Math Camp 2019 - Statistics*

Created by James Banovetz
Modified by Woongchan Jeon

Department of Economics, UC Santa Barbara

September 9, 2019

1. Random Samples

(a) Definition. $\mathbb{X}\mathcal{X}$ The random variables $X_1, \ldots, X_n$ are called a **random sample** of size n from the population $f_X(x)$ if $X_1, \ldots X_n$ 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 $X_1, \ldots, X_n$ are **independentan identically distributed** (or i.i.d.).

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

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

(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:

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

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

Then the joint distributions is:

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

(d) Aside. For the rest of math camp, we'll be assuming that we're working with a random sample. 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 note 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 $Y_n = T(X_1, \ldots, X_n)$ is a **statistic**.

(b) Example. 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 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_n is called the **sampling distribution** of Y_n.

(d) Example. 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 $x_1, \ldots, x_n$ 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 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:

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

(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 $\bar{X}_n = \frac{1}{n}\sum X_i$ 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

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

$$X_i = \mu + U_i \text{ where } U_i \overset{i.i.d.}{\sim} N(0, \sigma^2) \ \forall i = 1, \cdots, n$$

or equivalently

$$\vec{\mathbf{X}} = \boldsymbol{\iota}\mu + \vec{\mathbf{U}} \text{ where } \vec{\mathbf{U}} \sim N(\mathbf{0}, \sigma^2 I_n)$$

- The first point is easy to prove using MGFs and is a common result.

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = (\boldsymbol{\iota}^T\boldsymbol{\iota})^{-1}\boldsymbol{\iota}^T\vec{\mathbf{X}} = \mu + (\boldsymbol{\iota}^T\boldsymbol{\iota})^{-1}\boldsymbol{\iota}^T\vec{\mathbf{U}}$$

$$\mathbb{E}[\hat{\mu}_n] = \mu + (\boldsymbol{\iota}^T\boldsymbol{\iota})^{-1}\boldsymbol{\iota}^T\mathbb{E}[\vec{\mathbf{U}}] = \mu$$

$$\text{Var}(\hat{\mu}_n) = (\boldsymbol{\iota}^T\boldsymbol{\iota})^{-1}\boldsymbol{\iota}^T(\sigma^2 I_n)\boldsymbol{\iota}(\boldsymbol{\iota}^T\boldsymbol{\iota})^{-1} = \sigma^2(\boldsymbol{\iota}^T\boldsymbol{\iota})^{-1} = \frac{1}{n}\sigma^2$$

- Suppose that $\vec{\mathbf{U}} \sim N(\mathbf{0}, \sigma^2 I_n)$ and M is an $n \times n$ symmetric idempotent matrix. Then

$$\frac{1}{\sigma^2}\vec{\mathbf{U}}^T M\vec{\mathbf{U}} \sim \chi^2(\text{rank}(M)) = \chi^2(\text{tr}(M))$$

$$\frac{(n-1)S^2}{\sigma_n^2} = \frac{1}{\sigma^2}\vec{\mathbf{U}}^T M_{\boldsymbol{\iota}}\vec{\mathbf{U}} \sim \chi^2(n-1) \quad \text{where} \quad M_{\boldsymbol{\iota}} = I_n - \boldsymbol{\iota}(\boldsymbol{\iota}^T\boldsymbol{\iota})^{-1}\boldsymbol{\iota}^T$$

- Suppose that $\vec{\mathbf{U}} \sim N(\mathbf{0}, \sigma^2 I_n)$ and M: $n \times n$ symmetric idempotent matrix, L: $m \times n$ matrix and $LM = \mathbf{0}$. Then $M\mathbf{u}$ and $L\mathbf{u}$ are independent.

$$\begin{bmatrix} M \\ L \end{bmatrix}\vec{\mathbf{U}} \text{ is jointly normally distributed.}$$

$$LM = 0 \Rightarrow \text{Cov}(M\vec{\mathbf{U}}, L\vec{\mathbf{U}}) = 0$$

$$g(M\vec{\mathbf{U}}) \text{ and } h(L\vec{\mathbf{U}}) \text{ are independent.}$$

$$M = I_n - \boldsymbol{\iota}(\boldsymbol{\iota}^T\boldsymbol{\iota})^{-1}\boldsymbol{\iota}^T \text{ and } L = (\boldsymbol{\iota}^T\boldsymbol{\iota})^{-1}\boldsymbol{\iota}^T$$

(c) Aside. Why do we care about $\bar{X}_n$ and S_n^2 from normal distribution? 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. $$\frac{\bar{X}_n - \mu}{\sqrt{S_{x,n}^2/n}} = \frac{\frac{\bar{X}_n - \mu}{\sigma/n}}{\sqrt{\frac{(n-1)S_{x,n}^2/\sigma^2}{n-1}}} \sim t_{n-1}$$

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

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 easily 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 $X_1, \ldots X_n$ 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:

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

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:

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

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

5. **Three Corner Stones of Econometrics**

(a) Identification. Given a statistical model, relating a parameter of interests to an estimand is called **identification**.
Definition An estimand is a real number which is a function of the probability distribution of the random variables we will get to observe.

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

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

(b) Estimation. Proposing an estimator for an estimand is called **estimation**.
Definition An estimator is a function of random variables we will get to observe.

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

$$\hat{\boldsymbol{\beta}}_n = \left(\frac{1}{n}\sum_{i=1}^{n}\vec{\mathbf{X}}_i\vec{\mathbf{X}}_i^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\vec{\mathbf{X}}_i Y_i\right) = \left(\mathcal{X}^T\mathcal{X}\right)^{-1}\mathcal{X}^T\vec{\mathbf{Y}}$$

(c) Inference. Using an estimator to infer plausible values of an estimand is called **inference**.
Definition An estimate is a realized value of the estimator given a realized sample.

$$H_0 : \mu = 0$$

Given a **realized** sample $\{y_i\}_{i=1}^n$,

$$T_n(\alpha) = \mathbb{I}\left\{\left|\frac{\bar{Y}_n}{S_n^2/n}\right| > q_{1-\frac{\alpha}{2}}\right\}$$

$$H_0 : \beta_j = 0$$

Given a **realized** sample $\{\mathbf{x}_i, y_i\}_{i=1}^n$,

$$T_n(\alpha) = \mathbb{I}\left\{\left|\frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}\right| > q_{1-\frac{\alpha}{2}}\right\}$$

6. **Point Estimation**

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

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

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:

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

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

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:

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

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

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

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

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

(f) Aside. Note that we're making a methodological change here: we're treating the values of $x_1, \ldots, x_n$ are fixed, and we're varying the values $\theta_1, \ldots, \theta_n$. 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 $X_1, \ldots, X_n$ 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

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

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 $l(\cdot)$:

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

Differentiating with respect to our parameters:

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

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) 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.

(j) 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) Definition. 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**.

(b) Definition. 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:

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

(c) Example.

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \quad \text{and} \quad \hat{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \sim \chi^2(n-1)$$

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

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

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

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

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

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

(d) 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.

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

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

Alternatively, we write $\vec{\mathbf{U}}_n \xrightarrow{p} \vec{\mathbf{V}}$.
Remark For convergence in probability, the individual convergence of the entries of the vector is necessary and sufficient for their joint convergence.

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

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

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

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

(e) Theorem. Suppose $Y_n \xrightarrow{p} Y$ and $Z_n \xrightarrow{p} Z$. Then

i. $cY_n \xrightarrow{p} cY$, where $c \in \mathbb{R}$.

ii. $Y_n + Z_n \xrightarrow{p} Y + Z$

iii. $Y_n Z_n \xrightarrow{p} YZ$

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

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

(g) 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 $\mu = \mathbb{E}X < \infty$.

i. $\hat{\theta}_n(X_1, \cdots, X_n) = X_1$ is unbiased for μ but not consistent.

ii. $\hat{\theta}_n(X_1, \cdots, X_n) = \frac{1}{n}\sum_{i=1}^n X_i + \frac{1}{n}$ is biased but consistent.

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

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

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

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

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

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

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

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

(k) Theorem.

i. (Slutsky Theorem) If $\vec{\mathbf{Y}}_n \xrightarrow{d} \vec{\mathbf{Y}}$ and $\vec{\mathbf{Z}}_n \xrightarrow{p} \mathbf{c}$ where $\mathbf{c}$ is constant, then

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

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

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

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

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

(l) Definition. Let $\{\vec{\mathbf{X}}_1, \cdots, \vec{\mathbf{X}}_n\}$ be a random sample. Let $\hat{\boldsymbol{\theta}}_n(\vec{\mathbf{X}}_1, \ldots, \vec{\mathbf{X}}_n)$ 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.

(m) Cramér-Rao Lower Bound.

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

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

If these hold, then the following matrix is positive semidefinite.

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

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,

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

$$\begin{aligned}
\frac{d}{d\boldsymbol{\theta}}\mathbb{E}\big[W(\vec{\mathbf{X}})\big] &= \int_{\mathbf{R}^n} \frac{\partial}{\partial \boldsymbol{\theta}} \Big[W(\vec{\mathbf{X}}) f_{\vec{\mathbf{X}}}(\mathbf{x}|\boldsymbol{\theta})\Big] d\mathbf{x} \\
&= \mathbb{E}\left[W(\vec{\mathbf{X}}) \frac{\frac{\partial}{\partial \boldsymbol{\theta}} f_{\vec{\mathbf{X}}}(\vec{\mathbf{X}}|\boldsymbol{\theta})}{f_{\vec{\mathbf{X}}}(\vec{\mathbf{X}}|\boldsymbol{\theta})}\right] \\
&= \mathbb{E}\left[W(\vec{\mathbf{X}}) \frac{\partial}{\partial \boldsymbol{\theta}} \log[f_{\vec{\mathbf{X}}}(\vec{\mathbf{X}}|\boldsymbol{\theta})]\right] \\
&= \mathrm{Cov}\left(W(\vec{\mathbf{X}}),\ \frac{\partial}{\partial \boldsymbol{\theta}} \log[f_{\vec{\mathbf{X}}}(\vec{\mathbf{X}}|\boldsymbol{\theta})]\right)
\end{aligned}$$

ii. 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.

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

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

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

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

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:

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

iv. Asymptotic Normality of MLE.

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

v. 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 $\mathbb{E}[\bar{X}] = \mu$

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

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

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

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

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

$$\begin{aligned}
\ln\left[f(x|\mu)\right] &= -\frac{1}{2}\ln\left[2\pi\sigma^2\right] - \frac{1}{2\sigma^2}(x_i-\mu)^2 && \text{(from above)}\\
\frac{\partial}{\partial\sigma^2}\ln\left[f(\cdot)\right] &= -\frac{1}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(x_i-\mu)^2 && \text{(differentiating w.r.t. } \sigma^2\text{)}\\
\frac{\partial^2}{\left(\partial\sigma^2\right)^2}\ln\left[f(\cdot)\right] &= \frac{1}{2(\sigma^2)^2} - \frac{(x-\mu)^2}{(\sigma^2)^3} && \text{(the second derivative)}\\
\mathbb{E}\left[\frac{\partial^2}{\left(\partial\sigma^2\right)^2}\ln\left[f(\cdot)\right]\right] &= \frac{1}{2\sigma^4} - \frac{\sigma^2}{\sigma^6} && \text{(taking the expectation)}\\
&= -\frac{1}{2\sigma^4} && \text{(simplifying)}
\end{aligned}$$

$$\begin{aligned}
\ln\left[f(x|\mu)\right] &= -\frac{1}{2}\ln\left[2\pi\sigma^2\right] - \frac{1}{2\sigma^2}(x_i-\mu)^2 && \text{(from above)}\\
\frac{\partial^2}{\partial\mu\partial\sigma^2}\ln\left[f(\cdot)\right] &= -\frac{(x-\mu)}{(\sigma^2)^2} && \text{(taking the cross partial derivative)}\\
\mathbb{E}\left[\frac{\partial^2}{\partial\mu\partial\sigma^2}\ln\left[f(\cdot)\right]\right] &= 0 && \text{(taking the expectation)}
\end{aligned}$$

$$\sqrt{n}\left(\begin{bmatrix}\hat{\mu}_n\\ \hat{\sigma}_n^2\end{bmatrix} - \begin{bmatrix}\mu\\ \sigma^2\end{bmatrix}\right) \xrightarrow{d} N\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}\sigma^2 & 0\\ 0 & 2\sigma^4\end{bmatrix}\right)$$

(n) Definition. 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

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

(o) Theorem

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

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

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

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

9. **Hypothesis Testing**

(a) Example. Let $\{X_1, \ldots, X_n\}$ 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.

(b) Definition. 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 $\mathbb{E}[X] = \mu$.

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

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

(c) Definition. 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.

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

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

(d) Definition. A statistical test is always attached to two mutually exclusive hypotheses on an estimand of interests, $\mathbb{E}[X]$ here. A **hypothesis** is a set of values for that estimand. The two complementary hypotheses are the **null hypothesis**, denoted H_0, and the **alternative hypothesis**, denoted H_1.

$$\begin{aligned} H_0 &= \{0\} & H_1 &= \mathbb{R} \setminus \{0\} \\ H_0 &= \{0\} & H_1 &= (-\infty, -3] \cup [3, \infty) \end{aligned}$$

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

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

(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 the null hypothesis when it is true; **Type II** error is when we fail to reject the null hypothesis when it is not true.

(f) Definition. 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.

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

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

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

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

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

i. $H_0 = \{0\}$ vs. $H_1 = \mathbb{R} \setminus \{0\}$
   - Level of $T_{500}(.05)$ is .05
   - Power of $T_{500}(.05)$ is .05

   (Trivial test) This is a negative result. 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.

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

(h) Asymptotic Test with MLE.

$$\hat{\boldsymbol{\theta}}_n \in arg \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^k} \mathcal{L}(\boldsymbol{\theta} : \vec{\mathbf{X}}_1, \cdots, \vec{\mathbf{X}}_n)$$

It follows from asymptotic normality of MLE that

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

and

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

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

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

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

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

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

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

(i) Wald Test.

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

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

and it follows from WLLN that

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

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