Math Camp 2017 - Statistics[*]

James Banovetz

Department of Economics, UC Santa Barbara

September 15, 2017

1. **Random Samples**

(a) Definition. The random variables $X_1, \dots, X_n$ are called a **random sample** of size n from the population $f(x)$ if $X_1, \dots X_n$ are mutually independent random variables and the marginal PDF (or PMF) of each X_i is the same $f_X(x)$. Alternatively, we say that $X_1, \dots, X_n$ are **independent an identically distributed** (or i.i.d.).

(b) Definition. From our probability sections, recall that the **joint PDF** (or **PMF**) of a random sample $X_1, \dots, X_n$ is given by

$$f_{\mathbf{X}}(x_1, \dots, x_n) = \prod_{i=1}^{n} f_X(x_i)$$

(c) Example. Suppose a coin flip lands on heads with probability p. If we flip a coin n times, we can find the joint distribution be first defining the individual RV and PMF:

$$X_i = \begin{cases} 1 & \text{if heads} \\ 0 & \text{if tails} \end{cases} \qquad \text{(defining the RV)}$$

$$f_{X_i}(x) = \begin{cases} p^{x_i}(1-p)^{1-x_i} & \text{if } x_i \in \{0,1\} \\ 0 & \text{else} \end{cases} \qquad \text{(the PMF of } X_i\text{)}$$

Then the joint distributions is:

$$f_{\mathbf{X}}(x_1, \dots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i}, \qquad x_i \in \{0,1\} \text{ for } i = 1, \dots n$$

(d) Aside. For the rest of math camp, we'll be assuming that we're working with i.i.d. samples. In reality, you'll virtually never see a random sample with real data. Many of the results and principles we use here, however, will still hold with somewhat weaker assumptions. You'll touch on the weaker assumptions come winter quarter (Econ 241B).

[*] These lecture notes are drawn principally from *Statistical Inference*, 2nd ed., by George Casella and Roger L. Berger. The material posted on this website is for personal use only and is not intended for reproduction, distribution, or citation.

2. **Statistics**

   (a) Definition. Let $X_1, \ldots, X_n$ be a random sample. Let $T(x_1, \ldots, x_n)$ be a real-valued or vector valued function. Then the random variable $Y = T(X_1, \ldots, X_n)$ is a **statistic**.

   (b) Example. The most common statistic we see is the sample mean:

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

   Another extremely common statistic is the sample variance

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

   Note that there are infinitely many statistics that could come up with, including trivial statistics like X_1 or $X_1, \ldots, X_n$, i.e., the whole sample itself.

   (c) Definition. Suppose we have a statistic $Y = T(X_1, \ldots, X_n)$. Then the probability distribution of the statistic Y is called the **sampling distribution** of Y.

   (d) Example. Recall the distribution of coin flips from before. Suppose we're interested in the sum, $Y = \sum X_i$ (i.e., the number of heads observed). Intuitively, for a particular observed sample $x_1, \ldots, x_n$ that produces $y = \sum x_i$, the probability of the sample is

$$p^{\sum x_i}(1-p)^{n-\sum x_i} = p^y(1-p)^{n-y}$$

   There are potentially many different samples however, that would produce $y = \sum x_i$. In fact, there are $\binom{n}{y}$ (i.e., n coin flips and y heads are observed). Thus, the PMF of Y is

$$f_Y(y) = \binom{n}{y} p^y (1-p)^{n-y}$$

   Which is the binomial distribution. Note that we could prove that the sum of independent Bernoulli RVs is distributed binomial using MGFs.

   (e) Theorem. Let $X_1, \ldots, X_n$ be i.i.d from a distribution with mean μ and variance $\sigma^2 < \infty$. Then:

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

   (f) Aside. These are very useful theorems, and are fairly easy to prove. Throughout the first year, you'll take these as given (unless explicitly told otherwise).

3. **Sampling from a Normal Distribution**.

   (a) Theorem. Let $X_1, \ldots, X_n$ be i.i.d. from a $N(\mu, \sigma^2)$ distribution. Let $\bar{X} = \frac{1}{n} \sum X_i$ and let $S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$. Then:

   - $\bar{X}$ is distributed $N(\mu, \sigma^2/n)$
   - $\frac{(n-1)S^2}{\sigma^2}$ is distributed $\chi^2_{(n-1)}$
   - $\bar{X}$ and S^2 are independent

(b) Aside. We won't spend the time to prove all of these now. There is some intuition, however, behind these results.

- The first point is easy to prove using MGFs and is a common result.
- Notice that
$$\frac{(n-1)S^2}{\sigma^2} = \frac{(n-1)}{\sigma^2(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sigma}\right)^2$$
While not technically the same thing, this is very close to the sum of squared standard normal RVs: $Z = \left(\frac{X_i - \mu}{\sigma}\right)$. Note we lose a degree of freedom because we're estimating μ with $\bar{X}$.
- The third point is not immediately apparent. With normal distributions, however, the mean and variance independently describe the distribution (e.g., μ tells us where the distribution is centered, while σ^2 tells us how spread out the distribution is). The sample moments do the same.

(c) Aside. Why do we care about $\bar{X}$ and S^2 from normal distribution so much? First off, much of our econometrics makes the assumption that error terms are i.i.d. normal. Second, we have the CLT, where the mean of a distribution behaves like a normally distributed random variable in the limit.

(d) Theorem. Let $X_1, \ldots, X_n$ be i.i.d. from a $N(\mu_x, \sigma_x^2)$ and $Y_1, \ldots, Y_m$ be i.i.d. from a $N(\mu_y, \sigma_y)$. Consider the following statistics:

i. $\dfrac{\bar{X} - \mu}{\sqrt{S^2/n}}$ has a student's t distribution with $n-1$ degrees of freedom.

ii. $\dfrac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2}$ has an F distribution with $n-1$ and $m-1$ degrees of freedom.

(e) Aside. We saw these previously in our discussion of transformations. While they aren't used daily in econometrics, they are well-known results that you are expected to have seen in undergraduate statistics.

4. **Order Statistics**

(a) Definition. The **order statistics** of a random sample $X_1, \ldots, X_n$ are the sample values placed in ascending order, denoted by
$$X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \cdots \leq X_{(n)}$$
$X_{(1)}$ is known as the **sample minimum**. $X_{(n)}$ is the **sample maximum**. Another common value is the **sample median:**
$$M = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(X_{(n/2)} + X_{(n/2+1)}\right) & \text{if } n \text{ is even} \end{cases}$$

(b) Aside. The median is occasionally very interesting to us, as it is not as easy skewed by extreme observations as the mean. For example, the mean of an income distribution may not be very enlightening if there are a small number of people who earn extremely high incomes.

(c) Example. Suppose we have a random sample $X_1, \dots X_n$ from a Uniform (0,1) distribution. We can find the CDF and PDF of $X_{(n)}$ using the PDF and CDF of X_i:

$$f_X(x_i) = \begin{cases} 1 & \text{if } x_i \in (0,1) \\ 0 & \text{else} \end{cases} \qquad F_X(x_i) = \begin{cases} 0 & \text{if } x_i \leq 0 \\ x_i & \text{if } 0 < x_i < 1 \\ 1 & \text{if } 1 \leq x_i \end{cases}$$

How to find the CDF? Think about the probability that $X_n < k$. If the maximum is less than k, then every value of X_i is also less than k:

$$\begin{aligned} P(X_{(n)} \leq x) &= P(X_1 \leq x, X_2 \leq x, \dots, X_n \leq x) && \text{(by def. of the max)} \\ &= P(X_1 \leq x)P(X_2 \leq x) \dots (X_n \leq x) && \text{(by independence)} \\ &= F_{X_1}(x)F_{X_2}(x) \dots F_{X_n}(x) && \text{(by def. of the CDF)} \\ &= \prod_{i=1}^{n} F_X(x) && \text{(by identically dist.)} \\ &= x^n && \text{(plugging in the CDFs)} \end{aligned}$$

If we want to be complete:

$$F_{X_n}(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x^n & \text{if } 0 < x < 1 \\ 1 & \text{if } 1 \leq x \end{cases} \qquad \text{(defining over } \mathbb{R}\text{)}$$

To find the PDF, we simply need to take the derivative:

$$f_{X_{(n)}}(x) = \frac{d}{dx} F_{X_n}(x) \qquad (F_{X_{(n)}}(x) \text{ is differentiable})$$

$$f_{X_{(n)}}(x) = \begin{cases} nx^{n-1} & \text{if } 0 < x < 1 \\ 0 & \text{else} \end{cases}$$

(d) Theorem. Let $X_1, \dots, X_n$ be a random sample, with the order statistics $X_{(1)}, \dots, X_{(n)}$, from a continuous distribution with CDF $F_X(x)$ and PDF $f_X(x)$. Then the PDF of $X_{(j)}$ is

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!(n-j)!} f_X(x) \big[F_X(x)\big]^{j-1} \big[1 - F_X(x)\big]^{n-j}$$

(e) Aside. The intuition (which is technically incorrect), can be thought through by parsing from right to left:

- "Group 3": The $n - j$ values that are greater than $X_{(j)} = x$
- "Group 2": The $j - 1$ values that are less than $X_{(j)} = x$:
- "Group 1": The value that is "equal": $X_{(j)} = x$

Now, there are $N = \frac{n!}{(j-1)!1!(n-j)!}$ ways to select the three groups. There $n!$ ways to organize n elements. The ordering within each group, however, does not matter. For each sample, there are $(j-1)!1!(n-j)!$ ways of organizing elements, so we divide to avoid over counting.

Then we have the "probability" that $X_{(j)} = x$ is $f_X(x)$ (note that it's actually zero); the probability that $j-1$ elements are less than x; and the probability that $n-j$ elements are greater than x.

Note that: you will almost certainly use the sample maximum and minimum, but otherwise you won't use this PDF much.

(f) Example. Let $X_1, \dots X_9$ be a random sample from an exponential distribution with PDF

$$f_X(x_i) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}$$

Find the distribution of the sample median.

Finding the CDF of X_i:

$$F_X(x_i) = \int_0^x \lambda e^{-\lambda t}\, dt = 1 - e^{-\lambda x}$$

The sample median will be $X_{(5)}$ (since we have an odd number of observations). Employing the theorem, we have $n = 9$, $j = 5$:

$$f_X(x_i) = \begin{cases} \frac{9!}{4!4!}\left(\lambda e^{-\lambda x}\right)\left(1 - e^{-\lambda x}\right)^4 \left(e^{-\lambda x}\right)^4 & \text{if } x \geq 0 \\ 0 & \text{else} \end{cases}$$

As you can see, this isn't very intuitive. Medians, while they have some nice properties, can be difficult to work with, particularly compared to the mean.

5. Estimation

(a) Definition. Given a sample collected from a population, **statistical interference** consists of the methods used to make inferences or generalizations about a population and its parameters based on statistics calculated from the sample data. There are a few basic categories:

- **Point Estimation**, i.e., a particular value that best approximates some parameter.
- **Interval Estimation**, i.e., an interval constructed, such that if the sample were repeated, the true population parameter would fall in the interval a certain percentage of the time.
- **Hypothesis Testing**, i.e., the process where a statistic and its distribution are observed and used to evaluate a statement made about the unknown value of the parameter.

(b) Aside. Again, we try to be clear about our notation. The parameter of interest is frequently denoted θ (which may be vector-valued), while an estimator is given by $\hat{\theta} = T(X_1, \dots, X_n)$. Again, we denote unrealized estimates (i.e., random variables) with capital letters: $\bar{X}$ or S^2; observed values of the statistic are denoted $\bar{x}$ and s^2.

6. Point Estimation

(a) Definition. Let $X_1, \dots, X_n$ be a sample from a population with $\theta_1, \dots, \theta_k$ parameters. We define the jt population (non-central) moment as

$$M_j(\theta_1, \dots, \theta_k) = \mathbb{E}[X^j]$$

and the jth (non-central) sample moment as

$$m_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j$$

Then the **method of moments** estimator $(\hat{\theta}_1, \ldots, \hat{\theta}_k)$ for $(\theta_1, \ldots, \theta_k)$ is the solution to the system

$$\begin{aligned} m_1 &= M_1(\hat{\theta}_1, \ldots, \hat{\theta}_k) \\ m_2 &= M_2(\hat{\theta}_1, \ldots, \hat{\theta}_k) \\ \vdots &= \vdots \\ m_k &= M_k(\hat{\theta}_1, \ldots, \hat{\theta}_k) \end{aligned}$$

That is, we set the sample moments equal to the population moments, then solve for θ_1 through θ_k (note that we have k equations sand k unknowns). Note that when we set them equal, the θ's become $\hat{\theta}$'s.

(b) Example. Suppose we have a random sample $X_1, \ldots, X_n$ from a normal distribution $N(\mu, \sigma^2)$. Note that $\mathbb{E}[X] = \mu$ and $\mathbb{E}[X^2] = \sigma^2 + \mu^2$. Then we can find the method of moments estimator (MME) by solving the sytem:

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \hat{\mu} \qquad \text{(first moment condition)}$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^n = \hat{\sigma}^2 + \hat{\mu}^2 \qquad \text{(second moment condition)}$$

Solving this relatively trivial system:

$$\boxed{\hat{\mu}_{mm} = \bar{X}} \qquad \text{(simplifying notation)}$$

$$\begin{aligned} \hat{\sigma}^2_{mm} &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \hat{\mu}^2 && \text{(from the second cond.)} \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \bar{X}^2 && \text{(plugging in for } \hat{\mu}\text{)} \end{aligned}$$

Now, we can play around with some algebra:

$$\begin{aligned} &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \frac{1}{n}\sum_{i=1}^{n} \bar{X}^2 && (\bar{X} \text{ is a ``constant''}) \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \frac{2}{n}\sum_{i=1}^{n} \bar{X}^2 + \frac{1}{n}\sum_{i=1}^{n} \bar{X}^2 && \text{(adding zero)} \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \left(2\bar{X}\right)\frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n} \bar{X}^2 && \text{(by def. of } \bar{X}\text{)} \\ &= \frac{1}{n}\sum_{i=1}^{n} \left(X_i^2 - 2\bar{X}X_i + \bar{X}^2\right) && \text{(writing as a single sum)} \end{aligned}$$

$$\boxed{\hat{\sigma}^2_{mm} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2} \qquad \text{(simplifying)}$$

Although it is not technically necessary, this simplification is how we typically see the method of moments estimator presented in textbooks. This simplification is not obvious to most, but comes up occasionally in statistics-based classes.

(c) Aside. While the method of moments is not used all that frequently, it is an intuitive way to begin the construction of estimators. further, although it is not used that much in economics (beyond teaching OLS in undergraduate and first-year Ph.D. programs), it does serve as the basis of *generalized method of moments* estimators, which are used heavily in the field.

(d) Definition. Let $X_1, \dots, X_n$ be a random sample with PDF (or PMF) $f_X(x_i|\theta_1, \dots, \theta_k)$. The the **likelihood function** is defined by

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{x}) = \mathcal{L}(\theta_1, \dots, \theta_k|x_1, x_2, \dots x_n) = \prod_{i=1}^{n} f_X(x_i|\theta_1, \dots, \theta_k)$$

(e) Definition. For each sample point $x_1, \dots, x_n$, let $\hat{\boldsymbol{\theta}}(x_1, \dots, x_n)$ be a parameter value at which $\mathcal{L}(\boldsymbol{\theta}|\mathbf{x})$ attains its maximum as a function of $\boldsymbol{\theta}$, holding $x_1, \dots, x_n$ fixed. A **maximum likelihood estimator** (MLE) of $\boldsymbol{\theta}$ based on sample $X_1, \dots, X_n$ is $\hat{\boldsymbol{\theta}}(X_1, \dots, X_n)$.

(f) Aside. Note that we're making a methodological change here: we're treating the values of $x_1, \dots, x_n$ are fixed, and we're varying the values $\theta_1, \dots, \theta_n$. Essentially, the intuition is, "assuming that our data comes from a particular distribution, what parameters are *most likely* given the data we observe?"

When we can use calculus (as we'll see in a few examples), this can be a straightforward exercise for well-behaved likelihoods. Although MLE is extremely popular, in practice (i.e., with real data), this can be extremely difficult and will require numerical simulations on a computer. That said, MLEs have some very nice properties and you're quite likely to use them quite a bit in research.

(g) Example. Let $X_1, \dots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. We can find the MLEs $\hat{\mu}_{mle}$ and $\hat{\sigma}^2_{mle}$ using calculus. The likelihood function is

$$\begin{aligned}
\mathcal{L}(\mu, \sigma^2|\mathbf{x}) &= \prod_{i=1}^{n} f_X(x|\mu, \sigma^2) && \text{(by def. of the likelihood func.)} \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} && \text{(plugging in PDFs)} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right\} && \text{(multiplying)}
\end{aligned}$$

Note that frequently, logarithmic transformations can make problems easier to solve. In the context of MLE problems, we refere to these as log-likelihood functions, and usually denote them $l(\cdot)$:

$$l(\mu, \sigma^2|\mathbf{x}) = -\frac{n}{2} \ln\left(2\pi\sigma^2\right) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \qquad \text{(using a log transform)}$$

Differentiating with respect to our parameters:

$$\frac{\partial l(\cdot)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \qquad \text{(differentiating w.r.t } \mu\text{)}$$

$$\frac{\partial l(\cdot)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i - \mu)^2 \qquad \text{(differentiating w.r.t. } \sigma^2\text{)}$$

The first-order conditions give us a system of two equations and two unknowns:

$$0 = -\frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(x_i - \hat{\mu}) \qquad \text{(the first FOC)}$$

$$0 = -\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 \qquad \text{(the second FOC)}$$

Solving the first FOC for $\hat{\mu}$:

$$0 = \sum_{i=1}^{n}(x_i - \hat{\mu}) \qquad \text{(multiplying by } -\hat{\sigma}^2\text{)}$$

$$0 = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \hat{\mu} \qquad \text{(distributing the sum)}$$

$$\boxed{\hat{\mu}_{mle} = \frac{1}{n}\sum_{i=1}^{n} x_i} \qquad \text{(solving for } \hat{\mu}\text{)}$$

Thus, the MLE for μ is our usual $\bar{X}$. Considering the second FOC:

$$0 = -n\hat{\sigma}^2 + \sum_{i=1}^{n}(x_i - \hat{\mu})^2 \qquad \text{(multiplying by } 2(\hat{\sigma}^2)^2\text{)}$$

$$\boxed{\hat{\sigma}^2_{mle} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{(solving for } \hat{\sigma}^2\text{)}$$

Thus, the MLE for σ^2 is the same as the estimator from method of moments (this is, more or less, a coincidence).

(h) Aside. Technically, we'd need to check second order conditions. For the first year, however, we won't do it unless explicitly told to do so. While calculus helps with some problems, there are quite a few distributions of interest where we can't use FOCs from calculus to find the MLE.

(i) Example. Let $U_1, \ldots, U_n$ be a random sample from a uniform distribution on $[0, \theta]$. Again, we can find the MLE of θ by examining the likelihood function, which here will be:

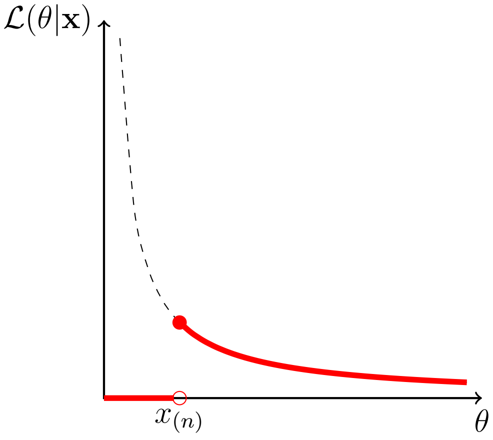
$$\mathcal{L}(\theta|\mathbf{u}) = \prod_{i=1}^{n}\left(\frac{1}{\theta}\right)\mathbb{1}_{\{0 \le x_i \le \theta\}} \qquad \text{(by def. of the likelihood func.)}$$

$$= \frac{1}{\theta^n}\prod_{i=1}^{n}\mathbb{1}_{\{0 \le x_i \le \theta\}} \qquad \text{(multiplying)}$$

Note that we can simplify the n indicator functions using the sample maximum. Every indicator will take a value of 1 so long as the sample maximum $X_{(n)} < \theta$. Thus:

$$= \frac{1}{\theta^n} \mathbb{1}_{\{0 \leq x_{(n)} \leq \theta\}} \qquad \text{(simplifying)}$$

Unfortunately, we can't take a derivative of this function w.r.t. θ (at least near $x_{(n)}$, as the function is discontinuous). We can, however, see that the likelihood is decreasing in θ, with a discontinuity at $x_{(n)}$ (displayed as the red function):



Thus, the likelihood function is maximized at $x_{(n)}$, so

$$\boxed{\hat{\theta}_{mle} = x_{(n)}}$$

(j) Theorem (CB THM 2.7.10). If $\hat{\theta}$ is the MLE for θ, then for any function $\tau(\theta)$, the MLE for $\tau(\theta)$ is $\tau(\hat{\theta})$. This is known as the invariance property of MLEs.

(k) Aside. This is an extremely useful property in certain contexts, e.g., trying to find estimators for transformations of parameters. It can save you quite a bit of time on exams, but it probably won't come up more than a handful of times in the first year.

7. **Evaluating Estimators**.

(a) Aside. MLE and MME are two common methods to obtain estimators (note that there are others). Among these and others, how should you go about choosing which estimator to use? We have a few criteria to help pick the right estimator for the right context.

(b) Definition. The **bias** of a point estimator $\hat{\theta}$ of a parameter θ is the difference between the expected value of $\hat{\theta}$ and θ:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta$$

If $\text{Bias}(\hat{\theta}) = 0$, then the estimator $\hat{\theta}$ is **unbiased**.

(c) Example. Recall the MLE of θ for a uniform $[0, \theta]$ distribution: $\hat{\theta}_{mle} = X_{(n)}$. We can figure out if this is an unbiased estimator; if it isn't we can also calculate the bias. To do so, we need to find the expected value of $X_{(n)}$. Without deriving it, the PDF of $X_{(n)}$ is given by:

$$f_{X_{(n)}}(x) = \frac{nx^{n-1}}{\theta}$$

where $x \in [0, \theta]$. Finding the expected value:

$$\begin{aligned}
\mathbb{E}[X_{(n)}] &= \int_0^\theta x \left(\frac{nx^{n-1}}{\theta^n} \right) \, dx && \text{(by def. of expected value)} \\
&= \int_0^\theta n \left(\frac{x}{\theta} \right)^n \, dx && \text{(simplifying)} \\
&= \frac{n}{\theta^n} \left[\frac{1}{n+1} x^{n+1} \right]_0^\theta && \text{(taking the integral)} \\
&= \left(\frac{n}{\theta^n} \right) \left[\frac{\theta^{n+1}}{n+1} \right] && \text{(evaluating)} \\
&= \left(\frac{n}{n+1} \right) \theta && \text{(simplifying)}
\end{aligned}$$

Thus, the estimator is not unbiased; unsurprisingly, it underestimates θ (since we can't actually observe anything greater than θ). Thus, the bias is

$$\begin{aligned}
\text{Bias}\left(\hat{\theta}_{mle} \right) &= \left(\frac{n}{n+1} \right) \theta - \theta && \text{(by def. of bias)} \\
&= - \left(\frac{\theta}{n+1} \right) && \text{(simplifying)}
\end{aligned}$$

(d) Definition. The **mean squared error** (MSE) of an estimator $\hat{\theta}$ of a parameter θ is defined as

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta} - \theta)^2\big]$$

Alternatively, this may be stated in the form:

$$\text{MSE}(\hat{\theta}) = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2$$

(e) Example. To find the MSE of $\hat{\theta}_{mle}$ from above, we need both the variance and the bias. We have the bias, so we simply need to find the variance:

$$\begin{aligned}
\text{Var}(\hat{\theta}_{mle}) &= \mathbb{E}[X_{(n)}^2] - \mathbb{E}[X_{(n)}]^2 && \text{(by def. of } \hat{\theta}_{mle} \text{ and Var)} \\
&= \int_0^\theta x^2 \frac{nx^{n-1}}{\theta^n} dx - \left[\frac{n}{n+1} \theta \right]^2 && \text{(by def. of the exp. values)} \\
&= \frac{n}{\theta^n} \int_0^\theta x^{n+1} dx - \left[\frac{n}{n+1} \theta \right]^2 && \text{(simplifying)} \\
&= \frac{n}{\theta^n} \left[\frac{x^{n+2}}{n+2} \right]_0^\theta - \left[\frac{n}{n+1} \theta \right]^2 && \text{(taking the integral)} \\
&= \frac{n\theta^2}{(n+1)^2(n+2)} && \text{(simplifying)}
\end{aligned}$$

Then, to find the MSE, we simply employ our formula from above:

$$\begin{aligned}
\text{MSE}(\hat{\theta}_{mle}) &= \left[\frac{n\theta^2}{(n+1)^2(n+2)} \right] + \left[- \frac{\theta}{n+1} \right]^2 && \text{(by def. of the MSE)} \\
&= \frac{2\theta^2}{(n+1)(n+2)} && \text{(simplifying)}
\end{aligned}$$

(f) Aside. The second formula is a more intuitive way to see why the MSE may be a good evaluator– it includes both the variance and the bias of an estimator. We want to minimize bias, to avoid sysematically over- or under-estimating our parameters; but we also want to limit variance, to avoid a wide spread of our estimators relative to the parameter.

(g) Definition. An estimator T^* is a **uniform minimum variance unbiased estimator** (UMVUE) of $\tau(\theta)$ if it satisfies $\mathbb{E}[T^*] = \tau(\theta)$ for all θ and for any other estimator T with $\mathbb{E}[T] = \tau(\theta)$, $\text{Var}(T^*) \leq \text{Var}(T)$.

(h) Aside. While this is great in theory, it would require us to evaluate every estimator that as an expected value $\tau(\theta)$. Obviously, this is not possible, so we have to develop other criteria to establish which estimators are "best." We'll start with the ideas of sufficiency and completeness.

(i) Definition. A statistic $T(X_1, \ldots, X)n)$ is a **sufficient statistic** for θ if the conditional distribution of the sample $X_1, \ldots, X_n$ given $T(X_1, \ldots, X_n)$ does not depend on θ.

(j) Aside. Intuitively, $T(X_1, \ldots, X_n)$ contains all relevant information about θ in the sample. For example, if we're considering a normal distribution, $\mathbb{E}[\bar{X}] = \mu$. The sample mean gives us all of the relevant information contained in the sample about μ.

In other words, we know all the sample can tell us about μ from just $\bar{X}$; even the whole sample $x_1, \ldots, x_n$ can't give us any more information.

(k) Theorem. If $f(\mathbf{x_1}, \ldots, \mathbf{x_n}|\theta)$ is the joint PDF or PMF of $X_1, \ldots, X_n$, and $q(t|\theta)$ is the PDF or PMF of $T(X_1, \ldots, X_n)$, then $T(X_1, \ldots, X_n)$ is a sufficient statistic for θ if for every $x_1, \ldots, x_n$ in the sample space,

$$\frac{f(x_1, \ldots, x_n|\theta)}{q(t|\theta)}$$

is constant as a function of θ.

(l) Example. Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli (p) distribution. We can show that $S = \sum_{i=1}^{n} X_i$ is a sufficient statistic for p.

We know that the distribution of the sum of i.i.d. Bernoulli RVs is distributed binomially. Thus, the ratio in question is:

$$\frac{f(x_1, \ldots, x_n|p)}{q(s|p)} = \frac{p^{\sum x_i}(1-p)^{n-\sum x_i}}{\binom{n}{s} p^s (1-p)^{n-s}}$$

Since $s = \sum x_i$ by definition, this simplifies easily:

$$= \binom{n}{s}^{-1}$$

This ratio is clearly independent of p, implying that S is a sufficient statistic for p.

(m) Theorem. Let $f(x_1, \ldots, x_n|\theta)$ be the joint PDF or PMF of a sample $X_1, \ldots, X_n$. A statistic $T(X_1, \ldots, X_n)$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(x_1, \ldots, x_n)$ such that for all points $x_1, \ldots, x_n$ and all parameter values

$$f(x_1, \ldots, x_n|\theta) = h(x_1, \ldots, x_n) g\big(T(x_1, \ldots, x_n|\theta)\big)$$

This is known as the **factorization theorem**.

(n) <u>Example</u>. Let $U_1, \ldots, U_n$ be a random sample from a uniform $[0, \theta]$ distribution. We can show that $X_{(n)}$ is a sufficient statistic for θ. The joint PDF of $U_1, \ldots, U_n$ is

$$\begin{aligned} f_{\mathbf{U}}(u_1, \ldots, u_n|\theta) &= \prod_{i=1}^{n} \left(\frac{1}{\theta}\right) \mathbb{1}\{0 \le u_i \le \theta\} && \text{(the joint PDF)} \\ &= \left(\frac{1}{\theta}\right)^n \mathbb{1}\{0 \le u_{(1)}\}\mathbb{1}\{u_{(n)} \le \theta\} && \text{(multiplying/simplifying)} \\ &= \left(\mathbb{1}\{0 \le u_{(1)}\}\right)\left(\frac{1}{\theta^n}\mathbb{1}\{u_{(n)} \le \theta\}\right) && \text{(rearranging)} \end{aligned}$$

If we name our functions according to the factorization theorem:

$$h(u_1, \ldots, u_n) = \mathbb{1}\{0 \le u_{(1)}\} \quad \text{and} \quad g(T(u_1, \ldots, u_n)|\theta) = \theta^{-n}\mathbb{1}\{u_{(n)} \le \theta\}$$

thus, $T(U_1, \ldots, U_n) = U_{(n)}$ is the sufficient statistic.

(o) <u>Example</u>. Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution. What are the sufficient statistics for μ and σ^2? Note that our sufficient statistic will be two-dimensional in this case:

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\mu, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\} && \text{(the joint PDF)} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} && \text{(multiplying)} \\ &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2 - \frac{\mu}{\sigma^2}\sum_{i=1}^{n}x_i - \frac{n\mu^2}{2\sigma^2}\right\} && \text{(expanding the exponent)} \end{aligned}$$

Again, naming our functions according to the factorization theorem:

$$g(T(X_1, \ldots, X_n)|\mu, \sigma^2) = f_{\mathbf{X}}(\mathbf{x}|\mu, \sigma^2) \quad \text{and} \quad h(X_1, \ldots, X_n) = 1$$

Then $T(\cdot)$ is a two dimensional sufficient statistic:

$$\begin{pmatrix} \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i^2 \end{pmatrix} \quad \text{are jointly sufficient for} \quad \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$$

(p) <u>Aside</u>. There may be some confusion from making $h(x_1, \ldots, x_n) = 1$. Can't this be done for every distribution? Yes, it can. Remember that the entire sample is also a sufficient statistic–the entire sample contains all relevant information in the sample.

That said, we're trying to find a way to "summarize" our data. In the case of the normal, we have a very clear $T(X_1, \ldots, X)$ in the function $g\big(T(X_1, \ldots, X_n)|\mu, \sigma^2\big)$. There are definite examples where making $h(x_1, \ldots, x_n)$ is not helpful to us.

(q) <u>Example</u>. Let $X_1, \ldots, X_n$ be a random sample from a Laplace distribution with PDF

$$f_X(x) = \frac{1}{2}\exp\{-|x - \lambda|\}$$

The joint PDF of the sample will be

$$f(x_1, \ldots, x_n|\lambda) = \frac{1}{2^n} \exp\left\{-\sum_{i=1}^{n} \left|x_i - \lambda\right|\right\}$$

If we "factor" this joint PDF so that $h(x_1, \ldots, x_n) = 1$ and $g\big(T(x_1, \ldots, x_n)|\lambda\big)$ is the PDF, what is our function $T(x_1, \ldots, x_n)$? Because of the absolute value, we can't isolate a simple function of the sample. Thus, the trick doesn't always produce something easy to use. In this case, the best we can do for a sufficient statistic here is the order statistics:

$$f(x_1, \ldots, x_n|\lambda) = \frac{1}{2^n} \exp\left\{-\sum_{i=1}^{n} \left|x_{(i)} - \lambda\right|\right\}$$

(r) <u>Definition</u>. Let $f(t|\theta)$ be a family of PDFs or PMFs for a statistic $T(X_1, \ldots, X_n)$. The family of probability distributions is **complete** if $\mathbb{E}[g(T)] = 0$ implies $P(g(T) = 0) = 1$ for all θ. Alternatively, $T(X_1, \ldots, X_n)$ is called a **complete statistic**.

(s) <u>Aside.</u> When we say "family" of distributions, we mean something like $N(\theta, 1)$ where $-\infty < \theta < \infty$. If you end up doing more statistics, completeness will be of more importance. In layman's terms, we can think of "completeness" as meaning there is no extraneous information in a statistic. For us, however, it's really just a stepping stone that we mention in passing.

(t) <u>Definition.</u> Let $X_1, \ldots X_n$ be a random sample from a distribution with PDF $f_X(x|\boldsymbol{\theta})$. The distribution belongs to the **exponential family** if we can write the PDF of X as

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta}) \exp\left\{\sum_{j=1}^{k} w_j(\boldsymbol{\theta})t_j(x)\right\}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.

(u) <u>Theorem.</u> If $X_1, \ldots X_n$ is a random sample from an exponential family, then the statistic

$$T(\mathbf{X}) = \left(\sum_{i=1}^{n} t_1(X_i), \sum_{i=1}^{n} t_2(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i)\right)$$

is a complete statistic.

(v) <u>Example.</u> Let $X_1, \ldots, X_n$ be a random sample from a binomial (n, p) distribution. Then the PDF of X is given by

$$\begin{aligned} f(x|p) &= \binom{n}{x} p^x (1-p)^{n-x} && \text{(the PDF)} \\ &= \binom{n}{x} p^x (1-p)^{-x} (1-p)^n && \text{(breaking up the exponent)} \\ &= \binom{n}{x} \left(\frac{p}{1-p}\right)^x (1-p)^n && \text{(rearranging)} \\ &= \binom{n}{x} (1-p)^n \exp\left\{x \ln\left(\frac{p}{1-p}\right)\right\} && \text{(rearranging again)} \end{aligned}$$

Now, we can name our functions using the definition of an exponential family:

$$h(x) = \binom{n}{x} \qquad c(p) = (1-p)^n \qquad w(p) = \ln[p/(1-p) \qquad t(x) = x$$

Thus, we this PDF fits the formulation of a an exponential family, i.e.,

$$f(x|p) = h(x)c(p)\exp\left\{w(p)t(x)\right\}$$

Finally, employing our theorem above we know that $\sum_{i=1}^{n} X_i$ is a complete statistic for p.

(w) Aside. We now have an idea of a "best" estimator, the UMVUE. Further, we have sufficiency (summarizes all relevant information) and completeness (no extraneous information). Luckily for us, the ideas have been united with one theorem.

(x) Theorem. Let U be a complete, sufficient statistic for θ, and let $W(U)$ be any estimator based ONLY on U. Then $W(U)$ is the unique UMVUE of $\tau(\theta) = \mathbb{E}[W(U)]$. This is the **Lehmann-Scheffé** theorem.

(y) Example. Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with PDF

$$f(x|\beta) = \frac{1}{\beta}e^{-x/\beta}, \quad x \in [0, \infty)$$

Find the uniformly minimum variance unbiased estimator of β. We can rewrite the PDF of X:

$$f(x|\beta) = 1 \cdot \frac{1}{\beta} \cdot \exp\left\{-\frac{1}{\beta} \cdot x\right\}$$

This is clearly from the exponential family:

$$h(x) = 1 \qquad c(\beta) = 1/\beta \qquad w(\beta) = -1/\beta \qquad t(x) = x$$

Then $\sum_{i=1}^{n} X_i$ is a complete statistic. Further, consider the joint distribution:

$$\begin{aligned} f(x_1, \ldots, x_n|\beta) &= \prod_{i=1}^{n} \frac{1}{\beta}e^{-x_i/\beta} && \text{(the joint PDF)} \\ &= \frac{1}{\beta^n}\exp\left\{-\frac{1}{\beta}\sum_{i=1}^{n} x_i\right\} && \text{(multiplying)} \end{aligned}$$

Then by the factorization theorem, $\sum_{i=1}^{n} X_i$ is a sufficient statistic as well. Note that:

$$\begin{aligned} \mathbb{E}[X_i] &= \beta && \text{(the expected value)} \\ \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] &= n\beta && \text{(the expected value of the sum)} \end{aligned}$$

We can easily make this an unbiased estimator for β by dividing by n:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \beta \qquad \text{(un-biasing)}$$

Thus, we have a complete and sufficient statistic $T(X_1, \ldots, X_n) = \sum X_i$. Further, the function $\bar{X} = \frac{1}{n}T(\cdot)$ has expected value β, and is based *only* on our statistic T. Then by Lehmann-Scheffé, $\bar{X}$ is the UMVUE for β.

(z) Aside. How to find the UMVUE? Here's a road map thus far:

i. State the Problem

   A. What distribution are we working with?

   B. What parameters are we trying to estimate?

ii. Sufficiency Principle: Good Estimators use Complete/Sufficient Statistics

   A. Do we have an exponential family?

   - Yes? Find a complete/sufficient statistic
   - No? Use the factorization theorem, appeal to definition of completeness

iii. Unbiasedness

   A. Is the statistic unbiased?

   - Yes? Then it's the UMVUE
   - No? Can we make it unbiased easily (e.g., divide by n)?
     - Yes? Do it and you have the UMVUE.
     - No? Take another statistics class.

This is a useful way to think about good estimators. That said, we have another theorem that we will use somewhat more frequently in first-year econometrics:

Cramér-Rao Lower Bound

(a) Theorem. Let $X_1, \dots, X_n$ be a sample (not necessarily random) with pdf $f(\mathbf{x}|\theta)$, and let $W(X_1, \dots, X_n)$ be an estimator satisfying "regularity" conditions

$$\frac{d}{d\theta}\mathbb{E}\big[W(\mathbf{X})\big] = \int_{\mathcal{X}} \frac{\partial}{\partial\theta}\Big[W(\mathbf{x})f(\mathbf{x}|\theta)\Big]d\mathbf{x} \qquad \text{and} \qquad \text{Var}\big[W(\mathbf{X})\big] < \infty$$

If these hold, then

$$\text{Var}(W(\mathbf{X})) \geq \frac{\left(\frac{d}{d\theta}\mathbb{E}\big[W(\mathbf{X})\big]\right)^2}{\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log\left[f(\mathbf{X}|\theta)\right]\right)^2\right]}$$

This is known as the **Cramér-Rao Lower Bound**. This inequality exists if the conditions hold, even when we have a biased estimator with non-i.i.d. data.

(b) Aside. We won't worry too much about the conditions. The second one says that the variance of the estimator must be finite; if it's not, there's not a big reason to set a lower bound on the variance anyway. The first condition states that we need to be able to switch an integral and a derivative. This is important theoretically, but we'll always be able to do it in first-year econometrics.

More frequently (virtually always in the first year), we'll be dealing with a random sample and an unbiased estimator, which simplifies our condition.

(c) Theorem. If $X_1, \dots, X_n$ is an i.i.d. sample and $\mathbb{E}\big[W(\mathbf{X})\big] = \psi(\theta)$ (i.e., W is an unbiased estimator for some function of θ), then

$$\text{Var}(W(\mathbf{X})) \geq \frac{\psi'(\theta)^2}{I(\theta)}$$

Where $I(\theta)$ is the **Fisher information**, given by:

$$I(\theta) = n\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log\left[f(X_i|\theta)\right]\right)^2\right]$$

Further, the Fisher information (under certain regularity conditions) can be simplified:

$$I(\theta) = -n\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\Big(\log\big[f(X_i|\theta)\big]\Big)\right]$$

(d) Aside. While this may appear rather complicated, it can be extremely useful to us in practice. If an unbiased estimator attains the Cramer-Rao lower bound, then we know that it must be the UMVUE.

(e) Example. Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu, \sigma^2)$ distribution, where we will assume that μ and σ^2 are unknown. Show that $\bar{X}$ attains the CRLB, but S^2 does not.

Note that $\mathbb{E}[\bar{X}] = \psi(\mu) = \mu$

$$f(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\} \qquad \text{(the PDF of } X\text{)}$$

$$\ln\big[f(x|\mu)\big] = -\frac{1}{2}\ln\big[2\pi\sigma^2\big] - \frac{1}{2\sigma^2}(x_i-\mu)^2 \qquad \text{(taking a log transform)}$$

$$\frac{\partial}{\partial\mu}\ln\big[f(\cdot)\big] = \frac{1}{\sigma^2}(x_i-\mu) \qquad \text{(differentiating w.r.t. } \mu\text{)}$$

$$\frac{\partial^2}{\partial\mu^2}\ln\big[f(\cdot)\big] = -\frac{1}{\sigma^2} \qquad \text{(the second derivative)}$$

$$\mathbb{E}\left[\frac{\partial^2}{\partial\mu^2}\ln\big[f(\cdot)\big]\right] = -\frac{1}{\sigma^2} \qquad \text{(taking the expected value)}$$

$$I(\mu) = -n\left[-\frac{1}{\sigma^2}\right] \qquad \text{(the Fisher information)}$$

This is the denominator. We can now find the numerator:

$$\psi(\mu) = \mu \qquad \text{(the estimator is unbiased)}$$

$$\frac{d}{d\mu}\psi(\mu) = 1 \qquad \text{(differentiating)}$$

$$\frac{d}{d\mu}\psi(\mu)^2 = 1 \qquad \text{(squaring)}$$

Now we can put together the pieces to find the CLRB:

$$Var(\bar{X}) \geq \frac{1}{-n\left[-\frac{1}{\sigma^2}\right]}$$

$$Var(\bar{X}) \geq \frac{\sigma^2}{n}$$

We know from before that $Var(\bar{X}) = \sigma^2/n$. Thus, it attains the CLRB, making it the UMVUE for μ. What about S^2?

$$\ln\big[f(x|\mu)\big] = -\frac{1}{2}\ln\big[2\pi\sigma^2\big] - \frac{1}{2\sigma^2}(x_i-\mu)^2 \qquad \text{(from above)}$$

$$\frac{\partial}{\partial\sigma^2}\ln\big[f(\cdot)\big] = -\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} \qquad \text{(differentiating w.r.t. } \sigma^2\text{)}$$

$$\frac{\partial^2}{\big(\partial\sigma^2\big)^2}\ln\big[f(\cdot)\big] = \frac{1}{2(\sigma^2)^2} - \frac{(x-\mu)^2}{(\sigma^2)^3} \qquad \text{(the second derivative)}$$

Taking the expectation:

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial^2}{\left(\partial\sigma^2\right)^2}\ln\left[f(\cdot)\right]\right] &= \frac{1}{2\sigma^4} - \frac{\sigma^2}{\sigma^6} && \text{(the expectation)} \\
&= -\frac{1}{2\sigma^4} && \text{(simplifying)} \\
I(\sigma^2) &= -n\left(-\frac{1}{2\sigma^4}\right) && \text{(the Fisher information)}
\end{aligned}$$

We know that S^2 is unbiased, so the numerator in the CRLB is equal to one. Thus, the CRLB for S^2 is given by:

$$\mathrm{Var}(S^2) \geq \frac{2\sigma^2}{n}$$

Now, we know that the variance of S^2 is $\mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}$ so S^2 does not attain the CRLB.

8. **Convergence Concepts**

(a) Aside. Although there is definitely more to discuss with finite-sample estimators, we'll move very briefly into a discussion of large-sample statistics, i.e., the properties estimators have when our sample size goes to infinity.

Hopefully, we're familiar with the notion of convergence for sequences. While convergence is useful in analysis topics, we have several other weaker forms of convergence that are extremely useful in large-sample statistics. We very often don't have data drawn from known, simple distributions; instead, we tend to rely on large-sample results quite a bit in practice.

(b) Definition. Let $X_1, X_2, \ldots,$ be a sequence of random variables. This sequence **converges almost surely** to a random variable X if for every $\varepsilon > 0$:

$$\mathbb{P}\left(\lim_{n\to\infty}\left|X_n - X\right| < \varepsilon\right) = 1$$

Alternatively, we write $X_n \overset{a.s.}{\to} X$.

(c) Definition. This is akin to saying that the probability we draw a convergent sequence is one (not exactly correct, but "intuitively" close). A convenient way to think about this is to imagine a pseudo-random number generator on a computer. If our seed is itself a random variable, then the sequence of draws we make depend on the seed. Almost sure convergence implies that if our sequence is long enough, X_n will be *extremely* close to X regardless of the seed.

(d) Example. Let $S \sim \mathrm{Unif}[0,1]$. Then define the sequence of random variables $\{X_n\}$ to be:

$$\begin{aligned}
&X_1(s) = s + \mathbb{1}\{0 \leq s \leq 1\} && X_2(s) = s + \mathbb{1}\{0 \leq s \leq 1/2\} && X_3(s) = \mathbb{1}\{0 \leq 1/3\} \\
&X_4(s) = s + \mathbb{1}\{0 \leq s \leq 1/4\} && \ldots
\end{aligned}$$

Note that this is a random sequence: we don't know what value S will take on ex ante; further, we don't know how many terms of the sequence will be equal to $S+1$ before the sequence becomes equal to S. This sequence, however, converges almost surely, i.e., $X_n(S) \overset{a.s}{\to} S$. The value s determines the amount of "time" before $x_n(s) = s$:

- $s = 0.2$, the first five elements will be $x_1(0.2) = 0.2 + 1, \ldots, x_5(0.2) = 0.2 + 1$; but every element thereafter equals $x_j(0.2) = 0.2$, $j > 5$.

- $s = 0.001$, the first 1000 elements, $X_1(0.001) = 0.001 + 1, \dots, X_{1000}(0.001) = 0.001 + 1$, but every element thereafter equals 0.001.

The point is, for finite a n, the probability we draw a value within ε is not zero; if n goes out to infinity, however, *eventually* our sequence $X_n(S)$ *will* equal S.

(e) Theorem. Let $X_1, X_2, \dots$ be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then for every $\varepsilon > 0$

$$P\left(\lim_{n\to\infty} \left|\bar{X}_n - \mu\right| < \varepsilon\right) = 1$$

That is, $\bar{X}_n$ converges almost surely to μ. This is known as the **strong law of large numbers**.

(f) Aside. While we say $X_n \overset{a.s.}{\to} X$ where X is a random variable, we virtually always have X be a non-random population parameter, such as μ. Further, almost sure convergence is a stronger concept than we require in any of our econometric settings during the first year.

(g) Definition. Let $X_1, X_2 \dots$ be a sequence of random variables. This sequence **converges in probability** to a random variable X if for every $\varepsilon > 0$:

$$\lim_{n\to\infty} P\Big(\left|X_n - X\right| < \varepsilon\Big) = 1$$

Alternatively, we write $X_n \overset{p}{\to} X$.

(h) Aside. Note the difference in the definitions between this and almost sure convergence. Almost sure convergence says something akin to "the probability we draw a convergent sequence is 1." Convergence in probability says, "the probability our nth draw is close to X goes to one."

(i) Example. Consider once again a random variable S drawn from an interval $[0, 1]$. Define the sequence X_n to be:

$$X_1(s) = s + \mathbb{1}\{0 \le s \le 1\}$$

$$X_2(s) = s + \mathbb{1}\{0 \le s < 1/2\} \quad X_3(s) = s + \mathbb{1}\{1/2 < s \le 1\}$$

$$X_4(s) = s + \mathbb{1}\{0 \le s < 1/3\} \quad X_5(s) = s + \mathbb{1}\{1/3 < s \le 2/3\} \quad X_6(s) = s + \mathbb{1}\{2/3 < s < 1\}$$

$$X_7(s) = s + \mathbb{1}\{0 \le s < 1/4\} \quad X_8(s) = s + \mathbb{1}\{1/4 < s \le 2/4\} \qquad \dots$$

This sequence converges in probability to S, but not almost surely. For example, if we're considering x_8, x_8 only differs from s if $s \in [1/4, 1/2]$—only a 25% chance. Following this pattern, as n gets large, for any particular value of n, the probability that $x_n \neq s$ is clearly going to zero. That said, as we continue to list elements in the sequence, for *any* value of s, we will get an *infinite* number of observations of $\{x_n\}$ that do not equal s; *but* they will become less and less common.

(j) Theorem. Let $X_1, X_2, \dots$ be i.i.d. random variables with $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then for every $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\bar{X}_n - \mu\right| < \varepsilon\right) = 1$$

That is, $\bar{X}_n$ converges in probability to μ. This is known as the **weak law of large numbers**.

(k) Aside. It can be a bit tricky to think about almost sure convergence v.s. convergence in distribution. One potential way to think about it is in terms of a machine making widgets. When the machine makes a mistake, it produces a broken widget. Think about mistakes as a random variable:

- Almost sure convergence: the machine makes mistakes randomly, but only makes a *finite* number of broken widgets.
- Convergence in probability: the machine makes mistakes randomly, but it makes an *infinite* number of mistakes, but mistakes get less likely over time.

Almost sure convergences does not allow *any* "large-n" values of X_n to be far away from X; convergence in probability does, but the *probability* thant X_n is far from X is very, very small.

(l) Theorem. Suppose that $X_1, X_2, \dots$ converges in probability to a random variable X and that h is a continuous function. Then $h(X_1), h(X_2), \dots$ converges in probability to $h(X)$. This is known as the **continuous mapping theorem**.

(m) Example. If $X_1, \dots, X_n$ are i.i.d. from an exponential distribution with $\mathbb{E}[X_i] = \beta$, then

$$\bar{X}_n \xrightarrow{p} \beta \qquad \text{(by the WLLN)}$$

$$\Longrightarrow \frac{1}{\bar{X}_n} \xrightarrow{p} \frac{1}{\beta} \qquad \text{(by the CMT)}$$

(n) Definition. Let $X_1, X_n, \dots$ be a sequence of random variables. Let $\hat{\theta}_n(X_1, \dots, X_n)$ be an estimator for the parameter θ, based on a sample size n. Then $\hat{\theta}_n$ is a **consistent estimator** for θ if $\hat{\theta}_n \xrightarrow{p} \theta$.

(o) Aside. We use these three together (WLLN, CMT, consistency) very frequently during the first year econometric sequence. Generally, we have some function of $\bar{X}$ that we're using to estimate a parameter, then we employ the continuous mapping theorem to show it's consistent.

During second-quarter econometrics, we will spend a great deal of time working with the concept of consistency. Basically, much of our OLS and other econometric techniques rely on large-sample results.

(p) Definition. A sequence of random variables $X_1, X_2, \dots$ **converges in distribution** to a random variable X if

$$\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$$

at all points x where $F_X(x)$ is continuous. Alternatively, we say $X_n \xrightarrow{d} X$.

(q) Theorem. Let $X_1, X_2, \dots$ be a sequence of i.i.d. random variables. Assume:

- The MGFs exist in a neighborhood of 0.
- $\mathbb{E}[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2 < \infty$.
- $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$.

Let $G_n(x)$ denote the CDF of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then for any x,

$$\lim_{n \to \infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{2\pi} \exp\left\{-\frac{y^2}{2}\right\} \, dy$$

that is, $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ follows a standard normal distribution in the limit. This is known as the **Central Limit Theorem**.

(r) Aside. This is the formal definition. More frequently, we simply state that

$$\sqrt{n}(\bar{X} - \mu) \rightsquigarrow N(0, \sigma^2)$$

where $\rightsquigarrow$ is short-hand for "distributed in the limit." Note that from our WLLN, $\bar{X} - \mu$ will converge in probability to zero. It converges at rate $\sqrt{n}$, however, so by multiplying by $\sqrt{n}$, we "grow" this value at the same rate it "shrinks," thus ensuring we get a distribution instead of a simply zero.

The last convergence concept we will discuss relates functions of random variables and convergence in distribution.

(s) Definition. Given that a function $g(x)$ has derivatives of order r (that is, the rth derivative $g^{(r)}(x)$ exists), then for any constant a, the **Taylor Polynomial** of order r around a is

$$T_r(x) = \sum_{i=0}^{r} \frac{g^{(i)}(a)}{i!}(x-a)^i$$

(t) Aside. We won't dwell on it here, but there is also Taylor's theorem, which says that we can approximate an $r - times$ differentiable function around a point using the Taylor polynomial. This is extremely useful in numerical methods (among other things). Basically:

$$\lim_{x \to a} \frac{g(x) - T_r(x)}{(x-a)^r} = 0$$

Generally, the remainder goes to zero "fast enough" for us to approximate things using the Taylor series:

$$g(x) \approx \sum_{i=0}^{r} \frac{g^{(i)}(a)}{i!}(x-a)^i$$

(u) Theorem (THM 5.5.24). Let Y_n be a sequence of random variables such that

$$\sqrt{n}(Y_n - \theta) \rightsquigarrow N(0, \sigma^2)$$

Then given a differentiable function g:

$$\sqrt{n}\big[g(Y_n) - g(\theta)\big] \rightsquigarrow N(0, \sigma^2[g'(\theta)]^2)$$

This is known as the **Delta Method**

(v) Example. Consider a distribution such that:

$$\mathbb{E}[x_i] = \beta \quad \text{and} \quad \text{Var}[X_i] = \beta^2$$

Where $\sqrt{n}(\bar{X} - \beta) \rightsquigarrow N(0, \beta^2)$. We can find the limiting distribution of $\ln(\bar{X})$ using the delta method:

$$\begin{aligned}
\ln(\bar{x}) &= \frac{\ln(\beta)(\bar{x}-\beta)^0}{0!} + \frac{\frac{1}{\beta}(\bar{x}-\beta)^1}{1!} + \frac{-\frac{1}{\beta^2}(\bar{x}-\beta)^2}{2!} + \ldots && \text{(the Taylor expansion)} \\
\big[\ln(\bar{x}) - \ln(\beta)\big] &= \frac{1}{\beta}(\bar{x}-\beta) - \frac{(\bar{x}-\beta)^2}{2\beta^2} + \ldots && \text{(rearranging)} \\
\sqrt{n}\big[\ln(\bar{x}) - \ln(\beta)\big] &= \frac{1}{\beta}\sqrt{n}(\bar{x}-\beta) - \frac{\sqrt{n}(\bar{x}-\beta)^2}{2\beta^2} + \ldots && \text{(multiplying by } \sqrt{n})
\end{aligned}$$

Although it is a bit "hand-wavy," we know that $\bar{x} - \beta$ converges in probability to zero at rate $\sqrt{n}$; thus, $(\bar{x} - \beta)^2$ converges at rate n. Since we're only multiplying by $\sqrt{n}$, our second- and higher-order terms go to zero "fast enough" for this approximation to work.

Thus, we can think only about the term $\frac{1}{\beta}\sqrt{n}(\bar{x}-\beta)$:

$$\begin{aligned}
\sqrt{n}(\bar{x}-\beta) &\rightsquigarrow N\big[0,\beta^2\big] && \text{(by assumption)}\\
\frac{1}{\beta}\sqrt{n}(\bar{x}-\beta) &\rightsquigarrow N\left[0,\left(\frac{1}{\beta}\right)^2\beta^2\right] && \text{(by rules of Var)}
\end{aligned}$$

Thus, because our remainder goes to zero "faster" than our first term converges in distribution:

$\sqrt{n}\big[\ln(\bar{x})-\ln(\beta)\big] \rightsquigarrow N\big[0,1\big]$

Note that if we had simply employed the Delta Method theorem, we would have gotten the same answer:

$$\begin{aligned}
g(\beta) &= \ln(\beta) && \text{(defining } g\text{)}\\
g'(\beta) &= \frac{1}{\beta} && \text{(differentiating)}\\
g'(\beta)^2 &= \frac{1}{\beta^2} && \text{(squaring)}\\
\sqrt{n}\big[\ln(\bar{x})-\ln(\beta)\big] &\rightsquigarrow N\left[0,\left(\frac{1}{\beta^2}\right)\beta^2\right] && \text{(by the Delta Method)}
\end{aligned}$$

9. **Hypothesis Testing**

(a) Definition. A **hypothesis** is a statement about a population parameter θ. The two complementary hypotheses are the **null hypothesis**, denoted $H_0 : \theta \in \Theta_0$, and the **alternative hypothesis**, denoted $H_1 : \Theta_0^c$ (or H_A).

(b) Example. From our undergraduate statistics courses, we very frequently have something like

$$\begin{aligned}
&H_0 : \beta = 0 && H_A : \beta \neq 0 && \text{(two sided)}\\
&H_0 : \beta = 0 && H_A : \beta \geq 0 && \text{(one sided)}
\end{aligned}$$

(c) Definition. A **test function**, or decision rule, maps $\mathbb{R}^n$ to $\{0,1\}$:

$$\psi(\mathbf{x}) = \begin{cases} 1 & \text{if we reject } H_0 \\ 0 & \text{if we don't reject} H_0 \end{cases}$$

A **test statistic** $T(\mathbf{X})$ and a critical value c can be combined with a test function such that:

$$\psi(\mathbf{x}) = \begin{cases} 1 & \text{if } T(\mathbf{x}) > c \\ 0 & \text{if } T(\mathbf{x}) \leq c \end{cases}$$

The set of values that lead to rejection of the null is known as the **critical region**.

(d) Example. In the case of a standard normal hypothesis test, we almost always use the critical value 1.96 to test $H_0 : \mu = 0$ vs $H_A : \mu \neq 0$:

$$\psi(\mathbf{x}) = \begin{cases} 1 & \text{if } |z| > 1.96 \\ 0 & \text{if } |z| \leq 1.96 \end{cases}$$

In this case, the critical region is $C = \{z \in \mathbb{R} | z > 1.96\}$

(e) Definition. There are two types of errors that can be made if we use such a procedure. **Type I** error is when we reject a true null hypothesis; **Type II** error is when we fail to reject a false null hypothesis:

		True	
		H_0	H_1
"Accept"	H_0	Correct	Type II Error
	H_1	Type I Error	Correct

(f) Example. Consider drug testing of distance cyclists. There are two hypotheses associated with whether or not an athlete is taking recombinant human erythropoietin (rhEPO) to boost their red blood cell count to try and win le Tour. Our Errors are then:

- Type I Error: False Positive (someone is not a cheater, gets busted for cheating)
- Type II Error: False Negative (Lance Armstrong is cheating, doesn't' get caught)

(g) Definition. Let $\psi(\mathbf{X})$ be a test function for the hypothesis $H_0 : \theta \in \Theta_0$ against the alternative $H_A : \theta \in \Theta_0^c$. Define:

$$B(\theta) = \mathbb{E}_\theta[\psi(\mathbf{X})] = P\big(\psi(\mathbf{x}) = 1|\theta\big)$$

This is known as the **power function**, which gives the probability our test statistic equals one.

(h) Example. Suppose we have $X \sim$Poisson(λ). Let $H_0 : \lambda = 1$, $H_A : \lambda > 1$. Let our parameter space be $\lambda \in [1, \infty)$ (ignore $\lambda \in (0, 1)$ for now). Let our critical region be $X \geq 3$. Then our power function is

$$B(\lambda) = P[X \geq 3] = 1 - P[X \leq 2] = 1 - e^{-\lambda}\left(1 + \lambda + \frac{\lambda^2}{2}\right)$$

(this is 1 - $F_X(2)$). We can list a couple of values of λ:

λ	$B(\lambda)$
1	0.08
1.25	0.13
1.5	0.19
1.75	0.26
2	0.32
5	0.88

If we were drawing from distributions with a larger value of λ, we're much more likely to reject the null. That is, the further we get from $\lambda = 1$, the more likely we are to reject.

(i) Definition. Let $\psi(\mathbf{X})$ be a test function for the hypothesis $H_0 : \theta \in \Theta_0$ against the alternative $H_A : \theta \in \Theta_0^c$. The probability of error α is the **size** of the test if

$$\sup_{\theta \in \Theta_0} B(\theta) = \alpha$$

If we have a single-point null hypothesis (e.g., $\lambda = 1$ or $\beta = 0$)

$$B(\theta_0) = \alpha$$

In social science, we typically use a value of $\alpha = 0.05$ (e.g., type I error). Note that this is not always true, or even desirable. It really depends on context.

(j) Aside. While some people prefer to use Type I and Type II error, many others prefer to use the terms "size" and "power." Size is the probability of rejecting the null when the null is true; Power is is the probability of rejecting the null when the alternative is true. Both are defined with the power function, making it much easier to remember.

(k) Definition. Let $\psi^*(\mathbf{X})$ be a test function for the hypothesis $H_0 : \theta \in \Theta_0$ with the alternative $H_A : \theta \in \theta_0^c$. The Test $\psi^*(\mathbf{X})$ is the **most powerful** test of level α if for every test $\psi(\mathbf{X})$ with $\sup_{\theta \in \Theta_0} B(\theta) \leq \alpha$:

$$B^*(\theta) = P\big(\psi^*(\mathbf{X}) = 1 | \theta \in \Theta_0^c\big) \geq B(\theta) = P\big(\psi(\mathbf{X}) = 1 | \theta \in \Theta_0^c\big)$$

where $B^*(\theta_0) \leq \alpha$ and the power of every other test $B(\theta_0) \leq \alpha$. for all other test functions and for every $\theta \in \Theta_0^c$.

(l) Aside. In other words, let $\psi^*(\mathbf{X})$ have size α. It is the most powerful test of size α if it rejects the null at least as often as every other size-α test, when the null is false.

(m) Theorem. Consider a hypothesis test comparing a simple null against a simple alternative: $H_0 : \theta = \theta_0$ vs $H_A : \theta = \theta_1$. The test of the form

$$\psi(\mathbf{X}) = \begin{cases} 1 & \text{if } \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)} > k \\ 0 & \text{if } \frac{f(\mathbf{x}|\theta_1)}{f(\mathbf{x}|\theta_0)} \leq k \end{cases}$$

Is the most powerful test.

(n) Example. Consider a hypothesis test for the mean of a normal distribution: $H_0 : \mu = 0$ vs. $H_1 : \mu = 1$. Assume the variance $\sigma^2 = 1$ is known. Let's assume we have one draw from the normal distribution. Then the relevant ratio is

$$\begin{aligned} \frac{f(\mathbf{x}|\mu = 1)}{f(\mathbf{x}|\mu = 0)} &= \frac{(2\pi)^{-1/2} \exp\left\{-\frac{1}{2}(x-1)^2\right\}}{(2\pi)^{-1/2} \exp\left\{\frac{1}{2}x^2\right\}} && \text{(the ratio)} \\ &= \frac{\exp\left\{-\frac{1}{2}(x^2 - 2x + 1)\right\}}{\exp\left\{-\frac{1}{2}x^2\right\}} && \text{(simplifying)} \\ &= \exp\left\{x - \frac{1}{2}\right\} && \text{(simplifying further)} \end{aligned}$$

The most powerful test, then will be of the form "reject when $\exp\{x - 1/2\} > c$". Note that this is a strictly increasing function of x; thus, if $x \geq k$, then $\exp\{x - 1/2\} > c(k)$. The one-sided 95% critical value is 1.64, e.g., reject H_0 when $x \geq 1.64$. This is the most powerful test.