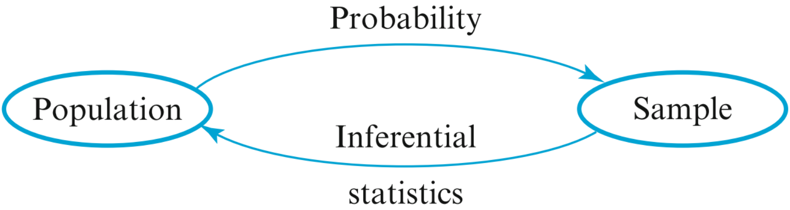
EC 303: Empirical Economic Analysis

Chapter 7: Point Estimation

Alex Hoagland, Boston University

October 21, 2019

Recall the goal of **sample collection**:



Recall the goal of **sample collection**:

- We want to move from a sample to information about the population
- Specifically, we care about parameters θ (e.g., μ, p, etc.)
- Can our sample give us a **number** for θ?

Recall the goal of **sample collection**:

- We want to move from a sample to information about the population
- Specifically, we care about parameters θ (e.g., μ, p, etc.)
- Can our sample give us a **number** for θ?

YES. This is the goal of **point estimation**.

SECTION 7.1: INTRODUCTION TO ESTIMATION

Definition. For a scalar (vector) parameter θ, a **point estimate**, $\hat{\theta}$, is another scalar (vector) calculated as a **relevant statistic from a suitable sample**.

Point Estimates

Definition. For a scalar (vector) parameter θ, a **point estimate**, $\hat{\theta}$, is another scalar (vector) calculated as a **relevant statistic from a suitable sample**.

We've seen this before!

- The mean of a sample $\overline{x}$ can tell us something about μ
- The sample variance s^2 tells us something about σ^2
- Formally, $(\overline{x}, s^2) = \hat{\theta}$ for $\theta = (\mu, \sigma^2)$

Point Estimates

Definition. For a scalar (vector) parameter θ, a **point estimate**, $\hat{\theta}$, is another scalar (vector) calculated as a **relevant statistic from a suitable sample**.

We've seen this before!

- The mean of a sample $\overline{x}$ can tell us something about μ
- The sample variance s^2 tells us something about σ^2
- Formally, $(\overline{x}, s^2) = \hat{\theta}$ for $\theta = (\mu, \sigma^2)$
- But different estimators exist! For example:

$$s_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \text{ or } s_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

- How can we differentiate these? What are good **qualities** of an estimator?

Consider the following data:
$\{24.46, 26.25, 27.15, 27.31, 27.74, 28.28, 28.49, 28.87, 29.13, 30.88\}$

- **Q:** If we think this data comes from a normal distribution, what is a sensible value of $\hat{\mu}$?

Multiple Estimators: Example

Consider the following data:
$\{24.46, 26.25, 27.15, 27.31, 27.74, 28.28, 28.49, 28.87, 29.13, 30.88\}$

- **Q:** If we think this data comes from a normal distribution, what is a sensible value of $\hat{\mu}$?
- **A:** It depends on who you ask!

Multiple Estimators: Example

Consider the following data:
$\{24.46, 26.25, 27.15, 27.31, 27.74, 28.28, 28.49, 28.87, 29.13, 30.88\}$

- **Q:** If we think this data comes from a normal distribution, what is a sensible value of $\hat{\mu}$?
- **A:** It depends on who you ask!

1. The trusty **mean**: $\overline{x} = \frac{1}{n}\sum_i x_i = 27.856$

Multiple Estimators: Example

Consider the following data:
$\{24.46, 26.25, 27.15, 27.31, 27.74, 28.28, 28.49, 28.87, 29.13, 30.88\}$

- **Q:** If we think this data comes from a normal distribution, what is a sensible value of $\hat{\mu}$?
- **A:** It depends on who you ask!

1. The trusty **mean**: $\overline{x} = \frac{1}{n}\sum_i x_i = 27.856$
2. The **median**: $\tilde{x} = (27.74 + 28.28)/2 = 28.01$

Multiple Estimators: Example

Consider the following data:
$\{24.46, 26.25, 27.15, 27.31, 27.74, 28.28, 28.49, 28.87, 29.13, 30.88\}$

- **Q:** If we think this data comes from a normal distribution, what is a sensible value of $\hat{\mu}$?
- **A:** It depends on who you ask!

1. The trusty **mean**: $\overline{x} = \frac{1}{n}\sum_i x_i = 27.856$
2. The **median**: $\tilde{x} = (27.74 + 28.28)/2 = 28.01$
3. The **midrange**: $\overline{x}_c = (\min(x_i) + \max(x_i))/2 = 27.67$

Multiple Estimators: Example

Consider the following data:
$\{24.46, 26.25, 27.15, 27.31, 27.74, 28.28, 28.49, 28.87, 29.13, 30.88\}$

- **Q:** If we think this data comes from a normal distribution, what is a sensible value of $\hat{\mu}$?
- **A:** It depends on who you ask!

1. The trusty **mean**: $\overline{x} = \frac{1}{n}\sum_i x_i = 27.856$
2. The **median**: $\tilde{x} = (27.74 + 28.28)/2 = 28.01$
3. The **midrange**: $\overline{x}_c = (\min(x_i) + \max(x_i))/2 = 27.67$
4. The **trimmed mean**: throw away the top/bottom 10% of observations: $\overline{x}_t = 27.90$

Multiple Estimators: Example

Consider the following data:
$\{24.46, 26.25, 27.15, 27.31, 27.74, 28.28, 28.49, 28.87, 29.13, 30.88\}$

- **Q:** If we think this data comes from a normal distribution, what is a sensible value of $\hat{\mu}$?
- **A:** It depends on who you ask!

1. The trusty **mean**: $\overline{x} = \frac{1}{n}\sum_i x_i = 27.856$
2. The **median**: $\tilde{x} = (27.74 + 28.28)/2 = 28.01$
3. The **midrange**: $\overline{x}_c = (\min(x_i) + \max(x_i))/2 = 27.67$
4. The **trimmed mean**: throw away the top/bottom 10% of observations: $\overline{x}_t = 27.90$

Which should you choose?

What Makes a Good Estimator?

If possible, we would like $\hat{\theta} = \theta$ always

- We cannot know this since we don't know θ!
- $\hat{\theta}$ is a statistic, so it is **sample dependent**.

What Makes a Good Estimator?

If possible, we would like $\hat{\theta} = \theta$ always

- We cannot know this since we don't know θ!
- $\hat{\theta}$ is a statistic, so it is **sample dependent**.
- We'll settle for an **accurate estimator**. Formally, if

$$\hat{\theta} = \theta + \epsilon,$$

then $\hat{\theta}$ is more **accurate** as ϵ gets smaller

- Want to consider average error across numerous samples
- That is, accurate estimators minimize $\mathbb{E}\left[f(\hat{\theta} - \theta)\right]$

There are several competing notions of close estimators $f(\hat{\theta} - \theta)$. Two big ones:

1. **Squared Error**: $f = (\hat{\theta} - \theta)^2$
2. **Absolute Error**: $f = |\hat{\theta} - \theta|$

There are several competing notions of close estimators $f(\hat{\theta} - \theta)$. Two big ones:

1. **Squared Error**: $f = (\hat{\theta} - \theta)^2$
2. **Absolute Error**: $f = |\hat{\theta} - \theta|$

When we take expectations, these become **Mean Squared Error (MSE)** and **Mean Absolute Deviations (MAD)**.

Mean Squared Error

People generally prefer the **MSE** for calculations.

1. More straightforward mathematically
2. Punishes larger deviations more heavily

Mean Squared Error

People generally prefer the **MSE** for calculations.

1. More straightforward mathematically
2. Punishes larger deviations more heavily
3. Has the convenient representation:

$$
\begin{aligned}
\mathbb{V}(Y) &= \mathbb{E}(Y^2) - (\mathbb{E}[Y])^2 \\
\Rightarrow \mathbb{E}[Y^2] &= \mathbb{V}[Y] + (\mathbb{E}[Y])^2 \\
\Rightarrow \mathbb{E}[(\hat{\theta} - \theta)^2] &= \mathbb{V}[\hat{\theta} - \theta] + \left(\mathbb{E}[\hat{\theta} - \theta]\right)^2 \\
\Rightarrow MSE &= \mathbb{V}[\hat{\theta}] + \left(\mathbb{E}[\hat{\theta}] - \theta\right)^2
\end{aligned}
$$

Mean Squared Error

People generally prefer the **MSE** for calculations.

1. More straightforward mathematically
2. Punishes larger deviations more heavily
3. Has the convenient representation:

$$\mathbb{V}(Y) = \mathbb{E}(Y^2) - (\mathbb{E}[Y])^2$$
$$\Rightarrow \mathbb{E}[Y^2] = \mathbb{V}[Y] + (\mathbb{E}[Y])^2$$
$$\Rightarrow \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{V}[\hat{\theta} - \theta] + \left(\mathbb{E}[\hat{\theta} - \theta]\right)^2$$
$$\Rightarrow MSE = \mathbb{V}[\hat{\theta}] + \left(\mathbb{E}[\hat{\theta}] - \theta\right)^2$$

$$MSE = (\textbf{Estimator Variance}) + (\textbf{Estimator bias})^2$$

Example: Binomial Experiment

Consider a binomial experiment with n trials. Our parameter of interest is $\theta = p$, the probability of success

- Let $\hat{p} = X/n$ for the number of successes X in a trial

Example: Binomial Experiment

Consider a binomial experiment with n trials. Our parameter of interest is $\theta = p$, the probability of success

- Let $\hat{p} = X/n$ for the number of successes X in a trial
- The LLN shows that $\hat{p} \to p$ as $n \to \infty$, so this seems like a good estimator
- **Q:** What is its MSE?

Example: Binomial Experiment

Consider a binomial experiment with n trials. Our parameter of interest is $\theta = p$, the probability of success

- Let $\hat{p} = X/n$ for the number of successes X in a trial
- The LLN shows that $\hat{p} \to p$ as $n \to \infty$, so this seems like a good estimator
- **Q:** What is its MSE?

First we calculate the bias:

$$\begin{aligned}
\mathbb{E}[\hat{p}] - p &= \mathbb{E}\left[\frac{X}{n}\right] - p \\
&= \frac{1}{n}\mathbb{E}[X] - p \\
&= \frac{1}{n}np - p \text{ (since } X \text{ has a binomial distribution)} \\
&= 0.
\end{aligned}$$

This estimator is **unbiased** on average.

Example: Binomial Experiment

Second, we calculate the variance:

$$\begin{aligned}
\mathbb{V}[\hat{p}] &= \mathbb{V}\left[\frac{X}{n}\right] \\
&= \frac{1}{n^2}\mathbb{V}[X] \\
&= \frac{np(1-p)}{n^2} \\
&= \frac{p(1-p)}{n}
\end{aligned}$$

Example: Binomial Experiment

Second, we calculate the variance:

$$\begin{aligned}\mathbb{V}[\hat{p}] &= \mathbb{V}\left[\frac{X}{n}\right] \\ &= \frac{1}{n^2}\mathbb{V}[X] \\ &= \frac{np(1-p)}{n^2} \\ &= \frac{p(1-p)}{n}\end{aligned}$$

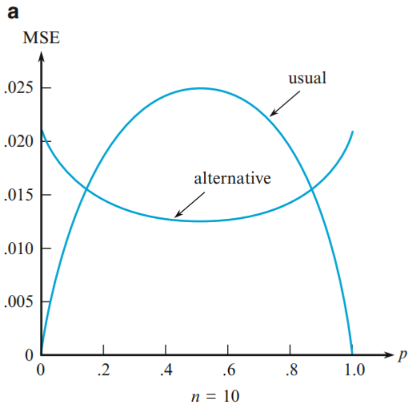
Hence, the MSE is given by

$$MSE = \frac{p(1-p)}{n}$$

- This depends on something we don't know!
- **How are we supposed to compare estimators?**

Example: Binomial Experiment

To compare estimators, fix n and look at how MSEs range over all values of p



A Desirable Estimator

In general, we prefer **unbiased** estimators to **biased** ones

A Desirable Estimator

In general, we prefer **unbiased** estimators to **biased** ones

- Centered around the truth
- Examples: $\overline{x}, \tilde{x}, s^2, \dots$

A Desirable Estimator

In general, we prefer **unbiased** estimators to **biased** ones

- Centered around the truth
- Examples: $\overline{x}, \tilde{x}, s^2, ...$

When choosing between unbiased estimators, we look for **minimum variance**

A Desirable Estimator

In general, we prefer **unbiased** estimators to **biased** ones

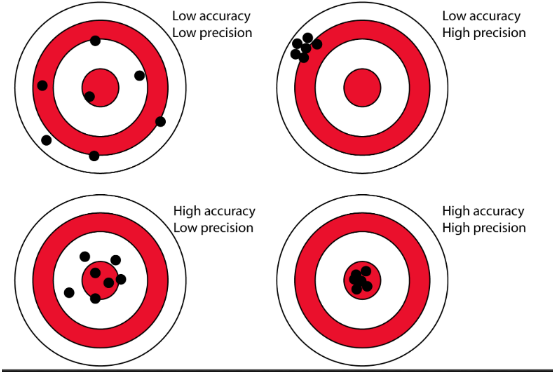
- Centered around the truth
- Examples: $\overline{x}, \tilde{x}, s^2, \ldots$

When choosing between unbiased estimators, we look for **minimum variance**

- If Bias $= 0$ and variance is minimized, then so is MSE

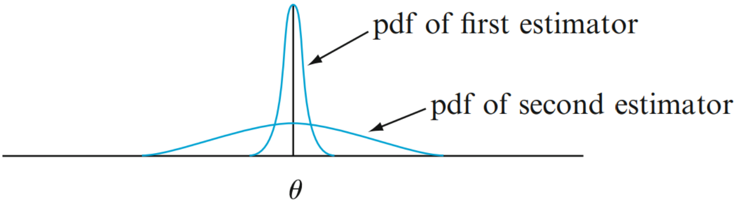
A Desirable Estimator

Bias and **minimum variance** are related to concepts of **accuracy** and **precision**:



A Desirable Estimator

Bias and **minimum variance** are related to concepts of **accuracy** and **precision**:



Example: Estimating a Mean

Suppose that TFP shocks to an economy are centered around a mean μ. Given a sample of shocks $\{X_1, ..., X_n\}$, consider two **unbiased** estimators: $\overline{x}$ and the midrange $\overline{x}_c$.

- Easy to see that both are unbiased since $\mathbb{E}[X_i] = \mu$ for all i.
- What about **variances**?

Example: Estimating a Mean

Suppose that TFP shocks to an economy are centered around a mean μ. Given a sample of shocks $\{X_1, ..., X_n\}$, consider two **unbiased** estimators: $\overline{x}$ and the midrange $\overline{x}_c$.

- Easy to see that both are unbiased since $\mathbb{E}[X_i] = \mu$ for all i.
- What about **variances**?

For the mean:

$$\begin{aligned}
\mathbb{V}[\overline{x}] &= \frac{1}{n^2} \sum \mathbb{V}[X_i] \\
&= \frac{n\sigma^2}{n^2} \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

Example: Estimating a Mean

Suppose that TFP shocks to an economy are centered around a mean μ. Given a sample of shocks $\{X_1, ..., X_n\}$, consider two **unbiased** estimators: $\overline{x}$ and the midrange $\overline{x}_c$.

- Easy to see that both are unbiased since $\mathbb{E}[X_i] = \mu$ for all i.
- What about **variances**?

For the midrange:

$$\begin{aligned}
\mathbb{V}[\overline{x}_c] &= \mathbb{V}\left[\frac{\min(x_i) + \max(x_i)}{2}\right] \\
&= \frac{1}{4}\left(\mathbb{V}[\min(x_i)] + \mathbb{V}[\max(x_i)]\right) \\
&= \frac{2}{4}\sigma^2 \\
&= \frac{\sigma^2}{2}.
\end{aligned}$$

Suppose that TFP shocks to an economy are centered around a mean μ. Given a sample of shocks $\{X_1, ..., X_n\}$, consider two **unbiased** estimators: $\overline{x}$ and the midrange $\overline{x}_c$.

- Easy to see that both are unbiased since $\mathbb{E}[X_i] = \mu$ for all i.
- What about **variances**?

Therefore, whenever $n > 2$, the mean has a *lower* variance than the midrange.

- Why does this make sense **intuitively**?
- In fact, the sample mean is the minimum variance unbiased estimator **(MVUE)** for the population mean of a normal distribution

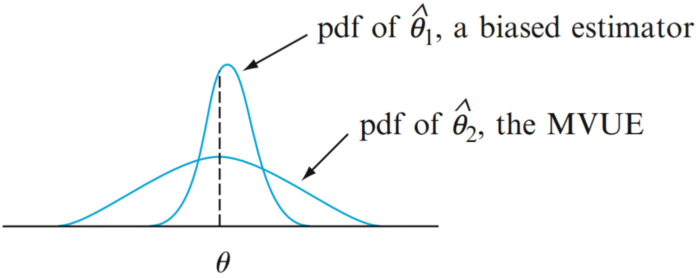
Estimation Tradeoffs: Which is Worse?

How do we choose between **accuracy** and **precision**?

Estimation Tradeoffs: Which is Worse?

How do we choose between **accuracy** and **precision**?

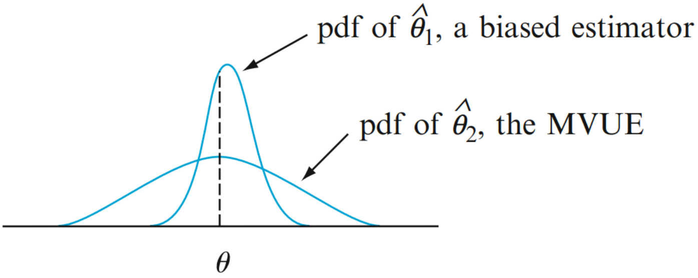
- For example, which of these distributions do you prefer?



Estimation Tradeoffs: Which is Worse?

How do we choose between **accuracy** and **precision**?

- For example, which of these distributions do you prefer?



- Depends on your application
- Often difficult to know your bias exactly, so unbiasedness generally preferred, even at cost of higher variance
- Variance can be controlled more (e.g., selecting larger n)

Standard Errors

To communicate information about an estimate's **precision**, we report its **standard error**:

$$\sigma_{\hat{\theta}} = \sqrt{V\left(\hat{\theta}\right)}$$

Standard Errors

To communicate information about an estimate's **precision**, we report its **standard error**:

$$\sigma_{\hat{\theta}} = \sqrt{V\left(\hat{\theta}\right)}$$

- When this contains values that need to be estimated: $\hat{\sigma}_{\hat{\theta}}$
- Example: $\overline{x}$ has variance σ^2/n, so its estimated standard error is $\hat{\sigma}/\sqrt{n} = s/\sqrt{n}$
- Standard errors will be **very** important in traditional hypothesis testing, confidence intervals, and inference

Bootstrapped Standard Errors

Frequently, $\hat{\theta}$ may be so complicated that $\sigma_{\hat{\theta}}$ isn't easily obtained. In this case, we use a simulation called **bootstrapping** to obtain valid SEs:

Frequently, $\hat{\theta}$ may be so complicated that $\sigma_{\hat{\theta}}$ isn't easily obtained. In this case, we use a simulation called **bootstrapping** to obtain valid SEs:

1. Begin with your sample $\{X_1, ..., X_n\}$
2. From that sample, **sample again with replacement** to get $\{X_1^1, ..., X_n^1\}$

Frequently, $\hat{\theta}$ may be so complicated that $\sigma_{\hat{\theta}}$ isn't easily obtained. In this case, we use a simulation called **bootstrapping** to obtain valid SEs:

1. Begin with your sample $\{X_1, ..., X_n\}$
2. From that sample, **sample again with replacement** to get $\{X_1^1, ..., X_n^1\}$
3. Calculate $\hat{\theta}^1$ using that sample

Bootstrapped Standard Errors

Frequently, $\hat{\theta}$ may be so complicated that $\sigma_{\hat{\theta}}$ isn't easily obtained. In this case, we use a simulation called **bootstrapping** to obtain valid SEs:

1. Begin with your sample $\{X_1, ..., X_n\}$
2. From that sample, **sample again with replacement** to get $\{X_1^1, ..., X_n^1\}$
3. Calculate $\hat{\theta}^1$ using that sample
4. Repeat steps (2) and (3) B times to obtain B different estimates of θ, $\{\hat{\theta}^1, ..., \hat{\theta}^B\}$.

Bootstrapped Standard Errors

Frequently, $\hat{\theta}$ may be so complicated that $\sigma_{\hat{\theta}}$ isn't easily obtained. In this case, we use a simulation called **bootstrapping** to obtain valid SEs:

1. Begin with your sample $\{X_1, ..., X_n\}$
2. From that sample, **sample again with replacement** to get $\{X_1^1, ..., X_n^1\}$
3. Calculate $\hat{\theta}^1$ using that sample
4. Repeat steps (2) and (3) B times to obtain B different estimates of θ, $\{\hat{\theta}^1, ..., \hat{\theta}^B\}$.
5. Calculate the mean $\overline{\theta}^*$ of these estimates, and standard errors:

$$S_{\hat{\theta}} = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}^i - \overline{\theta}^*)^2}$$

TO STATA!

SECTION 7.2: METHODS OF ESTIMATION

So far, our estimators have been “educated guesses". How can we **formalize** estimation?

Constructing an Estimator

So far, our estimators have been “educated guesses". How can we **formalize** estimation?

Today: two main classes of estimation:

1. **Method of Moments (MM)**: sample characteristics should match population values
2. **Maximum Likelihood Estimation (MLE)**: mathematically optimize likelihood of data

Method of Moments

Intuitively, we construct estimators that **match** sample & population characteristics:

$$\text{Population Moments} \Rightarrow \mathbb{E}[X^k] \Leftrightarrow \frac{1}{n}\sum_i X_i^k \Leftarrow \text{Sample Moments}$$

Method of Moments

Intuitively, we construct estimators that **match** sample & population characteristics:

$$\text{Population Moments} \Rightarrow \mathbb{E}[X^k] \Leftrightarrow \frac{1}{n}\sum_i X_i^k \Leftarrow \text{Sample Moments}$$

For a random sample $\{X_1, ..., X_n\}$ from $f(x; \theta_1, ..., \theta_m)$, the **moment estimators** $\vec{\hat{\theta}} = (\hat{\theta}_1, ..., \hat{\theta}_m)$ are obtained by equating the first m sample and population moments

Method of Moments

Intuitively, we construct estimators that **match** sample & population characteristics:

$$\text{Population Moments} \Rightarrow \mathbb{E}[X^k] \Leftrightarrow \frac{1}{n}\sum_i X_i^k \Leftarrow \text{Sample Moments}$$

For a random sample $\{X_1, ..., X_n\}$ from $f(x; \theta_1, ..., \theta_m)$, the **moment estimators** $\vec{\hat{\theta}} = (\hat{\theta}_1, ..., \hat{\theta}_m)$ are obtained by equating the first m sample and population moments

- Need m equations to solve for m unknowns

Example: Gamma Distribution

Suppose that $\{X_1, ..., X_n\}$ come from a Gamma distribution with parameters (α, β). We can solve for population moments:

$$\mathbb{E}[X] = \alpha\beta \tag{1}$$

$$\mathbb{E}[X^2] = \beta^2(\alpha + 1)\alpha \tag{2}$$

Example: Gamma Distribution

Suppose that $\{X_1, ..., X_n\}$ come from a Gamma distribution with parameters (α, β). We can solve for population moments:

$$\mathbb{E}[X] = \alpha\beta \tag{1}$$

$$\mathbb{E}[X^2] = \beta^2(\alpha + 1)\alpha \tag{2}$$

To find the moment estimators, we equate these to sample moments:

$$\overline{X} = \frac{1}{n}\sum_i X_i = \alpha\beta$$

$$\frac{1}{n}\sum_i X_i^2 = \beta^2(\alpha + 1)\alpha$$

We can **solve** this system for $(\hat{\alpha}, \hat{\beta})$.

Example: Gamma Distribution

After solving, we find

$$\hat{\alpha} = \frac{\overline{X}^2}{\frac{1}{n}\sum_i X_i^2 - \overline{X}^2}$$
$$\hat{\beta} = \frac{\frac{1}{n}\sum_i x_i^2 - \overline{X}^2}{\overline{X}}$$

After solving, we find

$$\hat{\alpha} = \frac{\overline{X}^2}{\frac{1}{n}\sum_i X_i^2 - \overline{X}^2}$$

$$\hat{\beta} = \frac{\frac{1}{n}\sum_i x_i^2 - \overline{X}^2}{\overline{X}}$$

Now we have an estimator for **any sample!** For example, let $X = \{152, 115, 109, 94, 101\}$. Then

$$\overline{X} \approx 114 \text{ and } \frac{1}{n}\sum_i X_i^2 \approx 13450$$

$$\Rightarrow \hat{\alpha} \approx 32 \text{ and } \hat{\beta} \approx 4.$$

Method of Moments: Takeaways

1. **Method of Moments** estimators are typically easy to compute

Method of Moments: Takeaways

1. **Method of Moments** estimators are typically easy to compute
2. They are usually **consistent** and **asymptotically normal**

Method of Moments: Takeaways

1. **Method of Moments** estimators are typically easy to compute
2. They are usually **consistent** and **asymptotically normal**
3. However, there are caveats:
   - Frequently biased (ex: $\hat{\sigma}^2$)
   - May give estimates that don't match data (ex: estimating $[\hat{A}, \hat{B}]$ for a uniform distribution may give estimates that don't catch all data)
   - Using first m moments doesn't use full information in distribution $\Rightarrow$ development of MLE

Method of Moments: Takeaways

1. **Method of Moments** estimators are typically easy to compute
2. They are usually **consistent** and **asymptotically normal**
3. However, there are caveats:
   - Frequently biased (ex: $\hat{\sigma}^2$)
   - May give estimates that don't match data (ex: estimating $[\hat{A}, \hat{B}]$ for a uniform distribution may give estimates that don't catch all data)
   - Using first m moments doesn't use full information in distribution $\Rightarrow$ development of MLE
4. More recently, this method has been generalized **(GMM)** by L.P. Hansen. This is incredibly popular in econometrics today.

Maximum Likelihood Estimation

- Using MM, we wanted to make the population match the sample for m characteristics
- For MLE, we want to choose a population that **maximizes the chance** that we would obtain a particular sample

Maximum Likelihood Estimation

- Using MM, we wanted to make the population match the sample for m characteristics
- For MLE, we want to choose a population that **maximizes the chance** that we would obtain a particular sample
- To do that, we specify a **likelihood** function for the creation of data
- For example, suppose we run a Bernoulli experiment with $n = 5$ and observe data $X = \{1, 0, 0, 1, 0\}$. What is the likelihood of this given p?

- Using MM, we wanted to make the population match the sample for m characteristics
- For MLE, we want to choose a population that **maximizes the chance** that we would obtain a particular sample
- To do that, we specify a **likelihood** function for the creation of data
- For example, suppose we run a Bernoulli experiment with $n = 5$ and observe data $X = \{1, 0, 0, 1, 0\}$. What is the likelihood of this given p?

$$\begin{aligned} f(x_1, x_2, ..., x_5; p) &= p(1-p)(1-p)p(1-p) \\ &= p^2(1-p)^3 \end{aligned}$$

Once we've specified f, we choose θ to **maximize** it.

Maximum Likelihood Estimation

Once we've specified f, we choose θ to **maximize** it.

- The value θ^* that maximizes f is our estimator $\hat{\theta}$
- It is typically easier to maximize $\ln(f)$ rather than f directly
  - Why is this an okay transformation?

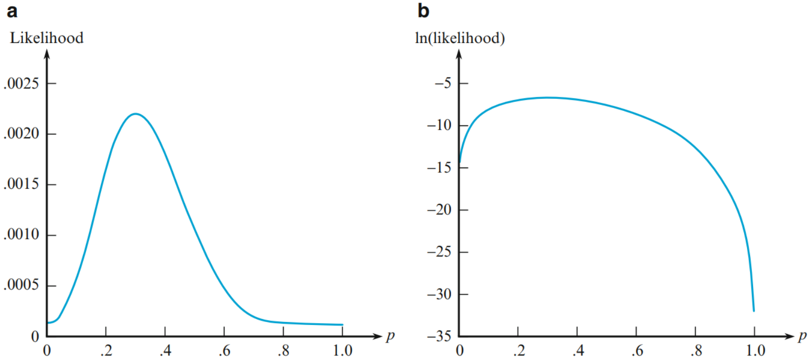
Once we've specified f, we choose θ to **maximize** it.

- The value θ^* that maximizes f is our estimator $\hat{\theta}$
- It is typically easier to maximize $\ln(f)$ rather than f directly
  - Why is this an okay transformation?
- In our example:

$$
\begin{aligned}
\ln(f) &= 2\ln(p) + 3\ln(1-p) \\
\Rightarrow \frac{d\ln(f)}{dp} &\equiv 0 \\
\Rightarrow \frac{2}{p} - \frac{3}{1-p} &\equiv 0 \\
\Rightarrow \frac{2}{p} &\equiv \frac{3}{1-p} \\
\Rightarrow 2 - 2p &\equiv 3p \Rightarrow p^* = \frac{2}{5}
\end{aligned}
$$

Why do We Prefer Log-Likelihoods?

This transformation generally **smooths** the function



Example 2: Exponential Distribution

Consider a more arbitrary example: $\{X_i\}_{i=1}^n$ is an i.i.d. sample from an **exponential** distribution with parameter λ. We want to estimate $\hat{\lambda}$ by MLE.

Example 2: Exponential Distribution

Consider a more arbitrary example: $\{X_i\}_{i=1}^n$ is an i.i.d. sample from an **exponential** distribution with parameter λ. We want to estimate $\hat{\lambda}$ by MLE.

1. **Write the likelihood function.** By independence, we can take the product of each pdf:

$$f(x_1, ..., x_n; \lambda) = \Pi_{i=1}^n \lambda e^{-\lambda x_i}$$

Example 2: Exponential Distribution

Consider a more arbitrary example: $\{X_i\}_{i=1}^{n}$ is an i.i.d. sample from an **exponential** distribution with parameter λ. We want to estimate $\hat{\lambda}$ by MLE.

2. **Transform to log-likelihood.** The log of a product is the sum of the logs:

$$
\begin{aligned}
\ln(f(x_1, ..., x_n; \lambda)) &= \ln\left(\Pi_{i=1}^{n} \lambda e^{-\lambda x_i}\right) \\
&= \sum_{i=1}^{n} \ln\left(\lambda e^{-\lambda x_i}\right) \\
&= \sum_{i=1}^{n} \ln(\lambda) - \sum_{i=1}^{n} \lambda x_i \\
&= n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i .
\end{aligned}
$$

Consider a more arbitrary example: $\{X_i\}_{i=1}^{n}$ is an i.i.d. sample from an **exponential** distribution with parameter λ. We want to estimate $\hat{\lambda}$ by MLE.

3. **Maximize.** Taking the first derivative and equating to 0:

$$\frac{d\ln(f)}{d\lambda} = \frac{n}{\lambda} - \sum_i x_i \equiv 0$$

$$\Rightarrow \lambda^* \equiv \left(\frac{1}{n}\sum_i x_i\right)^{-1} = \frac{1}{\overline{X}}$$

- Note that this is a **biased** estimator since $\mathbb{E}(1/\overline{X}) \neq 1/\mathbb{E}(\overline{X})$
- This is the same as the MM estimator in this case.

Example 2: Exponential Distribution

Consider a more arbitrary example: $\{X_i\}_{i=1}^n$ is an i.i.d. sample from an **exponential** distribution with parameter λ. We want to estimate $\hat{\lambda}$ by MLE.

4. **Check for a maximum.** Remember that the second derivative should be negative!

$$\frac{d^2 \ln(f)}{d\lambda^2} = -\frac{n}{\lambda^2} < 0 \text{ for all } \lambda$$

Properties of MLE

Maximum Likelihood is **popular** for a lot of reasons:

1. Very **tractable**—if your sample is independent, math isn't too bad

Properties of MLE

Maximum Likelihood is **popular** for a lot of reasons:

1. Very **tractable**—if your sample is independent, math isn't too bad
2. MLE is **invariant** to transformations:

$$\hat{\theta}^* \text{ maximizes } \ln(f(x;\theta)) \Leftrightarrow h(\hat{\theta}^*) \text{ maximizes } \ln(f(x;h(\theta)))$$

Properties of MLE

Maximum Likelihood is **popular** for a lot of reasons:

1. Very **tractable**—if your sample is independent, math isn't too bad
2. MLE is **invariant** to transformations:

$$\hat{\theta}^* \text{ maximizes } \ln(f(x;\theta)) \Leftrightarrow h(\hat{\theta}^*) \text{ maximizes } \ln(f(x; h(\theta)))$$

3. For large samples, the MLE $\hat{\theta}$ of any parameter is **approximately** the MVUE of θ:
   - $\hat{\theta} \xrightarrow{p} \theta$ (**consistency**)
   - Approximately **unbiased**: $\mathbb{E}[\hat{\theta}] - \theta \approx 0$
   - Nearly minimum variance among unbiased estimators

Properties of MLE

Maximum Likelihood is **popular** for a lot of reasons:

1. Very **tractable**—if your sample is independent, math isn't too bad
2. MLE is **invariant** to transformations:

$$\hat{\theta}^* \text{ maximizes } \ln(f(x;\theta)) \Leftrightarrow h(\hat{\theta}^*) \text{ maximizes } \ln(f(x;h(\theta)))$$

3. For large samples, the MLE $\hat{\theta}$ of any parameter is **approximately** the MVUE of θ:
   - $\hat{\theta} \xrightarrow{p} \theta$ (**consistency**)
   - Approximately **unbiased**: $\mathbb{E}[\hat{\theta}] - \theta \approx 0$
   - Nearly minimum variance among unbiased estimators
4. MLE is also **asymptotically normal**:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, I(\theta)^{-1}), \ I \text{ is the } \textbf{Fischer information matrix}$$

When does MLE fail?

MLE relies on us being able to use **calculus**.

- In cases of non-differentiability, we may have trouble

When does MLE fail?

MLE relies on us being able to use **calculus**.

- In cases of non-differentiability, we may have trouble

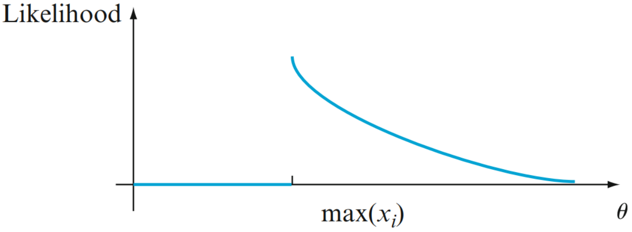
Example:
Suppose that initial endowments for agents in an economy is uniformly distributed on interval $[0, \theta]$. The likelihood function, given our data, is

$$f(x_1, ..., x_n; \theta) = \begin{cases} \frac{1}{\theta^n} & x_i \in [0, \theta] \text{ for all } i \\ 0 & \text{Otherwise.} \end{cases}$$

MLE relies on us being able to use **calculus**.

- In cases of non-differentiability, we may have trouble

Example:



The **discontinuity** contains the maximum value, but calculus wouldn't find that!

What if I'm not sure about my likelihood function?

There are estimation techniques that are **flexible** for multiple f's:

1. **Robust** estimators, or those that work for many pdfs.
   - Can handle small measurement errors and outliers well
   - Example: trimmed means or Winsorised estimators

What if I'm not sure about my likelihood function?

There are estimation techniques that are **flexible** for multiple f's:

1. **Robust** estimators, or those that work for many pdfs.
2. ***M*-estimation** generalizes MLE:
   - Instead of maximizing a likelihood function f, choose an "objective function" $\rho(x_i; \theta)$
   - Examples of ρ: MSE, MAD, etc. Ensures robustness of $\hat{\theta}$
   - The M-estimation problem is $\theta^* = \text{argmax} \sum_i \rho(x_i; \theta)$

What if I'm not sure about my likelihood function?

There are estimation techniques that are **flexible** for multiple f's:

1. **Robust** estimators, or those that work for many pdfs.
2. ***M*-estimation** generalizes MLE:

We won't cover these in depth in this course.

SECTION 7.3: EVALUATING ESTIMATORS

Can I improve my estimator?

So far when discussing estimators, we've **restricted attention** to specific **classes** of estimators

- Unbiased estimators
- Linear estimators
- From there, we aim for the **minimum variance** estimator

Can I improve my estimator?

So far when discussing estimators, we've **restricted attention** to specific **classes** of estimators

- Unbiased estimators
- Linear estimators
- From there, we aim for the **minimum variance** estimator

This section asks related questions:

1. How do I know if my estimator is good enough? (**Sufficiency**)
2. How much information am I getting from my sample? (**Information**)
3. Can I make my estimate better? (**Efficiency**)

Sufficiency

Suppose we are after θ and are considering an estimator $T = T(x_1, ..., x_n)$.

- We know that T tells us *nothing* about θ if they are independent
  - Example: X_1, X_2 come from a normal distribution $\mathcal{N}(\mu, \sigma^2)$.
  - The statistic $T = X_1 - X_2$ has a mean of 0 and variance of $2\sigma^2$
  - Since T's distribution does not depend on μ, T is **uninformative**

Sufficiency

Suppose we are after θ and are considering an estimator $T = T(x_1, ..., x_n)$.

- We know that T tells us *nothing* about θ if they are independent
- Conversely, it is possible for T to give us **all** the information about θ we want
    - Consider the conditional joint distribution $f(x_1, ..., x_n | T(\theta))$
    - If $T(\theta)$ contains information about θ but $f(\{x_i\}|T(\theta))$ doesn't, then there is no information from the sample left unused by T

Suppose we are after θ and are considering an estimator $T = T(x_1, ..., x_n)$.

- We know that T tells us *nothing* about θ if they are independent
- Conversely, it is possible for T to give us **all** the information about θ we want
- This is the notion of **sufficiency**

Definition. A statistic $T(X_1, ..., X_n)$ is **sufficient** for θ if the joint distribution of $(X_1, ..., X_n)$ given $T = t$ does not depend on θ for all possible values of $t \in \text{Supp}(T)$.

Sufficiency: Example

We are examining major defects in automobiles. Our data for number of defects in each sampled car X is $\{1, 0, 3\}$.

- You think X has a Poisson distribution and want to estimate λ
- Instead of seeing the whole sample, you're only told that $T = \sum_i x_i = 4$.
- **Q:** What can you infer?

Sufficiency: Example

We are examining major defects in automobiles. Our data for number of defects in each sampled car X is $\{1, 0, 3\}$.

- You think X has a Poisson distribution and want to estimate λ
- Instead of seeing the whole sample, you're only told that $T = \sum_i x_i = 4$.
- **Q:** What can you infer?

We argue that $T = \sum_i x_i = 4$ is **sufficient** to estimate $\hat{\lambda}$.

1. Consider joint distribution $f(x_1, x_2, x_3 | \sum_i x_i = 4)$

Sufficiency: Example

We are examining major defects in automobiles. Our data for number of defects in each sampled car X is $\{1, 0, 3\}$.

- You think X has a Poisson distribution and want to estimate λ
- Instead of seeing the whole sample, you're only told that $T = \sum_i x_i = 4$.
- **Q:** What can you infer?

We argue that $T = \sum_i x_i = 4$ is **sufficient** to estimate $\hat{\lambda}$.

1. Consider joint distribution $f(x_1, x_2, x_3 | \sum_i x_i = 4)$
2. Since each $x_i \in \mathbb{Z}_+$, the support of this is limited:

$$P\left(x_1, x_2, x_3 | \sum_i x_i = 4\right) = 0 \text{ unless } x_1 + x_2 + x_3 = 4$$

Sufficiency: Example

We are examining major defects in automobiles. Our data for number of defects in each sampled car X is $\{1, 0, 3\}$.

- You think X has a Poisson distribution and want to estimate λ
- Instead of seeing the whole sample, you're only told that $T = \sum_i x_i = 4$.
- **Q:** What can you infer?

We argue that $T = \sum_i x_i = 4$ is **sufficient** to estimate $\hat{\lambda}$.

1. Consider joint distribution $f(x_1, x_2, x_3 | \sum_i x_i = 4)$
2. Since each $x_i \in \mathbb{Z}_+$, the support of this is limited:
3. Each of these probabilities is fixed since T is Poisson(3λ):

$$P(X = (2,1,1) | T = 4) = \frac{P[X = (2,1,1)]}{P(T = 4)}$$
$$= \frac{\frac{e^{-\lambda}\lambda^2}{2!} \frac{e^{-\lambda}\lambda^1}{1!} \frac{e^{-\lambda}\lambda^1}{1!}}{\frac{e^{-3\lambda}(3\lambda)^4}{4!}} = \frac{4}{81}$$

We are examining major defects in automobiles. Our data for number of defects in each sampled car X is $\{1, 0, 3\}$.

- You think X has a Poisson distribution and want to estimate λ
- Instead of seeing the whole sample, you're only told that $T = \sum_i x_i = 4$.
- **Q:** What can you infer?

We argue that $T = \sum_i x_i = 4$ is **sufficient** to estimate $\hat{\lambda}$.

1. Consider joint distribution $f(x_1, x_2, x_3 | \sum_i x_i = 4)$
2. Since each $x_i \in \mathbb{Z}_+$, the support of this is limited:
3. Each of these probabilities is fixed since T is Poisson(3λ):
4. The conditional pdf is determined so T is **sufficient**

Intuitively, think of the above setup in two steps:

1. You first observe the value of $T = \sum_i x_i$ given a Poisson distribution
2. Given T, you then assign the probability of each combination of (x_1, x_2, x_3)

Sufficiency: Example

Intuitively, think of the above setup in two steps:

1. You first observe the value of $T = \sum_i x_i$ given a Poisson distribution
2. Given T, you then assign the probability of each combination of (x_1, x_2, x_3)

Since the second step **does not depend on** λ, T is sufficient.

Sufficiency: Example

Intuitively, think of the above setup in two steps:

1. You first observe the value of $T = \sum_i x_i$ given a Poisson distribution
2. Given T, you then assign the probability of each combination of (x_1, x_2, x_3)

Since the second step **does not depend on** λ, T is sufficient.

- It is always possible to find an MLE estimator that is just a function of sufficient statistic(s)!
- Sufficiency is very handy when you trust the distribution in your head
  - If you want to be flexible, need to use more **robust** options

As we've seen before, the asymptotic variance of an MLE estimator is the inverse of something called the Fisher Information Matrix:

$$\begin{aligned} I_n(\theta) &= \mathbb{V}\left[\frac{\partial}{\partial\theta}\ln(f(\vec{x};\theta))\right] \\ &= \mathbb{V}\left[s(\vec{x};\theta)\right] \end{aligned}$$

Information

As we've seen before, the asymptotic variance of an MLE estimator is the inverse of something called the **Fisher Information Matrix**:

$$
\begin{aligned}
I_n(\theta) &= \mathbb{V}\left[\frac{\partial}{\partial\theta}\ln(f(\vec{x};\theta))\right] \\
&= \mathbb{V}\left[s(\vec{x};\theta)\right]
\end{aligned}
$$

- $s(\cdot)$ is called the **score** operator—how sensitive your log-likelihood is to θ
  - Note that we set this equal (close) to 0 in MLE.
  - Hence $s(\cdot)$ is a random variable of mean 0

Information

As we've seen before, the asymptotic variance of an MLE estimator is the inverse of something called the **Fisher Information Matrix**:

$$
\begin{aligned}
I_n(\theta) &= \mathbb{V}\left[\frac{\partial}{\partial\theta}\ln(f(\vec{x};\theta))\right] \\
&= \mathbb{V}\left[s(\vec{x};\theta)\right]
\end{aligned}
$$

- $s(\cdot)$ is called the **score** operator—how sensitive your log-likelihood is to θ
  - Note that we set this equal (close) to 0 in MLE.
  - Hence $s(\cdot)$ is a random variable of mean 0
- If the sample is i.i.d., this can be simplified to a multiple of a single information matrix:

$$
I_n(\theta) = n\mathbb{V}\left[\frac{\partial}{\partial\theta}\ln(f(x_1;\theta))\right] = nI_1(\theta)
$$

The Cramer-Rao Inequality

Theorem: Cramer-Rao

If $T(X_1, ..., X_n)$ is an unbiased estimator for θ, then

$$\mathbb{V}(T) \geq \frac{1}{I_n(\theta)}$$

Theorem: Cramer-Rao

If $T(X_1, ..., X_n)$ is an unbiased estimator for θ, then

$$\mathbb{V}(T) \geq \frac{1}{I_n(\theta)}$$

- Won't cover the proof here, but it's in the text
- This makes $I_n(\theta)^{-1}$ the lower bound for an estimator's variance
- An estimator is efficient if its variance achieves this bound

Efficiency of MLE

As shown before, MLE estimators are asymptotically normal with distribution $\mathcal{N}(0, I_n(\theta)^{-1})$

- Hence, the MLE is **asymptotically efficient!**
- We may prove this if we have time/energy.

QUESTIONS?