Math Camp 2024 – Statistics*

Seonmin Will Heo†

Department of Economics, UC Santa Barbara

September 5, 2024

1 Random Samples

Definition 1.1. The random variables $X_1, \cdots, X_n$ are called a **random sample** of size n from the population $f_X(x)$ if $X_1, \cdots X_n$ 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 $X_1, \cdots, X_n$ are **independent and identically distributed** (or i.i.d.).

Definition 1.2. From our probability sections, recall that the **joint PDF** (or **PMF**) of a random sample $X_1, \cdots, X_n$ is given by

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

Example 1.3. 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:

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

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

Then the joint distributions is:

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

For the rest of math camp, we'll be assuming that we're working with a random sample. However, 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).

†This lecture note was originally written by James Banovetz.

2 Statistics

Definition 2.1. Let $X_1, \cdots, X_n$ be a random sample. Let $T(X_1, \cdots, X_n)$ be a real-valued or vector valued function. Then the random variable $Y_n = T(X_1, \cdots, X_n)$ is a **statistic**.

Example 2.2. The most common statistic we see is the sample mean:

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

Another extremely common statistic is the sample variance:

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

Note that there are infinitely many statistics that one could come up with, including trivial statistics like X_1 or $X_1, \cdots, X_n$, i.e., the whole sample itself. However, we choose and discuss a few notable statistics such as the sample mean, because they are evaluated to be better than other statistics that we can use. Can you think of reasons why some statistics could be better than others?

Definition 2.3. Suppose we have a statistic $Y_n = T(X_1, \cdots, X_n)$. Then the probability distribution of the statistic Y_n is called the **sampling distribution** of Y_n.

Example 2.4. Recall the distribution of coin flips from before. Suppose we're interested in the sum, $Y_n = \sum_{i=1}^n X_i$ (i.e., the number of heads observed). Intuitively, for a particular observed sample $x_1, \cdots, x_n$ that produces $y = \sum_{i=1}^n x_i$, 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 $y = \sum_{i=1}^n x_i$. In fact, there are $\binom{n}{y}$ (i.e., n coin flips and y heads are observed). Thus, the PMF of Y_n is

$$f_{Y_n}(y) = \binom{n}{y} p^y (1-p)^{n-y},$$

which is the binomial distribution. Note that we could prove that the sum of independent Bernoulli RVs follows a binomial distribution using MGFs:

- MGF of Bernoulli $= e^t p + (1-p)$
- The sum of n *i.i.d.* random variables X_i has a MGF of $(\mathbb{E}(e^{tx}))^n$.
- MGF of Binomial $= (pe^t + 1 - p)^n$

Theorem 2.5. *Let $X_1, \cdots, X_n$ be i.i.d. from a distribution with mean μ and variance $\sigma^2 < \infty$. Then:*

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

3 Sampling from a Normal Distribution

Definition 3.1. Here are some important distributions in hypothesis testing:

1. The **chi-squared distribution** with k degrees of freedom, denoted χ^2_k or $\chi^2(k)$, is the distribution of a sum of the squares of k independent standard normal random variables. For example, suppose we have indepedent, standard normal random variables $Z_1, \ldots, Z_k$. Then the sum of their squares is distributed according to the chi-squared distribution with k degrees of freedom:

$$Q = \sum_{i=1}^{k} Z_i^2 \sim \chi^2_k.$$

   Note that the chi-squared distribution takes only one parameter, which is the degrees of freedom, k, *which* is the number of standard normal random variables being summed.

2. The **t-distribution** is the generalized distribution of the standard normal distribution. Though it resembles the normal distribution, it has heavier tails than the standard normal distribution.[1] The t-distribution with ν degrees of freedom can be defined as the distribution of the random variable T with

$$T = \frac{Z}{\sqrt{V/\nu}},$$

   where Z is a standard normal random variable, $V \sim \chi^2_\nu$, and Z and V are independent.

3. The **F-distribution** with d_1 and d_2 degrees of freedom is the distribution of

$$F = \frac{U_1/d_1}{U_2/d_2}$$

   where $U_1 \sim \chi^2_{d_1}$ and $U_2 \sim \chi^2_{d_2}$ and U_1 and U_2 are independent.

Theorem 3.2. *(CB Thm 5.3.1) Let $X_1, \cdots, X_n$ be i.i.d. from a $N(\mu, \sigma^2)$ distribution. Let $\bar{X}_n = \frac{1}{n}\sum X_i$ and let $S_n^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$. Then:*

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

Theorem 3.3. *Let $X_1, \cdots, X_n$ be i.i.d. from a $N(\mu_x, \sigma_x^2)$ and $Y_1, \cdots, Y_m$ be i.i.d. from a $N(\mu_y, \sigma_y)$. Consider the following statistics:*

1. $\dfrac{\bar{X}_n - \mu}{\sqrt{S^2_{x,n}/n}} = \dfrac{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2_{x,n}/\sigma^2}{n-1}}} \sim t_{n-1}$

2. $\dfrac{S^2_{x,n}/\sigma_x^2}{S^2_{y,m}/\sigma_y^2} = \dfrac{\frac{(n-1)S^2_{x,n}/\sigma_x^2}{n-1}}{\frac{(m-1)S^2_{y,m}/\sigma_y^2}{m-1}} \sim F_{n-1,m-1}$

[1] The t-distribution has the following probability density function:

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

where ν is the number of degrees of freedom and Γ is the gamma function.

4 Order Statistics

Definition 4.1. The **order statistics** of a random sample $X_1, \cdots, X_n$ 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**:

$$M = \begin{cases} X_{((n+1)/2)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(X_{(n/2)} + X_{(n/2+1)}\right) & \text{if } n \text{ is even} \end{cases}$$

The median is occasionally very interesting to us, especially in the presence of outliers, skewness, or non-normality. 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.

Example 4.2. Suppose we have a random sample $X_1, \cdots X_n$ from a Uniform (0,1) distribution. We can find the CDF and PDF of $X_{(n)}$ using the PDF and CDF of X_i:

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

How can we find its 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:

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

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:

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

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

5 Three Crucial Concepts in Econometrics

Much of statistics is concerned with descriptive tasks that ask "what is" questions. On the other hand, econometrics cares about causality and causal inference that answer "what if" questions. Suppose you have found a parameter of interest that would answer the research question you had. The following processes below are crucial steps for conducting good research in applied microeconometrics.

Definition 5.1. Given a statistical model, relating a parameter of interest to an estimand is called **identification**. Identification deals with the ability to uniquely determine the true values of the model parameters from the available data and model structure. Identification is concerned with whether it is theoretically possible to recover the true parameters from the data.

Definition 5.2. An **estimand** is a real number, which is a function of the probability distribution of the random variables we will get to observe.

$$Y = \mu + U \quad \text{where} \quad \mathbb{E}[U] = 0 \quad \Rightarrow \quad \mu = \mathbb{E}[Y]$$

$$Y = \mathbf{X}^T\boldsymbol{\beta} + U \quad \text{where} \quad \mathbb{E}\big(\mathbf{X}U\big) = \mathbf{0}, \mathbb{E}\big(\mathbf{X}\mathbf{X}^T\big) \text{ positive definite} \quad \Rightarrow \quad \boldsymbol{\beta} = \Big(\mathbb{E}\big[\mathbf{X}\mathbf{X}^T\big]\Big)^{-1}\mathbb{E}\big[\mathbf{X}Y\big]$$

Definition 5.3. Proposing an estimator for an estimand is called **estimation**. It is the process of determining the values of unknown parameters (e.g., coefficients in a regression model) using sample data.

Definition 5.4. An **estimator** is a function of random variables we will get to observe.

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$\widehat{\boldsymbol{\beta}}_n = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_iY_i\right)$$

Definition 5.5. Using an estimator to infer plausible values of an estimand is called **inference**. It involves drawing conclusions about a population based on sample data.

Definition 5.6. An **estimate** is a realized value of the estimator given a realized sample.

Identification should logically come prior to inference. This is because if we cannot recover a parameter when we know the population distribution, we definitely cannot recover it with a sample distribution.

6 Point Estimation

Definition 6.1. Let $X_1, \cdots, X_n$ be a sample from a population with $\theta_1, \cdots, \theta_k$ parameters and let X be a random variable with the same probability distribution as X_i's. We define the j^{th} population (non-central) moment as

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

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 $(\hat{\theta}_1, \cdots, \hat{\theta}_k)$ for $(\theta_1, \cdots, \theta_k)$ is the solution to the system

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

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 and k unknowns). Note that when we set them equal, the θ's become $\hat{\theta}$'s.

Example 6.2. Suppose we have a random sample $X_1, \cdots, X_n$ 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:

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

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

Solving this relatively trivial system:

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

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

Now, we can play around with some algebra:

$$= \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \frac{1}{n}\sum_{i=1}^{n} \bar{X}_n^2 \qquad (\bar{X}_n \text{ is a "constant"})$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \frac{2}{n}\sum_{i=1}^{n} \bar{X}_n^2 + \frac{1}{n}\sum_{i=1}^{n} \bar{X}_n^2 \qquad \text{(adding zero)}$$

$$= \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \left(2\bar{X}_n\right)\frac{1}{n}\sum_{i=1}^{n} X_i + \frac{1}{n}\sum_{i=1}^{n} \bar{X}_n^2 \qquad \text{(by def. of } \bar{X}_n)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \left(X_i^2 - 2\bar{X}_n X_i + \bar{X}_n^2\right) \qquad \text{(writing as a single sum)}$$

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

Aside. While the method of moments is not used all that frequently, it is an intuitive way to begin the construction of estimators. It also serves as the basis for *generalized method of moments* estimators, which are used heavily in the field.

Definition 6.3. Let $X_1, \cdots, X_n$ be a random sample with PDF (or PMF) $f_X(x_i|\theta_1, \cdots, \theta_k)$. The **likelihood function** is defined as

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

The likelihood function measures how well a model explains observed data by calculating the probability of seeing that data under different parameter values of the model.

Definition 6.4. For each sample point $x_1, \cdots, x_n$, let $\hat{\boldsymbol{\theta}}(x_1, \cdots, x_n)$ be a parameter value at which $\mathfrak{L}(\boldsymbol{\theta}|\mathbf{x})$ attains its maximum as a function of $\boldsymbol{\theta}$, holding $x_1, \cdots, x_n$ fixed. A **maximum likelihood estimator** (MLE) of $\boldsymbol{\theta}$ based on sample $X_1, \cdots, X_n$ is $\hat{\boldsymbol{\theta}}(X_1, \cdots, X_n)$.

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

Example 6.5. Let $X_1, \cdots, X_n$ 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

$$\mathfrak{L}(\mu, \sigma^2|\mathbf{x}) = \prod_{i=1}^{n} f_X(x_i|\mu, \sigma^2) \qquad \text{(by def. of the likelihood func.)}$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i - \mu)^2}{2\sigma^2}\right\} \qquad \text{(plugging in PDFs)}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}\right\} \qquad \text{(multiplying)}$$

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 $l(\cdot)$:

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

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 MM estimator (this is, more or less, a coincidence).

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.

Theorem 6.6. *(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****.*

7 Evaluating Estimators

We often take estimators as given and simply try to compute them, but how do we check whether the estimators we use are better than other estimators? There are several properties that make estimators better than others, and the concept that we discuss frequently is the bias-variance tradeoff. On average (in expectation), does our estimator estimate the parameter of interest, or does it overestimate or underestimate it? Does our estimator usually stay close to the parameter, or does it have a wide spread? We will look into these concepts below.

Definition 7.1. The **bias** of a point estimator $\hat{\theta}_n$ of a parameter θ is the difference between the expected value of $\hat{\theta}_n$ and θ:

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

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

Definition 7.2. The **mean squared error** (MSE) of an estimator $\hat{\theta}_n$ of a parameter θ is defined as

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

Alternatively, this may be stated in the form:

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

Example 7.3. Suppose we have two estimators for the parameter σ^2, which are $S_n^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ and $\hat{\sigma}_n^2 = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \bar{X})^2$. Then from $\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \sim \chi^2(n-1)$, we obtain the following:

$$\mathbb{E}\left[\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right] = n-1 \qquad \text{and} \qquad \text{Var}\left(\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right) = 2(n-1)$$

$$\mathbb{E}[S_n^2] = \sigma^2 \qquad \text{and} \qquad \mathbb{E}[\hat{\sigma}_n^2] = \frac{n-1}{n}\sigma^2$$

$$\text{Bias}(S_n^2) = 0 \qquad \text{and} \qquad \text{Bias}(\hat{\sigma}_n^2) = \frac{1}{n}\sigma^2$$

$$\text{Var}(S_n^2) = \frac{2}{n-1}\sigma^4 \qquad \text{and} \qquad \text{Var}(\hat{\sigma}_n^2) = \frac{2(n-1)}{n^2}\sigma^4$$

$$\text{MSE}(S_n^2) = \frac{2}{n-1}\sigma^4 \qquad \text{and} \qquad \text{MSE}(\hat{\sigma}_n^2) = \frac{2n-1}{n^2}\sigma^4$$

$\hat{\sigma}_n^2$ is biased towards zero, but it is less dispersed from its true value, σ^2 compared to S_n^2.

The mean-squared error is a good evaluator, because it provides a single measure of estimator quality, evaluating the bias-variance tradeoff. We want to avoid systematically over- or under-estimating our parameters (smaller bias), but we also want to avoid a wide spread of our estimators relative to the parameter (smaller variance).

8 Convergence

In this section, we'll discuss briefly about large-sample statistics, i.e., the properties estimators have when our sample size goes to infinity. 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.

Definition 8.1. Let $\mathbf{U}_1, \mathbf{U}_2, \cdots$ be a sequence of random vectors. This sequence **converges in probability** to a random vector $\mathbf{V}$ if for any $\varepsilon > 0$:

$$\lim_{n\to\infty} P\Big(||\mathbf{U}_n - \mathbf{V}|| < \varepsilon\Big) = 1.$$

Alternatively, we write $\mathbf{U}_n \xrightarrow{p} \mathbf{V}$.

Remark For convergence in probability, the individual convergence of the entries of the vector is necessary and sufficient for their joint convergence.

Theorem 8.2. *Let $\{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$ be a random sample and let $\mathbf{X}$ be a random vector with the same probability distribution as $\mathbf{X}_i$'s. Assume that $\mathbb{E}[\mathbf{X}] < \infty$. Define $\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{X}_i$. Then for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} P\big(||\bar{\mathbf{X}}_n - \mathbb{E}[\mathbf{X}]|| < \varepsilon\big) = 1.$$

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

Theorem 8.3. *Suppose $Y_n \xrightarrow{p} Y$ and $Z_n \xrightarrow{p} Z$. Then*

1. *$cY_n \xrightarrow{p} cY$ where $c \in \mathbb{R}$*
2. *$Y_n + Z_n \xrightarrow{p} Y + Z$*
3. *$Y_n Z_n \xrightarrow{p} YZ$*

Definition 8.4. Let $\{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$ be a random sample. Let $\hat{\theta}_n(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ be an estimator for the parameter θ, based on a sample size n. Then $\hat{\theta}_n$ is a **consistent estimator** for θ if

$$\hat{\theta}_n \xrightarrow{p} \theta$$

Aside. Suppose $\{X_1, \cdots, X_n\}$ is a random sample and let X be a random variable with the same probability distribution as X_i's with mean $\mu = \mathbb{E}[X] < \infty$.

1. $\hat{\theta}_n(X_1, \cdots, X_n) = X_1$ is unbiased for μ but not consistent.
2. $\hat{\theta}_n(X_1, \cdots, X_n) = \frac{1}{n}\sum_{i=1}^n X_i + \frac{1}{n}$ is biased but consistent.

Definition 8.5. A sequence of random vectors $\mathbf{U}_1, \mathbf{U}_2, \cdots$ **converges in distribution** to a random vector $\mathbf{V}$ if for any $\mathbf{x} \in \mathbb{R}^k$ at which the function $\mathbf{x} \to P(\mathbf{V} \le \mathbf{x})$ is continuous,

$$\lim_{n\to\infty} P\Big(\mathbf{U}_n \le \mathbf{x}\Big) = P\Big(\mathbf{V} \le \mathbf{x}\Big)$$

Alternatively, we say $\mathbf{U}_n \xrightarrow{d} \mathbf{V}$.
Remark For convergence in distribution, the individual convergence of the entries of the vector is necessary but not sufficient for their joint convergence.

Theorem 8.6. ***Central Limit Theorem***
Let $\{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$ *be a random sample and let* $\mathbf{X}$ *be a random vector with the same probability distribution as* $\mathbf{X}_i$*'s. If* $\mathbb{E}|\mathbf{X}\mathbf{X}^T| < \infty$*,*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i - \mathbb{E}[\mathbf{X}]\right) \rightsquigarrow N(\mathbf{0}, \Sigma)$$

where $\Sigma = \mathbb{E}\Big[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^T\Big]$ *and* $\rightsquigarrow$ *is short-hand for "distributed in the limit." Note that from our WLLN,* $\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i - \mathbb{E}[\mathbf{X}]$ *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.*

Theorem 8.7. *Suppose that* $\mathbf{U}_1, \mathbf{U}_2, \cdots$ *converges in probability/distribution to a random vector* $\mathbf{V}$ *and that* h *is a continuous function. Then* $h(\mathbf{U}_1), h(\mathbf{U}_2), \cdots$ *converges in probability/distribution to* $h(\mathbf{V})$*. This is known as the* ***continuous mapping theorem****.*

In the first-year sequence, you'll learn theorems and tools that discuss how convergence properties hold or change when two sequences interact with each other, such as Slutsky Theorem and Delta Method. I included them in Appendix 11.4 for your reference, but you'll learn them in detail later.

9 Bounded in Probability

Definition 9.1. The sequence of random variables $\{X_n\}$ is said to be **bounded in probability**, if there exists a constant $B_\varepsilon > 0$ and an integer N_ε such that

$$n \geq N_\varepsilon \Rightarrow P(|X_n| \leq B_\varepsilon) \geq 1 - \varepsilon$$

for all $\varepsilon > 0$.

Example 9.2. Suppose $X_n \sim \mathcal{N}(\sin(n\pi/2), 1)$. We can see that the mean is bounded in $[-1, 1]$. Since it follows the normal distribution, the amount of mass on the extremes does not grow out of control. This is why we can intuitively know that this sequence is bounded in probability. On the other hand, suppose $Z_n \sim \mathcal{N}(0, n)$. This sequence is not bounded in probability, because it has a growing standard deviation that will eventually be bigger than B_ε.

Definition 9.3. For sequences of random variables $\{X_n\}$ and $\{Y_n\}$, we write

1. $Y_n = o_p(X_n)$ if and only if $\frac{Y_n}{X_n} \xrightarrow{p} 0$, as $n \to \infty$.
2. $Y_n = O_p(X_n)$ if and only if $\frac{Y_n}{X_n}$ is bounded in probability, as $n \to \infty$.

Theorem 9.4. *Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. If $X_n \xrightarrow{d} X$, then $\{X_n\}$ is bounded in probability, i.e., $X_n = O_p(1)$.*

Theorem 9.5. *Let $\{X_n\}$ be a sequence of random variables bounded in probability and let $\{Y_n\}$ be a sequence of random variables which converge to 0 in probability. Then*

$$X_n Y_n \xrightarrow{p} 0$$

1. *If $X_n = O_p(1)$ and $Y_n = o_p(1)$, then $X_n Y_n = o_p(1)$.*
2. *$O_p(1)o_p(1) = o_p(1)$.*

Theorem 9.6. *Suppose $\{Y_n\}$ be a sequence of random variables bounded in probability. Suppose $X_n = o_p(Y_n)$. Then $X_n \xrightarrow{p} 0$, as $n \to \infty$.*

1. *$o_p(O_p(1)) = o_p(1)$.*
2. *Similarly, $O_p(o_p(1)) = o_p(1)$*

Example 9.7. If $\sqrt{n}(X_n - \theta) \xrightarrow{d} N(0, \sigma^2)$, then $\sqrt{n}(X_n - \theta) = O_p(1)$ and $(X_n - \theta) = O_p(1/\sqrt{n}) = o_p(1)$ since $1/\sqrt{n} = o(1) = o_p(1)$.

Aside. Another notation that is used to describe the limiting behavior of a function is called the Big O notation. Big O notation is useful when analyzing algorithms for efficiency. For example, suppose you wrote a code that performs mathematical operations $T(n) = 3n^2 + 6n + 5$ times (think of a loop that iterates through all elements of a vector of size n). Then as n grows, the first term ($3n^2$) will dominate the other term. We then write $T(n) = O(n^2)$ and say that the algorithm has order of n^2 time complexity.

10 Hypothesis Testing

Example 10.1. Let $\{X_1, \cdots, X_n\}$ be a random sample and let X be a random variable with the same probability distribution as X_i's. Assume that $X \sim N(\mathbb{E}[X], \sigma^2)$ and the variance σ^2 is known.

Definition 10.2. For any event A involving random sample $\{X_1, \cdots, X_n\}$ and/or X, let $P_\mu(A)$ denote the probability of the event A when $\mathbb{E}[X] = \mu$.

$$P_0\left(\frac{X}{\sigma} \leq 0\right) = P\left(\frac{X - \mathbb{E}[X]}{\sigma} \leq 0 \ : \ \mathbb{E}[X] = 0\right) = \Phi(0) = .5$$

$$P_{1.96\sigma}\left(\frac{X}{\sigma} \leq 0\right) = P\left(\frac{X - \mathbb{E}[X]}{\sigma} \leq -1.96 \ : \ \mathbb{E}[X] = 1.96\sigma\right) = \Phi(-1.96) = .025$$

Definition 10.3. A **test function**, or decision rule, maps $\mathbb{R}^n$ to $\{0, 1\}$: it is the indicator function of an event involving only $\{X_1, \cdots, X_n\}$ and known real numbers.

$$T_n = \mathbb{1}\{X_1 + X_n \geq 3\} \text{ is a statistical test.}$$

$T_n = \mathbb{I}\{X_1 + \mathbb{E}[X] \geq 3\}$ is not a statistical test, as it involves an unknown $\mathbb{E}[X]$.

Definition 10.4. A statistical test is always attached to two mutually exclusive hypotheses on an estimand of interests, μ here. A **hypothesis** is a set of values for that estimand. The statement being tested in a test of statistical significance is the **null hypothesis**, denoted H_0, and the statement that is being tested against the null hypothesis is the **alternative hypothesis**, denoted H_1.

$$\begin{array}{ll} H_0 : \mu = \{0\} & H_1 : \mu = \mathbb{R} \setminus \{0\} \\ H_0 : \mu = \{0\} & H_1 : \mu = (-\infty, -.12] \cup [.12, \infty) \end{array}$$

A T_n is a decision rule to choose between $\mu \in H_0$ and $\mu \in H_1$ once the $\{X_1, \cdots, X_n\}$ gets realized.

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } T_n = 1 \\ \text{Do not reject} H_0 & \text{if } T_n = 0 \end{array}$$

Definition 10.5. There are two types of errors that can be made if we use such a procedure.
Type I error is when we reject the null hypothesis when it is true; **Type II error** is when we fail to reject the null hypothesis when it is not true.

Definition 10.6. Let T_n be a test function for the hypothesis H_0 against the alternative H_1. Define:

$$\mu \mapsto P_\mu\big(T_n = 1\big).$$

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

$$Level = \sup_{\mu \in H_0} P_\mu(T_n = 1)$$

'*Level*' measures the worst case probability that T_n leads us to make a Type I error.

$$Power = \inf_{\mu \in H_1} P_\mu(T_n = 1)$$

'$1 - Power$' measures the worst case probability that T_n leads us to make a Type II error.

Example 10.7. For any $\alpha \in (0,1)$, let

$$T_n(\alpha) = \mathbb{1}\left\{\left|\frac{\sqrt{n}(\bar{X}_n - \mathbb{E}[X])}{\sigma}\right| > q_{1-\frac{\alpha}{2}}\right\} \quad \text{where} \quad \Phi\left(q_{1-\frac{\alpha}{2}}\right) = 1 - \frac{\alpha}{2}$$

Then $T_n(\alpha)$ is a statistical test. A sampling distribution is given by

$$\bar{X}_n \sim N\left(\mathbb{E}[X], \frac{\sigma^2}{n}\right) \quad \text{or equivalently} \quad \frac{\sqrt{n}\left(\bar{X}_n - \mathbb{E}[X]\right)}{\sigma} \sim N(0,1)$$

Assume that $n = 500$ and $\alpha = .05$

1. $H_0 : \mu = \{0\}$ vs. $H_1 : \mu = \mathbb{R} \setminus \{0\}$

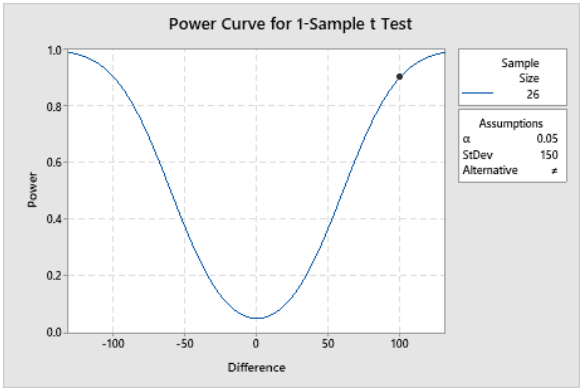
   - Level of $T_{500}(.05)$ is .05
   - Power of $T_{500}(.05)$ is .05

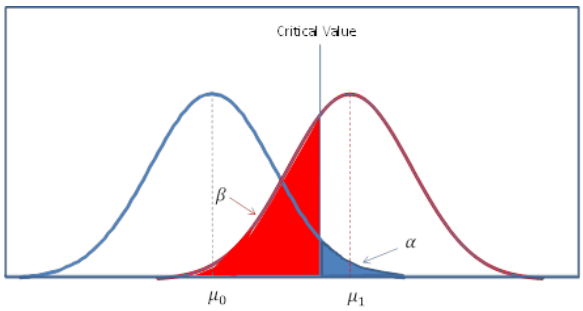
   (Trivial test) Is this statistical test a good test of $E(Y) = 0$ versus $E(Y) \neq 0$? No. As we would like to have small chances of making Type I error, we would like to pick up a small α. But if we do so, our test will have a low power too, because its power is equal to α. Therefore, we will have high chances of making a Type II error.

2. $H_0 : \mu = \{0\}$ vs. $H_1 : \mu = (-\infty, -.12] \cup [.12, \infty)$

   - Level of $T_{500}(.05)$ is .05
   - Power of $T_{500}(.05)$ is .8

   (Power Calculation, Minimum Detectable Difference from 0) We can test H_0 against H_1 with 5% probability of making a Type I error and 20% probability of making a Type II error.





Aside. Bayes Estimation.
If we have data x and parameters θ,

$$f(\theta|x) = \frac{f(x|\theta) \times f(\theta)}{f(x)}$$

where $f(\theta|x)$: posterior
$f(x|\theta)$: likelihood function
$f(\theta)$: prior
$f(x)$: marginal likelihood

11 Appendices

11.1 Proof of Theorem

Sketch of Proof.

- $\bar{X}_n \sim (\mu, \frac{\sigma^2}{n})$ from Theorem 2.5. We check the mgf of the sample mean and check whether it follows the mgf of the normal distribution. (CB Exmp 5.2.8, p.215)
-

$$\frac{(n-1)S_n^2}{\sigma^2} = \sum \left(\frac{X_i - \bar{X}}{\sigma}\right)^2$$

$$\sum \left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum \left(\frac{X_i - \bar{X}_n}{\sigma} + \frac{\bar{X}_n - \mu}{\sigma}\right)^2$$

$$= \sum \left(\frac{X_i - \bar{X}_n}{\sigma}\right)^2 + 2\sum \left(\frac{X_i - \bar{X}_n}{\sigma}\right)\left(\frac{\bar{X}_n - \mu}{\sigma}\right) + \sum \left(\frac{\bar{X}_n - \mu}{\sigma}\right)^2$$

$$= \sum \left(\frac{X_i - \bar{X}_n}{\sigma}\right)^2 + n \cdot \left(\frac{\bar{X}_n - \mu}{\sigma}\right)^2$$

$$\sum \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_{(n)}, n \cdot \left(\frac{\bar{X}_n - \mu}{\sigma}\right)^2 \sim \chi^2_{(1)}$$

$$\therefore \sum \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \chi^2_{(n-1)}$$

- Applying CB Thm 4.6.12, we show that $\bar{X}_n$ and S^2 are functions of independent random vectors.

$$S^2 = \frac{1}{n-1}\left(\left[\sum_{i=2}^{n}(X_i - \bar{X})\right]^2 + \sum_{i=2}^{n}(X_i - \bar{X})^2\right)$$

since $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$. Therefore, S^2 is the function of $(X_2 - \bar{X}), \cdots, (X_n - \bar{X})$. The joint pdf of the sample $X_1, \cdots, X_n$ can be rewritten as the multiplication of the pdf of $\bar{X}$ and the joint pdf of $(X_2 - \bar{X}), \cdots, (X_n - \bar{X})$. Therefore, $\bar{X}_n$ and S^2 are independent. (CB p.219)

11.2 Cauchy Schwartz Inequality

Given $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$,

$$\mathbf{a}^T\mathbf{b} = ||\mathbf{a}||\ ||\mathbf{b}||\ cos(\theta)$$

$$\left(\mathbf{a}^T\mathbf{b}\right)^2 = \left(||\mathbf{a}||\right)^2\ \left(||\mathbf{b}||\right)^2\ \left(cos(\theta)\right)^2$$

$$\left(\mathbf{a}^T\mathbf{b}\right)^2 \leq \left(||\mathbf{a}||\right)^2\ \left(||\mathbf{b}||\right)^2 \qquad (cos(\theta) \in [-1, 1])$$

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right)$$

$$\left(\frac{1}{n}\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\frac{1}{n}\sum_{i=1}^{n} a_i^2\right)\left(\frac{1}{n}\sum_{i=1}^{n} b_i^2\right) \qquad \text{(Divide by } n^2\text{)}$$

Now, let $a_i = Y_i - \mathbb{E}[Y]$ and let $b_i = Z_i - \mathbb{E}[Z]$. Then

$$\begin{aligned}
\mathrm{Cov}(Y,Z)^2 &\leq \mathrm{Var}(Y)\mathrm{Var}(Z) \\
\mathrm{Var}(Y) &\geq \Big[\mathrm{Var}(Z)\Big]^{-1}\mathrm{Cov}(Y,Z)^2 \\
\mathrm{Var}(Y) - \Big[\mathrm{Var}(Z)\Big]^{-1}&\mathrm{Cov}(Y,Z)^2 \ \text{ is PSD.}
\end{aligned}$$

11.3 Matrix Algebra

$$\begin{pmatrix} a_{11}+\cdots+a_{1n} \\ a_{21}+\cdots+a_{2n} \\ \vdots \\ a_{k1}+\cdots+a_{kn} \end{pmatrix} \begin{pmatrix} a_{11}+\cdots+a_{1n} & a_{21}+\cdots+a_{2n} & \cdots & a_{k1}+\cdots+a_{kn} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^n\sum_{j=1}^n a_{1i}a_{1j} & \sum_{i=1}^n\sum_{j=1}^n a_{1i}a_{2j} & \cdots & \sum_{i=1}^n\sum_{j=1}^n a_{1i}a_{kj} \\ \sum_{i=1}^n\sum_{j=1}^n a_{2i}a_{1j} & \sum_{i=1}^n\sum_{j=1}^n a_{2i}a_{2j} & \cdots & \sum_{i=1}^n\sum_{j=1}^n a_{2i}a_{kj} \\ \vdots & \vdots & & \vdots \\ \sum_{i=1}^n\sum_{j=1}^n a_{ki}a_{1j} & \sum_{i=1}^n\sum_{j=1}^n a_{ki}a_{2j} & \cdots & \sum_{i=1}^n\sum_{j=1}^n a_{ki}a_{kj} \end{pmatrix}$$

$$= \sum_{i=1}^n\sum_{j=1}^n \begin{pmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{ki} \end{pmatrix} \begin{pmatrix} a_{1j} & a_{2j} & \cdots & a_{kj} \end{pmatrix}$$

11.4 Additional Theorems and Concepts

Theorem 11.1. *Suppose that $\mathbf{U}_1, \mathbf{U}_2, \cdots$ converges in distribution to a random vector $\mathbf{V}$ and that h is a continuous function. Then $h(\mathbf{U}_1), h(\mathbf{U}_2), \cdots$ converges in distribution to $h(\mathbf{V})$. This is known as the* ***continuous mapping theorem****.*

Theorem 11.2. ***Slutsky Theorem***

1. *If $\mathbf{Y}_n \xrightarrow{d} \mathbf{Y}$ and $\mathbf{Z}_n \xrightarrow{p} \mathbf{c}$ where $\mathbf{c}$ is constant, then*

$$\begin{bmatrix} \mathbf{Y}_n \\ \mathbf{Z}_n \end{bmatrix} \xrightarrow{d} \begin{bmatrix} \mathbf{Y} \\ \mathbf{c} \end{bmatrix}$$

(Slutsky lemma) *Assume that* $Y_n \xrightarrow{d} Y$ *and that* $Z_n \xrightarrow{p} c$. *Then it follows from continuous mapping theorem that*

(a) $Z_n Y_n \xrightarrow{d} cY$

(b) $Z_n + Y_n \xrightarrow{d} c + Y$

2. $\mathbf{Y}_n \xrightarrow{d} \mathbf{Y}$ *and* $\mathbf{Z}_n \xrightarrow{d} \mathbf{Z} \;\nrightarrow\; \mathbf{Y}_n + \mathbf{Z}_n \xrightarrow{d} \mathbf{Y} + \mathbf{Z}$ *where everything is conformable.*

3. $\mathbf{Y}_n \xrightarrow{p} \mathbf{c} \;\Leftrightarrow\; \mathbf{Y}_n \xrightarrow{d} \mathbf{c}$ *where* $\mathbf{c}$ *is a constant.*

4. $\mathbf{Y}_n \xrightarrow{p} \mathbf{Y} \;\Rightarrow\; \mathbf{Y}_n \xrightarrow{d} \mathbf{Y}$

Definition 11.3. Let $\{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$ be a random sample. Let $\hat{\boldsymbol{\theta}}_n(\mathbf{X}_1, \cdots, \mathbf{X}_n)$ be an estimator for the parameter $\boldsymbol{\theta}$, based on a sample size n. Then $\hat{\boldsymbol{\theta}}_n$ is a $\sqrt{n}$**-consistent estimator** for $\boldsymbol{\theta}$ if

$$\sqrt{n}\,(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} Z$$

If $Z \sim N(\mathbf{0}, \Sigma)$, then $\hat{\boldsymbol{\theta}}_n$ is said to be asymptotically normally distributed.

Theorem 11.4. ***Cramer-Rao Lower Bound****. Let* $\{X_1, \cdots, X_n\}$ *be a sample (not necessarily random) with a joint pdf* $f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$ *where* $\boldsymbol{\theta} \in \mathbb{R}^k$, *and let* $W(X_1, \cdots, X_n)$ *be an estimator satisfying "regularity" conditions*

$$\frac{d}{d\boldsymbol{\theta}}\mathbb{E}\big[W(\mathbf{X})\big] = \int_{\mathbb{R}^n} \frac{\partial}{\partial \boldsymbol{\theta}}\Big[W(\mathbf{X}) f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})\Big] d\mathbf{x} \qquad \textit{and} \qquad Var\big[W(\mathbf{X})\big] < \infty$$

If these hold, then the following matrix is positive semi-definite.

$$Var(W(\mathbf{X})) - \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log\left[f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})\right] \frac{\partial}{\partial \boldsymbol{\theta}^T} \log\left[f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})\right]\right]^{-1} \left(\frac{d}{d\boldsymbol{\theta}}\mathbb{E}\big[W(\mathbf{X})\big]\right) \left(\frac{d}{d\boldsymbol{\theta}^T}\mathbb{E}\big[W(\mathbf{X})\big]\right)$$

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.*

Remark Under regularity conditions,

$$\mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})\right] = \mathbf{0} \quad \text{and} \quad \text{Var}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \log[f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})]\right) = \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log\left[f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})\right] \frac{\partial}{\partial \boldsymbol{\theta}^T} \log\left[f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})\right]\right]$$

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})\right] &= \int \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})\right] f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) dx \\
&= \int \frac{1}{f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})} \frac{\partial}{\partial \boldsymbol{\theta}} f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) dx \\
&= \int \frac{\partial}{\partial \boldsymbol{\theta}} f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) dx \\
&= \frac{\partial}{\partial \boldsymbol{\theta}} \int f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) dx = 1
\end{aligned}$$

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.

Theorem 11.5. *Let $\{X_1, \cdots, X_n\}$ be a random sample and let X be a random variable with the same probability distribution as X_i's. If $\mathbb{E}\big[W(\mathbf{X})\big] = \boldsymbol{\theta}$, then*

$$\frac{d}{d\boldsymbol{\theta}}\boldsymbol{\theta} = I_k$$

$$\begin{aligned}\mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log\left[f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})\right]\frac{\partial}{\partial\boldsymbol{\theta}^T}\log\left[f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta})\right]\right] &= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log\left[f_X(X_i|\boldsymbol{\theta})\right]\frac{\partial}{\partial\boldsymbol{\theta}^T}\log\left[f_X(X_j|\boldsymbol{\theta})\right]\right] \\ &= n\ \mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log\left[f_X(X|\boldsymbol{\theta})\right]\frac{\partial}{\partial\boldsymbol{\theta}^T}\log\left[f_X(X|\boldsymbol{\theta})\right]\right] \\ &= n\ \mathcal{I}(\boldsymbol{\theta}:X)\end{aligned}$$

$$Var(\hat{\boldsymbol{\theta}}) - \Big(n\ \mathcal{I}(\boldsymbol{\theta}:X)\Big)^{-1} \text{ is PSD.}$$

Where $\mathcal{I}(\boldsymbol{\theta}:X)$ is the ***Fisher information****, given by:*

$$\mathcal{I}(\boldsymbol{\theta}:X) = \mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\log\left[f_X(X|\boldsymbol{\theta})\right]\frac{\partial}{\partial\boldsymbol{\theta}^T}\log\left[f_X(X|\boldsymbol{\theta})\right]\right]$$

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

$$\mathcal{I}(\boldsymbol{\theta}:X) = -\mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\log\left[f_X(X|\boldsymbol{\theta})\right]\right]$$

Aside. Asymptotic Normality of MLE.

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N\left(\mathbf{0},\ \mathcal{I}(\boldsymbol{\theta}_0:X)^{-1}\right)$$

Example 11.6. Let $\{X_1, \cdots, X_n\}$ 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.

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

$$\begin{aligned}
f(x_i|\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\} && \text{(the PDF of } X\text{)} \\
\ln\big[f(x_i|\mu)\big] &= -\frac{1}{2}\ln\big[2\pi\sigma^2\big] - \frac{1}{2\sigma^2}(x_i-\mu)^2 && \text{(taking a log transform)} \\
\frac{\partial}{\partial\mu}\ln\big[f(\cdot)\big] &= \frac{1}{\sigma^2}(x_i-\mu) && \text{(differentiating w.r.t. } \mu\text{)} \\
\frac{\partial^2}{\partial\mu^2}\ln\big[f(\cdot)\big] &= -\frac{1}{\sigma^2} && \text{(the second derivative)} \\
\mathbb{E}\left[\frac{\partial^2}{\partial\mu^2}\ln\big[f(\cdot)\big]\right] &= -\frac{1}{\sigma^2} && \text{(taking the expected value)}
\end{aligned}$$

Therefore,

$$CRLB = \frac{1}{-n\cdot\mathbb{E}\left[\frac{\partial^2}{\partial\mu^2}\ln\big[f(\cdot)\big]\right]} = \frac{\sigma^2}{n}$$

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

Definition 11.7. Given that a function $g(x)$ has derivatives of order r (that is, the r^{th} derivative $g^{(r)}(x)$ exists), then for any constant a, the **Taylor Polynomial** of order r around a is

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

Theorem 11.8.

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{d} \mathbf{Z}$$

Then given a differentiable function $g : \mathbb{R}^k \to \mathbb{R}^q$:

$$\sqrt{n}\left(g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta}_0)\right) \xrightarrow{d} \frac{\partial g(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T}\mathbf{Z}$$

*This is known as the **Delta Method**.*

Aside. Asymptotic Test with MLE.

$$\hat{\boldsymbol{\theta}}_n \in arg \max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}\subset\mathbb{R}^k} \mathfrak{L}(\boldsymbol{\theta} : X_1, \cdots, X_n)$$

It follows from asymptotic normality of MLE that

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\Big(\mathbf{0},\ \mathcal{I}(\boldsymbol{\theta}_0 : X)^{-1}\Big)$$

and

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$$

1. Linear test $H_0 : R\boldsymbol{\theta}_0 = \mathbf{0}$ where R is a $q \times k$ matrix.

$$\sqrt{n}\left(R\hat{\boldsymbol{\theta}}_n - R\boldsymbol{\theta}_0\right) \xrightarrow{d} N\Big(\mathbf{0},\ R\,\mathcal{I}(\boldsymbol{\theta}_0 : X)^{-1}R^T\Big)$$

$$n\left(R\hat{\boldsymbol{\theta}}_n - R\boldsymbol{\theta}_0\right)^T\Big(R\,\mathcal{I}(\hat{\boldsymbol{\theta}}_n : X)^{-1}R^T\Big)^{-1}\Big(R\hat{\boldsymbol{\theta}}_n - R\boldsymbol{\theta}_0\Big) \xrightarrow{d} \chi^2(q)$$

2. Nonlinear test $H_0 : g(\boldsymbol{\theta}_0) = \mathbf{0}$ where $g : \mathbb{R}^k \to \mathbb{R}^q$ is differentiable.

$$\sqrt{n}\left(g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta}_0)\right) \xrightarrow{d} N\left(\mathbf{0},\ \frac{\partial g(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}^T}\mathcal{I}(\boldsymbol{\theta}_0 : X)^{-1}\frac{\partial g(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\right)$$

$$n\left(g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta}_0)\right)^T\left(\frac{\partial g(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}^T}\mathcal{I}(\hat{\boldsymbol{\theta}}_n : X)^{-1}\frac{\partial g(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}}\right)^{-1}\left(g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta}_0)\right) \xrightarrow{d} \chi^2(q)$$

Aside. Wald Test. Let $\{\mathbf{X}_1, \cdots, \mathbf{X}_n\}$ be a random sample and let $\mathbf{X}$ be a random vector in $\mathbb{R}^k$ with the same probability distribution as $\mathbf{X}_i$'s where $\mathbb{E}[\mathbf{X}\mathbf{X}^T] < \infty$. Then it follows from CLT that

$$\sqrt{n}\left(\bar{\mathbf{X}}_n - \mathbb{E}[\mathbf{X}]\right) \xrightarrow{d} N(\mathbf{0}, \Sigma)$$

and it follows from WLLN that

$$\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}}_n)(\mathbf{X}_i - \bar{\mathbf{X}}_n)^T = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i^T - \bar{\mathbf{X}}_n\bar{\mathbf{X}}_n^T \xrightarrow{p} \mathbb{E}[\mathbf{X}\mathbf{X}^T] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]^T = \Sigma$$

$$n\Big(\bar{\mathbf{X}}_n - \mathbb{E}[\mathbf{X}]\Big)\hat{\Sigma}_n^{-1}\Big(\bar{\mathbf{X}}_n - \mathbb{E}[\mathbf{X}]\Big)^T \xrightarrow{d} \chi^2(k)$$

Glossary