3: Introduction to Estimation and Inference

Bertille Antoine
(adapted from notes by Brian Krauth and Simon Woodcock)

Typically, the data we observe consist of repeated measurements on one or more variables of interest. We usually think of these as being the outcome of a DGP. Underlying the DGP are probability distributions such as those we discussed in the last two lectures. The goal of (parametric) econometric inference is to use the observed data to learn about the DGP. That is, to construct an empirical model of an economic process.

Random Samples

Classical statistical inference uses the observed data (the **sample**) to learn about the **population** from which the sample is drawn. A sample consists of n observations $\mathbf{X}_1, \mathbf{X}_2 ..., \mathbf{X}_n$ on one or more random variables. If certain conditions are met, we call these a random sample.

Definition 1 *The random variables $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ are called a **random sample of size n from the population $f_{\mathbf{X}}(x)$** (or simply a random sample) if $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ are mutually independent random variables, and the pdf of each $\mathbf{X}_i$ is the same function $f_{\mathbf{X}}(\mathbf{x})$. Sometimes $\mathbf{X}_1, ..., \mathbf{X}_n$ are called **independent and identically distributed (iid) random variables with pdf $f_{\mathbf{X}}(x)$.***

Since the observations in a random sample are independent, the joint pdf of $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ is

$$f_{\mathbf{X}}(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n) = f_{\mathbf{X}}(\mathbf{x}_1) f_{\mathbf{X}}(\mathbf{x}_2) \cdots f_{\mathbf{X}}(\mathbf{x}_n) = \prod_{i=1}^{n} f_{\mathbf{X}}(\mathbf{x}_i).$$

Typically, we assume the pdf of each $\mathbf{x}_i$ is a member of a parametric family like those introduced in earlier lectures. Denote parameters of the pdf of each $\mathbf{x}_i$ by θ. We write the pdf of each $\mathbf{x}_i$ as $f_{\mathbf{X}}(\mathbf{x}, \theta)$ or $f_{\mathbf{X}}(\mathbf{x}|\theta)$. The joint pdf is

$$f_{\mathbf{X}}(\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n, \theta) = \prod_{i=1}^{n} f_{\mathbf{X}}(\mathbf{x}_i, \theta).$$

Notice that the parameters are the same for each observation – this is the "identical" part of the iid assumption. We typically think of parameters θ as being unknown. The goal of parametric statistical inference is to use the observed data to learn about θ.

Example 2 *Let $X_1, ..., X_n$ be a random sample from an exponential population with parameter β. Suppose these are a sample of n unemployed individuals, and each X_i measures how long it takes person i to find a job (measured in weeks). The iid assumption requires that the process governing unemployment duration is the same for everyone in the sample. Since*

$E[X] = \beta$ when X has an exponential distribution, the parameter β measures the average unemployment duration. The joint pdf of the sample is

$$f_X(x_1, ..., x_n, \beta) = \prod_{i=1}^{n} f_X(x_i, \beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-(x_1+\cdots+x_n)/\beta}.$$

We can use the joint pdf to answer questions about the sample. For example, what is the probability that all unemployment spells last longer than 4 weeks? We can compute

$$\begin{aligned}
& \Pr(X_1 > 4, ..., X_n > 4) \\
= & \int_4^\infty \cdots \int_4^\infty \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} dx_1 \cdots dx_n && (1) \\
= & e^{-4/\beta} \int_4^\infty \cdots \int_4^\infty \frac{1}{\beta} e^{-x_i/\beta} dx_2 \cdots dx_n \qquad \textit{integrate out } x_1 && (2) \\
& \vdots \qquad \textit{successively integrate out the remaining } x_i && (3) \\
= & \left(e^{-4/\beta}\right)^n \\
= & e^{-4n/\beta}.
\end{aligned}$$

If β, the average unemployment duration, is large, we see that this probability is near 1. How could we use observed data to learn about the population? Since $E[X] = \beta$, we could estimate β using the average observed unemployment duration in the sample.

Statistics

A statistic is just a function of the data. Formally,

Definition 3 *Let $X_1, ..., X_n$ be a random sample of size n from a population, and let $T(x_1, ..., x_n)$ be a real-valued (or vector-valued) function whose domain includes the sample space of $X_1, ..., X_n$ (i.e., the set of possible values for the X_i). Then the random variable (or random vector) $Y = T(X_1, ..., X_n)$ is called a **statistic**. The probability distribution of a statistic Y is called the **sampling distribution of Y**.*

This definition is pretty broad. The only real restriction it imposes is that a statistic is not a function of parameters. Chances are you're pretty familiar with a number of common statistics. Definitions of some of the most common follow.

Definition 4 *The **sample mean** is the arithmetic average of values in a random sample. It is usually denoted $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.*

Definition 5 *The **sample variance** is the statistic defined by $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. The **sample standard deviation** is defined by $s = \sqrt{s^2}$.*

Definition 6 *The **sample covariance** is the statistic defined by $s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$. The **sample correlation** is the statistic defined by*

$$r_{XY} = \frac{s_{XY}}{s_x s_y}.$$

Note $-1 \leq r_{XY} \leq 1$.

Properties of the Sample Mean and Variance

Theorem 7 (Sampling Distribution of the Sample Mean) *Let $X_1, ..., X_n$ be a random sample of size n from a population with mean μ and variance $\sigma^2 < \infty$. Then $E[\bar{x}] = \mu$ and $Var[\bar{x}] = \frac{\sigma^2}{n}$.*

Proof. Recall that the expectation operator is a linear operator. Thus

$$E[\bar{x}] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}nE[X_1] = \mu.$$

Using $Var[aX + bY] = a^2Var[X] + b^2Var[Y] + 2abCov[X, Y]$, and recalling that the X_i are iid, we have

$$\begin{aligned} Var[\bar{x}] &= Var\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}Var\left[\sum_{i=1}^{n} x_i\right] \\ &= \frac{1}{n^2}\left\{\sum_{i=1}^{n} Var[x_i] + 2\sum_{i=1}^{n}\sum_{j\neq i} Cov[x_i, x_j]\right\} \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var[x_i] = \frac{1}{n^2}nVar[X_1] = \frac{\sigma^2}{n}. \end{aligned}$$

■

Now a useful algebraic result.

Theorem 8 *Let $x_1, ..., x_n$ be any numbers and $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then*

$$(n-1)s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

Proof. Expand the left hand side to get

$$\begin{aligned} \sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n}\left(x_i^2 - 2x_i\bar{x} + \bar{x}^2\right) = \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2 \\ &= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2. \end{aligned}$$

■

A similar proof (try this!) can be used to show:

$$(n-1)s_{XY} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}.$$

Theorem 9 (Sample Variance is Unbiased) *Let $X_1, ..., X_n$ be a random sample of size n from a population with mean μ and variance $\sigma^2 < \infty$. Then $E[s^2] = \sigma^2$.*

Proof. Using Theorem 8 and $E[X^2] = Var[X] + E[X]^2$, we have

$$\begin{aligned} E\left[s^2\right] &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right] = \frac{1}{n-1}\left(\sum_{i=1}^{n}E\left[x_i^2\right] - nE\left[\bar{x}^2\right]\right) \\ &= \frac{1}{n-1}\left(n\left(\sigma^2+\mu^2\right) - n\left(\frac{\sigma^2}{n}+\mu^2\right)\right) = \sigma^2. \end{aligned}$$

■

Theorem 10 (Sampling Distribution of $\bar{x}$ and s^2 Under Normality) *Let $X_1, ..., X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. Let $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$. Then*
a. $\bar{x}$ and s^2 are independent
b. $\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
c. $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$.

Proof. Left for an exercise. ■
Part b of Theorem 10 implies

$$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

In the first lecture, we saw that if $Z \sim N(0,1)$, and $X \sim \chi^2_\nu$ is independent of Z, then $t = Z/\sqrt{X/\nu}$ has a t distribution with ν degrees of freedom. Given the independence result (part a) and the sampling distribution of $(n-1)s^2/\sigma^2$ (part c), we see that

$$\frac{\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\left((n-1)s^2/\sigma^2\right)/(n-1)}} = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1}.$$

This is in fact how "Student" (his real name was William Gosset) derived the t distribution in the first place. It is the basis of the usual "t test of significance."

Data Reduction

Any statistic T defines a kind of data reduction or data summary. It consequently entails discarding some sample information. In fact, this is usually point of constructing a statistic in the first place: to summarize sample information about a parameter of interest, θ. One objective in doing so is to construct statistics that do not discard *valuable* information about θ. In principle there is little cost to discarding information that does not contribute to our knowledge of θ. There are three commonly adopted principles for data reduction: the sufficiency principle, the likelihood principle, and the invariance principle. We'll discuss the first two.

The Sufficiency Principle

Intuitively speaking, a **sufficient statistic** for a parameter θ is a statistic that captures all the information about θ contained in the sample. This leads to a data reduction principle called the sufficiency principle. To conserve notation somewhat, we'll use $\mathbf{X} = \mathbf{X}_1, ..., \mathbf{X}_n$ to denote a sample, and $\mathbf{x} = \mathbf{x}_1, ..., \mathbf{x}_n$ to denote a realization of the sample.

Definition 11 (The Sufficiency Principle) *If $T(\mathbf{X})$ is a sufficient statistic for θ then any inference about θ should depend on the sample $\mathbf{X}$ only through the value $T(\mathbf{X})$. That is, if $\mathbf{x}$ and $\mathbf{y}$ are two realizations of the sample such that $T(\mathbf{x}) = T(\mathbf{y})$ then the inference about θ should be the same whether $\mathbf{X} = \mathbf{x}$ or $\mathbf{X} = \mathbf{y}$ is observed.*

This principle motivates a formal definition of a sufficient statistic.

Definition 12 *A statistic $T(\mathbf{X})$ is a **sufficient statistic for** θ if the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{X})$ does not depend on θ.*

What exactly does this mean? Consider the following, which applies in the discrete case (an analogous argument for the continuous case requires a more sophisticated notion of conditional probability).

Suppose t is a possible value of $T(\mathbf{X})$, i.e., a value such that $\Pr(T(\mathbf{X}) = t|\theta) > 0$. Definition 12 concerns conditional probabilities of the form $\Pr(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t, \theta)$. If $\mathbf{x}$ is a sample value such that $T(\mathbf{x}) \neq t$, then $\Pr(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = t, \theta) = 0$. Hence the interesting case is where $\Pr(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}), \theta) > 0$. In this case, $T(\mathbf{X})$ is a sufficient statistic if $\Pr(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}), \theta)$ is the same for all values of θ, i.e., $\Pr(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}), \theta) = \Pr(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$ for all θ.

Consider the following example (again, for the discrete case). Suppose Researcher 1 observes $\mathbf{X} = \mathbf{x}$ and computes $T(\mathbf{X}) = T(\mathbf{x})$. To make an inference about θ, she can use the information that $\mathbf{X} = \mathbf{x}$ and $T(\mathbf{X}) = T(\mathbf{x})$. Now suppose that Researcher 2 does not observe the sample directly, but only observes $T(\mathbf{X}) = T(\mathbf{x})$. Researcher 2 can use this information and knowledge of the joint distribution of $\mathbf{X}$ and $T(\mathbf{X})$ – specifically, knowledge of $\Pr(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$ – to make an inference about θ. How? Using some random number generator (e.g., a computer), Researcher 2 can simulate a sample $\mathbf{Y}$ such that $\Pr(\mathbf{Y} = \mathbf{y}|T(\mathbf{X}) = T(\mathbf{x})) = \Pr(\mathbf{X} = \mathbf{y}|T(\mathbf{X}) = T(\mathbf{x}))$ by sampling from the conditional distribution of the data given the value of the statistic. If $T(\mathbf{X})$ is a sufficient statistic for θ, then the simulated sample $\mathbf{Y}$ contains all the same information about θ that the real sample $\mathbf{X}$ does – so both researchers have the sample information about θ.

To complete this argument, we need to show that $\mathbf{X}$ and $\mathbf{Y}$ have the same probability distribution, that is, $\Pr(\mathbf{X} = \mathbf{x}|\theta) = \Pr(\mathbf{Y} = \mathbf{x}|\theta)$ for all $\mathbf{x}$ and θ. First note that the events

$\{\mathbf{X}=\mathbf{x}\}$ and $\{\mathbf{Y}=\mathbf{x}\}$ are subsets of the event $\{T(\mathbf{X})=T(\mathbf{x})\}$. Therefore:

$$\begin{aligned}
\Pr(\mathbf{X}=\mathbf{x}|\theta) &= \Pr(\mathbf{X}=\mathbf{x}, T(\mathbf{X})=T(\mathbf{x})|\theta) \\
&= \Pr(\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x}),\theta)\Pr(T(\mathbf{X})=T(\mathbf{x})|\theta) \\
&= \Pr(\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x}))\Pr(T(\mathbf{X})=T(\mathbf{x})|\theta) \quad \text{sufficiency} \quad (4) \\
&= \Pr(\mathbf{Y}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x}))\Pr(T(\mathbf{X})=T(\mathbf{x})|\theta) \\
&= \Pr(\mathbf{Y}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x}),\theta)\Pr(T(\mathbf{X})=T(\mathbf{x})|\theta) \quad \text{sufficiency} \quad (5) \\
&= \Pr(\mathbf{Y}=\mathbf{x}, T(\mathbf{X})=T(\mathbf{x})|\theta) \\
&= \Pr(\mathbf{Y}=\mathbf{x}|\theta)
\end{aligned}$$

which confirms that $\mathbf{X}$ and $\mathbf{Y}$ have the same probability distribution given θ, and hence sample realizations $\mathbf{X}=\mathbf{x}$ and $\mathbf{Y}=\mathbf{y}$ contain the same information about θ.

How do we verify whether a statistic is sufficient? We need to verify that $\Pr(\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x}),\theta)$ is the same for all values of θ. Notice that

$$\Pr(\mathbf{X}=\mathbf{x}|T(\mathbf{X})=T(\mathbf{x}),\theta)=\frac{\Pr(\mathbf{X}=\mathbf{x}, T(\mathbf{X})=T(\mathbf{x})|\theta)}{\Pr(T(\mathbf{X})=T(\mathbf{x})|\theta)}=\frac{\Pr(\mathbf{X}=\mathbf{x}|\theta)}{\Pr(T(\mathbf{X})=T(\mathbf{x})|\theta)}.$$

So one way to verify whether T is sufficient is examine this ratio: it should be constant as θ varies. This gives us the following theorem.

Theorem 13 *If $f_{\mathbf{X}}(\mathbf{x}|\theta)$ is the joint pdf of the sample $\mathbf{X}$ and $f_T(T(\mathbf{x})|\theta)$ is the sampling distribution of $T(\mathbf{X})$, then $T(\mathbf{X})$ is a sufficient statistic for θ if and only if for every possible realization $\mathbf{x}$ the ratio $f_{\mathbf{X}}(\mathbf{x}|\theta)/f_T(T(\mathbf{x})|\theta)$ is constant as a function of θ.*

Example 14 *Let $X_1,...,X_n$ be an iid sample from a $N(\mu,\sigma^2)$ distribution where σ^2 is known. We'll show that the sample mean is a sufficient statistic for μ. The joint pdf of the sample is*

$$\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}|\mu) &= \prod_{i=1}^{n}\left(2\pi\sigma^2\right)^{-1/2}\exp\left(-(x_i-\mu)^2/2\sigma^2\right) \\
&= \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2\right) \\
&= \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-\mu)^2/2\sigma^2\right) \\
&= \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\left(\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2\right)/2\sigma^2\right)
\end{aligned}$$

where the last equality follows because the cross product term is

$$2\sum_{i=1}^{n}(x_i-\bar{x})(\bar{x}-\mu)=2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\bar{x})=0.$$

Now recall the sampling distribution of the sample mean under normality: $\bar{x} \sim N(\mu, \sigma^2/n)$. *Therefore*

$$\begin{aligned}\frac{f_{\mathbf{X}}(\mathbf{x}|\mu)}{f_{\bar{X}}(\bar{x}|\mu)} &= \frac{(2\pi\sigma^2)^{-n/2}\exp\left(-\left(\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2\right)/2\sigma^2\right)}{(2\pi\sigma^2/n)^{-1/2}\exp\left(-n(\bar{x}-\mu)^2/2\sigma^2\right)} \\ &= n^{-1/2}\left(2\pi\sigma^2\right)^{-(n-1)/2}\exp\left(-\sum_{i=1}^{n}(x_i-\bar{x})^2/2\sigma^2\right)\end{aligned}$$

which does not depend on μ. *Hence* $\bar{x}$ *is a sufficient statistic for* μ.

This can be a pretty cumbersome way of determining whether a statistic is sufficient. The following theorem (see if you can prove it!) gives us an easier way to verify sufficiency.

Theorem 15 (Factorization) *Let* $f_{\mathbf{X}}(\mathbf{x}|\theta)$ *denote the joint pdf of the sample* $\mathbf{X}$. *A statistic* $T(\mathbf{X})$ *is a sufficient statistic for* θ *if and only if there exist functions* $g(t|\theta)$ *and* $h(\mathbf{x})$ *such that for all possible realizations of the sample* $\mathbf{x}$ *and all possible parameter values* θ,

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = g(T(\mathbf{X})|\theta)\,h(\mathbf{x})$$

Example 16 *Return to the case of the sample mean under normality. We saw*

$$\begin{aligned}f_{\mathbf{X}}(\mathbf{x}|\mu) &= \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\left(\sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-\mu)^2\right)/2\sigma^2\right) \\ &= \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\sum_{i=1}^{n}(x_i-\bar{x})^2/2\sigma^2\right)\exp\left(-n(\bar{x}-\mu)^2/2\sigma^2\right).\end{aligned}$$

Define

$$\begin{aligned}h(x) &= \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\sum_{i=1}^{n}(x_i-\bar{x})^2/2\sigma^2\right) \\ g(\bar{x}|\mu) &= \exp\left(-n(\bar{x}-\mu)^2/2\sigma^2\right)\end{aligned}$$

then the factorization theorem implies $\bar{x}$ *is a sufficient statistic for* μ.

The Likelihood Principle

Here we introduce a very important statistic that you've probably seen before: the likelihood function. We'll see it often in this course. Here, we'll use it as the basis of a data reduction principle.

Definition 17 *Let* $f_{\mathbf{X}}(\mathbf{x}|\theta)$ *denote the joint pdf of the sample* $\mathbf{X}$. *Then, given that* $\mathbf{X} = \mathbf{x}$ *is observed, the* ***likelihood function*** *is the function of* θ *defined by* $L(\theta|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|\theta)$.

The distinction between the joint density of the sample and the likelihood function is a subtle one: $f_{\mathbf{X}}(\mathbf{x}|\theta)$ is a function of the sample data conditional on parameters θ, whereas $L(\theta|\mathbf{x})$ is regarded as a function of the parameters for given data. To say that $L(\theta_1|\mathbf{x}) > L(\theta_2|\mathbf{x})$ is to say that the observed sample $\mathbf{X} = \mathbf{x}$ is more likely to have occurred if $\theta = \theta_1$ than if $\theta = \theta_2$. This can be interpreted as information that θ_1 is a more plausible value for θ than θ_2 is. You've probably seen this used to motivate the maximum likelihood estimator before (the parameter value that maximizes the likelihood function for the observed sample). It also motivates a data reduction principle.

Definition 18 (The Likelihood Principle) *If* $\mathbf{x}$ *and* $\mathbf{y}$ *are two possible realizations of the sample such that* $L(\theta|\mathbf{x})$ *is proportional to* $L(\theta|\mathbf{y})$*, that is, if there exists a* $C(\mathbf{x}, \mathbf{y})$ *such that*

$$L(\theta|\mathbf{x}) = C(\mathbf{x}, \mathbf{y}) L(\theta|\mathbf{y}) \qquad \text{for all } \theta, \tag{6}$$

then the conclusions drawn from $\mathbf{x}$ *and* $\mathbf{y}$ *should be identical.*

The intuition is straightforward. Suppose we find that $L(\theta_2|\mathbf{x}) = 2L(\theta_1|\mathbf{x})$ in some sample $\mathbf{x}$. This tells us that the parameter value θ_2 is twice as plausible as θ_1 in some sense. If the condition of the likelihood principle holds, i.e., equation (6) is satisfied, then $L(\theta_2|\mathbf{y}) = 2L(\theta_1|\mathbf{y})$ also. Thus whether the sample realization is $\mathbf{x}$ or $\mathbf{y}$, we conclude that θ_2 is twice as plausible as θ_1.

Finite Sample Inference

As mentioned previously, the goal of statistical inference is to use sample data to infer the value of unknown parameters θ. Most of the time we are interested in what's called a **point estimate**: a statistic that gives a single value for an unknown parameter. However, because a point estimate is based only on the observed sample (not the population) there is always some probability that the true parameter value differs from the estimate. More precisely, any statistic $T(\mathbf{X})$ is a random variable, and hence has a probability distribution. We call this the **sampling distribution** of T. It is distinct from the distribution of the population, that is, the marginal distribution of each $\mathbf{X}_i$. We call the standard deviation of the sampling distribution the **standard error** of the point estimate (its square is called the **sampling variance**). Sometimes we are interested in an **interval estimate** rather than a point estimate. An interval estimate is an interval that contains the true parameter value with known probability.

Point Estimation

An **estimator** is a rule for using the sample data to estimate the unknown parameter. We use point estimators to obtain point estimates, and interval estimators to obtain interval estimates. A **point estimator** of θ is any function $\hat{\theta}(X_1, ..., X_n)$ of the sample. Thus any statistic is a point estimator.

The definition of a point estimator is very general. As a consequence, there are typically many candidate estimators for a parameter of interest. Of course some are better than

others. What does it mean for one estimator to be "better" than another? We use a variety of criteria to evaluate them. Some are based on the **finite sample properties** of the estimator: attributes that can be compared regardless of sample size. Other criteria are based on the **asymptotic properties** of the estimator: attributes of the estimator in (infinitely) large samples. For the moment we'll restrict attention to finite sample properties. The finite sample properties we are most often concerned with are **bias**, **efficiency**, and **mean-squared error**.

Definition 19 *The **bias** of a point estimator $\hat{\theta}$ of parameter θ is $E\left[\hat{\theta}-\theta\right]$. We call a point estimator **unbiased** if $E\left[\hat{\theta}-\theta\right]=0$, so that $E\left[\hat{\theta}\right]=\theta$.*

Example 20 *Let $X_1,...,X_n$ be a random sample from a population of size n with mean μ and variance $\sigma^2<\infty$. The statistics $\bar{x}$ and s^2 are unbiased estimators of μ and σ^2 since $E[\bar{x}]=\mu$ and $E[s^2]=\sigma^2$ (see Theorems 7 and 9 above). An alternative estimator for σ^2 is the maximum likelihood estimator,*

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

This estimator is biased, since

$$E\left[\hat{\sigma}^2\right]=E\left[\frac{n-1}{n}s^2\right]=\frac{n-1}{n}\sigma^2<\sigma^2.$$

Unbiasedness is not a particularly strong criterion. In particular, there are many unbiased estimators that make poor use of the data. For example, in a random sample $X_1,...,X_n$, from a population with mean μ, each of the observed x_i is an unbiased estimator of μ. However, this estimator wastes a lot of information. We need another criterion to compare unbiased estimators.

Definition 21 *An unbiased estimator $\hat{\theta}_1$ is more **efficient** than another unbiased estimator $\hat{\theta}_2$ if $Var\left[\hat{\theta}_1\right]<Var\left[\hat{\theta}_2\right]$. If θ is vector-valued, then $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if $Var\left[\hat{\theta}_2\right]-Var\left[\hat{\theta}_1\right]$ is positive definite.*

We prefer more efficient estimators because they are more precise. Figure 1 makes this clear.

Definition 22 *The **mean-squared error** (MSE) of an estimator $\hat{\theta}$ of parameter θ is*

$$\begin{aligned}MSE\left[\hat{\theta}\right] &= E\left[\left(\hat{\theta}-\theta\right)^2\right]\\ &= Var\left[\hat{\theta}\right]+Bias\left[\hat{\theta}\right]^2.\end{aligned}$$

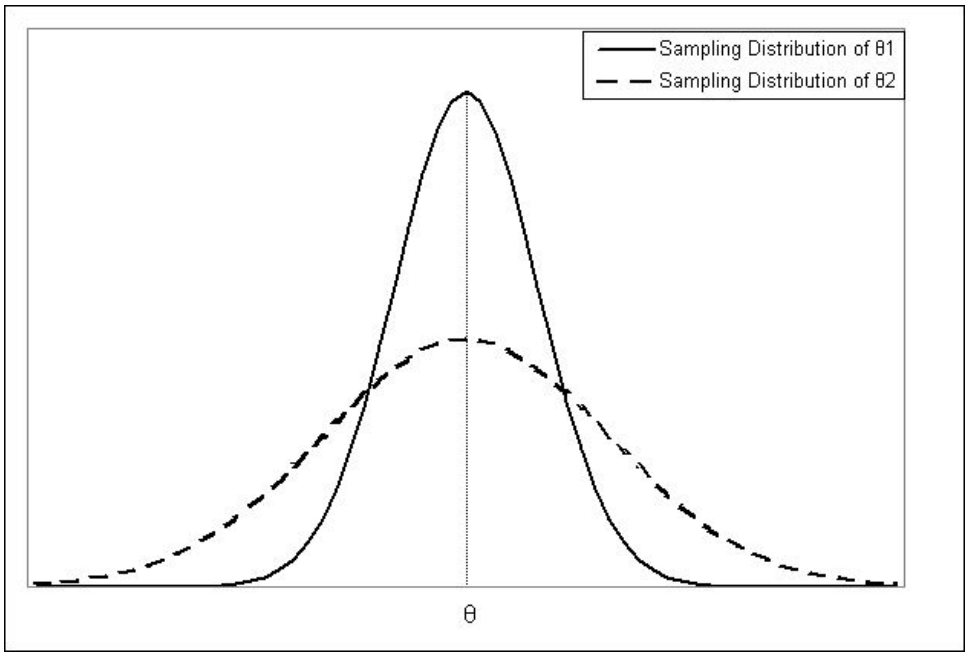

Figure 1: $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$

Example 23 *How do the MSE of s^2 and $\hat{\sigma}^2$ compare? We know from Theorem 10 that $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ under normality. We also know that s^2 is unbiased. Thus $MSE[s^2] = Var[s^2] = 2\sigma^4/(n-1)$. (This uses the result that if $X \sim \chi^2_\nu$, then $Var[X] = 2\nu$. We saw this in Lecture 1). We know from Example 20 that $E[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$, so $Bias[\hat{\sigma}^2] = -\frac{1}{n}\sigma^2$. Since $\hat{\sigma}^2 = \frac{n-1}{n}s^2$, we know $Var[\hat{\sigma}^2] = \left(\frac{n-1}{n}\right)^2 Var[s^2] = \left(\frac{n-1}{n}\right)^2 2\sigma^4/(n-1)$. Therefore*

$$\begin{aligned} MSE\left[\hat{\sigma}^2\right] - MSE\left[s^2\right] &= \left(\frac{n-1}{n}\right)^2 \frac{2\sigma^4}{n-1} + \frac{\sigma^4}{n^2} - \frac{2\sigma^4}{n-1} \\ &= \sigma^4\left(\frac{2n-1}{n^2} - \frac{2}{n-1}\right) < 0 \end{aligned}$$

and the biased estimator has a smaller MSE.

Interval Estimation

The idea behind interval estimation is as follows: use the sample data $\mathbf{X}$ to construct an interval $[L(\mathbf{X}), U(\mathbf{X})]$, that contains the true parameter value with some known probability. The **coverage probability** of an interval estimate is the probability that the interval $[L(\mathbf{X}), U(\mathbf{X})]$ contains the true parameter value. This is sometimes called a **confidence level**.

Example 24 *We know that when sampling from the normal distribution,*

$$z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$

We can use this to construct a confidence interval around the true mean μ. Given that we know the probability distribution of z, we can always make statements like

$$\Pr\left(L \leq \frac{\bar{x}-\mu}{s/\sqrt{n}} \leq U\right) = 1-\alpha. \tag{7}$$

If we know the desired values of L and U, then we can just look up the appropriate α from a table of the "critical values" for the t distribution. Conversely, if we choose α in advance, we can look up appropriate values of L and U. Suppose we are interested in a symmetric interval around μ, so that $L = -U$. Then we can rearrange (7) as follows

$$\Pr\left(\bar{x} - \frac{Us}{\sqrt{n}} \leq \mu \leq \bar{x} + \frac{Us}{\sqrt{n}}\right) = 1-\alpha.$$

This is a probability statement about the interval, not the parameter. It is the interval that is random, not the parameter. If we drew many repeated samples from this same population, the interval would be different each time, but μ would be the same. Consequently, we attach a probability (the coverage probability or confidence level) to the interval itself. In repeated sampling, we would expect an interval constructed this way to contain μ in 100$(1-\alpha)$ percent of the samples. Here, $L(\mathbf{X}) = \bar{x} - Us/\sqrt{n}$ and $U(\mathbf{X}) = \bar{x} + Us/\sqrt{n}$.

Example 25 *Suppose that in a sample of 36 observations we compute $\bar{x} = 1.81$ and $s = 0.42$. Assume the sample is drawn from a normal distribution. Let's construct a symmetric 95 percent confidence interval for μ. To do this, we look up the values L and U such that $\Pr[T_{35} \geq L] = 0.025$ and $\Pr[T_{35} \leq U] = 0.975$, where $T_{35} \sim t_{35}$. These are called critical values, and you can verify that $-L = U = 2.03$. Thus,*

$$\Pr\left(-2.03 \leq \frac{1.81-\mu}{0.42/6} \leq 2.03\right) = 0.95$$

and the 95 percent confidence interval is $1.81 \pm [2.03\,(0.42)/6] = 1.81 \pm 0.1421 = [1.6679, 1.9521]$.

Example 26 *Now suppose we want to construct a 95 percent confidence interval around σ^2 from the preceding example. Since $(n-1)\,s^2/\sigma^2 \sim \chi^2_{n-1}$, we know*

$$\Pr\left(20.57 \leq \frac{35\,(0.42)^2}{\sigma^2} \leq 53.2\right) = 0.95$$

where 20.57 and 53.2 are the 0.025 and 0.975 critical values from the χ^2_{35} distribution. The 95 percent confidence interval is thus $[0.116, 0.300]$.

Hypothesis Testing

The goal of classical hypothesis testing is to determine, with some degree of confidence, whether our econometric model is consistent with the observed data. That is, whether our sample could have been generated by the hypothesized population. We do so by constructing

a statistic, called a **test statistic**, and comparing its value to the set of values we could reasonably expect it to take if the data were in fact generated by the hypothesized population. We formalize this by means of two hypotheses: the null (or maintained) hypothesis H_0, and the alternative hypothesis H_1. A hypothesis test is a rule, stated in terms of the data, that dictates whether or not the null hypothesis should be rejected. For example, the null hypothesis might state that a parameter equals a specific value, e.g., $\theta = 0$. The alternative is then $\theta \neq 0$. We would then reject the null hypothesis if a sample estimate $\hat{\theta}$ differs greatly from zero. The classical approach is to divide the sample space into two regions: the **acceptance region** and the **rejection region**. If the test statistic falls in the rejection region then we reject the null hypothesis. If it falls in the acceptance region, we do not reject the null.

Like all statistics, a test statistic is a random variable. Thus it has a sampling distribution, and there is always some possibility that the result of our hypothesis test is erroneous. Two kinds of error are possible:

Definition 27 *A **type I error** occurs if we reject the null hypothesis when it is true. A **type II error** occurs if we fail to reject the null hypothesis when it is false.*

Definition 28 *The probability of a type I error is called the **size** of a test, sometimes called the **significance level** of the test. We usually denote it α.*

The analyst can control the size of the test by adjusting the decision rule to reject the null. Of course there is a trade-off – as we reduce the probability of making a type I error by making the rejection region smaller, we increase the probability of making a type II error. Typically, we look for tests that minimize the probability of a type II error for a given probability of a type I error.

Definition 29 *The **power** of a test is the probability that it leads to rejection of the null hypothesis when it is in fact false. Thus,*

$$power = 1 - \beta = 1 - \Pr\left(type\ II\ error\right).$$

Definition 30 *A test is **most powerful** if it has greater power than any other test of the same size.*

Powerful tests are good, but power depends on the alternative hypothesis. We may not always be able to find a most powerful test. Instead, we frequently look for tests that are **unbiased** and/or **consistent.** We say a test is unbiased if $(1 - \beta) \geq \alpha$ for all values of the unknown parameter. That is, we are more likely to reject the null when it is false than when it is true (a pretty weak requirement). If a test is biased, then for some values of the parameter we are more likely to reject the null when it is true than when it is false. We say a test is consistent if its power approaches 1 as the sample size goes to infinity. We'll discuss these ideas in greater detail when we get into asymptotic theory.