ECON 6190

Econometrics I: Chen Qiu

Gabe Sekeres

Fall 2024

Contents

Class Information This is a really exciting time to learn Econometrics, especially at Cornell. There's a long history at Cornell of treating Econometrics essentially as a decision problem. For Chen, Econometrics is useful and powerful because it helps people make decisions with their data.

The course material will be divided into three parts:

1. Introduction to statistical inference
2. Large-sample approaches to statistical inference
3. Classical theory of estimation and inference

There will be in-class midterms on Tuesday October 8 and Tuesday November 5.

1 Introduction to Statistical Inference

1.1 Sampling Models

Economists often collect data that consist of some observations on variables of interest. Statistically, this is a random sample from a large population, from which we can learn about the population.

We call X a random variable / vector of interest (*e.g.* , wage and education in the US). We say that $X \sim F$ where F is the true distribution of wage and education in the US. $X = (X_w, X_e)$. We might be interested in the joint CDF of X, denoted $F(x_w, x_e) = P\{X_w \leq x_w,\ X_e \leq x_e\}$. We could extend these notions to the discrete case, where there is a joint PMF, $f(x_w, x_e) = P\{X_w = x_w,\ X_e = x_e\}$. We can also define the joint PDF in the continuous case,

$$f(x_w, x_e) = \frac{\partial^2 F(x_w, x_e)}{\partial x_w \partial x_e}$$

and note that $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(u, w)dudw = 1$. Therefore, we have that

$$F(x_w, x_e) = P\{X_w \leq x_w,\ X_e \leq x_e\} = \int_{-\infty}^{x_w}\int_{-\infty}^{x_e} f(u, w)dwdu$$

We may also be interested in the *marginal distribution*, which says that given some $\{X_w, X_e\}$, we might be interested in the distribution of the wage (X_w) irrespective of the education (X_e). We say that the marginal distribution (or marginal CDF) is

$$F_{X_w}(x_w) = P\{X_w \leq x_w\} = P\{X_w \leq x_w,\ X_e \leq \infty\}$$

In the continuous case, we have the (marginal) pdf of X_W which is $f_{X_w}(x_w) = \int_{-\infty}^{\infty} f(x_w, x_e)dx_e$.

In the discrete case, the (marginal) pmf of X_W is $f_{X_w}(x_w) = P\{X_w = x_w\} = \sum_{t\in\mathbb{R}} P\{X_w = x_w,\ X_e = t\}$.

Now think about conditional distributions. Consider, for example, the distribution of wage conditional on education being a certain level. In the discrete case, we have $P\{X_w \leq x_w \mid X_e = \overline{x_e}\}$. If we are in the continuous case, we can define the conditional pdf $f_{X_w|X_e}(x_w \mid x_e) = \frac{f(x_w, x_e)}{f_{X_e}(x_e)}$ (*i.e.* , the joint pdf over the marginal pdf).

In summary: starting from the population distribution $X \sim F$, we have different aspects that are themselves distributions – the joint, marginal, and conditional distributions. We are interested in them for different reasons.

In the sampling model, we observe n repeated observations from the distribution X, which are $\{X_1, \ldots, X_n\}$. The central question is, given these n observations, how can we make inferences about the population. First, we need to be precise about how $\{X_1, \ldots, X_n\}$ are generated from the population.

Definition. The *random sampling model* assumes that $\{X_1, \ldots, X_n\} \overset{\text{i.i.d.}}{\sim} F$. *i.e.* , they are

- Independent: $\{X_1, \ldots, X_n\}$ are mutually independent
- Identically distributed: $X_1, \ldots, X_n$ have the same marginal distribution F.

Why do we call F their marginal distribution? Because $\{X_1, \ldots, X_n\} \sim \mathcal{F}$, which is the joint distribution of $\{X_1, \ldots, X_n\}$. We can define their joint CDF as follows

$$F(x_1, \ldots, x_n) = P\{X_1 \leq x_1, \ldots, X_n \leq x_n\} =^{\text{ind}} P\{X_1 \leq x_1\} \cdot P\{X_2 \leq x_2\} \cdots P\{X_n \leq x_n\}$$

$$= \prod_{i=1}^{n} P\{X \leq x_i\}$$

We can also construct the joint pdf / pmf:

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdot f_{X_2}(x_2) \cdots f_{X_n}(x_n) = f_X(x_1) \cdot f_X(x_2) \cdots f_X(x_n) = \prod_{i=1}^{n} f_X(x_i)$$

Importantly, if $\mathcal{F}$ is known, everything is known about the random sample.
Definition. The statistical approach to the sampling model is to consider the parameter $\theta = \theta(\mathcal{F})$, which is a function of the distribution. We have that $\theta = \mathbb{E}[X_w] = \int x f_w(x) dx$. We construct the test statistic $T(X_1, \ldots, X_n)$, which is any function of the data $\{X_1, \ldots, X_n\}$, (*e.g.* , $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} X_{w_i}$). Importantly, statistics are also random variables – they are just (weighted) sums of random variables.

We will use the statistic to infer on parameters – use a function of the data to infer what a function of the true distribution would look like.
Example. Estimation of θ. We want to form a guess of θ based on data $\{X_1, \ldots, X_n\}$.

$$\theta = \int x dF_w(x) \qquad , \qquad \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_{w_i}$$

Our goal is to pick a statistic as close as possible to θ.
Example. Hypothesis testing. $H_0 : \theta = 1$ versus $H_1 : \theta \neq 1$. For example, we might reject H_0 if $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_{w_i} \geq \frac{2}{n}$, and accept H_0 otherwise.

A statistic that implements the above procedure is $\mathbb{1}\{\hat{\theta} > \frac{2}{n}\}$. Our goal is to pick a statistic that makes fewer mistakes, which we call a 'high-quality' statistic. How would we find that? It's a random variable and a function of data, we are looking to choose the random variable to use. We need to study the distribution of $T(X_1, \ldots, X_n)$, which we call the *sampling distribution*. However, even under the random sampling model, the sampling distribution can be highly complicated.
Example. (Judging a coin) Want to know whether you have a fair coin by flipping it 10 times and recording 0 for each tail and 1 for each head. Our sample is $\mathcal{X} = \{X_1, \ldots, X_n\}$, where X_i is the result of the ith experiment. Note that $X_i \overset{\text{i.i.d.}}{\sim} Bernoulli(p)$, so the pmf of each X_i is $f(x_i) = p^{x_i}(1-p)^{1-x_i}$, and the pmf of $\mathcal{X}$ is $f_{\mathcal{X}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$. The goal is to make some judgement about p. A statistic is a function of $\mathcal{X}$, such as the number of heads, the order number of the first experiment to return heads, etc.
Example. (Estimating average income) Suppose you want to estimate the average income of a worker aged between 25 and 65 who lives in Ithaca. A sample of n workers $\mathcal{X}$, where $X_i \overset{\text{i.i.d.}}{\sim} F(\cdot)$, and $F(\cdot)$ is the unknown distribution of income. The parameter of interest if $\mu = \int x dF(x)$, and you could try the average of the sample, or the average of the 80% of middle values, or something else.
Definition. We will use four approaches to studying sampling distributions:

- A finite-sample approach. You could impose a 'nice' class of distributions $\mathcal{F}$, which hopefully makes the distribution of $\hat{\theta}$ tractable (*e.g.* , $\mathcal{X} \sim \mathcal{N}(\mu, \sigma^2)$). However, this can be a very strong assumption!

- Simulation. Instead of imposing normality or some other form, say that $\mathcal{X} \sim \mathcal{F}$, where $\mathcal{F}$ is a distribution that you think is true, say log normal. However, the functional forms are really complicated here. Instead, you could simulate data. For $b = 1, \ldots, B$, draw $\{X_{1b}, \ldots, X_{nb}\}$, where $X_{ib} \overset{\text{i.i.d.}}{\sim} \mathcal{F} \ \forall \ i = 1, \ldots, n$. Can calculate $\hat{\theta}_b = \frac{1}{n}\sum_{i=1}^{n} X_{ib}$. You now have B realizations of $\hat{\theta}$, for B very large. Then, we have that $\mathbb{E}[\hat{\theta}] \Rightarrow \frac{1}{B}\sum_{i=1}^{B} \hat{\theta}_i$, and $P\{\hat{\theta} \leq t\} \Rightarrow \frac{1}{B}\sum_{i=1}^{B} \mathbb{1}\{\hat{\theta} \leq t\}$
- Asymptotic approach. $\hat{\theta} = T(X_1, \ldots, X_n)$. The finite distribution is hard to track. Let $n \to \infty$. As n becomes large, the distribution of $\hat{\theta}$ is easier to track, *e.g.* , Central Limit Theorem:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right) \approx \mathcal{N}(0, Var(X)) \Rightarrow \frac{1}{n}\sum_{i=1}^{n} X_i \approx \mathcal{N}(\mathbb{E}[X], Var(X))$$

- Bootstrap. (See notes below). This is an alternative sampling model. There are two forms:
  - *Bootstrap with replacement*. We have $\{x_1, \ldots, x_N\}$, a finite population of N values. We draw one, then replace, then redraw again, n times. We get $\{X_1, \ldots, X_n\}$, where $P\{X = x_i\} = \frac{1}{N} \ \forall \ i$. The joint pmf of $\{X_1, \ldots, X_n\}$ is

$$P\{X_1 = t_1, X_2 = t_2, \ldots, X_n = t_n\} = \left(\frac{1}{n}\right)^n, \ \forall \ t_j \in \{x_1, \ldots, x_N\}, \ j = 1, \ldots, n$$

    Note that this is also an iid model, which is important.
  - *Bootstrap without replacement*. We have $\{x_1, \ldots, x_N\}$, a finite population of N values. We draw X_1, with probability $\frac{1}{N}$. We then draw X_2, with probability $\frac{1}{N-1}$, and so on. However, note that the sample we have drawn, $\{X_1, \ldots, X_n\}$ does *not* satisfy the iid assumption. They are not independent, but they are identically distributed.

Note that boostrap without replacement is identically distributed. This seems weird, because they have different probabilities – $\frac{1}{N}$ for X_1, $\frac{1}{N-1}$ for X_2, etc. However, we care about the *marginal* probability, so we are taking an *ex ante* perspective. The probability of X_1 being x for any $x \in \Omega$ is the same for all x, so they are identically distributed. Here is a quick proof of identically distributed:

Proof. $P\{X_1 = x\} = \frac{1}{N} \ \forall \ x \in \{x_1, \ldots, x_N\}$. How do we derive the marginal distribution of X_2? We have that $P\{X_2 = x\} = P\{X_2 = x\}P\{X_1 \neq x\} = \frac{1}{N-1}\frac{N-1}{N} = \frac{1}{N} \ \forall \ x \in \{x_1, \ldots, x_N\}$. □

Chen's Proof:

Proof. Law of total probability, where we note that the possible realizations of X_1 partition the sample space. We have that

$$\begin{aligned}
P\{X_2 = x\} &= \sum_{j=1}^{N} P\{X_2 = x, X_1 = x_j\} \\
&= \sum_{j=1}^{N} P\{X_1 = x_j\}P\{X_2 = x \mid X_1 = x_j\} \\
&= \sum_{j=1}^{N} \begin{cases} \frac{1}{N}\frac{1}{N-1} & x_j \neq x \\ \frac{1}{N}0 & x_j = x \end{cases} \\
&= (N-1)\frac{1}{N}\frac{1}{N-1} = \frac{1}{N}
\end{aligned}$$

□

1.2 Some Common Statistics

We have $X \in \mathbb{R} \sim F$, from which we draw iid data $\{X_1, \ldots, X_n\}$. We will define some common statistics, and think about their sampling distributions.

Definition. The *sample mean* $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is the arithmetic mean of the sample

Definition. The *sample variance* $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$

Definition. The *sample standard deviation* $s = \sqrt{s^2}$

$\bar{X}, s^2, s$ are all statistics, and all random variables. Our goal is to study their sampling distribution. This class, we will just look at their moments – their mean and variance. Let's state some simple facts: $\mathbb{E}[\bar{X}] = \mathbb{E}[X] = \mu$, where the left side is the expectation over the data, and the right is the expectation with respect to F. Also, $Var(\bar{X}) = \frac{Var(X)}{n} = \frac{\sigma^2}{n}$, and $\mathbb{E}[s^2] = Var(X) = \sigma^2$. To establish these results, we need some auxiliary results.

Theorem 1.1. *The following are true:*

- $\min_a \sum_{i=1}^n (X_i - a)^2 = \sum_{i=1}^n (X_i - \bar{X})^2$
- $(n-1)s^2 = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n(\bar{X})^2$

Proof.

$$\begin{aligned}
\sum_{i=1}^n (X_i - a)^2 &= \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - a)^2 \\
&= \sum_{i=1}^n \left[(X_i - \bar{X})^2 + (\bar{X} - a)^2 + 2(X_i - \bar{X})(X_i - a)\right] \\
&= \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - a)^2 + 0 \\
&= \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - a)^2
\end{aligned}$$

and since the right term is squared and always non-negative, the entire equation is minimized at precisely $a = \bar{X}$, and in that case we have that

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

Based on the above, let $a = 0$. Then we have that

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

which leads to

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

□

Theorem 1.2. *Let $\{X_1, \ldots, X_n\}$ be a random sample from the population. Let $g(x)$ be a function such that $\mathbb{E}\, g(X_1)$ and $var(X_1)$ exist. Then*

- $\mathbb{E}[\sum_{i=1}^n g(X_i)] = n\,\mathbb{E}[g(X_1)]$
- $Var(\sum_{i=1}^n g(X_i)) = nVar(g(X_1))$

Proof.

$$\begin{aligned}\mathbb{E}[\sum_{i=1}^{n} g(X_i)] &= \sum_{i=1}^{n} \mathbb{E}[g(X_i)] \\ &=_{iid} \sum_{i=1}^{n} \mathbb{E}[g(X_1)] = n\,\mathbb{E}[g(X_1)]\end{aligned}$$

$$\begin{aligned}Var(\sum_{i=1}^{n} g(X_i)) &= \mathbb{E}\left[\left(\sum_{i=1}^{n} g(X_i) - \mathbb{E}[\sum_{i=1}^{n} g(X_i)]\right)^2\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n} \left(g(X_i) - \mathbb{E}[g(X_i)]\right)^2\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^{n} (g(X_i) - \mathbb{E}[g(X_i)])\right) \cdot \left(\sum_{i=1}^{n} (g(X_i) - \mathbb{E}[g(X_i)])\right)\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{n} (g(X_i) - \mathbb{E}[g(X_i)])^2 + \sum_{i\neq j} (g(X_i) - \mathbb{E}[g(X_i)]) \cdot (g(X_j) - \mathbb{E}[g(X_j)])\right] \\ &= \sum_{j=1}^{n} \mathbb{E}\left[(g(X_i) - \mathbb{E}[g(X_i)])^2\right] + \sum_{i\neq j} \mathbb{E}\left[(g(X_i) - \mathbb{E}[g(X_i)]) \cdot (g(X_j) - \mathbb{E}[g(X_j)])\right] \\ &= \sum_{i=1}^{n} Var(g(X_i)) + \sum_{i\neq j} Cov(g(X_i), g(X_j)) \\ &= \sum_{i=1}^{n} Var(g(X_i)) \qquad \text{by independence} \\ &= nVar(g(X_1)) \qquad \text{by identical distribution}\end{aligned}$$

□

Theorem 1.3. *Let $\{X_1, \ldots, X_n\}$ be a random sample from a population with mean μ and variance σ^2. Then*

- $\mathbb{E}[\hat{X}] = \mu$
- $Var(\hat{X}] = \frac{\sigma^2}{n}$
- $\mathbb{E}[s^2] = \sigma^2$

1.3 Sampling from Normal Distribution

Assumption 1.1. $X \sim \mathcal{N}(\mu, \sigma^2)$.

The key thing we are studying is the distribution of $\{X_1, \ldots, X_n\} \sim_{\text{iid}} X \sim \mathcal{N}(\mu, \sigma^2)$. We say that $\{X_1, \ldots, X_n\} \sim$ (jointly normal / multivariate normal). This is important because any affine combination of $(X_1, \ldots, X_n)$ are jointly normal. Even stronger, any marginal or conditional distribution of the sample is similarly jointly normal.

Definition. A random variable Z has the *standard normal distribution*, written as $Z \sim \mathcal{N}(0, 1)$ if it has the density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \; x \in \mathbb{R}$$

The cdf of a standard normal does not have a closed form, but is written as

$$\Phi(x) = \int_{-\infty}^{x} \phi(u)du$$

Some key properties:

- $\int_{-\infty}^{\infty} \phi(u)du = 1$
- $\phi(x) = \phi(-x) \Rightarrow \Phi(-x) = 1 - \Phi(x)$
- If $Z \sim \mathcal{N}(0,1)$ and $X = \mu + \sigma Z$ for $\mu \in \mathbb{R}$ and $\sigma \geq 0$, then $X \sim \mathcal{N}(\mu, \sigma^2)$
- If $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\sigma > 0$, then X has the density

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \ x \in \mathbb{R}$$

Definition. All positive integer moments of the standard normal distribution are finite, because the tails of the density decline exponentially. Formally, if $Z \sim \mathcal{N}(0,1)$, then $\mathbb{E}[Z] = 0$ and $Var(Z) = 1$. For any $m \in \mathbb{N}$,

$$\mathbb{E}[Z^m] = \begin{cases} 0 & m \text{ odd} \\ 2^{-\frac{m}{2}} \frac{m!}{(m/2)!} & m \text{ even} \end{cases}$$

Definition. Let $X \sim \mathcal{N}(0,1)$. Then

$$f(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} \phi(X_i) = \left(\frac{1}{(2\pi)^{\frac{n}{2}}}\right) e^{-\frac{x'x}{2}}$$

where $x = (x_1, x_2, \ldots, x_n)'$. We call this the *multivarate standard normal* density. We say that $x \in \mathbb{R}^n \sim \mathcal{N}(0, I_n)$.

Definition. The expectation of $X \in \mathbb{R}^m$ is

$$\mathbb{E}[X] = \begin{bmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix}$$

Definition. The $m \times m$ covariance matrix of $X \in \mathbb{R}^m$ is

$$\Sigma = Var(X) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_m^2 \end{bmatrix}$$

where on the diagonal, $\sigma_j^2 = Var(X_j) \ \forall \ j = 1, \ldots, m$, and off the diagonal $\sigma_{ij} = Cov(X_i, X_j) \ \forall \ i \neq j$.

Theorem 1.4. *For $X \in \mathbb{R}^m$, Σ is symmetric and positive semi-definite.*

Proof. Symmetry follows from the fact that $Cov(X_i, X_j) = Cov(X_j, X_i)$. PSD follows from the fact that variances are weakly positive. □

Theorem 1.5. *If $X \in \mathbb{R}^m$ has expectation μ and covariance matrix Σ, and A is some matrix in $\mathbb{R}^{q \times m}$, then AX is a random vector with mean $A\mu$ and variance $A\Sigma A'$.*

Definition. If $Z \sim \mathcal{N}(0, I_m)$ and $X = \mu + BZ$, then $X \sim \mathcal{N}(\mu, \Sigma)$, where $\Sigma = B'B$. We say that X is *multivariate normal* If $X \sim \mathcal{N}(\mu, \Sigma)$ where Σ is invertible, then X has pdf

$$f(x) = \frac{1}{(2\pi)^{\frac{m}{2}} (\det \Sigma)^{\frac{1}{2}}} \exp\left(-\frac{(x-\mu)'\Sigma^{-1}(x-\mu)}{2}\right)$$

Theorem 1.6. *If X, Y are multivariate normal with $Cov(X,Y) = 0$, then $X \perp\!\!\!\perp Y$.*
Theorem 1.7. *If $X \sim \mathcal{N}(\mu, \Sigma)$, then $Y = a + BX \sim \mathcal{N}(a + B\mu, B\Sigma B')$.*
Theorem 1.8. *If (X,Y) are multivariate normal*

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_Y \\ \mu_X \end{pmatrix}, \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{XY} & \Sigma_{XX} \end{pmatrix} \right)$$

where $\Sigma_{XX}, \Sigma_{YY} > 0$, then the conditional distributions $Y \mid X$ and $X \mid Y$ are also normal:

$$Y \mid X \sim \mathcal{N}\left(\mu_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \mu_X), \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\right)$$

$$X \mid Y \sim \mathcal{N}\left(\mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(Y - \mu_Y), \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\right)$$

Wrong Statement: Assume that

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$$
$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

Then $X + Y \coloneqq T \sim$ Normal.
Remark. We have that $\mathbb{E}[T] = \mu_X + \mu_Y$, $var(T) = \sigma_X^2 + \sigma_Y^2 + 2\sigma_{XY}$. However, the joint distribution is not necessarily normal. That only holds if they are jointly normal.
Proposition 1.1. *If X is a multivariate normal distribution, then any of the marginal or conditional distributions are also multivariate normal.*
Theorem 1.9. *If $\{X_1, \dots, X_n\}$ are i.i.d $\mathcal{N}(\mu, \sigma^2)$, then $\hat{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.*

However, what is the variance? We have that the sample variance is

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

To study its distribution, we need to introduce a new distribution.
Definition. Let $\{Z_1, \dots, Z_r\}$ be $r > 0$ i.i.d $\mathcal{N}(0,1)$ random variables. Then $\sum_{i=1}^{r} X_i^2$ follows a *chi square distribution* with degrees of freedom r, denoted χ_r^2.
Theorem 1.10. *If $\{X_1, \dots, X_n\}$ are i.i.d $\mathcal{N}(\mu, \sigma^2)$, then*

1. *$\hat{X}_n$ and s^2 are independent*
2. *$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$.*

Proof. Statement 1: Define the residual $\hat{e}_i = X_i - \bar{X}_n \ \forall\ i$. Note that $\hat{e}_i$ is a linear combination of $\{X_1, \dots, X_n\}$, so it is also multivariate normal like they are. Also, $\mathbb{E}[\hat{e}_i] = \mathbb{E}[X_i] - \mathbb{E}[\bar{X}_n] = \mu - \mu = 0$, and

$$\begin{aligned} cov(\hat{e}_i, \bar{X}_n) &= \mathbb{E}[\hat{e}_i(\bar{X}_n - \mu)] \\ &= \mathbb{E}[(X_i - \mu + \mu - \bar{X}_n)(\bar{X}_n - \mu)] \\ &= \mathbb{E}[(X_i - \mu)(\bar{X}_n - \mu)] - \mathbb{E}[(\bar{X}_n - \mu)^2] \\ &= \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0 \end{aligned}$$

Since $\hat{e}_i$ and $\bar{X}_n$ are jointly normal, the fact that they are uncorrelated means that they are independent, and any function of $\hat{e}_i$ (including s^2) is also independent with $\bar{X}_n$. □
Definition. Let $Z \sim \mathcal{N}(0,1)$ by independent. Then $T = \frac{Z}{\sqrt{Q/r}}$ has a *student's t distribution with r degrees of freedom*, written as $T \sim t_r$

Theorem 1.11. *If X_i, $i = 1, \ldots, n$ are i.i.d. $\mathcal{N}(\mu, \sigma^2)$, then*

$$\frac{\bar{X}_n - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

Remark. Some facts about the t distribution:

- The pdf of t_r is symmetric around 0
- The pdf of t_r has heavier tails than $\mathcal{N}(0, 1)$
- Only the first $r - 1$ moments exist (as opposed to $\mathcal{N}(0, 1)$, where all moments exist)
- As $r \to \infty$, $t_r \to \mathcal{N}(0, 1)$.

1.4 Sufficient Statistics

Suppose we want to estimate a parameter $\theta \coloneqq \theta(\mathcal{F})$ from a population $X \sim \mathcal{F}$, where $X \coloneqq \{X_1, \ldots, X_n\}$ are drawn i.i.d. Ultimately, our goal is to pick a good statistic $T(X)$ to learn about θ. We should really think about all possible functions on the data, and choose the statistic that gives us the most information about θ. However, there are way too many candidate statistics to choose from! The concept of sufficient statistics lets us separate information from X into two parts – one containing useful information about θ, and one containing no useful information about θ. Formally:
Definition. A statistic $T(X)$ is *sufficient* for θ if the conditional distribution of X given $T(X)$ does not depend on θ.

Intuition. A sufficient statistic $T(X)$ contains all useful information about θ in the following sense. Researcher 1 is provided with X and can learn about θ from the pair $(X, T(X))$. Researcher 2 is provided with only $T(X)$, but since $T(X)$ is sufficient, the distribution of X given $T(X)$ is known to Researcher 2. Researcher 2 can back out the joint distribution of $(X, T(X))$ without knowing X, so the two researchers have the same information about θ. We have that if $f_{X|T}(x \mid t)$ does not depend on θ, then T is a sufficient statistic. Think about this as: 'once conditioned on T, X has no more useful information on θ.' Researcher 1 has $(X, T(X))$ and can infer the joint distribution $f(X, T(X))$. Researcher 2 can back out the conditional pdf/pmf of $T(X)$, denoted $f_T(t)$. They have the same information because $f(X, T(X)) = f_T(t) f_{X|T}(x \mid t)$. Since T is a sufficient statistic, $f_{X|T}(x \mid t)$ does not depend on θ, so it is completely known to the researcher.
Question. Given a sufficient statistic, what can we learn about the parameter being estimated?
Answer. Chen: In general, nothing – take as an example the sample mean, which is sufficient for the mean of a normal distribution OR the probability of a Bernoulli process. If we make assumptions about the underlying distribution, you can approach the question – one framework might be to think about learning a modeler's assumptions from the statistic they choose, assuming the underlying distribution is normal.
Theorem 1.12. *If $p(x \mid \theta)$ is the joint pdf or pmf of X and $q(t \mid \theta)$ is the pdf or pmf of a statistic $T(X)$, then $T(X)$ is a sufficient statistic for X if $\frac{p(x|\theta)}{q(t)}$ does not depend on θ for all x in the sample space.*

Proof. (Intuitive, discrete case) Consider

$$\begin{aligned}
\mathbb{P}\{X = x | T(X) = t\} &= \frac{\mathbb{P}\{X = x, T(X) = t\}}{\mathbb{P}\{T(X) = t\}} \\
&= \frac{\mathbb{P}\{X = x\}\mathbb{P}\{T(X) = t \mid X = x\}}{\mathbb{P}\{T(X) = t\}} \\
&= \frac{f_X(x \mid \theta)}{f_T(t)} \cdot \mathbb{1}_{T(X)=t} \\
&= \begin{cases} 0 & \text{if } T(X) \neq t \\ \frac{f_X(x|\theta)}{f_T(t)} & \text{if } T(X) = t \end{cases}
\end{aligned}$$

The first case of course does not depend on θ. Thus, it suffices to show that the ratio does not depend on θ to show that $\mathbb{P}\{X = x|T(X) = t\}$ does not depend on θ. □

Example. Let $X = \{X_1, \ldots, X_n\}$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$, with σ^2 known. We show that the sample mean $T(X) = \bar{X}$ is a sufficient statistic for μ. The joint pdf of the sample X is

$$\begin{aligned} f(x \mid \mu) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\sum_{i=1}^{n} \frac{(x_i-\bar{x}+\bar{x}-\mu)^2}{2\sigma^2}\right) \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right) \end{aligned}$$

Remember that in a normal sampling model, $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. Thus, we have that

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

which does not depend on μ. Thus, $\bar{X}$ is a sufficient statistic for μ.

Remark. The main drawback of this method is that you need to choose a T and write down its pdf. If you already have a good candidate statistic, this is a great method to use. However, if you are completely clueless about the ideal candidate statistic, this will be intensive. What should we do in that case?

Theorem 1.13. (Factorization Theorem) *Let $f(x \mid \theta)$ be the joint pdf or pmf of X. A statistic $T(X)$ is a sufficient statistic for θ if and only if there exist functions $g(t \mid \theta)$ and $h(x)$ such that, for all sample points x and all parameter points θ,*

$$f(x \mid \theta) = g(T(x) \mid \theta)h(x)$$

Proof. (Only for discrete case) ($\Rightarrow$): We have that $T(X)$ is sufficient. Choose $g(t \mid \theta) = \mathbb{P}_\theta\{T(X) = t\}$ and $h(x) = \mathbb{P}\{X = x \mid T(X) = T(x)\}$. Since $T(X)$ is sufficient, $h(x)$ does not depend on θ. For this choice, we have

$$\begin{aligned} f(x \mid \theta) &= \mathbb{P}_\theta\{X = x\} \\ &= \mathbb{P}_\theta\{X = x, T(X) = T(x)\} \\ &= \mathbb{P}_\theta\{T(X) = T(x)\}\mathbb{P}\{X = x \mid T(X) = T(x)\} \\ &= g(T(X) \mid \theta)h(x) \end{aligned}$$

($\Leftarrow$): Suppose that the factorization exists. Let $q(t)$ be the pmf of $T(X)$. To show that $T(X)$ is sufficient,

it suffices to examine the ratio $\frac{f(x|\theta)}{q(T(x))}$ for each x. Define $A_{T(x)} \coloneqq \{y \mid T(y) = T(x)$. Then we have:

$$\begin{aligned}
\frac{f(x \mid \theta)}{q(T(x)} &= \frac{g(T(x) \mid \theta)h(x)}{q(T(x))} \\
&= \frac{g(T(x) \mid \theta)h(x)}{\sum_{A_{T(x)}} f(x \mid \theta)} \\
&= \frac{g(T(x) \mid \theta)h(x)}{\sum_{A_{T(x)}} g(T(y) \mid \theta)h(y)} \\
&= \frac{g(T(x) \mid \theta)h(x)}{g(T(x) \mid \theta) \sum_{A_{T(x)}} h(y)} \\
&= \frac{h(x)}{\sum_{A_{T(x)}} h(y)}
\end{aligned}$$

Since this does not depend on θ, $T(x)$ is a sufficient statistic for θ. □

Example. Let $X = \{X_1, \ldots, X_n\}$ be iid $\mathcal{N}(\mu, \sigma^2)$, where σ^2 is known. We start by writing the joint pdf:

$$f(x) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\sigma}\right) \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right)$$

We will factorize this into a part including θ, and a part not including θ. Recall that we can split the joint pdf:

$$f(x) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2}{2\sigma^2}\right)$$

which becomes

$$f(x) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{2\sigma^2}\right) \exp\left(-\frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right)$$

Only the second exponential depends on $\theta \equiv \mu$. We say that

$$g(\bar{x} \mid \theta) = \exp\left(-\frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right)$$

and

$$h(x) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{2\sigma^2}\right)$$

Thus, we can conclude that $T(x) = \bar{x}$ is a sufficient statistic for μ in $\mathcal{N}(\mu, \sigma^2)$ model when σ^2 known.

Example. Let $X = \{X_1, \ldots, X_n\}$ be iid $\mathcal{N}(\mu, \sigma^2)$, where σ^2 is *unknown*. Now the parameters are (μ, σ^2). We can now write the joint pdf as

$$f(x \mid \mu, \sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{2\sigma^2}\right) \exp\left(-\frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right)$$

Since everything here depends on σ, this entire equation will be $g(T_1(x), T_2(x) \mid \mu, \sigma^2)$. We set $h(x) = 1$, $T_1(x) = \bar{x}$, and $T_2(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2$. Thus, we have that

$$f(x) = g(\bar{x}, s^2 \mid \mu, \sigma^2)h(x)$$

Remark. The entire sample $X = \{X_1, \ldots, X_n\}$ is always a sufficient statistic.

Remark. Any one-to-one function of a sufficient statistic is also a sufficient statistic (exercise)

Definition. A sufficient statistic $T^\star(X)$ is a *minimal sufficient statistic* if for any sufficient statistic $T(X)$, there exists a function r such that

$$T^\star(X) = r(T(X))$$

This definition implies that for any sufficient statistic $T(X)$, if $T(x) = T(y)$, then $T^\star(x) = T^\star(y)$. Intuitively, the minimal sufficient statistic achieves the most dimensional reduction without a loss of information about the parameters.

Theorem 1.14. *Let $f(x \mid \theta)$ be the joint pdf or pmf of X. Suppose that there exists a $T(X)$ such that, for any $x, y \in X$, the ratio*

$$\frac{f(x \mid \theta)}{f(y \mid \theta)}$$

does not depend on θ if and only if $T(x) = T(y)$, then $T(X)$ is a minimal sufficient statistic.

Proof. Left to reader □

Remark. Note that minimal sufficient statistics are not necessarily unique.

Example. Finding a minimal sufficient statistic for $X \sim \mathcal{N}(\mu, \sigma^2)$, with σ^2 unknown. Let x and y be two sample points, and let $(\bar{x}, s_x^2)$ and $(\bar{y}, s_y^2)$ be the sample means and variances respectively. It follows that

$$\begin{aligned}\frac{f(x \mid \theta)}{f(y \mid \theta)} &= \frac{(2\pi\sigma)^{-\frac{n}{2}} \exp\left(-\frac{(n-1)s_x^2 + n(\bar{x}-\mu)^2}{2\sigma^2}\right)}{(2\pi\sigma)^{-\frac{n}{2}} \exp\left(-\frac{(n-1)s_y^2 + n(\bar{y}-\mu)^2}{2\sigma^2}\right)} \\ &= \exp\left(\frac{(n-1)(s_x^2 - s_y^2) + n(\bar{y} - \bar{x}) + 2n\mu(\bar{x} - \bar{y})}{2\sigma^2}\right)\end{aligned}$$

This ratio does not depend on (μ, σ) if and only if $\bar{x} = \bar{y}$, and when $s_x^2 = s_y^2$. Thus, $(\bar{x}, s^2)$ is a minimal sufficient statistic.

Example. Let $\{X_1, \ldots, X_n\}$ be a random sample from the discrete uniform distribution on $\{1, 2, \ldots, \theta\}$. That is, the pmf for X_i is

$$f(x \mid \theta) = \begin{cases} \frac{1}{\theta} & x = 1, 2, \ldots, \theta \\ 0 & \text{otherwise} \end{cases}$$

Show that $\max_i X_i$ is a sufficient statistic for θ.

Solution. We will use the factorization theorem. First, we write down the joint pmf of the data:

$$f_X(x) = \prod_{i=1}^{n} f(x_i \mid \theta) = \begin{cases} \frac{1}{\theta^n} & x_i \in \{1, 2, \ldots, \theta\} \; \forall \; i = 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

This could also be written as

$$f_X(x) = \left(\frac{1}{\theta^n}\right) \mathbb{1}\left\{x_i \in \{1, 2, \ldots, \theta\} \; \forall \; i\right\}$$

We can split this as follows:

$$f_X(x) = \left(\frac{1}{\theta^n}\right) \cdot \mathbb{1}\left\{x_i \in \mathbb{Z}_+\right\} \cdot \mathbb{1}\left\{\max_i x_i \leq \theta\right\}$$

So defining $h(x) = \mathbb{1}\left\{x_i \in \mathbb{Z}_+\right\}$ and $g(\max_i x_i \mid \theta) = \frac{1}{\theta^n} \cdot \mathbb{1}\left\{\max_i x_i \leq \theta\right\}$, we get that

$$f_X(x) = h(x) g(\max_i x_i \mid \theta)$$

So $\max_i x_i$ is a sufficient statistic for θ.

1.5 Examples of Estimators and Measures of Their Quality

Definition. An *estimator* $\hat{\theta}$ for a parameter θ is also a statistic, intended as a guess about θ. $\hat{\theta}$ is an estimate when it is a specific (or realized) value calculated in a specific sample.

Let the population parameter be $\mu = \mathbb{E}[X]$. The sample mean is $\bar{X}_n = \hat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i$.

Let the population parameter be $\theta = \mathbb{E}[g(X)]$ for some known function g. An estimator is a sample mean of g: $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n g(X_i)$.

Let the population parameter be $\beta = h(\mathbb{E}[g(X)])$ for known functions h and g. A plug-in estimator for β is $\hat{\beta} = h(\hat{\theta}) = h(\frac{1}{n}\sum_{i=1}^n g(X_i))$.

Definition. The *bias* of an estimator $\hat{\theta}$ of a parameter θ is $bias[\hat{\theta}] = \mathbb{E}[\hat{\theta}] - \theta$. An estimator is *unbiased* if the bias is 0. Note that bias depends on the population distribution F.

Definition. Let $\mathcal{F}$ be a family of possible distributions. An estimator $\hat{\theta}$ is *unbiased in* $\mathcal{F}$ if $bias[\hat{\theta}] = 0$ for every $F \in \mathcal{F}$.

Theorem 1.15. *$\bar{X}$ is unbiased for $\mu = \mathbb{E}[X]$ if $\mathbb{E}[X] < \infty$.*

One common criterion for a good estimator is the *mean squared error* where

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

Theorem 1.16. *If $\hat{\theta}$ has finite variance, then*

$$mse(\hat{\theta}) = (bias(\hat{\theta}))^2 + var(\hat{\theta})$$

Proof. We have that

$$mse(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\theta] + \mathbb{E}[\theta] - \theta)^2]$$

□

Given a set of estimators, we say that the estimator with the smallest mean squared error is said to be *more efficient*. We generally don't talk about the *most efficient* estimator, because it will only perform well for a particular θ – think of an estimator which is constant, so has 0 MSE but only works for a specific θ.

We can restrict the set of estimators to only consider *unbiased* estimators:

$$S_u \coloneqq \{\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k\}$$

where $\mathbb{E}[\hat{\theta}_i] = \theta \; \forall \; i = 1, \ldots, k$.

Among the class of unbiased estimators, the estimator with the lowest sampling variance also has the lowest mean squared error. This motivates finding the *best unbiased estimator* for estimating parameter θ.

Theorem 1.17. *If $\sigma^2 < \infty$, the sample mean $\bar{X}_n$ has the lowest variance among all linear unbiased estimators of μ.*

Proof. Consider a class of linear estimators $\tilde{\mu} = \sum_{i=1}^n w_i X_i$ with some weights $\{w_1, \ldots, w_n\}$. Unbiasedness requires

$$\mu = \mathbb{E}[\tilde{\mu}] = \sum_{i=1}^n w_i \, \mathbb{E}[X_i] = \sum_{i=1}^n w_i \mu$$

which holds if and only if $\sum_{i=1}^n w_i = 1$. The variance of $\tilde{\mu}$ is

$$var(\tilde{\mu}) = var\left(\sum_{i=1}^n w_i X_i\right) = \sum_{i=1}^n w_i^2 var(X_i) = \sigma^2 \sum_{i=1}^n w_i^2$$

Which is minimized with $w_i = \frac{1}{n}$ for all i. Thus, the sample mean is the best unbiased estimator for the population mean. □

We actually have a much stronger statement:

Theorem 1.18. *If $\sigma^2 < \infty$, the sample mean $\bar{X}_n$ has the lowest variance among all unbiased estimators of μ.*

Definition. The *variance of an estimator* $\hat{\theta}$, also called the *sampling variance*, is $var[\hat{\theta}] = \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2]$.

We already know that if $\mathbb{E}[X^2] < \infty$, then $var(\bar{X}) = \frac{\sigma^2}{n}$ where $\sigma^2 = var(X)$. Therefore, the variance of $\bar{X}$ declines with sample size at rate $\frac{1}{n}$.

Remark. Sampling variance is the variance of an estimator and is usually unknown!

To estimate $var(\bar{X})$, we need an estimator for $\sigma^2 = var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$. The plug-in estimator for σ^2 is

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

Theorem 1.19. *If* $\sigma^2 < \infty$, *then* $\mathbb{E}[\hat{\sigma}^2] = \left(1 - \frac{1}{n}\right)\sigma^2$.

Question. Is there an unbiased estimator for σ^2? Yes, sample variance s^2.

Definition. The *standard error* of an estimator $\hat{\theta}$ for parameter θ is

$$se(\hat{\theta}) = \hat{V}^{\frac{1}{2}}, \text{ where } \hat{V} \text{ is an estimator for } V = var(\hat{\theta})$$

Standard error can be interpreted as an estimator for $V^{1/2}$, the *standard deviation* of $\hat{\theta}$. Standard error is usually a biased estimator of $V^{1/2}$.

Example. Sample mean $\bar{X}_n$ is an estimator for μ. The exact variance of $\bar{X}_n$ is $\frac{\sigma^2}{n}$. If we estimate σ^2 with the plug-in estimator $\hat{\sigma}^2$, the standard error of $\bar{X}_n$ is $\sqrt{\frac{\hat{\sigma}^2}{n}}$.

Let $X \in \mathbb{R}^m$ be a random vector and $\mu = \mathbb{E}[X]$ be its mean. The sample mean estimator for μ is

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n}X_i = \begin{pmatrix} \bar{X}_{1n} \\ \bar{X}_{2n} \\ \vdots \\ \bar{X}_{mn} \end{pmatrix}$$

Most properties of the univariate sample mean extend to the multivariate sample mean. It is unbiased, so $\mathbb{E}[\bar{X}_n] = \mu$, its exact covariance matrix is

$$var(\bar{X}_n) = \mathbb{E}\left[(\bar{X}_n - \mathbb{E}[\bar{X}_n])(\bar{X}_n - \mathbb{E}[\bar{X}_n])'\right] = \frac{1}{n}var(X) = \frac{\Sigma}{n}$$

The mean squared error matrix of $\bar{X}_n$ is

$$mse(\bar{X}_n) = \mathbb{E}\left[(\bar{X}_n - \mu)(\bar{X}_n - \mu)'\right] = \frac{\Sigma}{n}$$

$\bar{X}_n$ is the best unbiased estimator for μ. An unbiased covariance estimator is

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}\left[(X_i - \bar{X}_n)(X_i - \bar{X}_n)'\right]$$

In the 1950s, the James-Stein estimator was developed. It is a shrinkage estimator, where

$$\tilde{\theta}_{J-S} = \left(1 - \frac{(m-2)\sigma^2}{\hat{\theta}\hat{\theta}'}\right)\hat{\theta}$$

It is a biased estimator, but $mse(\tilde{\theta}_{J-S}) < mse(\hat{\theta})$

Suppose we have a random sample $X = \{X_1, \ldots, X_n\}$ from a distribution F_θ where $\theta \in \mathbb{R}^k$ is the parameter of interest. Let $\hat{\theta} \coloneqq \hat{\theta}(X)$ be a candidate estimator for θ that we, as researchers, think is "good" (*i.e.* has some desirable MSE properties). Suppose also that we know that $T(X)$ is a sufficient statistic for θ. Can we do better than $\hat{\theta}$?

Theorem 1.20. (Rao-Blackwell) *Under the above setup, let*

$$\tilde{\theta}(X) = \mathbb{E}\left[\hat{\theta}(X)\middle|T(X)\right]$$

Then,

1. $mse(\tilde{\theta}(X)) \leq mse(\hat{\theta}(X))$
2. *If $\hat{\theta}(X)$ is an unbiased estimator, then so is $\tilde{\theta}(X)$.*

Proof. First, we need to verify that $\tilde{\theta}$ is indeed an estimator, meaning that it is only a function of data and *not* a function of anything unknown (namely, θ). We have that

$$\tilde{\theta}(x) = \int \hat{\theta}(S) f_{X|T}(s \mid t) ds$$

for $t = T(x)$. By the precise definition of a sufficient statistic, this does not depend on θ, only on known data, since T is a sufficient statistic.

Next, we will show that $\text{MSE}(\tilde{\theta}(X)) \leq \text{MSE}(\hat{\theta}(X))$. We have that

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\
&= \mathbb{E}[(\hat{\theta} - \tilde{\theta} + \tilde{\theta} - \theta)^2] \\
&= \mathbb{E}[(\hat{\theta} - \tilde{\theta})^2] + \mathbb{E}[(\tilde{\theta} - \theta)^2] + 2\,\mathbb{E}[(\hat{\theta} - \tilde{\theta})(\tilde{\theta} - \theta)] \\
&=_{\text{LIE}} \mathbb{E}[(\hat{\theta} - \tilde{\theta})^2] + \mathbb{E}[(\tilde{\theta} - \theta)^2] + 2\,\mathbb{E}[(\hat{\theta} - \tilde{\theta})(\tilde{\theta} - \theta) \mid T] \\
&= \mathbb{E}[(\hat{\theta} - \tilde{\theta})^2] + \text{MSE}(\tilde{\theta}) + \Big((\mathbb{E}[\hat{\theta} \mid T] - \tilde{\theta})(\tilde{\theta} - \theta)\Big) \\
&= \mathbb{E}[(\hat{\theta} - \tilde{\theta})^2] + \text{MSE}(\tilde{\theta}) + (\tilde{\theta} - \tilde{\theta})(\tilde{\theta} - \theta) \\
&= \mathbb{E}[(\hat{\theta} - \tilde{\theta})^2] + \text{MSE}(\tilde{\theta}) \\
&\geq \text{MSE}(\tilde{\theta})
\end{aligned}$$

Finally, we will show that if the original estimator is unbiased, the alternative estimator is as well. Consider:

$$\begin{aligned}
\mathbb{E}[\tilde{\theta}] &= \mathbb{E}\left[\mathbb{E}[\hat{\theta} \mid T]\right] \\
&=_{\text{LIE}} \mathbb{E}[\hat{\theta}]
\end{aligned}$$

Thus, if $\hat{\theta}$ is unbiased, then $\mathbb{E}[\hat{\theta}] = \theta$, which means that $\mathbb{E}[\tilde{\theta}] = \theta$, meaning that $\tilde{\theta}$ is unbiased. □

Intuition. We can project the estimator into the sufficient statistic, and always get improvement.

Example. (PS4 Q5) Suppose we have a random sample with a Poisson distribution. Further suppose that we are interested in estimating the probability of a count of zero so $\theta = \mathbb{P}\{X = 0\} = e^{-\lambda}$. We have an unbiased estimator that is

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{X_i = 0}$$

Note that a sufficient statistic for λ is $T = \sum_{i=1}^{n} X_i$. Since $\hat{\theta}$ is not a function of T, just of the data, we can

definitely find improvement by Blackwellizing the estimator. We say that

$$\begin{aligned}
\tilde{\theta}(x) &= \mathbb{E}\left[\hat{\theta}\middle| T(x) = t\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i=0}\middle|\sum_{i=1}^{n}X_i = t\right] \\
&= \mathbb{E}\left[\mathbb{1}_{X_1=0}\middle|\sum_{i=1}^{n}X_i = t\right] \\
&= \mathbb{P}\left\{X_1 = 0\middle|\sum_{i=1}^{n}X_i = t\right\} \\
&= \frac{\mathbb{P}\left\{X_1 = 0, \sum_{i=1}^{n}X_i = t\right\}}{\mathbb{P}\left\{\sum_{i=1}^{n}X_i = t\right\}} \\
&= \frac{\mathbb{P}\left\{X_1 = 0, \sum_{i=2}^{n}X_i = t\right\}}{\mathbb{P}\left\{\sum_{i=1}^{n}X_i = t\right\}} \\
&= \frac{\mathbb{P}\{X_1 = 0\}\mathbb{P}\{\sum_{i=2}^{n}X_i = t\}}{\mathbb{P}\{\sum_{i=1}^{n}X_i = t\}}
\end{aligned}$$

and we can calculate the probabilities directly, using the properties of the Poisson distribution. We will get that this is

$$\tilde{\theta}(x) = \left(\frac{n-1}{n}\right)^t$$

2 Asymptotic Theory

We derived the distribution of $\bar{X}_n$ under a normal distribution assumption. This can be quite restrictive: What happens when the population is not normal? What is the distribution of nonlinear transformations of $\bar{X}_n$?

Idea: Allow n to grow to infinity and investigate the behavior of estimators as this happens.

- Pros: provides useful approximations for the finite-sample case; simpler results; asymptotic properties preserved under continuous transformations
- Cons: never realistic

The main tools of asymptotic theory are the law of large numbers (LLN), central limit theorem (CLT), and continuous mapping theorem (CMT).

2.1 Convergence in Probability

Definition. A sequence of numbers a_n has the *limit* a, or *converges* to a as $n \to \infty$ if for all $\delta > 0$, there exists n_δ such that for all $n > n_\delta$, $|a_n - a| < \delta$.

We think about asymptotic properties as follows. We have data $X = \{X_1, X_2, \ldots, X_n\}$, and construct a statistic $T(X) \coloneqq T(X_1, X_2, \ldots, X_n) \coloneqq T_n$. We think about the sequence of statistics $\{T_n\}_{n=1}^{\infty}$, which is indexed by sample size n.

A non-random sequence can converge to a limit. What about a sequence of random variables? For example, consider $\bar{X}_n$. In what sense does $\bar{X}_n$ converge as n increases? Since $\bar{X}_n$ is random, we need to modify the definition of convergence and limit. There are different ways to define this.

Let $\{X_n, n = 1, 2, \ldots\}$ be a sequence of random variables and let X be another random variable (it may be degenerate).
Definition. We say that X_n *converges in probability* to X if for all $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}\{|X_n - X| > \delta\} = 0$$

and

$$\lim_{n\to\infty} \mathbb{P}\{|X_n - X| \leq \delta\} = 1$$

Or, equivalently, for all $\delta, \varepsilon > 0$ there exists $n_{\delta,\varepsilon}$ such that for all $n > n_{\delta,\varepsilon}$,

$$\mathbb{P}\{|X_n - X| > \delta\} < \varepsilon$$

and

$$\mathbb{P}\{|X_n - X| \leq \delta\} \geq 1 - \varepsilon$$

We say that $X_n \xrightarrow{p} X$ if X_n converges in probability to X.
Example. Consider a discrete random variable Z_n such that

$$\mathbb{P}\{Z_n = 0\} = 1 - \frac{1}{n} \quad \text{and} \quad \mathbb{P}\{Z_n = a_n\} = \frac{1}{n}$$

for some arbitrary sequence a_n. We can show that $Z_n \xrightarrow{p} 0$ since for each $\delta > 0$, $\exists\, n$ such that

$$\mathbb{P}\{|Z_n - 0| > \delta\} \leq \mathbb{P}\{Z_n = a_n\} = \frac{1}{n} \to 0$$

Next, let X_n, X be $k \times 1$ random vectors with the jth element denoted X_{nj}, $j = 1, 2, \ldots, k$. Then $X_n \xrightarrow{p} X$ if and only if $X_{nj} \xrightarrow{p} X_j$ for all $j \in \{1, \ldots, k\}$. Convergence in probability is equivalent to elementwise convergence in probability. The same holds for matrices.

Definition. An estimator $\hat{\theta}_n$ based on a sample size n for parameter θ is *(weakly) consistent* if $\hat{\theta}_n - \theta \xrightarrow{p} 0$, *i.e.* , $\hat{\theta}_n \xrightarrow{p} \theta$.

Remark. Consistency is an asymptotic property of an estimator, typically a minimum requirement for any estimator, and is a different notion than the finite sample property of unbiasedness. In fact, many consistent estimators are biased or asymptotically biased.

Definition. An estimator $\hat{\theta}_n$ based on a sample size n for parameter θ is *asymptotically unbiased (AU)* if

$$\lim_{n\to\infty}\{\mathbb{E}[\hat{\theta}_n] - \theta\} = \{\lim_{n\to\infty}\mathbb{E}[\hat{\theta}_n]\} - \theta = 0$$

Theorem 2.1. *Consistency and asymptotic unbiasedness do not imply each other.*

Proof. ($\not\Leftarrow$) Suppose that $X \sim \mathcal{N}(\mu, \sigma^2)$, and our estimator for μ is $\hat{\mu} = X_1$. This is unbiased since $\mathbb{E}[X_1] = \mu$ so $\lim_{n\to\infty}\{\mathbb{E}[\hat{\mu}_n] - \mu\} = 0$, but $\mathbb{P}\{|\hat{\mu}_n - \mu| > \delta\}$ is constant, so does not go to zero.

($\not\Rightarrow$) Consider the following artificial example. Suppose the true parameter is θ and $\hat{\theta}_n$ is binary such that

$$\mathbb{P}\{\hat{\theta}_n = \theta\} = 1 - \frac{1}{n} \quad \text{and} \quad \mathbb{P}\{\hat{\theta}_n = n\} = \frac{1}{n}$$

$\hat{\theta}_n$ is consistent since for all $\delta > 0$,

$$\mathbb{P}\{|\hat{\theta}_n - \theta| > \delta\} \leq \mathbb{P}\{\hat{\theta}_n = n\} = \frac{1}{n} \to 0 \text{ as } n \to \infty$$

However, $\hat{\theta}_n$ is not asymptotically unbiased since

$$\mathbb{E}[\hat{\theta}_n] = \theta\left(1 - \frac{1}{n}\right) + n\left(\frac{1}{n}\right) = \theta - \frac{\theta}{n} + \frac{n}{n}$$

which approaches $\theta + 1$ as n gets large. □

Theorem 2.2. (Continuous Mapping Theorem) *Let X_n, X be $k \times 1$ random vectors. If $X_n \xrightarrow{p} X$ and g is a real-valued continuous function, then*

$$g(X_n) \xrightarrow{p} g(X)$$

Corollary 2.1. (Slutsky's Theorem) *Let g be continuous at c. Then*

$$X_n \xrightarrow{p} c \Longrightarrow g(X_n) \xrightarrow{p} g(c)$$

Corollary 2.2. $X_n \xrightarrow{p} X \Longrightarrow \|X_n - X\| \xrightarrow{p} 0$ *where $\|\cdot\|$ is the Euclidean norm.*

2.2 Proving Convergence in Probability

Definition. Let X be a random variable and A be an event. An *indicator function* is

$$\mathbb{1}_{X\in A} = \begin{cases} 1 & X \in A \\ 0 & X \notin A \end{cases}$$

Note that $\mathbb{E}[\mathbb{1}_{X\in A}] = \mathbb{P}\{X \in A\}$

Theorem 2.3. (Markov Inequality) *For each $r > 0$,*

$$\mathbb{P}\{|X| > \delta\} \leq \frac{\mathbb{E}[|X|^r]}{\delta^r} \quad \text{for all } \delta > 0$$

provided that $\mathbb{E}[|X|^r] < \infty$

Proof.

$$\begin{aligned}\mathbb{P}\{|X| > \delta\} &= \mathbb{E}[\mathbb{1}_{|X|>\delta}] \\ &\leq \mathbb{E}\left[\mathbb{1}_{|X|>\delta}\frac{|X|^r}{\delta^r}\right] \\ &= \frac{1}{\delta^r}\mathbb{E}[\mathbb{1}_{|X|>\delta}|X|^r] \\ &\leq \frac{\mathbb{E}[|X|^r]}{\delta^r}\end{aligned}$$

□

Definition. Assuming $\mathbb{E}[|X|^r] < \infty$. Then X_n *converges in rth mean*, written as $X_n \to_r X$, if

$$\lim_{n\to\infty}\mathbb{E}[|X_n - X|^r] = 0$$

Theorem 2.4. *For any* $r > 0$

$$X_n \to_r X \implies X_n \xrightarrow{p} X$$

Proof. By the Markov Inequality:

$$\lim_{n\to\infty}\mathbb{P}\{|X_n - X| > \delta\} \leq \lim_{n\to\infty}\frac{\mathbb{E}[|X_n - X|^r]}{\delta^r} = 0$$

□

Example. *Mean square convergence* is convergence in rth mean for $r = 2$. We can show that $\hat{\theta}_n \xrightarrow{p} \theta$ if

$$\underbrace{\mathbb{E}[\hat{\theta}_n - \theta]^2}_{\text{Mean square error}} \xrightarrow{p} 0, \text{ as } n \to \infty$$

Since

$$\mathbb{E}[\hat{\theta}_n - \theta]^2 = \text{bias}(\hat{\theta}_n)^2 + \text{Var}(\hat{\theta}_n)$$

we can show that $\hat{\theta} \xrightarrow{p} \theta$ if $\text{bias}(\hat{\theta}_n) \to 0$ and $\text{Var}(\hat{\theta}_n) \to 0$ as $n \to \infty$.

Theorem 2.5. $\hat{\theta}_n \to_r \theta$ *for some* $r \geq 1$ *implies that* $\lim_{n\to\infty}\mathbb{E}[\hat{\theta}_n] = \theta$.

Proof. See that:

$$\begin{aligned}\mathbb{E}[\hat{\theta}_n] - \theta &\leq \left|\mathbb{E}[\hat{\theta}_n - \theta]\right| \\ &\leq \mathbb{E}\left[|\hat{\theta}_n - \theta|\right] && \text{(By Jensen's)} \\ &\leq \left\{\mathbb{E}\left[|\hat{\theta}_n - \theta|\right]^r\right\}^{1/r} && \text{(By Jensen's again)} \\ &\to 0 \text{ as } n \to \infty\end{aligned}$$

□

Remark. $\hat{\theta}_n \to_r \theta$, g continuous $\Rightarrow g(\hat{\theta}_n) \xrightarrow{p} g(\theta)$. However, it is not true that $g(\hat{\theta}_n) \to_r g(\theta)$. $\mathbb{E}[|g(\hat{\theta}_n)|^r]$ might not even exist.

By applying the Markov Inequality with $r = 2$ and replacing X with the demeaned version $X - \mathbb{E}[X]$, we get *Chebyshev's Inequality*:

Definition. We have that:

$$\mathbb{P}\{|X - \mathbb{E}[X]| > \delta\} \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^2]}{\delta^2} = \frac{\text{Var}(X)}{\delta^2} \text{ for all } \delta > 0$$

Remark. An estimator $\hat{\theta}_n \xrightarrow{p} \mathbb{E}[\hat{\theta}_n]$ if $\text{Var}(\hat{\theta}_n)$ is vanishing to 0.

Theorem 2.6. (Chebyshev's Weak Law of Large Numbers) *If $\{X_i, i = 1, \ldots, n\}$ are iid with mean μ and finite variance σ^2, then*

$$\bar{X}_n \xrightarrow{p} \mu$$

Proof. Recall that we've shown under iid that

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

Applying Chebyshev's Inequality yields

$$\mathbb{P}\{|\bar{X}_n - \mu| > \delta\} = \mathbb{P}\{|\bar{X}_n - \mathbb{E}[\bar{X}_n]| > \delta\} \leq \frac{\text{Var}(\bar{X}_n)}{\delta^2} = \frac{\sigma^2}{n\delta^2} \to 0 \text{ for all } \delta > 0$$

□

Theorem 2.7. (Khinchine's Weak Law of Large Numbers) *If $\{X_i, i = 1, \ldots, n\}$ are iid with $\mathbb{E}[X_i] < \infty$, then*

$$\bar{X}_n \xrightarrow{p} \mathbb{E}[X_i] = \mu$$

Proof. Technical, so omitted. Relies on showing that $\mathbb{E}[|\bar{X}_n - \mu|] \to 0$, which is convergence in rth mean when $r = 1$. □

Remark. This does not require finite variance, so is stronger than Chebyshev's Weak Law of Large Numbers. We often call this *the Weak Law of Large Numbers*.

We will now extend this result to the vector case.

Theorem 2.8. *Suppose $X_i \in \mathbb{R}^m, i = 1, \ldots, n$ are iid distributed and $\mathbb{E}\,\|X_i\| = \mathbb{E}\,\|X\| < \infty$, then*

$$\bar{X}_n \xrightarrow{p} \mathbb{E}\,X$$

as $n \to \infty$.

Proof omitted. Note that $\mathbb{E}\,\|X\| < \infty$ if and only if $\mathbb{E}\,|X_j| < \infty$ for all $j = 1, \ldots, m$.

2.3 Almost Sure Convergence

Convergence in probability is sometimes called *weak convergence*. Consider a stronger version: *almost sure convergence*, also called *strong convergence* or *convergence with probability one*

Definition. We say that X_n *converges almost surely* to X, denoted $X_n \xrightarrow{a.s.} X$, if

$$\mathbb{P}\left\{\lim_{n\to\infty} X_n = X\right\} = 1$$

or, equivalently, for all $\delta > 0$ and $\varepsilon > 0$,

$$\mathbb{P}\left\{|X_m - X| \leq \delta \text{ for all } m \geq n_{\delta,\varepsilon}\right\} > 1 - \varepsilon$$

Theorem 2.9. $X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{p} X$

Proof. Recall that if $C \Rightarrow D$, then $\mathbb{P}\{C\} \leq \mathbb{P}\{D\}$. Since we have that $X_n \xrightarrow{a.s.} X$, for all $\varepsilon > 0$, $\delta > 0$ there exists $n_{\delta,\varepsilon} > 0$ such that for all $m > n_{\delta,\varepsilon}$, we have that

$$\mathbb{P}\{|X_m - X| \leq \delta \;\forall\; m > n_{\delta,\varepsilon}\} > 1 - \varepsilon \iff \mathbb{P}\left\{\bigcap_{m=n_{\delta,\varepsilon}}^{\infty} \{|X_m - X| \leq \delta\}\right\} > 1 - \varepsilon$$

Take $D = |X_m - X| \leq \delta$ for any $m > n_{\delta,\varepsilon}$, and $C = \bigcap_{m=n_{\delta,\varepsilon}}^{\infty} \{|X_m - X| \leq \delta\}$. Clearly, $C \Longrightarrow D$. Thus, for any $m > n_{\delta,\varepsilon}$,

$$\begin{aligned}\mathbb{P}\{|X_m - X| \leq \delta\} &= \mathbb{P}\{D\} \\ &\geq \mathbb{P}\{C\} = \mathbb{P}\{\cap_{m=n_{\delta,\varepsilon}}^{\infty} \{|X_m - X| \leq \delta\}\} \\ &> 1 - \varepsilon\end{aligned}$$

□

2.4 Stochastic Orders of Magnitude

Definition. (*Nonstochastic Orders*) For nonstochastic sequences x_n and f_n, $n = 1, \ldots,$ we have that

1. $x_n = o(f_n)$ if $\frac{x_n}{f_n} \to 0$ as $n \to \infty$
2. $x_n = O(f_n)$ if $\frac{x_n}{f_n}$ is bounded for sufficiently large n, meaning that there exists M such that for all $n > n_M$, $|\frac{x_n}{f_n}| < M$

Definition. (*Stochastic Orders*) For X_n random variables and f_n constants, for $n = 1, \ldots,$ we have that

1. $X_n = o_p(f_n)$ if $\frac{X_n}{f_n} \xrightarrow{p} 0$
2. $X_n = O_p(f_n)$ if $\frac{X_n}{f_n}$ is bounded in probability, meaning that for all $\varepsilon > 0$, there exists a constant $M_\varepsilon < \infty$ and $n_{\varepsilon,M} > 0$ such that

$$\mathbb{P}\left\{\left|\frac{X_n}{f_n}\right| > M_\varepsilon\right\} < \varepsilon \quad \text{for all } n > n_{\varepsilon,M}$$

Remark. $X_n = o_p(1)$ means that $X_n \xrightarrow{p} 0$.

Theorem 2.10. *If $X_n \xrightarrow{p} c$ for some constant c, then $X_n = O_p(1)$.*

Proof. Fix $\varepsilon > 0$. It suffices to show that there exists a constant C_ε such that $\mathbb{P}\{|X_n| > C_\varepsilon\} < \varepsilon$. Since $X_n \xrightarrow{p} c$, we know that for each $\delta > 0$ there exists $n_{\delta,\varepsilon}$ such that $\mathbb{P}\{|X_n - c| > \delta\} < \varepsilon$ for all $n > n_{\delta,\varepsilon}$. By the Triangle Inequality, we have that $|X_n| \leq |X_n - c| + |c|$. Choose $C = |c| + \delta$. Then we have that

$$\begin{aligned}\mathbb{P}\{|X_n| > C\} &= \mathbb{P}\{|X_n| > |c| + \delta\} \\ &\leq \mathbb{P}\{|X_n - c| + |c| > |c| + \delta\} \\ &= \mathbb{P}\{|X_n - c| > \delta\} < \varepsilon\end{aligned}$$

since $X_n \xrightarrow{p} c$. □

Definition. Some algebraic definitions:

- If $X_n = O_p(f_n)$ and $Y_n = O_p(g_n)$, then $X_n Y_n = O_p(f_n g_n)$ and $X_n + Y_n = O_p(\max\{f_n, g_n\})$
- The same holds for o
- If $X_n = O_p(f_n)$ and $Y_n = o_p(g_n)$, then $X_n Y_n = o_p(f_n g_n)$
- If $X_n = O_p(f_n)$ and $\frac{f_n}{g_n} \to 0$, then $X_n = o_p(g_n)$

Example. (Using Stochastic Orders) Suppose $X \sim \{X_1, \ldots, X_n\}$ are iid with finite variance σ^2. We know from the weak law of large numbers that $\bar{X}_n \xrightarrow{p} \mu$. But how fast does $\bar{X}_n$ converge to μ? Recall that by Chebyshev's Inequality, $\mathbb{P}\{|\bar{X}_n - \mu| > \delta\} = \frac{\sigma^2}{n\delta^2}$. It also implies that for all $\delta > 0$,

$$\mathbb{P}\left\{\frac{|\bar{X}_n - \mu|}{\frac{1}{\sqrt{n}}} > \delta\right\} = \mathbb{P}\left\{|\bar{X}_n - \mu| > \frac{1}{\sqrt{n}} > \delta\right\} \leq \frac{\sigma^2}{\delta^2}$$

We can choose $C_\varepsilon = \frac{\sigma}{\sqrt{\varepsilon}}$ such that

$$\mathbb{P}\left\{\frac{|\bar{X}_n - \mu|}{\frac{1}{\sqrt{n}}} > C_\varepsilon\right\} \leq \varepsilon$$

Thus, $\bar{X}_n - \mu = O_p(1/\sqrt{n})$ or equivalently, $\bar{X}_n = \mu + O_p(1/\sqrt{n})$, so $\bar{X}_n$ converges to μ at a rate no slower than $\frac{1}{\sqrt{n}}$.

Theorem 2.11. *$X_n = O_p\left(\mathbb{E}\left[|X_n|^r\right]^{\frac{1}{r}}\right)$ for $r > 0$*

Proof. Fix some $\varepsilon > 0$, and pick $C_\varepsilon = \left(\frac{1}{\varepsilon}\right)^{\frac{1}{r}}$. It follows from Markov's Inequality that

$$\begin{aligned}
\mathbb{P}\left\{\left|\frac{X_n}{\mathbb{E}\left[|X_n|^r\right]^{\frac{1}{r}}}\right| > C_\varepsilon\right\} &= \mathbb{P}\left\{|X_n| > \mathbb{E}\left[|X_n|^r\right]^{\frac{1}{r}} C_\varepsilon\right\} \\
&\leq \frac{\mathbb{E}[|X_n|^r]}{\mathbb{E}[|X_n|^r] C_\varepsilon^r} \\
&= \frac{1}{C_\varepsilon^r} = \varepsilon
\end{aligned}$$

□

2.5 Convergence in Distribution

Let $F_X(x) = \mathbb{P}\{X \leq x\}$ be the distribution function of random variable X, and consider a sequence of random variables X_n with distribution $F_{X_n}(x) = \mathbb{P}\{X_n \leq x\}$.

Definition. X_n *converges in distribution* to X (denoted $X_n \xrightarrow{d} X$) if

$$F_{X_n}(a) \to F_X(a) \text{ as } n \to \infty \; \forall \; a, \text{ where } F_X(a) \text{ is continuous}$$

Remark. It's quite difficult to show $X_n \xrightarrow{d} X$ by working directly with the distributions. Instead, we can work with the characteristic function.

Theorem 2.12. *$X_n \xrightarrow{d} X \iff C_{X_n}(t) \xrightarrow{d} C_X(t)$ as $n \to \infty$ for all t, where $C_X(t) = \mathbb{E}[\exp(itX)]$ is the characteristic function of X.*

Theorem 2.13. *We have that:*

1. $X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$
2. $X_n \xrightarrow{p} c \iff X_n \xrightarrow{d} c$ *for some constant* c
3. $X_n \xrightarrow{d} X \implies X_n = O_p(1)$

Proof. (Just of statement 2):

$\Rightarrow$: Recall that the CDF of a constant variable X such that $P\{X = c\} = 1$ is degenerate: $\mathbb{P}\{X \leq x\} = \mathbb{1}_{x \geq c}$. We want to show that (i) For each $\delta > 0$, $\mathbb{P}\{X_n \leq c - \delta\} \to 0$ as $n \to \infty$, and (ii) For each $\delta > 0$, $\mathbb{P}\{X_n \leq c + \delta\} \to 1$ as $n \to \infty$. For (i), note that

$$\mathbb{P}\{X_n \leq c - \delta\} = \mathbb{P}\{X_n - c \leq -\delta\} \leq \mathbb{P}\{|X_n - c| \leq \delta\} \to 0 \text{ as } n \to \infty$$

from the definition of $X_n \xrightarrow{p} c$. We can see (ii) by a similar argument.

$\Leftarrow$: Fix some $\delta > 0$, and we have that

$$\begin{aligned}
\mathbb{P}\{|X_n - c| > \delta\} &= \mathbb{P}\{X_n - c > \delta\} + \mathbb{P}\{X_n - c < -\delta\} \\
&\leq 1 - F_{X_n}(c + \delta) + F_{X_n}(c - \delta) \\
&\to 1 - 1 + 0 = 0, \text{ as } n \to \infty
\end{aligned}$$

from the definition of $X_n \xrightarrow{d} c$. □

Example. We aim to approximate the distribution of $\bar{X}_n$ as $n \to \infty$. By the weak law of large numbers, $\bar{X}_n \xrightarrow{p} \mu$, meaning that $\bar{X}_n \xrightarrow{d} \mu$. Asymptotically, the distribution of $\bar{X}_n$ degenerates to μ.

In order to get more useful results, we need to rescale $\bar{X}_n$ so that it has a stable distribution. Since $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$, consider

$$Z_n = \sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right)$$

Note that $\mathbb{E}[Z_n] = 0$ and $\text{Var}(Z_n) = 1$. The distribution of Z_n is "stabilized." We aim to find the asymptotic distribution of Z_n.

Theorem 2.14. Lindeberg-Lévy Central Limit Theorem *If X_i for $i = 1, \ldots, n$ are i.i.d. and $\mathbb{E}[X_i^2] < \infty$, then $Z_n \xrightarrow{d} \mathcal{N}(0,1)$ or, equivalently, $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, where $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \text{Var}(X_i)$.*

Proof. WLOG, assume $\mu = 0$. We will show that $C_{Z_n}(t) = \exp\left(-\frac{t^2}{2}\right)$ as $n \to \infty$, since $\exp\left(-\frac{t^2}{2}\right)$ is the characteristic function of a standard normal. Note that $Z_n = \sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) = \sum_{j=1}^n x_{jn}$, where $x_{jn} = \frac{X_j - \mu}{\sigma\sqrt{n}} = \frac{X_j}{\sigma\sqrt{n}}$. We have that

$$\begin{aligned}
C_{Z_n}(t) &= \mathbb{E}\left[\exp\left(itZ_n\right)\right] = \mathbb{E}\left[\exp\left(it\sum_{j=1}^n x_{jn}\right)\right] \\
&= \prod_{j=1}^n \mathbb{E}\left[\exp\left(itx_{jn}\right)\right] && \text{by independence} \\
&= \left\{\mathbb{E}\left[\exp\left(itx_{1n}\right)\right]\right\}^n && \text{by identical distribution} \\
&= \left\{C_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right)\right\}^n
\end{aligned}$$

where $C_{X_1}(s) = \mathbb{E}[\exp(isX_i)]$ is the characteristic function of X_1. Since $\mathbb{E}[X_1^2] < \infty$, by Taylor's Theorem, we have that

$$C_{X_1}(s) = \underbrace{C_{X_1}(0)}_{1} + \underbrace{is\,\mathbb{E}[X_1]}_{0} + \frac{i^2 s^2}{2}\underbrace{\mathbb{E}[X_1^2]}_{\sigma^2} + o(s^2), \text{ as } s \to 0$$

Thus for each fixed t,

$$C_{X_1}\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + o\left(\frac{t^2}{\sigma^2 n}\right)$$

and for each fixed t, as $n \to \infty$,

$$C_{Z_n}(t) = \left\{1 - \frac{t^2}{2n} + o\left(\frac{t^2}{\sigma^2 n}\right)\right\}^n = e^{-\frac{t^2}{2}}$$

since $\left(1 + \frac{a}{n}\right)^n \to e^a$ as $n \to \infty$. □

Theorem 2.15. Cramér-Wold Device *For a sequence of random vectors $X_n \in \mathbb{R}^k$,*

$$X_n \xrightarrow{d} X \iff \lambda' X_n \xrightarrow{d} \lambda' X, \text{ for all } \lambda \in \mathbb{R}^k$$

Remark. This implies that to show that a random vector X_n is asymptotically univariate normal, it is necessary and sufficient to show that any linear combination of elements of X_n is asymptotically univariate normal.

Theorem 2.16. Multivariate Lindeberg-Lévy Central Limit Theorem *If $X_i \in \mathbb{R}^k$, for $i = 1, \ldots, n$ are i.i.d. normal and $\mathbb{E}\|X_i\|^2 < \infty$, then*

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

where $\mu = \mathbb{E}[X_i]$ and $\Sigma = \mathbb{E}[(X_i - \mu)(X_i - \mu)']$.

2.6 Delta Method

Remark. So far, we've used $\bar{X}$ to estimate $\mathbb{E}[X_i]$. The same idea applies to a transformation of X, say $g(X)$. We can obtain the law of large numbers

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} g(X_i) \xrightarrow{p} \mathbb{E}[g(X)] = \mu$$

and the central limit theorem

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \mathcal{N}(0, \text{Var}(g(X)))$$

What about functions of moments? Consider $\beta = h(\mu) = h(\mathbb{E}[g(x)])$, where $h(\cdot)$ is not necessarily linear. A natural estimator is the plug-in estimator

$$\hat{\beta} = h(\hat{\mu}) \qquad \text{where } \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

How would we derive the asymptotic distribution of $\hat{\beta}$?

Theorem 2.17. Continuous Mapping Theorem *For random vectors $X_n, X \in \mathbb{R}^k$,*

$$X_n \xrightarrow{d} X,\ g \text{ is continuous} \implies g(X_n) \xrightarrow{d} g(X)$$

Corollary 2.3. Slutsky Theorem *If $X_n \xrightarrow{d} X$ and $c_n \xrightarrow{p} c$, then*

(i) $X_n + c_n \xrightarrow{d} X + c$

(ii) $X_n \cdot c_n \xrightarrow{d} X \cdot c$

(iii) $\frac{X_n}{c_n} \xrightarrow{d} \frac{X}{c}$ *provided* $c \neq 0$

Example. $X_n \xrightarrow{d} X \sim \mathcal{N}(0, I_k) \implies X_n' X_n \xrightarrow{d} X'X \sim \chi^2_n$

Example. Suppose $\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} \mathcal{N}(0,1)$ and $\hat{\sigma}$ is a consistent estimator for $\sigma > 0$. Then

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\hat{\sigma}}\right) = \sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right)\left(\frac{\sigma}{\hat{\sigma}}\right) \xrightarrow{d} \mathcal{N}(0,1)$$

Now, let's derive the asymptotic distribution of $\hat{\beta} = h(\hat{\mu})$. Note that $\hat{\beta}$ is written as a function of $\hat{\mu}$, not $\sqrt{n}(\hat{\mu} - \mu)$, so the continuous mapping theorem is not directly applicable. The key step here is the first order Taylor expansion, assuming differentiability of $h(\cdot)$. We have that

$$\hat{\beta} = h(\hat{\mu}) = h(\mu) + \left.\frac{\partial h(u)}{\partial u}\right|_{u=\mu^\star} (\hat{\mu} - \mu)$$

where $\mu^\star$ is on the line joining μ and $\hat{\mu}$. Then

$$\sqrt{n}(\hat{\beta} - h(\mu)) = \left.\frac{\partial h(u)}{\partial u}\right|_{u=\mu^\star} \sqrt{n}(\hat{\mu} - \mu)$$

so we can use the asymptotic distribution of $\sqrt{n}(\hat{\mu} - \mu)$ and the continuous mapping theorem.

Theorem 2.18. Delta Theorem *If* $\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{d} \xi$ *and* $h(\cdot)$ *is a function continuously differentiable in a neighborhood around* μ*, then*

$$\sqrt{n}(h(\hat{\mu}) - h(\mu)) \xrightarrow{d} H'\xi$$

where $H' = \frac{\partial}{\partial u} h(u)\Big|_{u=\mu}$.

In particular, if $\xi \sim \mathcal{N}(0, V)$*, then* $\sqrt{n}(h(\hat{\mu}) - h(\mu)) \xrightarrow{d} \mathcal{N}(0, H'VH)$.

When μ *and* h *are scalar-valued, then*

$$\sqrt{n}(h(\hat{\mu}) - h(\mu)) \xrightarrow{d} \mathcal{N}\left(0, \left(\frac{\partial}{\partial u} h(u)\Big|_{u=\mu}\right)^2 V\right)$$

3 Estimation

We will cover two methods – Maximum Likelihood Estimation, and the Method of Moments. These cover basically everything – MLE is a complete probability model, and MoM is a partial probability model. Almost every estimation model, no matter how exotic, can be boiled down to either Maximum Likelihood or Method of Moments.

3.1 Maximum Likelihood Estimation

Motivation. Parameter estimation in complete probability models – useful for IO, structural modeling, etc. Maximum Likelihood Estimation is very popular for these *parametric* models. The main advantage is that it has wide applicability and can handle complicated data and models. The disadvantage is the strong distributional assumptions you need to make.
Model. *Parametric Model* We have a vector $X \in \mathbb{R}^d$, $X \sim F$, and we have a random sample $\{X_1, \ldots, X_n\}$. We will assume that X has a density or probability mass function $f(x \mid \theta)$ with *known* form of f but unknown parameter value $\theta \in \Theta \subseteq \mathbb{R}^k$, where Θ is the (known) parameter space.
Example. Assume that $X \sim \mathcal{N}(\mu, \sigma^2)$, which has density

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

The parameters are $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_{++}$.
Remark. Here, we will focus on unconditional distributions (so $f(x \mid \theta)$ does not depend on conditioning variables. Next semester, and in most economic modeling, we will focus on conditional distributions.
Definition. A model is *correctly specified* when there is a unique parameter value $\theta_0 \in \Theta$ such that $f(x \mid \theta_0)$ coincides with the true density or pmf of X (*i.e.* $f(x \mid \theta) = f(x)$).

This parameter value θ_0 is called the *true parameter value*.

The parameter θ_0 is *unique* if there is no other θ such that $f(x \mid \theta_0) = f(x \mid \theta)$.

A model is *mis-specified* if there is no parameter value $\theta \in \Theta$ such that $f(x \mid \theta)$ coincides with the true density or pmf of X.
Example. Suppose that the true model is $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$. Our model is

$$f(x \mid p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = p\frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right) + (1-p)\frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{1}{2}\left(\frac{x-\mu_2}{\sigma_2}\right)^2\right)$$

This model is "correct" since it includes $f(x)$ as a special case. However, the "true" parameter is not unique – it includes:

$$(p, 0, 1, 0, 1) \; \forall \; p$$
$$(1, 0, 1, \mu_2, \sigma_2^2) \; \forall \; \mu_2, \sigma_2^2$$
$$(0, \mu_1, \sigma_1^2, 0, 1) \; \forall \; \mu_1, \sigma_1^2$$

Thus, the model is not correctly specified.

Note that the joint density or pmf of i.i.d. $\{X_1, \ldots, X_n\}$ given θ is

$$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

Definition. The *likelihood function* is

$$L_n(\theta) = f(X_1, \ldots, X_n \mid \theta) = \prod_{i=1}^{n} f(X_i \mid \theta)$$

It is the joint density (or pmf) of the data, and is viewed always as a function of θ. It essentially describes the compatibility of different values of θ with the observed data.

Definition. A *maximum likelihood estimator* $\hat{\theta}$ is the value that maximizes $L_n(\theta)$:

$$\hat{\theta} \in \operatorname*{argmax}_{\theta \in \Theta} L_n(\theta)$$

Or equivalently,

$$\hat{\theta} \in \operatorname*{argmax}_{\theta \in \Theta} \ell_n(\theta)$$

where

$$\ell_n(\theta) = \log L_n(\theta) = \sum_{i=1}^{n} \log f(X_i \mid \theta)$$

is called the *log likelihood function*.

Remark. In practice (finite samples, basically) it is very possible for there to be multiple maximizers – it may even not be a global maximum, just a local maximum. In this class, when we impose correct specification and large (asymptotic) samples, there will always be exactly one.

Example. (Exponential Distribution) Assume $f(x \mid \lambda) = \frac{1}{\lambda} \exp\left(-\frac{x}{\lambda}\right)$ for $x \in \mathbb{R}_+, \lambda \in \mathbb{R}_{++}$. The log likelihood is

$$\ell_n(\lambda) = \sum_{i=1}^{n} \left(-\log \lambda - \frac{X_i}{\lambda} \right) = -n \log \lambda - n \frac{\bar{X}_n}{\lambda}$$

The first order condition is

$$\frac{\partial \ell_n}{\partial \lambda}(\lambda) = -n \frac{1}{\lambda} + n \frac{\bar{X}_n}{\lambda^2} = 0 \Longrightarrow \hat{\lambda} = \bar{X}_n$$

We can see that $\hat{\lambda}$ is a maximizer since

$$\frac{\partial^2 \ell_n}{\partial \lambda^2}(\hat{\lambda}) = n \frac{1}{\hat{\lambda}^2} - 2n \frac{\bar{X}_n}{\hat{\lambda}^3} = -\frac{n}{\bar{X}_n^2} < 0$$

Question. Why does MLE make sense?

Definition. Define the *expected log likelihood function* as

$$\ell(\theta) \coloneqq \mathbb{E}[\log f(X \mid \theta)]$$

Note that the expectation is with respect to the random vector X, not θ.

Theorem 3.1. (Analog Principle) *When the model is correctly specified, the true parameter θ_0 maximizes $\ell(\theta)$.*

Proof. For each $\theta \neq \theta_0$, we have that

$$\ell(\theta) - \ell(\theta_0) = \mathbb{E}\left[\log \left(\frac{f(X \mid \theta)}{f(X \mid \theta_0)} \right) \right] < \log \mathbb{E}\left[\frac{f(X \mid \theta)}{f(X \mid \theta_0)} \right]$$

where the inequality follows from Jensen's Inequality, and strict inequality holds since log is strictly concave and the argument is not a constant.

Let the true density of the data be $f(x)$. Since $f(X \mid \theta_0) = f(x)$ and $f(X \mid \theta)$ is a valid density,

$$\mathbb{E}\left[\frac{f(X \mid \theta)}{f(X \mid \theta_0)} \right] = \int \frac{f(x \mid \theta)}{f(x \mid \theta_0)} f(x) dx = \int f(x \mid \theta) dx = 1$$

Thus,

$$\ell(\theta) - \ell(\theta_0) < \log 1 = 0 \Longrightarrow \ell(\theta) < \ell(\theta_0)$$

□

Remark. The likelihood function of parametric models provides a way to evaluate their estimators. Recall that $\ell(\theta) = \mathbb{E}[\log f(X \mid \theta)]$ is the expected log likelihood.

Definition. The log likelihood at a single observation X and the true parameter θ_0 is $\log f(X \mid \theta_0)$. The *efficient score* is

$$S = \frac{\partial}{\partial \theta} \log f(X \mid \theta_0)$$

The *Fisher information* is

$$\mathscr{F}_{\theta_0} = \mathbb{E}[SS']$$

Theorem 3.2. *Assume that the model is correctly specified, the support of X does not depend on θ, and θ_0 lies in the interior of Θ. Then $\mathbb{E}[S] = 0$ and $\mathrm{Var}(S) = \mathscr{F}_{\theta_0}$*

Proof. By Leibniz Rule:

$$\begin{aligned}
\mathbb{E}[S] &= \mathbb{E}\left[\frac{\partial}{\partial \theta} \log f(X \mid \theta_0)\right] \\
&= \frac{\partial}{\partial \theta} \mathbb{E}\left[\log f(X \mid \theta_0)\right] \\
&= \frac{\partial}{\partial \theta} \ell(\theta_0) = 0
\end{aligned}$$

where the last equality follows from the fact that θ_0 maximizes $\ell(\cdot)$ and $\theta_0 \in \operatorname{int} \Theta$. Then,

$$\mathrm{Var}(S) = \mathbb{E}\left[(S - \mathbb{E}[S])(S - \mathbb{E}[S])'\right] = \mathbb{E}[SS'] = \mathscr{F}_{\theta_0}$$

□

Theorem 3.3. (Information Matrix Equality)

$$\underbrace{\mathbb{E}\left[\frac{\partial \log f(X \mid \theta_0)}{\partial \theta} \frac{\partial \log f(X \mid \theta_0)}{\partial \theta'}\right]}_{\text{Fisher Information}} = \underbrace{-\mathbb{E}\left[\frac{\partial^2}{\partial \theta \partial \theta'} \log f(X \mid \theta_0)\right]}_{\text{Curvature of } \ell(\theta_0)}$$

That is,

$$\mathscr{F}_{\theta_0} = \mathscr{H}_{\theta_0}$$

where

$$\mathscr{H}_{\theta_0} = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta \partial \theta'} \log f(X \mid \theta_0)\right] = -\frac{\partial^2}{\partial \theta \partial \theta'} \mathbb{E}[\log f(X \mid \theta_0)] = -\frac{\partial^2}{\partial \theta \partial \theta'} \ell(\theta_0)$$

is called the expected Hessian.

Proof. Left for homework. □

Remark. This result is useful for simplifying the formula for the asymptotic variance of the maximum likelihood estimator.

Theorem 3.4. *Assume that the model is correctly specified, that the support of X does not depend on θ, and that θ_0 lies in the interior of Θ. If $\tilde{\theta}$ is an unbiased estimator of θ, then*

$$\mathrm{Var}(\tilde{\theta}) \geq (n\mathscr{F}_{\theta_0})^{-1}$$

$(n\mathscr{F}_{\theta_0})^{-1}$ is called the Cramér-Rao Lower Bound (CRL). *An estimator $\tilde{\theta}$ is* Cramér-Rao efficient *if it is unbiased and $\mathrm{Var}(\tilde{\theta}) = (n\mathscr{F}_{\theta_0})^{-1}$*

Remark. If $\text{Var}(\tilde{\theta})$ is a matrix, $\text{Var}(\tilde{\theta}) \geq (n\mathscr{F}_{\theta_0})^{-1}$ means that $\text{Var}(\tilde{\theta}) - (n\mathscr{F}_{\theta_0})^{-1}$ is positive semidefinite.

Intuition. More curvature of the expected log likelihood $\implies$ more information $\implies$ lower variance bound.

Proof. We write $x = (x_1, \ldots, x_n)'$ and $X = (X_1, \ldots, X_n)'$, and we write the joint density of X as $f(x \mid \theta)$. Since $\tilde{\theta}$ is an estimator, it is a function of data (X), and since it is unbiased, it must hold that

$$\theta = \mathop{\mathbb{E}}_{\theta}[\tilde{\theta}(X)] = \int \tilde{\theta}(x) f(x \mid \theta) dx$$

for any θ. By taking derivatives of both sides, we get that

$$I = \int \tilde{\theta}(x) \frac{\partial}{\partial \theta'} f(x \mid \theta) dx = \int \tilde{\theta}(x) \left(\frac{\partial}{\partial \theta'} \log f(x \mid \theta) \right) f(x \mid \theta) dx$$

where I is the identity matrix. Evaluating at the true value θ_0, we get that

$$\begin{aligned}
I &= \int \tilde{\theta}(x) \left(\frac{\partial}{\partial \theta'} \log f(x \mid \theta_0) \right) f(x \mid \theta_0) dx \\
&= \mathbb{E}\left[\tilde{\theta}(X) \left(\frac{\partial}{\partial \theta'} \log f(X \mid \theta_0) \right) \right] \\
&= \mathbb{E}\left[\tilde{\theta}(X) \left(\frac{\partial}{\partial \theta'} \log f(X \mid \theta_0) \right) \right] - \underbrace{\mathbb{E}[\tilde{\theta}(X)]}_{\theta_0} \underbrace{\mathbb{E}\left[\frac{\partial}{\partial \theta'} \log f(X \mid \theta_0) \right]}_{0} \\
&= \text{cov}\left(\tilde{\theta}(X), \frac{\partial}{\partial \theta'} \log f(X \mid \theta_0) \right)
\end{aligned}$$

where the third equality follows from the fact that

$$\mathbb{E}\left[\left(\frac{\partial}{\partial \theta'} \log f(X \mid \theta_0) \right) \right] = \mathbb{E}\left[\left(\sum_{i=1}^{n} \frac{\partial}{\partial \theta'} \log f(X_i \mid \theta_0) \right) \right] = n\, \mathbb{E}[S'] = 0$$

Thus, we have that

$$\text{Var}\begin{pmatrix} \tilde{\theta} \\ \frac{\partial}{\partial \theta} \log f(X \mid \theta_0) \end{pmatrix} = \begin{pmatrix} \text{Var}(\tilde{\theta}) & I \\ I & n\mathscr{F}_{\theta_0} \end{pmatrix}$$

(note that $\frac{\partial}{\partial \theta} \log f(X \mid \theta_0) = n\mathscr{F}_{\theta_0}$, showing this is left for homework)

Since this matrix is positive semi-definite, for any matrix A we have that

$$A' \,\text{Var}\begin{pmatrix} \tilde{\theta} \\ \frac{\partial}{\partial \theta} \log f(X \mid \theta_0) \end{pmatrix} A \geq 0$$

Choosing $A = \begin{pmatrix} I \\ -(n\mathscr{F}_{\theta_0})^{-1} \end{pmatrix}$, we get that

$$\text{Var}(\tilde{\theta}) - (n\mathscr{F}_{\theta_0})^{-1} \geq 0$$

□

Asymptotic Properties of MLE If θ_0 uniquely maximizes $\ell(\theta) = \mathbb{E}[\log f(X \mid \theta)]$ and some technical conditions hold such that

$$\frac{1}{n} \sum_{i=1}^{n} \log f(X_i \mid \theta) \xrightarrow{p} \mathbb{E}[\log f(X \mid \theta)]$$

uniformly for all $\theta \in \Theta$, then $\hat{\theta} \xrightarrow{p} \theta_0$, where $\hat{\theta}$ is the maximum likelihood estimator. With more technical conditions, we have that

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathscr{F}_{\theta_0}^{-1})$$

The MLE estimator is: consistent, converging at rate $n^{-\frac{1}{2}}$, asymptotically normal, and *asymptotically Cramér-Rao efficient*!

Variance Estimation The asymptotic variance of $\sqrt{n}(\hat{\theta} - \theta_0)$ is $\mathscr{F}_{\theta_0}^{-1}$, which is unknown. Since

$$\mathscr{F}_{\theta_0} = \mathbb{E}\left[\frac{\partial \log f(X \mid \theta_0)}{\partial \theta} \frac{\partial \log f(X \mid \theta_0)}{\partial \theta'}\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta \partial \theta'} \log f(X \mid \theta_0)\right]$$

by the Information Matrix Equality, we can estimate $\mathscr{F}_{\theta_0}^{-1}$ by either

$$\left\{-\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial \theta \partial \theta'} \log f(X_i \mid \hat{\theta})\right\}^{-1}$$

or

$$\left\{\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f(X_i \mid \hat{\theta}) \frac{\partial}{\partial \theta'} \log f(X_i \mid \hat{\theta})\right\}^{-1}$$

3.2 Method of Moments

Introduction MLE is used for parametric models. *Method of moments* allows semi-parametric models: estimation of a finite dimensional parameter when the distribution is non-parametric. A distribution is called non-parametric if it cannot be described by a finite list of parameters.
Example. Estimation of the mean $\mu = \mathbb{E}[X]$ when the distribution of X is unspecified. $\hat{\mu}_{MME} = \frac{1}{n}\sum_{i=1}^{n} X_i$, so the sample mean is a method of moments estimator. By the central limit theorem, as long as $\mathbb{E}\|X\|^2 < \infty$,

$$\sqrt{n}(\hat{\mu}_{MME} - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

where $\Sigma = \mathrm{Var}(X)$. Meanwhile, Σ can be consistently estimated by the sample covariance matrix:

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})'$$

Algorithm. In method of moments, we match the "theoretic moment" with the "sample moment" – essentially, we take the mth moment of X, $\mu'_m = \mathbb{E}\|X\|^m$, and estimate it using

$$\hat{\mu}'_m = \frac{1}{n}\sum_{i=1}^{n} \|X_i\|^m$$

Example. We can take this even further – if the mean of some transformation $g(X)$ is $\theta = \mathbb{E}[g(X)]$, the MME for θ is

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

By CLT, if $\mathbb{E}\|g(X)\|^2 < \infty$, then

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, V_\theta), \qquad \text{where} V_\theta = \mathrm{Var}(g(X))$$

We can consistently estimate V_θ using

$$\hat{V} = \frac{1}{n-1}\sum_{i=1}^{n}(g(X)-\hat{\theta})(g(X)-\hat{\theta})'$$

Example. Suppose that we are interested in the CDF of X, $F(x)$. We have that the MME is

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{X_i \le x}$$

We can show (homework) that

$$\sqrt{n}(F_n(x) - F(x)) \xrightarrow{d} \mathcal{N}(0, F(x)(1-F(x)))$$

Smooth Functions of Moments. Now, let's be a bit more general. Suppose the parameter of interest is $\beta = h(\theta)$, where $\theta = \mathbb{E}[g(X)]$, where X, g, and h can all be vectors. By plugging in the MME $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$, β can be estimated by $\hat{\beta} = h(\hat{\theta})$. When h is continuously differentiable, we call it *smooth*. By applying Delta method, we get

$$\hat{\beta} - \beta \xrightarrow{d} \mathcal{N}(0, V_\beta)$$

where $V_\beta = H'V_\theta H$, $H' = \frac{\partial}{\partial \theta'}h(\theta)$, and $V_\theta = \text{Var}(g(X))$.

V_β can be consistently estimated by $\hat{V}_\beta = \hat{H}'\hat{V}_\theta\hat{H}$, where

$$\hat{H}' = \frac{\partial}{\partial \theta'}h(\hat{\theta})$$
$$\hat{V}_\theta = \frac{1}{n-1}\sum_{i=1}^{n}(g(X_i)-\hat{\theta})(g(X_i)-\hat{\theta})'$$

Example. The variance of a random variable X is

$$\sigma^2 = \mathbb{E}[(X-\mathbb{E}[X])^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

a smooth function of uncentered first and second moment. The MME for σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2$$

and we can find the asymptotic distribution from delta method.

Definition. In many problems, we can write moments as explicit functions of parameters called *moment functions*:

$$\mathbb{E}[m(X,\beta)] = 0$$

with parameter $\beta \in \mathbb{R}^k$ and m is a $k \times 1$ function. For each β, the sample moment of $\mathbb{E}[m(X,\beta)]$ is

$$\frac{1}{n}\sum_{i=1}^{n}m(X,\beta)$$

The MME $\hat{\beta}$ solves a system of k nonlinear equations:

$$\frac{1}{n}\sum_{i=1}^{n}m(X,\hat{\beta})$$

Remark. The functions m are chosen by researchers, guided by economic theory or statistical theory.
Example. If we are estimating $\mathbb{E}[X] = \mu$, our moment function is

$$\mathbb{E}[(X - \mu)] = 0 \Longrightarrow m(x, \mu) = x - \mu$$

Thus, we get that

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_{MM}) = 0 \Longrightarrow \hat{\mu}_{MM} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

This is very simple, because the first moment gives us the MME estimator. However, we could also consider the second moment. We have that

$$\mathbb{E}[X^2] = \sigma^2 + \mu^2$$

so we have that

$$\mathbb{E}\begin{bmatrix} X - \mu \\ X^2 - \sigma^2 - \mu^2 \end{bmatrix} = 0 \Longrightarrow \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} X_i - \hat{\mu}_{MM} \\ X_i^2 - \hat{\sigma}^2_{MM} - \hat{\mu}^2_{MM} \end{bmatrix} = 0$$

so we have that

$$\hat{\mu}_{MM} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad \hat{\sigma}^2_{MM} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 + \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

Example. Parametric models. This is the classical way of defining MME. We have $X \sim f(x \mid \theta), \theta \in \mathbb{R}^k$. The kth moment of this model is

$$\mu_k(\beta) = \int x^k f(x \mid \theta) dx$$

Hence, β satisfies

$$\mathbb{E}\begin{bmatrix} X - \mu_1(\beta) \\ X^2 - \mu_2(\beta) \\ \vdots \\ X^m - \mu_m(\beta) \end{bmatrix} = 0$$

We can set

$$m(x, \beta) = \begin{pmatrix} X - \mu_1(\beta) \\ X^2 - \mu_2(\beta) \\ \vdots \\ X^m - \mu_m(\beta) \end{pmatrix}$$

so the MME $\hat{\beta}$ solves

$$\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} X_i - \mu_1(\hat{\beta}) \\ X_i^2 - \mu_2(\hat{\beta}) \\ \vdots \\ X_i^m - \mu_m(\hat{\beta}) \end{bmatrix} = 0$$

Example. Euler equation. We have that the consumer's utility function is

$$U(C_t, C_{t+1}) = u(C_t) + \frac{1}{\beta}u(C_{t+1})$$

and their budget it

$$C_t + \frac{C_{t+1}}{R_{t+1}} \leq W_t$$

They want to maximize expected utility

$$\mathbb{E}\left[u(C_t) + \frac{1}{\beta}u((W_t - C_t)R_{t+1})\right]$$

which admits first order condition

$$0 = u'(C_t) - \mathbb{E}\left[\frac{R_{t+1}}{\beta}u'(C_{t+1})\right]$$

Assuming CRRA utility, so $u(c) = \frac{c^{1-\alpha}}{1-\alpha}$, the Euler equation is now

$$\mathbb{E}\left[R_{t+1}\left(\frac{C_{t+1}}{C_t}\right)^{-\alpha} - \beta\right] = 0$$

Suppose that β is known and we are interested in estimating α. Then, α satisfies $\mathbb{E}[m(R_{t+1}, C_{t+1}, C_t, \alpha)] = 0$, where

$$m(R_{t+1}, C_{t+1}, C_t, \alpha) = R_{t+1}\left(\frac{C_{t+1}}{C_t}\right)^{-\alpha} - \beta$$

the MME for α solves

$$\frac{1}{n}\sum_{i=1}^{n}\left[m(R_{t+1}, C_{t+1}, C_t, \hat{\alpha})\right] = 0$$

Asymptotic Theory of MME. If there is a unique β_0 that solves $\mathbb{E}[m(X, \beta)] = 0$, and further technical conditions hold so that

$$\frac{1}{n}\sum_{i=1}^{n} m(X_i, \beta) \xrightarrow{p} \sum_{i=1}^{n}\mathbb{E}[m(X_i, \beta)]$$

uniformly for all β in some set B, then the MME $\hat{\beta} \xrightarrow{p} \beta_0$. With more technical conditions we can show that

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(0, V)$$

where $V = (Q')^{-1}\Omega Q^{-1}$, $\Omega = \mathrm{Var}(m(X, \beta_0))$, and $Q' = \mathbb{E}\left[\frac{\partial}{\partial \beta'}m(X, \beta_0)\right]$

Efficiency of MME. We know that sample mean $\hat{\mu}$ is the best linear unbiased estimator for population mean μ, which might justify the use of MME, but the restriction to linear estimators is not particularly convincing. In fact, we can show that $\hat{\mu}$ has the lowest variance among *all unbiased estimators*.

Theorem 3.5. *Let X be a random vector and $\mathcal{F}$ be a set of distributions such that $\mathbb{E}\|X\|^2 < \infty$. If $\tilde{\mu}$ is an unbiased estimator for $\mu = \mathbb{E}[X]$ for all distributions in $\mathcal{F}$, then*

$$\mathrm{Var}(\tilde{\mu}) \geq \frac{1}{n}\Sigma$$

where $\Sigma = \mathrm{Var}(X)$.

Remark. Since sample mean $\hat{\mu}$ is unbiased and $\mathrm{Var}(\hat{\mu}) = \frac{1}{n}\Sigma$, we conclude that $\hat{\mu}$ has the lowest variance among all unbiased estimators.

Proof. (Not examinable, but interesting!) The basic framework: If X has a parametric pdf $f(x \mid \theta)$, we can apply Cramér-Rao theory to find a lower bound. However, the distribution of X is left unspecified – the space of possible distributions is too big. We will construct a smaller class of correctly specified parametric distributions $f(x \mid \alpha)$, so that when $\alpha = 0$, $f(x \mid \theta) = f(x)$. Since $\tilde{\mu}$ is unbiased for all distributions, it is

also unbiased for $f(x \mid \alpha)$. The variance lower bound among all distributions must be at least as large as the Cramér-Rao lower bound for the subclass of distributions $f(x \mid \alpha)$. Conclusion follows.

Focus on the case where X is continuous with $f(x)$. WLOG, assume that $\mu = 0$ and X is bounded so that $\|X\| \leq C$ for some $0 < C < \infty$ (extending to cases where $\mu \neq 0$ and unbounded X only involves some more technicality). Now let $\mathcal{F}$ be the set of distributions such that $\mathbb{E}\|X\| = 0$ and $\|X\| \leq C$ with probability 1. Note that the condition that $\mathbb{E}\|X\|^2 \leq \infty$ is automatically satisfied.

Step 1: Construct a parametric subclass of distributions

$$f(x \mid \alpha) = f(x)\left\{1 + \alpha'\Sigma^{-1}x\right\}$$

where $\alpha \in \left\{\alpha : \|\Sigma^{-1}\alpha\| \leq \frac{1}{C}\right\}$, and $\Sigma = \mathrm{Var}(X) = \mathbb{E}[XX']$. Note $\mathbb{E}[X] = 0, \|x\| \leq C$. Let $\mathbb{E}_\alpha[\cdot]$ denote expectation under $f(x \mid \alpha)$

Step 2: Verify that $f(x \mid \alpha) \in \mathcal{F}$:

1. $f(x \mid \alpha)$ is a valid pdf sharing support with $f(x)$: $f(x \mid \alpha) \geq 0$ since $|\alpha'\Sigma^{-1}x| \leq \|\Sigma^{-1}\alpha\|\|x\| \leq 1$, and
$$\int f(x \mid \alpha)dx = \int f(x)dx + \int f(x)\alpha'\Sigma^{-1}xdx = 1 + \alpha'\Sigma^{-1}\,\mathbb{E}[X] = 1$$
2. $f(x \mid \alpha)$ is correctly specified: when $\alpha = 0$, $f(x \mid \alpha) = f(x)$
3. Variance of X under $f(x \mid \alpha)$ is finite: the above implies that $f(x \mid \alpha) \leq 2f(x)$. Thus $\mathbb{E}_\alpha\|X\|^2 \leq 2\,\mathbb{E}\|X\| < \infty$
4. Expectation of X under $f(x \mid \alpha)$ is:
$$\int xf(x \mid \alpha)dx = \int xf(x)dx + \left(\int xx'f(x)dx\right)\Sigma^{-1}\alpha = 0 + \Sigma^{-1}\Sigma^{-1}\alpha = \alpha$$

Step 3: Apply Cramér-Rao Theorem for $f(x \mid \alpha)$. First, note that unbiasedness of $\tilde{\mu}$ implies that it is unbiased for all $f(x) \in \mathcal{F}$. Since $f(x \mid \alpha) \in \mathcal{F}$, $\tilde{\mu}$ is unbiased for $f(x \mid \alpha)$. By Cramér-Rao Theorem, $\mathrm{Var}(\tilde{\mu}) \geq n^{-1}\mathscr{I}_\alpha$, where

$$\mathscr{I}_\alpha = \mathbb{E}\left[\frac{\partial}{\partial\alpha}\log f(X \mid 0)\frac{\partial}{\partial\alpha'}\log f(X \mid 0)\right]$$

Note that

$$\frac{\partial}{\partial\alpha'}\log f(X \mid 0) = \frac{\Sigma^{-1}X}{1 + \alpha'\Sigma^{-1}X}$$

Hence, $\mathscr{I}_\alpha = \Sigma^{-1}\,\mathbb{E}[XX']\Sigma^{-1} = \Sigma^{-1}$ as desired.

□

4 Hypothesis Testing

4.1 Basic Concepts

A random vector X has distribution $F(x)$, where F is unknown. We have a parameter of interest θ, determined by $F \in \mathscr{F}$. The parameter space is Θ, and we consider $\theta \in \Theta$. We have a random sample $\{X_1, \ldots, X_n\}$ from F. In previous sections, we talked about estimating θ. Now, we talk about testing hypotheses about θ.

Definition. A *hypothesis* is a statement about population parameter θ. We call the hypothesis to be tested the *null hypothesis*, $\mathbb{H}_0$, which is the restriction of θ. Specifically, it is either a restriction of θ to some θ_0, or to some subset $\Theta_0 \subseteq \Theta$. We often write

$$\mathbb{H}_0 = \{\theta \in \Theta : \theta = \theta_0\} \quad \text{or} \quad \mathbb{H}_0 = \{\theta \in \Theta : \theta \in \Theta_0\}$$

The complement of $\mathbb{H}_0$ is the *alternative hypothesis* $\mathbb{H}_1$, defined as either

$$\mathbb{H}_1 = \{\theta \in \Theta : \theta \neq \theta_0\} \quad \text{or} \quad \mathbb{H}_1 = \{\theta \in \Theta : \theta \notin \Theta_0\}$$

Remark. In these notes, we will focus on only *point hypotheses*, *i.e.* $\mathbb{H}_0 = \{\theta \in \Theta : \theta \neq \theta_0\}$. The alternative could be one-sided ($\mathbb{H}_1 : \theta > \theta_0$ or $\theta < \theta_0$) or two-sided ($\mathbb{H}_1 : \theta \neq \theta_0$). The one-sided alternative is relevant when the null lies on the boundary of the parameter space ($\theta_0 \in \partial\Theta \equiv \Theta = \{\theta : \theta \geq \theta_0\}$). An example of this would be if a policy necessarily has a non-negative effect.

Definition. A hypothesis is a restriction on the underlying distribution. Define the *null distribution* as a set F_0 such that

$$F_0 = \{F \in \mathscr{F} : \mathbb{H}_0 \text{ is true}\}$$

F_0 can be a singleton, a parametric family, or a nonparametric family.

Example. Suppose $\mathbb{H}_0 = \{\mu = \mu_0\}$. Examples of F_0:

- Singleton: $X \sim \mathcal{N}(\mu, \sigma^2)$ with known σ^2
- Parametric: $X \sim \mathcal{N}(\mu, \sigma^2)$ with unknown σ^2
- Nonparametric: X has finite mean

Definition. A hypothesis $\mathbb{H}$ is *simple* if $\{F \in \mathscr{F} : \mathbb{H} \text{ is true}\}$ is a singleton, and *composite* if the set contains multiple distributions.

Example. In the example above, $X \sim \mathcal{N}(\mu, \sigma^2)$ with known σ^2 is the only simple hypothesis. The others are composite.

Definition. A *hypothesis test* is a decision based on data. The decision either accepts $\mathbb{H}_0$ or rejects $\mathbb{H}_0$ in favor of $\mathbb{H}_1$. The procedure is as follows:

1. Construct a real-valued function of the data called a *test statistic*

   $$T = T(X_1, \ldots, X_n) \in \mathbb{R}$$

   which is a random variable.

2. Pick a *critical region* C, where in a one-sided test $C \coloneqq \{x : x > c\}$ for *critical value* c, and in a two-sided test $C \coloneqq \{x : |x| > c\}$.

3. State the hypothesis test as a *decision rule*

   $$\begin{cases} \text{accept } \mathbb{H}_0 & \text{if } T \notin C \\ \text{reject } \mathbb{H}_0 & \text{if } T \in C \end{cases}$$

Evaluation of hypothesis tests. A decision could be correct or incorrect. We evaluate hypothesis tests through their probability of making mistakes. There are two types of error in hypothesis testing:

Truth: $\mathbb{H}_0$	Decision: Accept $\mathbb{H}_0$	Error: None
Truth: $\mathbb{H}_0$	Decision: Reject $\mathbb{H}_0$	Error: **Type I**
Truth: $\mathbb{H}_1$	Decision: Accept $\mathbb{H}_0$	Error: **Type II**
Truth: $\mathbb{H}_1$	Decision: Reject $\mathbb{H}_0$	Error: None

Definition. The *power function* of a hypothesis test is the probability of rejection:

$$\pi(F) = \mathbb{P}\left\{\text{reject } \mathbb{H}_0 \mid F\right\} = \mathbb{P}\{T \in C \mid F\}$$

Definition. The *size* of a hypothesis test is the probability of Type I error:

$$\mathbb{P}\{\text{reject } \mathbb{H}_0 \mid F_0\} = \pi(F_0)$$

for F_0 satisfying $\mathbb{H}_0$.
Definition. The *power* of a hypothesis test is the complement of the probability of Type II error:

$$\mathbb{P}\{\text{reject } \mathbb{H}_0 \mid F_1\} = \pi(F_1) = 1 - \mathbb{P}\{\text{accept } \mathbb{H}_0 \mid \mathbb{H}_1\}$$

for F_1 satisfying $\mathbb{H}_1$.
Remark. Size is the power function evaluated at null, power is the power function evaluated at alternative.
Theorem 4.1. *Type I and Type II errors cannot be reduced at the same time.*

Proof. (Intuition) Let $G(x \mid F) = \mathbb{P}\{T \leq x \mid F\}$ be the sampling distribution of T. $G(x \mid F_0)$ is called the null sampling distribution, and $G(x \mid F_1)$ is called the alternative sampling distribution. Consider a one-sided test with rejection rule $T > c$. Type I error is size $\pi(F_0) = \mathbb{P}\{T > c \mid F_0\} = 1 - G(c \mid F_0)$. Type 2 error is $1 - \pi(F_1) = \mathbb{P}\{T \leq c \mid F_1\} = G(c \mid F_1)$. Since any distribution function $G(x \mid F)$ is increasing in x, Type I error is decreasing in c while Type II error is increasing in c. □

4.2 Classical Approach

We will control size, and pick the test to maximize power subject to the size constraint.
Definition. The *significance level* $\alpha \in (0, 1)$ is the probability selected by the researcher to be the maximal acceptable size of the hypothesis test.
Example. Consider a one-sided test, where $\mathbb{H}_0 : \theta = \theta_0$ and $\mathbb{H}_1 : \theta > \theta_0$. Given test statistic T, consider the test taking form

$$\text{Decision } = \begin{cases} \text{accept } \mathbb{H}_0 & \text{if } T \leq c \\ \text{reject } \mathbb{H}_0 & \text{if } T > c \end{cases}$$

We choose c to control Type I error, so c solves

$$\pi(F_0) = \mathbb{P}\{T > c \mid F_0\} = 1 - G(c \mid F_0) = \alpha \Longrightarrow c = G^{-1}(1 - \alpha \mid F_0)$$

The Type I controlled decision rule is

$$\text{Decision } = \begin{cases} \text{accept } \mathbb{H}_0 & \text{if } T \leq G^{-1}(1 - \alpha \mid F_0) \\ \text{reject } \mathbb{H}_0 & \text{if } T > G^{-1}(1 - \alpha \mid F_0) \end{cases}$$

and it has a size equal to α.

Example. Consider a two-sided test, where $\mathbb{H}_0 : \theta = \theta_0$ and $\mathbb{H}_1 : \theta \neq \theta_0$. The test takes the form:

$$\text{Decision } = \begin{cases} \text{accept } \mathbb{H}_0 & \text{if } |T| \leq c \\ \text{reject } \mathbb{H}_0 & \text{if } |T| > c \end{cases}$$

We again choose c to control size:

$$\pi(F_0) = \mathbb{P}\{|T| > c \mid F_0\} = 1 - G(c \mid F_0) + G(-c \mid F_0) = \alpha$$

Assuming that G is symmetric about 0, we have that

$$1 - G(c \mid F_0) + G(-c \mid F_0) = 2(1 - G(c \mid F_0)) = \alpha \Longrightarrow c = G^{-1}\left(1 - \frac{\alpha}{2}\Big| F_0\right)$$

The test rule

$$\text{Decision } = \begin{cases} \text{accept } \mathbb{H}_0 & \text{if } |T| \leq G^{-1}\left(1 - \frac{\alpha}{2}\big| F_0\right) \\ \text{reject } \mathbb{H}_0 & \text{if } |T| > G^{-1}\left(1 - \frac{\alpha}{2}\big| F_0\right) \end{cases}$$

has size α.

Example. Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$ and we wish to test

$$\mathbb{H}_0 : \mu = \mu_0 \quad \text{versus} \quad \mathbb{H}_1 : \mu > \mu_0$$

We create a test statistic, using $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$:

$$T = \frac{\bar{X}_n - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

Under $\mathbb{H}_0$, $T \sim t_{n-1}$. Given α, set $c = q_{1-\alpha}$, where $q_{1-\alpha}$ is the $(1-\alpha)$th quantile of the t_{n-1} distribution. A one-sided t-test with size α is

$$\text{accept } \mathbb{H}_0 \text{ if } T \leq q_{1-\alpha} \qquad ; \qquad \text{reject } \mathbb{H}_0 \text{ if } T > q_{1-\alpha}$$

If σ^2 is known, replacing s^2 with σ^2, we get that T yields a z test that uses the quantiles of a standard normal. A two-sided test is similar to this entire process, with absolute values.

Theorem 4.2. *In the normal sampling model $X \sim \mathcal{N}(\mu, \sigma^2)$, let*

$$T = \frac{\bar{X}_n - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

1. *The t test of $\mathbb{H}_0 : \mu = \mu_0$ against $\mathbb{H}_1 : \mu > \mu_0$ rejects if $T > q_{1-\alpha}$*
2. *The t test of $\mathbb{H}_0 : \mu = \mu_0$ against $\mathbb{H}_1 : \mu < \mu_0$ rejects if $T < q_\alpha$*
3. *The t test of $\mathbb{H}_0 : \mu = \mu_0$ against $\mathbb{H}_1 : \mu \neq \mu_0$ rejects if $|T| > q_{1-\alpha/2}$*

These tests have exact size α.

Example. Suppose X has mean μ and finite variance. We wish to test

$$\mathbb{H}_0 : \mu = \mu_0 \qquad ; \qquad \mathbb{H}_1 : \mu > \mu_0$$

The t-statistic is

$$T = \frac{\bar{X} - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

where s^2 could be replaced by the plug-in variance estimator. Under $\mathbb{H}_0$, T is not normal, but is asymptotically normal – by CLT, $T \xrightarrow{d} \mathcal{N}(0,1)$. Thus, as $n \to \infty$,

$$\pi(F_0) = \mathbb{P}\{T > c \mid F_0\} \to \mathbb{P}\{\mathcal{N}(0,1) > c\} = 1 - \Phi(c)$$

Theorem 4.3. *If X has finite mean μ and variance σ^2, then*

1. *The asymptotic t-test of $\mathbb{H}_0 : \mu = \mu_0$ against $\mathbb{H}_1 : \mu > \mu_0$ rejects if $T > Z_{1-\alpha}$*
2. *The asymptotic t-test of $\mathbb{H}_0 : \mu = \mu_0$ against $\mathbb{H}_1 : \mu < \mu_0$ rejects if $T < Z_{\alpha}$*
3. *The asymptotic t-test of $\mathbb{H}_0 : \mu = \mu_0$ against $\mathbb{H}_1 : \mu \neq \mu_0$ rejects if $|T| > Z_{1-\alpha/2}$*

where $Z_{1-\alpha}$ is the $(1-\alpha)$th quantile of the standard normal distribution. These tests have asymptotic size α.

Again, consider a one-sided test which accepts $\mathbb{H}_0$ if $T \leq c$ and rejects $\mathbb{H}_0$ if $T > c$, where c is chosen to control size at α, so $\mathbb{P}\{T > c \mid F_0\} = 1 - G(c \mid F_0) = \alpha$.
Question. How should we report the results of this test?

We've used two methods thus far: (i) report size α and the decision to reject or accept $\mathbb{H}_0$; and (ii) report the critical value c and the value T at sample points.

Another method: report the value of a certain kind of statistic called the p-value.
Definition. Define the *p-value* as

$$p = 1 - G(T \mid F_0)$$

Since $G(T \mid F_0)$ is increasing, p is a decreasing function of T. Also note that $\alpha = 1 - G(c \mid F_0)$, so the decision to reject $\mathbb{H}_0$ if $T > c$ is equivalent to rejecting $\mathbb{H}_0$ if $p < \alpha$.

For each $\alpha \in (0, 1)$, the test 'accept $\mathbb{H}_0$ if $p > \alpha$, reject $\mathbb{H}_0$ if $p \leq \alpha$' is a size α test:

$$\begin{aligned}
\mathbb{P}\{p \leq \alpha \mid F_0\} &= \mathbb{P}\{1 - G(T \mid F_0) \leq \alpha \mid F_0\} \\
&= \mathbb{P}\{G^{-1}(1-\alpha \mid F_0) \leq T \mid F_0\} \\
&= 1 - G(G^{-1}(1-\alpha \mid F_0) \leq T \mid F_0) \\
&= \alpha
\end{aligned}$$

p is 'the degree of evidence against $\mathbb{H}_0$' (the smaller the p-value, the stronger the evidence against the null); p is the 'marginal significance level' (the lower bound of the range at size α that we would reject the null).
Remark. p is the transformation of a statistic rather than a probability – it transforms the T statistic to an easily interpretable universal scale in $[0, 1]$.

p allows inference to be continuous rather than dichotomous, which is more informative. If we had one statistic that had $p = 0.049$ and one that had $p = 0.051$, we would know that they are essentially the same. Otherwise, we'd just see reported 'reject' and 'accept'.

4.3 Power Analysis

So far, we've focused on the size of the tests. We know how to construct a test of (asymptotic) size α for mean. But a good test should also have good power. It's important to know the power of the test we construct.
Example. Suppose $X \sim \mathcal{N}(\mu, \sigma^2)$ with known σ^2. Consider the standard statistic

$$T = \frac{\bar{X}_n - \mu_0}{\sqrt{\frac{\sigma^2}{n}}}$$

for tests $\mathbb{H}_0 : \mu = \mu_0$ and $\mathbb{H}_1 : \mu > \mu_0$. We reject if $T > c$, where c is chosen to control size at level α. Since $\bar{X}$ is centered around the true mean μ,

$$\frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0, 1)$$

The power function of the test is

$$\pi(F) = \mathbb{P}\{T > c \mid F\} = \mathbb{P}\left\{ \frac{\bar{X}_n - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} > c \middle| F \right\}$$
$$= \mathbb{P}\left\{ \underbrace{\frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}}}_{Z \sim \mathcal{N}(0,1)} - \frac{\mu - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} > c \middle| F \right\}$$
$$= 1 - \Phi\left(c + \frac{\mu_0 - \mu}{\sqrt{\frac{\sigma^2}{n}}} \right)$$

The size is $\pi(F_0) = 1 - \Phi(c)$, since under F_0, $\mu = \mu_0$. Power is

$$\pi(\mu \mid F_1) = 1 - \Phi\left(c + \frac{\mu_0 - \mu}{\sqrt{\frac{\sigma^2}{n}}} \right)$$

where $\mu > \mu_0$. Note that $\pi(\mu \mid F_1)$ is increasing in n and μ, and decreasing in σ^2 and c.
Example. Suppose we want to choose n and c to achieve size of 0.1 and power of at least 0.8 if $\mu \geq \mu_0 + \sigma$. How should we proceed?

Step 1: selecting c such that $\pi(F_0) = 1 - \Phi(c) = 0.1$ yields $c = 1.28$.

Step 2: selecting n such that

$$1 - \Phi\left(1.28 - \frac{\mu_0 - \mu}{\sqrt{\frac{\sigma^2}{n}}} \middle| \mu = \mu_0 + \sigma \right) \geq 0.8$$

which yields $n \geq 4.49$.

Thus, choosing $c = 1.28$ and $n = 5$ yields the desired.

4.4 Likelihood Ratio Test

Recall the classical approach to testing, where we control size and then maximize power subject to the size constraint. So far, we've focused on the t-test, but another important class of tests is called the *likelihood ratio test*. We will show that it maximizes power subject to the size constraint for *simple* hypothesis tests.

Consider a parametric model $f(x \mid \theta)$ with likelihood $L_n(\theta) = \prod_{i=1}^n f(X_i \mid \theta)$. We want to test simple hypotheses $\mathbb{H}_0 : \theta = \theta_0$, $\mathbb{H}_1 : \theta = \theta_1$ for some hypothetical values θ_0, θ_1. The ratio $\frac{L_n(\theta_1)}{L_n(\theta_0)}$ compares the likelihood function under the two hypotheses. A decision rule could be

$$\text{accept } \mathbb{H}_0 \text{ if } \frac{L_n(\theta_1)}{L_n(\theta_0)} \leq c \qquad ; \qquad \text{reject } \mathbb{H}_0 \text{ if } \frac{L_n(\theta_1)}{L_n(\theta_0)} > c$$

for some critical value c. Note that the size is

$$\mathbb{P}\left\{ \frac{L_n(\theta_1)}{L_n(\theta_0)} > c \middle| \theta = \theta_0 \equiv \mathbb{H}_0 \right\} = \underbrace{a_0}_{\text{size}}$$

We can fix size by choosing c such that $a_0 = a^\star$ for some optimal size $a^\star$.

Definition. (for convenience) We define the *likelihood ratio statistic* as

$$LR_n = 2(\ell_n(\theta_1) - \ell_n(\theta_0))$$

where $\ell_n(\theta) = \log L_n(\theta)$. A *likelihood ratio test* is

$$\text{accept } \mathbb{H}_0 \text{ if } LR_n \leq c \quad ; \quad \text{reject } \mathbb{H}_0 \text{ if } LR_n > c$$

for some critical value c.

Example. For $X \sim \mathcal{N}(\mu, \sigma^2)$ with known σ^2, we have that

$$\ell_n(\mu) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

Suppose that $\mathbb{H}_0 : \mu = \mu_0$, $\mathbb{H}_1 : \mu = \mu_1 > \mu_0$. Then we have that

$$LR_n = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left((X_i - \mu_0)^2 - (X_i - \mu_1)^2\right) = \frac{n}{\sigma^2}\left[2\bar{X}_n(\mu_1 - \mu_0) + (\mu_0^2 - \mu_1^2)\right]$$

Rejecting $\mathbb{H}_0$ for some $LR_n > c$ is equivalent to rejecting if

$$T = \frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{n}} > c$$

Theorem 4.4. (Neyman-Pearson Lemma) *Suppose random variable X has a parametric pdf / pmf $f(X \mid \theta)$. Among all tests of a simple null hypothesis $\mathbb{H}_0 : \theta = \theta_0$ against a simple alternative hypothesis $\mathbb{H}_1 : \theta = \theta_1$ with size α, the likelihood ratio test has the greatest power.*

Remark. In a normal sampling model with known variance, the likelihood ratio test of simple hypotheses is identical to a t-test with known variance. By Neyman-Pearson, this is now the most powerful test for this hypothesis in this model.

Proof. Consider the likelihood ratio test:

$$\text{accept } \mathbb{H}_0 \text{ if } \frac{L_n(\theta_1)}{L_n(\theta_0)} \leq c \quad ; \quad \text{reject } \mathbb{H}_0 \text{ if } \frac{L_n(\theta_1)}{L_n(\theta_0)} > c$$

where c is chosen such that

$$\mathbb{P}\left\{\frac{L_n(\theta_1)}{L_n(\theta_0)} > c \mid \theta = \theta_0\right\} = \alpha$$

Let the joint density of observations be $f(x \mid \theta)$, meaning that $L_n(\theta) = f(X \mid \theta)$. Since the test is a binary decision, we can represent it with a binary test function. Define the likelihood ratio test function as

$$\psi_{LR} = \mathbb{1}_{f(X|\theta_1) > c f(X|\theta_0)}$$

so $\psi_{LR} = 1$ if the decision rejects $\mathbb{H}_0$ and 0 otherwise. Define some alternative test as ψ_a, where ψ_a also has size α. Since both tests have the same size, we have that

$$\mathbb{P}\{\psi_{LR} = 1 \mid \theta = \theta_0\} = \mathbb{P}\{\psi_a = 1 \mid \theta = \theta_0\} = \alpha$$

or, equivalently,

$$\int \psi_{LR} f(x \mid \theta_0)dx = \int \psi_a f(x \mid \theta_0)dx = \alpha$$

We have that the power of the likelihood ratio test is

$$\begin{aligned}
\mathbb{P}\left\{\frac{L_n(\theta_1)}{L_n(\theta_0)} > c \middle| \theta = \theta_1\right\} &= \mathbb{P}\{\psi_{LR} = 1 \mid \theta = \theta_1\} \\
&= \int \psi_{LR} f(x \mid \theta_1) dx \\
&= \int \psi_{LR} f(x \mid \theta_1) dx - c\left\{\int \psi_{LR} f(x \mid \theta_0) dx - \int \psi_a f(x \mid \theta_0) dx\right\} \\
&= \int \psi_{LR}(f(x \mid \theta_1) - cf(x \mid \theta_0)) dx + c\int \psi_a f(x \mid \theta_0) dx \\
&\geq \int \psi_a(f(x \mid \theta_1) - cf(x \mid \theta_0)) dx + c\int \psi_a f(x \mid \theta_0) dx \\
&= \int \psi_a f(x \mid \theta_1) dx = \mathbb{P}\{\psi_a = 1 \mid \theta = \theta_1\}
\end{aligned}$$

where the inequality holds because

$$\psi_a(f(x \mid \theta_1) - cf(x \mid \theta_0)) \leq \psi_{LR}(f(x \mid \theta_1) - cf(x \mid \theta_0)) = \begin{cases} f(x \mid \theta_1) - f(x \mid \theta_0) & f(x \mid \theta_1) - f(x \mid \theta_0) > 0 \\ 0 & f(x \mid \theta_1) - f(x \mid \theta_0) \leq 0 \end{cases}$$

Thus, the power of the likelihood ratio test is (weakly) greater than any other test with size α. □

Example. Consider a two-sided composite alternative, where $\mathbb{H}_0 : \theta = \theta_0$ and $\mathbb{H}_1 : \theta \neq \theta_0$. The log likelihood under $\mathbb{H}_1$ is the unrestricted maximum of the likelihood. Let $\hat{\theta}$ be the MLE that maximizes $L_n(\theta)$. The likelihood ratio statistic is

$$LR_n = 2\left(\ell_n(\hat{\theta}) - \ell_n(\theta_0)\right)$$

and the likelihood ratio test is

$$\text{accept } \mathbb{H}_0 \text{ if } LR_n \leq c \qquad ; \qquad \text{reject } \mathbb{H}_0 \text{ if } LR_n > c$$

for some critical value c.

Example. Consider the one-sided composite alternative, where $\mathbb{H}_0 : \theta = \theta_0$ and $\mathbb{H}_1 : \theta > \theta_0$. The log likelihood under $\mathbb{H}_1$ is the maximum of the likelihood on the set $\{\theta : \theta \geq \theta_0\}$, which we call $\ell_n(\hat{\theta}^+)$, where $\hat{\theta}^+ \in \text{argmax}_{\theta \geq \theta_0} \ell_n(\theta)$. The likelihood ratio statistic is

$$LR_n^+ = 2\left(\ell_n(\hat{\theta}^+) - \ell_n(\theta_0)\right)$$

and the likelihood ratio test is

$$\text{accept } \mathbb{H}_0 \text{ if } LR_n^+ \leq c \qquad ; \qquad \text{reject } \mathbb{H}_0 \text{ if } LR_n^+ > c$$

for some critical value c.

Example. Again, suppose $X \sim \mathcal{N}(\mu, \sigma^2)$, with σ^2 known. Consider testing $\mathbb{H}_0 : \mu = \mu_0$ against $\mathbb{H}_1 : \mu > \mu_0$. We showed earlier that the t-test is equivalent to the likelihood ratio test for simple hypotheses. This analysis did not depend on the specific value of the alternative μ_1, so it still holds here. The t-test is still the likelihood ratio test for the one-sided composite alternative.

Theorem 4.5. (Asymptotics) *For simple null hypotheses, under* $\mathbb{H}_0 : \theta = \theta_0$,

$$LR_n \xrightarrow{d} \chi^2_{\dim(\theta)}$$

Let $q_{1-\alpha}$ *be the* $(1-\alpha)$*th quantile of* $\chi^2_{\dim(\theta)}$. *The test*

$$\textit{accept } \mathbb{H}_0 \textit{ if } LR_n \leq q_{1-\alpha} \qquad ; \qquad \textit{reject } \mathbb{H}_0 \textit{ if } LR_n > q_{1-\alpha}$$

has asymptotic size α. *Moreover, likelihood ratio tests and* t*-tests are* asymptotically *equivalent tests.*

Proof. (Sketch) Note that $LR_n = 2\left(\ell_n(\hat{\theta}) - \ell_n(\theta_0)\right)$. Second-order Taylor expansion yields

$$\ell_n(\theta_0) \simeq \ell_n(\hat{\theta}) + \underbrace{\frac{\partial}{\partial\theta}\ell_n(\hat{\theta})'}_{0}(\hat{\theta} - \theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)' \underbrace{\frac{\partial^2}{\partial\theta\partial\theta}\ell_n(\hat{\theta})}_{-\hat{V}^{-1}}(\hat{\theta} - \theta_0)$$

where $\hat{V} = \left\{-\frac{\partial^2}{\partial\theta\partial\theta}\ell_n(\hat{\theta})\right\}^{-1}$ is the estimator of the expected Hessian of the asymptotic variance of $\hat{\theta}$. Thus,

$$2\left(\ell_n(\hat{\theta}) - \ell_n(\theta_0)\right) \simeq (\hat{\theta} - \theta_0)'\hat{V}^{-1}(\hat{\theta} - \theta_0)$$

as $n \to \infty$, the right hand side converges to $\chi^2_{\dim(\theta)}$ □

5 Confidence Intervals

5.1 Motivation

We've seen point estimation of a parameter θ, where we report a single value as a guess of θ. Now, we consider interval estimation as a tool to report estimation uncertainty. Essentially, we can now quantify how precise our estimates for θ are.

Definition. Given a sample $X = \{X_1, \ldots, X_n\}$, an *interval estimator* of a real-valued parameter θ is an interval $C = C(X) = [L(X), U(X)]$. Note the following:

(i) $L(X)$ and $U(X)$ are functions of X, so they are themselves random

(ii) For $X = x$, $[L(x), U(x)]$ are realized values of the interval

(iii) If $L(X) = -\infty$, we have a one-sided interval $(-\infty, U(X)]$

(iv) If $U(X) = \infty$, we have a one-sided interval $[L(X), \infty)$

Example. Consider a random sample $\{X_1, X_2, X_3, X_4\}$ where $X_i \sim \mathcal{N}(\mu, 1)$. An interval estimator for μ could be $[\bar{X} - 1, \bar{X} + 1]$, we will assert that μ is in this interval. Of course, reporting $[\bar{X} - 1, \bar{X} + 1]$ is less precise than reporting $\bar{X}$. So why would we do that? Well, because by giving up precision we gain confidence that our report is true. To see why, recall that $\mathbb{P}\{\bar{X} = \mu\} = 0$. However,

$$\begin{aligned}
\mathbb{P}\{\bar{X} - 1 \leq \mu \leq \bar{X} + 1\} &= \mathbb{P}\{-1 \leq \bar{X} - \mu \leq 1\} \\
&= \mathbb{P}\left\{-2 \leq \frac{\bar{X} - \mu}{\sqrt{1/4}} \leq 2\right\} \\
&= \mathbb{P}\{-2 \leq Z \leq 2\} \qquad \text{(where } Z \sim \mathcal{N}(0,1)\text{)} \\
&= 0.9544
\end{aligned}$$

So we have a 95% chance of covering μ

Definition. For an interval estimator $[L(X), U(X)]$ of parameter θ, the *coverage probability* of $[L(X), U(X)]$ is the probability that the random interval includes the true parameter θ, denoted by

$$\mathbb{P}\{L(X) \leq \theta \leq U(X)\} \qquad \text{or} \qquad \mathbb{P}\{\theta \in [L(X), U(X)]\}$$

Note that the probability statements depend on the distribution F of X.

Definition. A $1 - \alpha$ *confidence interval* for θ is an interval $[L(X), U(X)]$ with coverage probability $1 - \alpha$.

Remark. When the finite sample distribution is unknown, we can approximate its coverage probability by its asymptotic limit.

Definition. The *asymptotic coverage probability* of interval estimator $[L(X), U(X)]$ is

$$\liminf_{n \to \infty} \mathbb{P}\{\theta \in [L(X), U(X)]\}$$

A $1 - \alpha$ *asymptotic confidence interval* for θ is an interval estimator $[L(X), U(X)]$ with asymptotic coverage probability $1 - \alpha$.

5.2 Finding Confidence Interval by Pivotal Quantities

Definition. A random variable $Q(X, \theta) = Q(X_1, \ldots, X_n, \theta)$ is a *pivotal quantity* (or *pivot*) if the distribution of $Q(X, \theta)$ is independent of parameters θ. That is, if $X \sim F(x, \theta)$, then $Q(X, \theta)$ has the same distribution for all values of θ.

Example. Let $\{X_1, \ldots, X_n\}$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. Then the t-statistic $\frac{\bar{X} - \mu}{s/\sqrt{n}}$ is a pivot since it follows a t_{n-1} distribution and does not depend on μ or σ^2.

Remark. Once we have a pivot, finding a confidence interval is easy.

Example. Again let $\{X_1, \ldots, X_n\}$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$, and consider the same t-statistic, distributed t_{n-1}. Fix a coverage probability $1-\alpha$. Let $q_{1-\alpha/2}$ be the $(1-\alpha/2)$th quantile of t_{n-1}. Then

$$\mathbb{P}\left\{-q_{1-\alpha/2} \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq q_{1-\alpha/2}\right\} = 1-\alpha$$
$$\Rightarrow \mathbb{P}\left\{\bar{X} - q_{1-\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + q_{1-\alpha/2}\frac{s}{\sqrt{n}}\right\} = 1-\alpha$$

Thus, a $1-\alpha$ confidence interval for μ is

$$\left[\bar{X} - q_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + q_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$

Example. Again let $\{X_1, \ldots, X_n\}$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$. Now we consider the variance statistic

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

Specify a coverage probability $1-\alpha$, and let $c_{\alpha/2}$ and $c_{1-\alpha/2}$ be the $(\alpha/2)$th and $(1-\alpha/2)$th quantiles of χ^2_{n-1} respectively. Then

$$\mathbb{P}\left\{c_{\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq c_{1-\alpha/2}\right\} = 1-\alpha$$
$$\Rightarrow \mathbb{P}\left\{\frac{(n-1)s^2}{c_{1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{c_{\alpha/2}}\right\} = 1-\alpha$$

Thus, a $1-\alpha$ confidence interval for σ^2 is

$$\left[\frac{(n-1)s^2}{c_{1-\alpha/2}}, \frac{(n-1)s^2}{c_{\alpha/2}}\right]$$

Example. Let $\{X_1, \ldots, X_n\}$ be a random sample from some F with mean μ and variance σ^2. By the Central Limit Theorem, as $n \to \infty$

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

By the weak law of large numbers, s is a consistent estimator for σ, so by the continuous mapping theorem

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

This is an *asymptotic pivot*. Again specify a coverage probability $1-\alpha$ and let $z_{1-\alpha/2}$ be the $(1-\alpha/2)$th quantile of $\mathcal{N}(0,1)$. Then as $n \to \infty$,

$$\mathbb{P}\left\{-z_{1-\alpha/2} \leq \frac{\bar{X}-\mu}{s/\sqrt{n}} \leq z_{1-\alpha/2}\right\} \to 1-\alpha$$
$$\Rightarrow \mathbb{P}\left\{\bar{X} - z_{1-\alpha/2}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right\} \to 1-\alpha$$

Thus, an asymptotic $1-\alpha$ confidence interval for μ is

$$\left[\bar{X} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]$$

Example. Let $\hat{\theta}$ be an estimator of scalar-valued parameter θ satisfying $\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} \mathcal{N}(0,V)$ as $n \to \infty$ and $\hat{V}$ is a consistent estimator of V. Standard error for $\hat{\theta}$ is given by $s(\hat{\theta}) = \sqrt{\hat{V}/n}$. By continuous mapping theorem,

$$\frac{\hat{\theta}-\theta}{s(\hat{\theta})} \xrightarrow{d} \mathcal{N}(0,1)$$

which implies that an asymptotic $1-\alpha$ confidence interval is

$$\left[\hat{\theta} - z_{1-\alpha/2}s(\hat{\theta}), \hat{\theta} + z_{1-\alpha/2}s(\hat{\theta})\right]$$

5.3 Finding Confidence Interval by Test Inversion

Definition. A general way of getting a confidence interval is by *test inversion*. Consider testing $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta \neq \theta_0$ for some parameter $\theta \in \Theta$. Suppose we have a test statistic $T(\theta_0)$ and critical value c so that the decision rule

$$\text{accept } \mathbb{H}_0 \text{ if } T(\theta_0) \leq c \quad ; \quad \text{reject } \mathbb{H}_0 \text{ if } T(\theta_0) > c$$

has size α. Define the *test inversion set* as the set of all θ not rejected by the test:

$$C = \{\theta \in \Theta : T(\theta) \leq c\}$$

This test inversion set is a valid choice of confidence interval.

Theorem 5.1. *If $T(\theta_0)$ has exact size α for all $\theta_0 \in \Theta$, then*

$$C = \{\theta \in \Theta : T(\theta) \leq c\}$$

is a $1-\alpha$ confidence interval for θ. If $T(\theta_0)$ has asymptotic size α for all $\theta_0 \in \Theta$, then C is an asymptotic $1-\alpha$ confidence interval for θ.

Proof. Let the true value be θ_0. Then

$$\mathbb{P}\{\theta_0 \in C\} = \mathbb{P}\{T(\theta_0) \leq c\} = 1 - \mathbb{P}\{T(\theta_0) > c\} = 1-\alpha$$

If instead T has asymptotic size α, the same proof works by applying the limit. □

Example. Again, if $\frac{\hat{\theta}-\theta}{s(\hat{\theta})} \xrightarrow{d} \mathcal{N}(0,1)$, then an asymptotic size α test for $\mathbb{H}_0 : \theta = \theta_0$ versus $\mathbb{H}_1 : \theta \neq \theta_0$ is

$$\text{accept } \mathbb{H}_0 \text{ if } |T(\theta_0)| \leq z_{1-\alpha/2} \quad \text{and} \quad \text{reject } \mathbb{H}_0 \text{ if } |T(\theta_0)| > z_{1-\alpha/2}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$th quantile of $\mathcal{N}(0,1)$, and

$$T(\theta_0) = \frac{\hat{\theta}-\theta_0}{s(\hat{\theta})}$$

The test inversion confidence interval is

$$\begin{aligned} C &= \left\{\theta \in \Theta : |T(\theta)| \leq z_{1-\alpha/2}\right\} \\ &= \left\{\theta \in \Theta : -z_{1-\alpha/2} \leq \frac{\hat{\theta}-\theta_0}{s(\hat{\theta})} \leq z_{1-\alpha/2}\right\} \\ &= \left\{\theta \in \Theta : \hat{\theta} - z_{1-\alpha/2}s(\hat{\theta}) \leq \theta \leq \hat{\theta} + z_{1-\alpha/2}s(\hat{\theta})\right\} \end{aligned}$$

which is the same as the one derived in the previous section.

Remark. In fact, all confidence intervals derived using pivotal quantities rely on test inversion.

Example. Consider a parametric model $f(x \mid \theta)$ with log likelihood $\ell_n(\theta) = \sum_{i=1}^n \log f(X_i \mid \theta)$. The likelihood ratio statistic for testing $\mathbb{H}_0 : \theta = \theta_0$ against $\mathbb{H}_1 : \theta \neq \theta_0$ is

$$LR_n(\theta_0) = 2\left(\max_{\theta\in\Theta} \ell_n(\theta) - \ell_n(\theta_0)\right)$$

Since $LR_n(\theta_0) \xrightarrow{d} \chi^2_{\dim\theta}$, an asymptotic size α test is

$$\text{accept } \mathbb{H}_0 \text{ if } LR_n(\theta_0) \leq q_{1-\alpha} \quad ; \quad textreject\mathbb{H}_0 \text{ if } LR_n(\theta_0) > q_{1-\alpha}$$

where $q_{1-\alpha}$ is the $(1-\alpha)$th quantile of $\chi^2_{\dim\theta}$. Thus, a test inversion $(1-\alpha)$ confidence interval is

$$\{\theta \in \Theta : LR_n(\theta) \leq q_{1-\alpha}\}$$

5.4 Evaluation of Confidence Interval

Remark. For the same problem, we can find many different confidence intervals. Naturally, we want small length and large coverage probability. We can always increase coverage probability by increasing the size of the interval (for example, $(-\infty, \infty)$ has coverage probability 1, but is useless). One method is to minimize length subject to a specific coverage probability.

Example. Let $\{X_1, \ldots, X_n\}$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$ with known σ^2. Then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

is a pivot. Any numbers a and b such that

$$\mathbb{P}\left\{a \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq b\right\} = 1 - \alpha$$

gives a $1-\alpha$ confidence interval. Note that this is equivalent to:

$$\left\{\mu : \bar{X} - b\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - a\frac{\sigma}{\sqrt{n}}\right\}$$

The length of the confidence interval is $(b-a)\frac{\sigma}{\sqrt{n}}$. The constrained optimization problem we can consider is

$$\min_{a,b\in\mathbb{R}} (b-a) \text{ s.t. } \mathbb{P}\{a \leq Z \leq b\} = 1 - \alpha$$

We can examine some different intervals as follows:

a	b	Probability	$b-a$
-1.34	2.33	$\mathbb{P}\{Z < a\} = 0.09$, $\mathbb{P}\{Z > b\} = 0.01$	3.67
-1.44	1.96	$\mathbb{P}\{Z < a\} = 0.075$, $\mathbb{P}\{Z > b\} = 0.025$	3.40
-1.65	1.65	$\mathbb{P}\{Z < a\} = 0.05$, $\mathbb{P}\{Z > b\} = 0.05$	3.30

Table 1: Three 90% Normal Confidence Intervals

In this case, splitting α equally in the two tails results in the shortest interval.

Definition. A pdf $f(x)$ is *unimodal* if there exists $x^\star$ such that $f(x)$ is nondecreasing for $x \leq x^\star$ and $f(x)$ is nonincreasing for $x \geq x^\star$.

Theorem 5.2. *Let $f(x)$ be a unimodal pdf. If $[a, b]$ satisfies*

1. $\int_a^b f(x)dx = 1 - \alpha$
2. $f(a) = f(b) > 0$
3. $a \leq x^\star \leq b$ *where* $x^\star$ *is the mode of* $f(x)$

Then $[a, b]$ *is the shortest among all intervals that satisfy (1).*

Correspondence between hypothesis tests and confidence intervals. Recall that a hypothesis test where size is controlled at α will reject if

$$\left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| < z_{1-\alpha/2}$$

meaning that we 'accept' $\mathbb{H}_0$ if $\bar{X}$ falls into the following acceptance region:

$$\left[\mu_0 - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \mu_0 + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \right]$$

Meanwhile, a confidence interval for $(1 - \alpha)$ is

$$\left[\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \right]$$

Remark. Hypothesis testing asks: fixing a parameter, what values of the data are consistent with that fixed parameter?

Confidence interval asks: given realized values of the data, what parameter values make these realized data most plausible?

We can think about this relationship visually:

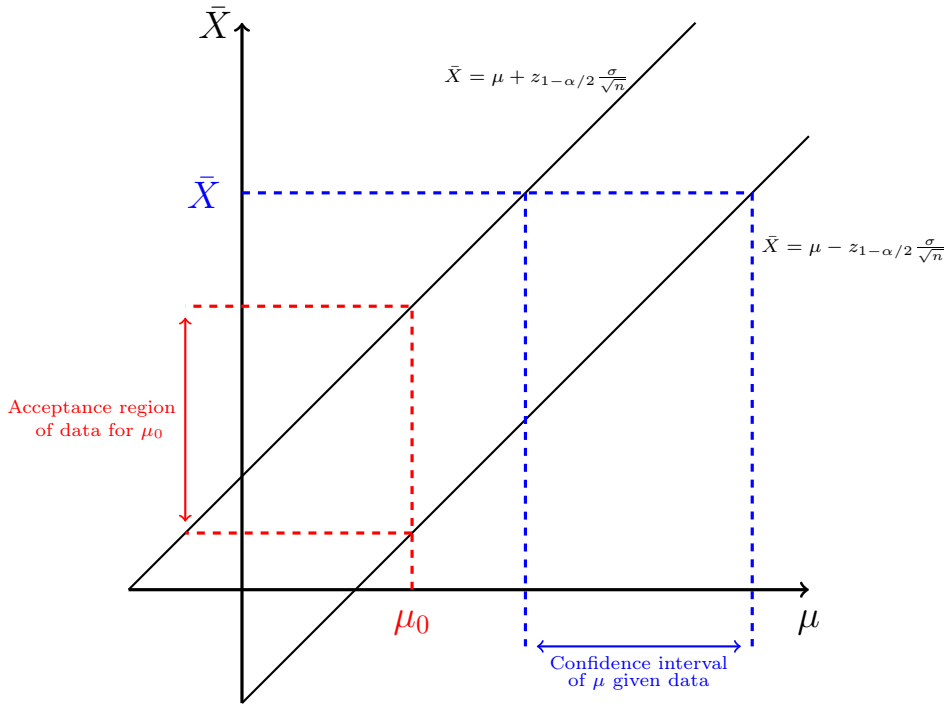

Figure 1: Confidence Intervals and Hypothesis Testing