabla_\theta\lambda(\theta) \equiv \int_{-\infty}^{+\infty} \nabla_\theta l(x,\theta) f(x,\theta_o)dx \equiv \mathbf{E}_{x|\theta_o}\nabla_\theta l(x,\theta)$$

$$\nabla_{\theta\theta}\lambda(\theta) \equiv \int_{-\infty}^{+\infty} \nabla_{\theta\theta} l(x,\theta) f(x,\theta_o)dx \equiv \mathbf{E}_{x|\theta_o}\nabla_{\theta\theta} l(x,\theta)$$

Starting from the identity

$$1 \equiv \int_{-\infty}^{+\infty} f(x,\theta)dx \equiv \int_{-\infty}^{+\infty} e^{l(x,\theta)}dx,$$

one obtains by differentiation

$$0 \equiv \int_{-\infty}^{+\infty} \nabla_\theta l(x,\theta)e^{l(x,\theta)}dx$$

$$0 \equiv \int_{-\infty}^{+\infty} [\nabla_{\theta\theta} l(x,\theta) + \nabla_\theta l(x,\theta)^2]e^{l(x,\theta)}dx$$

Evaluated at θ_o, these imply

$$0 = \nabla_\theta\lambda(\theta_o) \quad \text{and} \quad -\nabla_{\theta\theta}\lambda(\theta_o) = \mathbf{E}_{x|\theta_o}\nabla_\theta l(x,\theta)^2 = J \;.$$

Assumption A.5 requires further that $J \neq 0$, and that θ_o is the only root of $\nabla_\theta\lambda(\theta)$. Hence, $\lambda(\theta)$ has a unique maximum at θ_o, and at no other θ satisfies a first-order condition or boundary condition for a local maximum.

We argue next that any function which is close enough to $\nabla_\theta\lambda(\theta)$ will have at least one root near θ_o and no roots far away from θ_o. The figure on the following page graphs $\nabla_\theta\lambda(\theta)$, along with a "sleeve" which is a vertical distance δ from $\nabla_\theta\lambda$. Any function trapped in the sleeve must have at least one root between $\theta_o - \varepsilon_1$ and $\theta_o + \varepsilon_2$, where $[\theta_o-\varepsilon_1,\theta_o+\varepsilon_2]$ is the interval where the sleeve intersects the axis, and must have no roots outside this interval. Furthermore, the uniqueness of the root θ_o of $\nabla_\theta\lambda(\theta)$ plus the condition $\nabla_{\theta\theta}\lambda(\theta_o) < 0$ imply that as δ shrinks toward zero, so do ε_1 and ε_2.

The last step in the consistency argument is to show that $L_n(\mathbf{x},\theta)$ is with probability approaching one contained in a δ-sleeve around $\lambda(\theta)$. For fixed θ, $L_n(\mathbf{x},\theta) = \frac{1}{n}\sum_{i=1}^{n} l(x_i,\theta)$ is a sample average of i.i.d. random variables $l(x,\theta)$ with mean $\lambda(\theta)$. Then Kolmogorov's SLLN implies $L_n(\mathbf{x},\theta) \xrightarrow{as} \lambda(\theta)$. This is not quite enough, because there is a question of whether $L_n(\mathbf{x},\theta)$ could converge non-uniformly to $\lambda(\theta)$, so that for any n there are some values of θ where $L_n(\mathbf{x},\theta)$ is outside the sleeve. However, assumptions A.1, A.3, and A.4 imply $\max_{\theta\in\Theta} |L_n(\mathbf{x},\theta) - \lambda(\theta)| \xrightarrow{as} 0$. This follows in particular because the differentiability of $f(x,\theta)$ in θ from A.3 and the bound on $\nabla_\theta l(x,\theta)$ from A.4 imply that $l(\cdot,\theta)$ is almost surely continuous on the compact set Θ, so that the uniform SLLN in Chapter 4.4 applies. This establishes that $T_n \xrightarrow{as} \theta_o$.

We next demonstrate the asymptotic normality of T_n. A Taylor's expansion about θ_o of the first-order condition for maximization of the log likelihood function gives

(1) $$0 = \nabla_\theta L_n(T_n) = \nabla_\theta L_n(\theta_o) + \nabla_{\theta\theta}L_n(\theta_o)\cdot(T_n-\theta_o) + \nabla_{\theta\theta\theta}L_n(\tilde{T}_n)\cdot(T_n-\theta_o)^2/2 \ ,$$

where $\tilde{T}_n$ is some point between T_n and θ_o. Multiply this equation by $\sqrt{n}/(1+\sqrt{n}|T_n-\theta_o|)$ to obtain

(2) $$0 = \frac{B_n + C_n\sqrt{n}(T_n - \theta_o)}{1 + \sqrt{n}\cdot|T_n - \theta_o|} + \frac{D_n Z_n(T_n - \theta_o)}{2}$$

with

$$B_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \nabla_\theta l(y_i,\theta_o) \qquad\qquad C_n = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta\theta} l(y_i,\theta_o)$$

$$D_n = \frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta\theta\theta} l(y_i,\tilde{T}) \qquad\qquad Z_n = \sqrt{n}(T_n-\theta_o)/(l + \sqrt{n}\cdot|T_n-\theta_o|).$$

We make a limiting argument on each of the terms. First, $\nabla_\theta l(y_i,\theta_o)$ are iid random variables with $\mathbf{E}\nabla_\theta l(y_i,\theta_o) = \nabla_\theta\lambda(\theta_o) = 0$ and $\mathbf{E}[\nabla_\theta l(y_i,\theta_o)]^2 = -\nabla_{\theta\theta}\lambda(\theta_o) = J$. Hence the Lindeberg-Levy CLT implies $B_n \xrightarrow{d} W_O \sim N(O,J)$. Second, $\nabla_{\theta\theta} l(Y_i,\theta_o)$ are i.i.d. random variables with $\mathbf{E}\nabla_{\theta\theta} l(Y_i,\theta_o) = -J$. Hence the Khinchine WLLN implies $C_n \xrightarrow{p} -J$. Third,

$$|D_n| \le \frac{1}{n}\sum_{i=1}^{n} |\nabla_{\theta\theta\theta} l(y_i,\tilde{T}_n)| \le \frac{1}{n}\sum_{i=1}^{n} \beta(y_i)$$

and $\mathbf{E}\beta(Y) < +\infty$ by A.4. Hence, Khinchine's WLLN implies $|D_n|$ is bounded by an expression which converges in probability to $\mathbf{E}\beta(Y) < +\infty$. Thus, $D_n = O_p(1)$. Furthermore, $|Z_n| \le 1$, implying $Z_n = O_p(1)$. Since T_n is consistent, $(T_n - \theta_o) = o_p(1)$. Therefore, by rule 6 in Figure 4.3, $D_nZ_n(T_n- \theta_o)/2 = o_p(1)$.

Given $J/2 > \varepsilon > 0$, these arguments establish we can find n_o such that for $n > n_o$ with probability at least $1-\varepsilon$, we have $|D_nZ_n(T_n-\theta_o)/2| < \varepsilon$, $|C_n+J| < \varepsilon$ and $|B_n| < M$ for a large constant M (since $B_n \xrightarrow{d} W_o \Rightarrow B_n = O_p(1)$). In this event, $|C_n| > J-\varepsilon$, $|B_n + C_n\sqrt{n}(T_n-\theta_o)| < \varepsilon(1 + \sqrt{n}\cdot|T_n-\theta_o|)$, and $|B_n| \le M$ imply $|C_n|\sqrt{n}|T_n-\theta_o| - |B_n| \le |B_n+C_n\sqrt{n}|T_n-\theta_o)| < \varepsilon\cdot(1 - \sqrt{n}\cdot|T_n-\theta_o|)$, or $(J - 2\varepsilon)\sqrt{n}\cdot|T_n-\theta_o| < M + \varepsilon$. Therefore $\sqrt{n}(T_n-\theta^*) = O_p(1)$; i.e., it is stochastically bounded. Therefore, by rule 6 in Figure 3, multiplying (2) by $1 + \sqrt{n}\cdot|T_n-\theta_o|$ yields $0 = B_n + C_n \sqrt{n}\cdot|T_n-\theta_o| + o_p(1)$. But $C_n \xrightarrow{p} -J < 0$ implies $C_n^{-1} \xrightarrow{p} -J^{-1}$. By rule 6, $(C_n^{-1}+J^{-1})B_n = o_p(1)$ and $\sqrt{n}\,(T_n-\theta_o) = J^{-1}B_n + o_p(1)$. The limit rules in Figure 1 then imply $J^{-1}B_n \xrightarrow{d} Z_o \sim N(0,J^{-1})$, $\sqrt{n}\cdot|T_n-\theta_o| - J^{-1}B_n \xrightarrow{p} 0$, and hence $\sqrt{n}\cdot|T_n-\theta_o| \xrightarrow{d} Z_o$.

The third result in the theorem is that J is estimated consistently by

(3) $$J_n = \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta l(y_i,T_n)^2.$$

To show this, make a Taylor's expansion of this expression around θ_o,

(4) $$J_n = \frac{1}{n}\sum_{i=1}^{n} l_\theta(y_i,\theta_o)^2 + \frac{1}{n}\sum_{i=1}^{n} 2\,\nabla_\theta l(y_i,\tilde{T}_n)\cdot\nabla_{\theta\theta} l(y_i,\tilde{T}_n)(T_n-\theta_o).$$

We have already shown that the first term in (4) converges in probability to J. The second term is the product of $(T_n - \theta_o) \xrightarrow{p} 0$ and an expression which is bounded by $\frac{1}{n}\sum_{i=1}^{n} 2\beta(y_i)^2 \xrightarrow{p} 2\mathbf{E}_Y\beta(Y)^2 < +\infty$, by Khinchine's WLLN. Hence the second term is $o_p(1)$ and $J_n \xrightarrow{p} J$.

The final result in the theorem establishes that the MLE is efficient relative to any M-estimator T_n' satisfying $\sum_{i=1}^{n} g(y_i,T_n') = 0$, where g has the stated properties. The first conclusion in this result is that T_n' is consistent and $\sqrt{n}(T_n'-\theta_o)$ is asymptotically normal. This is actually of considerable independent interest, since many of the alternatives to MLE that are used in econometrics for reasons of computational convenience or robustness are M-estimators. Ordinary least squares is a leading example of an estimator in this class. The argument for the properties of $\bar{\theta}$ are exactly the same as for the MLE case above, with g replacing $\nabla_\theta l$. The only difference is that R and S are not necessarily equal, whereas for $g = \nabla_\theta l$ in the MLE case, we had R = S = J. To make the efficiency argument, consider together the Taylor's expansions used to get the asymptotic distributions of T_n and T_n',

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla_\theta \ell(y_i,T_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla_\theta \ell(y_i,\theta_o) + \frac{1}{n} \sum_{i=1}^{n} \nabla_{\theta\theta} \ell(y_i,\theta_o)\sqrt{n}(T_n-\theta_o) + o_p(1)$$

$$0 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(y_i,T_n') = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(y_i,\theta_o) + \frac{1}{n} \sum_{i=1}^{n} g_\theta(Y_i,\theta_o)\sqrt{n}(T_n'-\theta_o) + o_p(1)$$

Solving these two equations in $\sqrt{n}(T_n-\theta_o)$ and $\sqrt{n}(T_n'-\theta_o)$ yields

$$\sqrt{n}(T_n-\theta_o) = J^{-1}W_n + o_p(1)$$

$$\sqrt{n}(T_n'-\theta_o) = R^{-1}U_n + o_p(1)$$

with $W_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \nabla_\theta \ell(y_i,\theta_o)$ and $U_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(y_i,\theta_o)$. Consider any weighted average of these equations,

$$\sqrt{n}((1-\gamma)T_n + \gamma T_n' - \theta_o) = J^{-1}(1-\gamma)W_n + R^{-1}\gamma\, U_n + o_p(1) \ .$$

The Lindeberg-Levy CLT implies that this expression is asymptotically normal with mean zero and variance

$$\Omega = J^{-2}(1-\gamma)^2\mathbf{E}\ell_\theta(Y|\theta^*)^2 + R^{-2}\gamma^2\mathbf{E}g(Y,\theta^*)^2 + 2J^{-1}R^{-1}(1-\gamma)\gamma\mathbf{E}\ell_\theta(Y|\theta^*)g(Y,\theta^*) \ .$$

The condition $0 \equiv \int g(y,\theta)f(y|\theta)dy \equiv \int g(y,\theta)e^{\ell(y|\theta)}dy$, implies, differentiating under the integral sign,

$$0 \equiv \int g_\theta(y,\theta)e^{\ell(y,\theta)}dy + \int \nabla_\theta \ell(y,\theta)g(y,\theta)e^{\ell(y,\theta)}dy \ .$$

Evaluated at θ_o, this implies $0 \equiv -R + \mathbf{E}\nabla_\theta \ell(Y|\theta_o)g(Y,\theta_o)$. Hence,

$$\Omega = J^{-1}(1-\gamma)^2 + R^{-2}S\,\gamma^2 + 2(1-\gamma)\gamma\, J^{-1}R^{-1}R = J^{-1} + [R^{-2}S - J^{-1}]\gamma^2.$$

Since $\Omega \geq 0$ for any γ, this requires $V = R^{-2}S \geq J^{-1}$, and hence $\Omega \geq J^{-1}$. Further, note that

$$\Omega = \text{var}(Z_o+\gamma(Z_1-Z_o)) = \text{var}(Z_o) + \gamma^2 \text{var}(Z_1-Z_o) + 2\gamma \text{cov}(Z_o,Z_1-Z_o) ,$$

and $\text{var}(Z_o) = J^{-1}$, implying

$$2\gamma \text{cov}(Z_o, Z_1 - Z_o) \geq -\gamma^2 \text{var}(Z_1 - Z_o) .$$

Taking γ small positive or negative implies $\text{cov}(Z_o, Z_1 - Z_o) = 0$.