Statistics Review II

EC 320: Introduction to Econometrics

Winter 2022

Prologue

Housekeeping

Problem Set 1 available on Canvas.

Course GitHub page.

Statistics Review

Overview

Goal: Learn about a population.

- In particular, learn about an unknown population *parameter*.

Challenge: Usually cannot access information about the entire population.

Solution: Sample from the population and estimate the parameter.

- Draw n observations from the population, then use an estimator.

Sampling

There are myriad ways to produce a sample,* but we will restrict our attention to **simple random sampling**, where

1. Each observation is a random variable.

2. The n random variables are independent.

3. Life becomes much simpler for the econometrician.

* Only a subset of these can help produce reliable statistics.

Estimators

An **estimator** is a rule (or formula) for estimating an unknown population parameter given a sample of data.

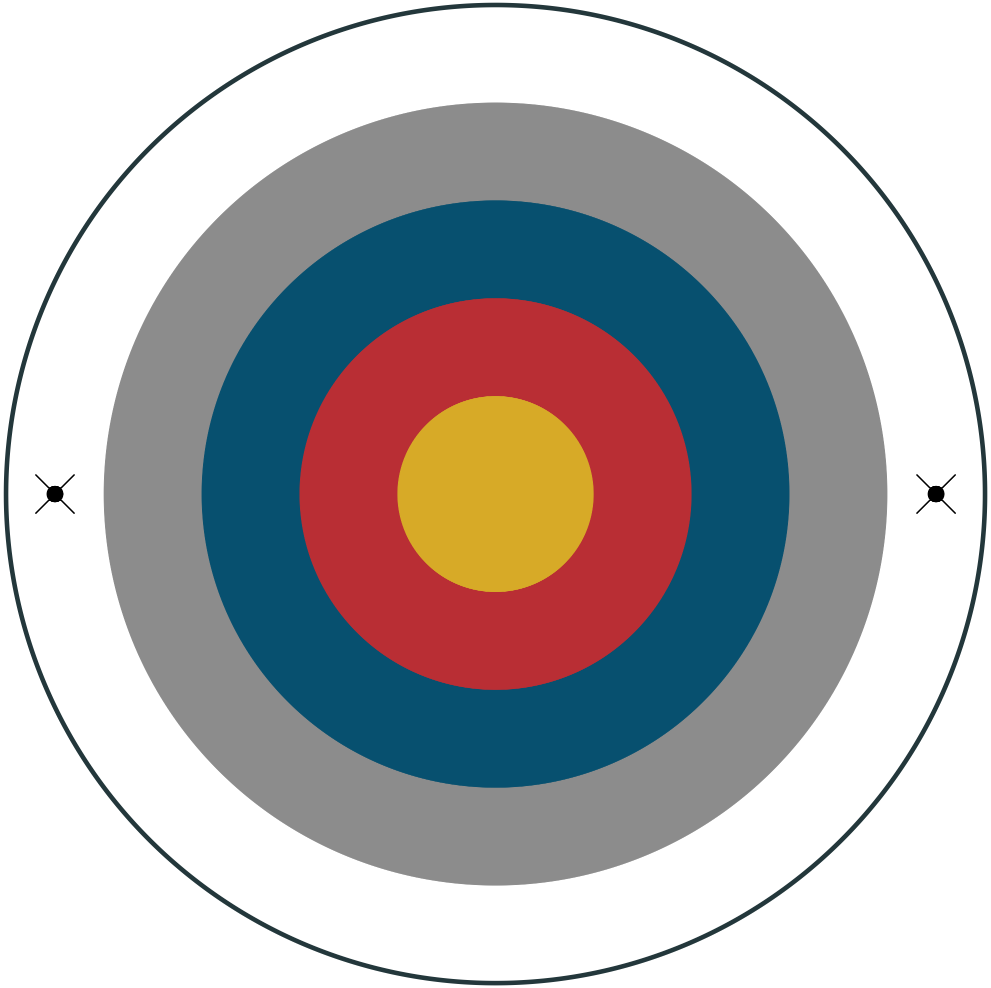
- Each observation in the sample is a random variable.
- An estimator is a combination of random variables $\implies$ it is a random variable.

Example: Sample mean

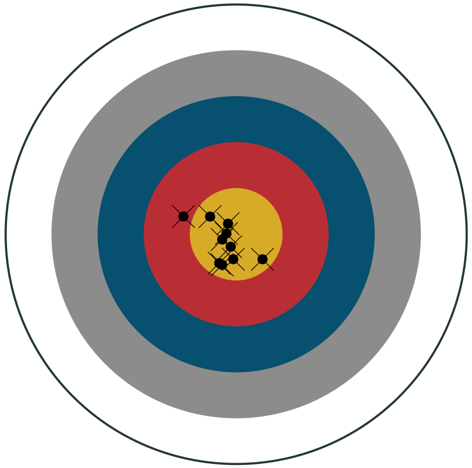
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- $\bar{X}$ is an estimator for the population mean μ.
- Given a sample, $\bar{X}$ yields an **estimate** $\bar{x}$ or $\hat{\mu}$, a specific number.

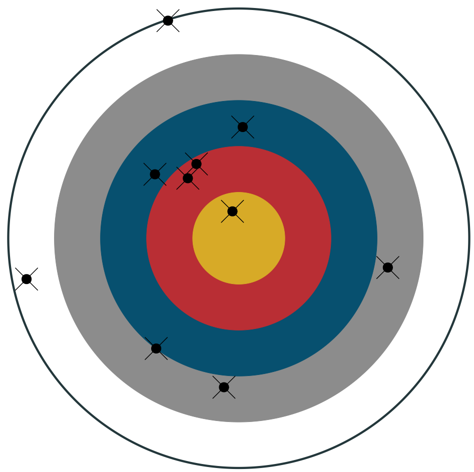
A physicist, a chemist, and an econometrician go to an archery range...



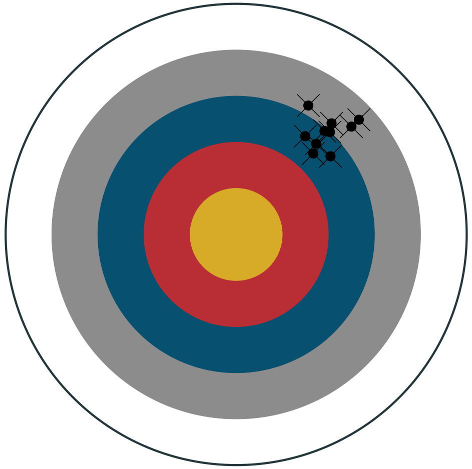
Archer 1



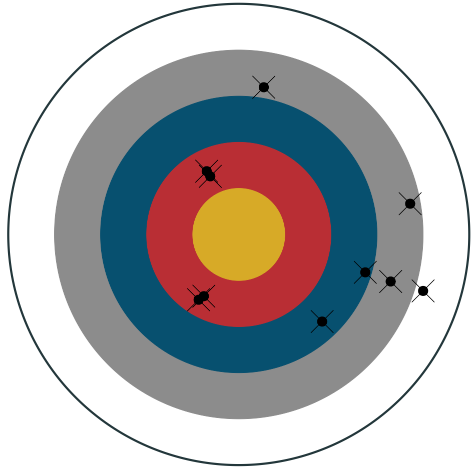
Archer 2



Archer 3

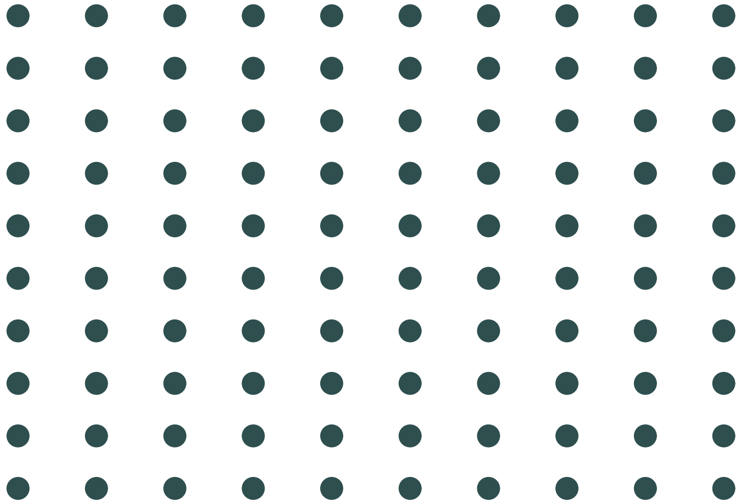


Archer 4

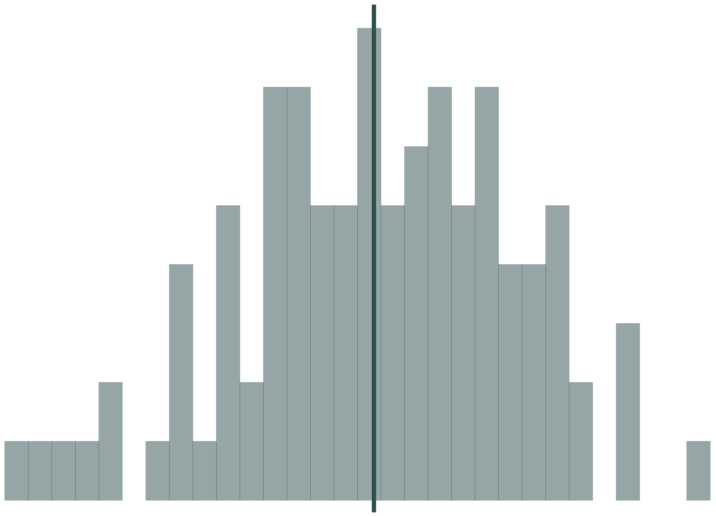


Population *vs.* Sample

Question: Why do we care about *population vs. sample*?



Population

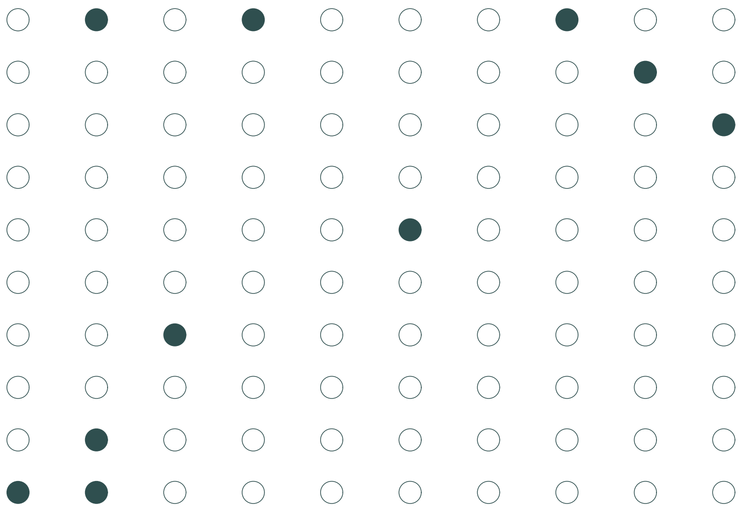


Population relationship

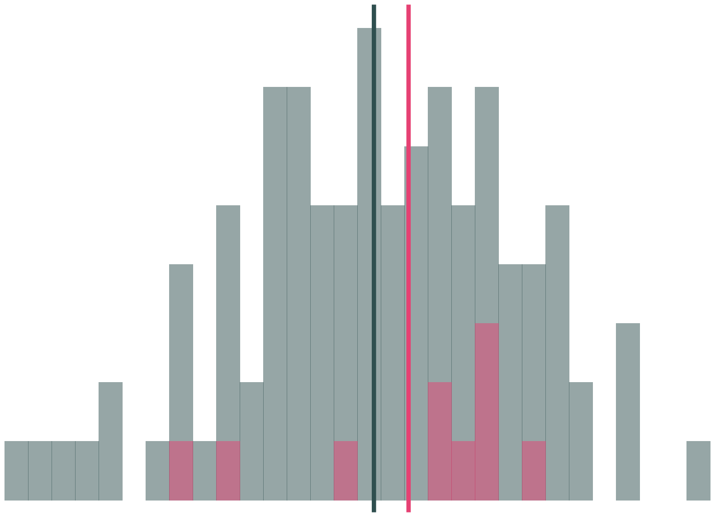
$$\mu = 3.75$$

Population *vs.* Sample

Question: Why do we care about *population vs. sample*?



Sample 1: 10 random individuals



Population relationship

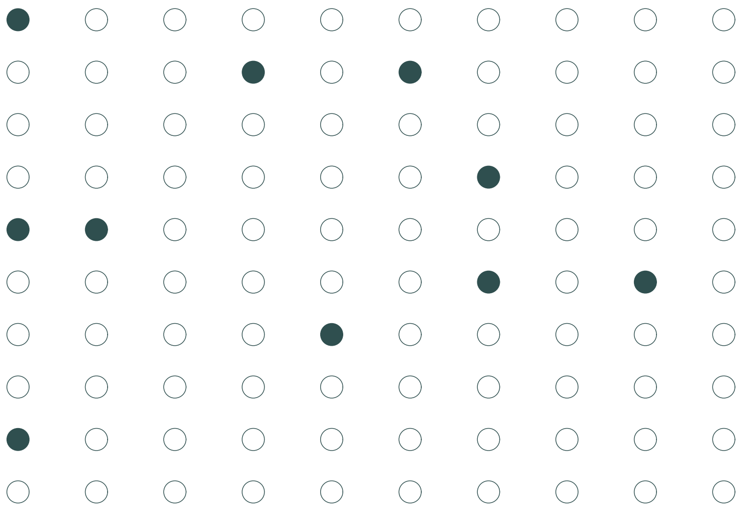
$$\mu = 3.75$$

Sample relationship

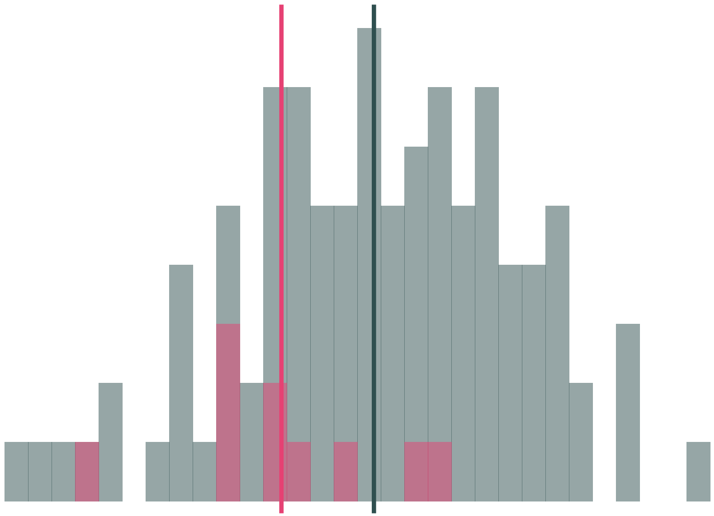
$$\hat{\mu} = 8.34$$

Population *vs.* Sample

Question: Why do we care about *population vs. sample*?



Sample 2: 10 random individuals



Population relationship

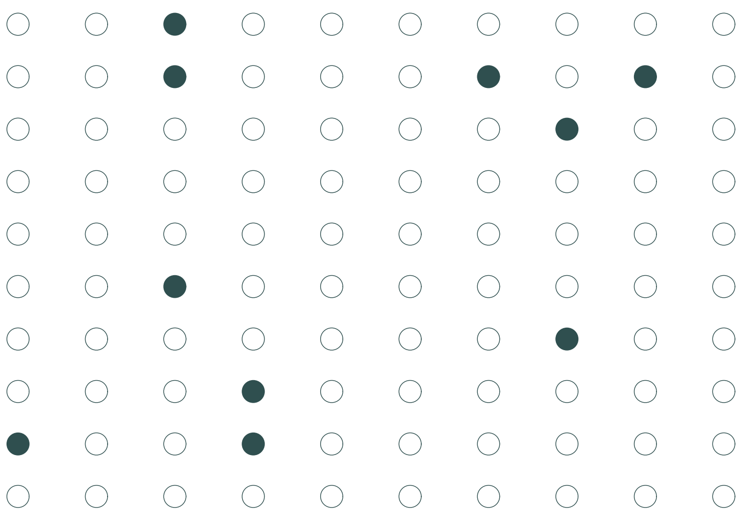
$$\mu = 3.75$$

Sample relationship

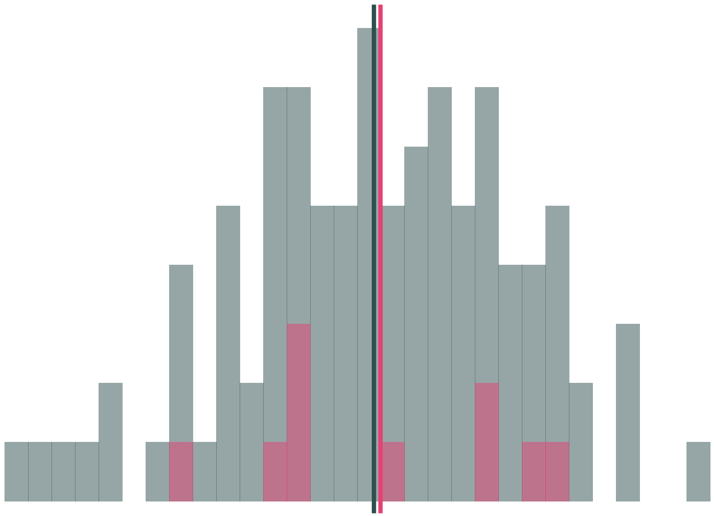
$$\hat{\mu} = -8.54$$

Population *vs.* Sample

Question: Why do we care about *population vs. sample*?



Sample 3: 10 random individuals



Population relationship

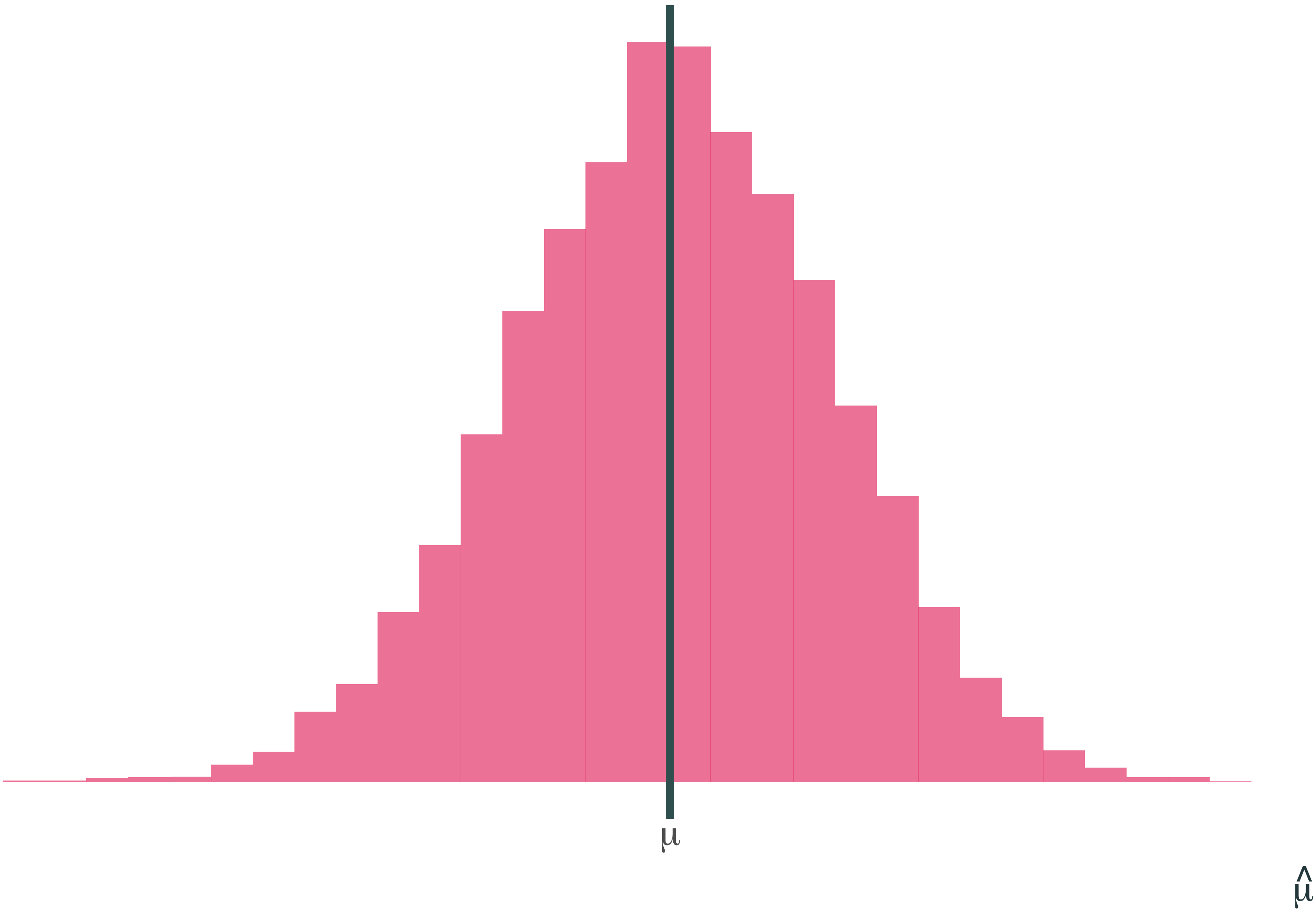
$$\mu = 3.75$$

Sample relationship

$$\hat{\mu} = 4.62$$

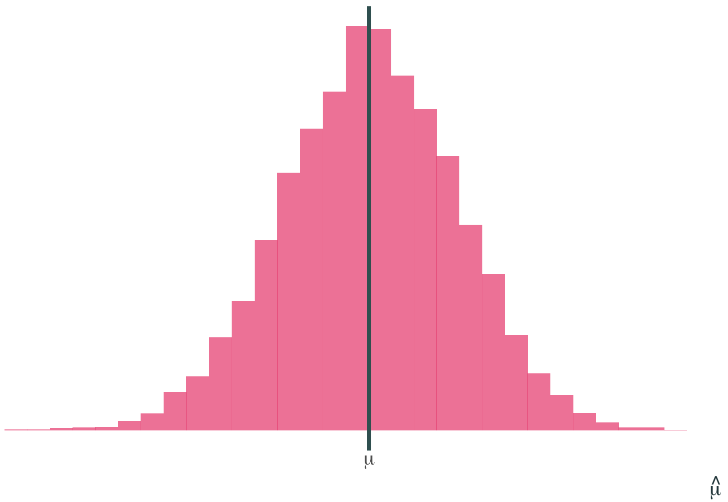
Let's repeat this **10,000 times** and then plot the estimates.

(This exercise is called a Monte Carlo simulation.)


μ
μ̂

Population *vs.* Sample

Question: Why do we care about *population vs. sample*?



- On average, the mean of the samples are close to the population mean.
- But…some individual samples can miss the mark.
- The difference between individual samples and the population creates **uncertainty**.

Population *vs.* Sample

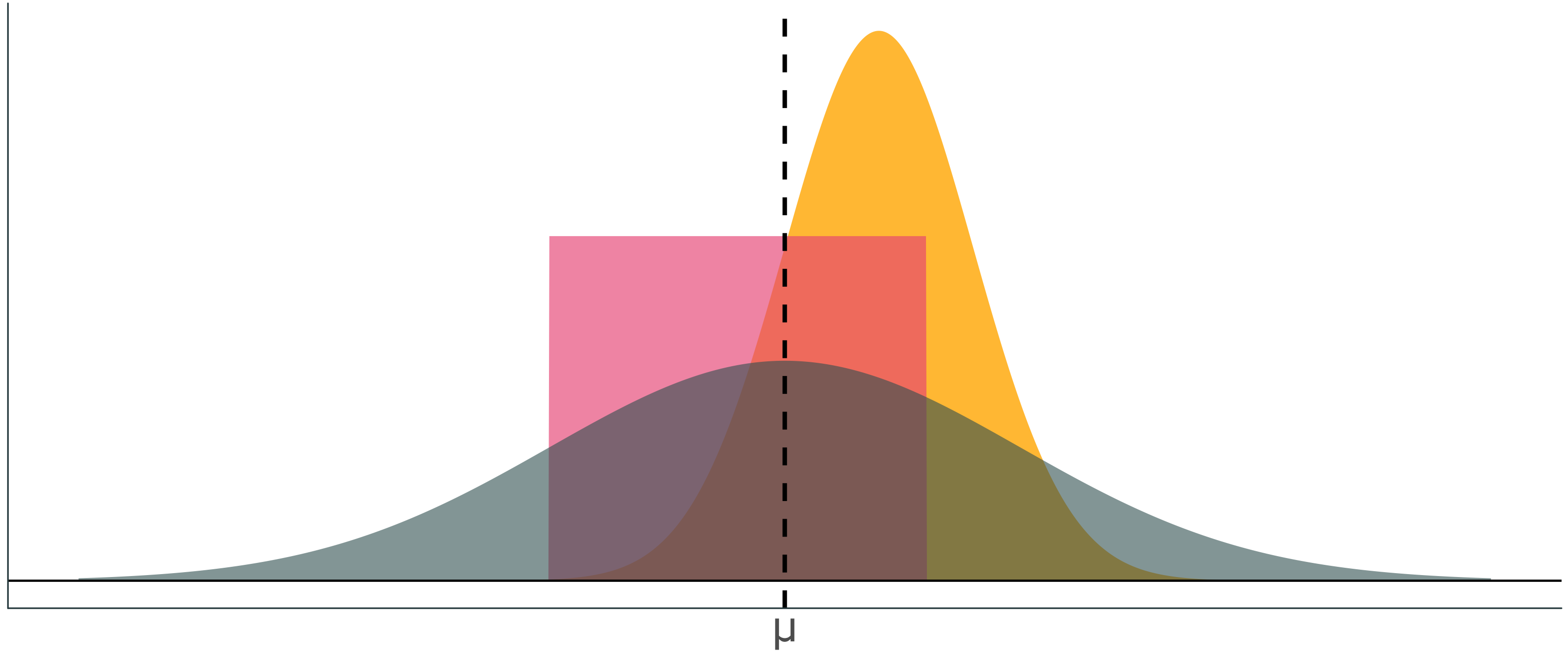
Question: Why do we care about *population vs. sample*?

Answer: Uncertainty matters.

- $\hat{\mu}$ is a random variable that depends on the sample.
- In practice, we don't know whether our sample is similar to the population or not.
- Individual samples may have means that differ greatly from the population.
- We will have to keep track of this uncertainty.

Properties of Estimators

Imagine that we want to estimate an unknown parameter μ, and we know the distributions of three competing estimators. **Which one should we use?**



Properties of Estimators

Question: What properties make an estimator reliable?

Answer 1: Unbiasedness.

On average (after *many* samples), does the estimator tend toward the correct value?

More formally: Does the mean of estimator's distribution equal the parameter it estimates?

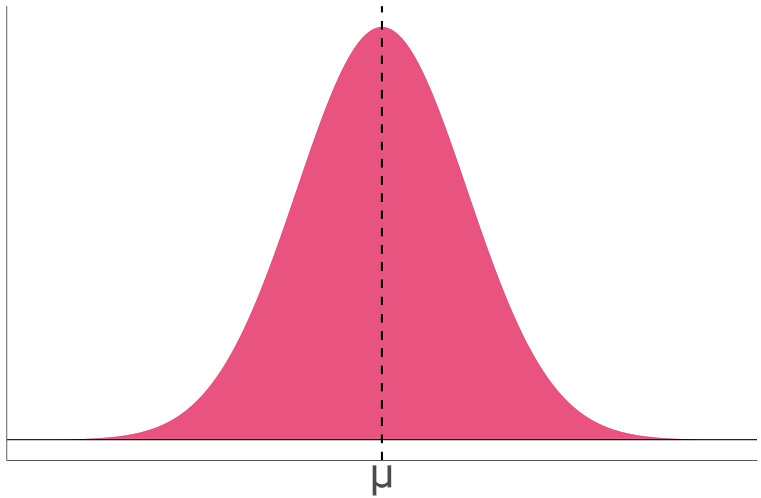
$$\mathrm{Bias}_\mu(\hat{\mu}) = \mathbb{E}[\hat{\mu}] - \mu$$

Properties of Estimators

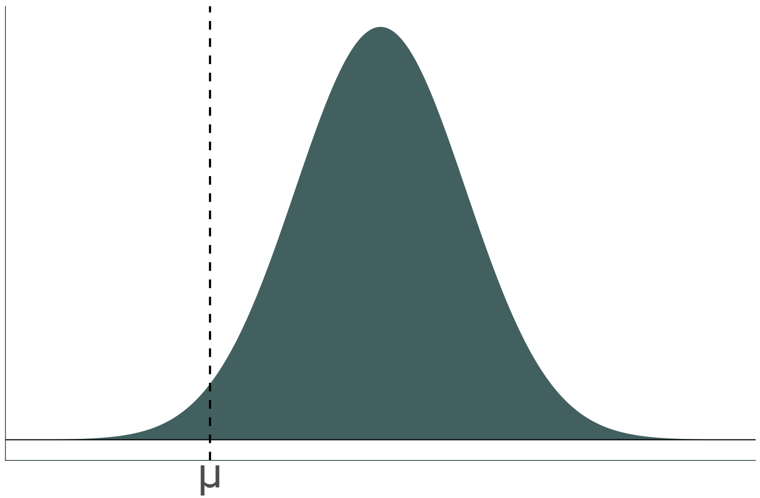
Question: What properties make an estimator reliable?

Answer 1: Unbiasedness.

Unbiased estimator: $\mathbb{E}[\hat{\mu}] = \mu$



Biased estimator: $\mathbb{E}[\hat{\mu}] \neq \mu$



Properties of Estimators

Question: What properties make an estimator reliable?

Answer 2: Low Variance (a.k.a. Efficiency).

The central tendencies (means) of competing distributions are not the only things that matter. We also care about the **variance** of an estimator.

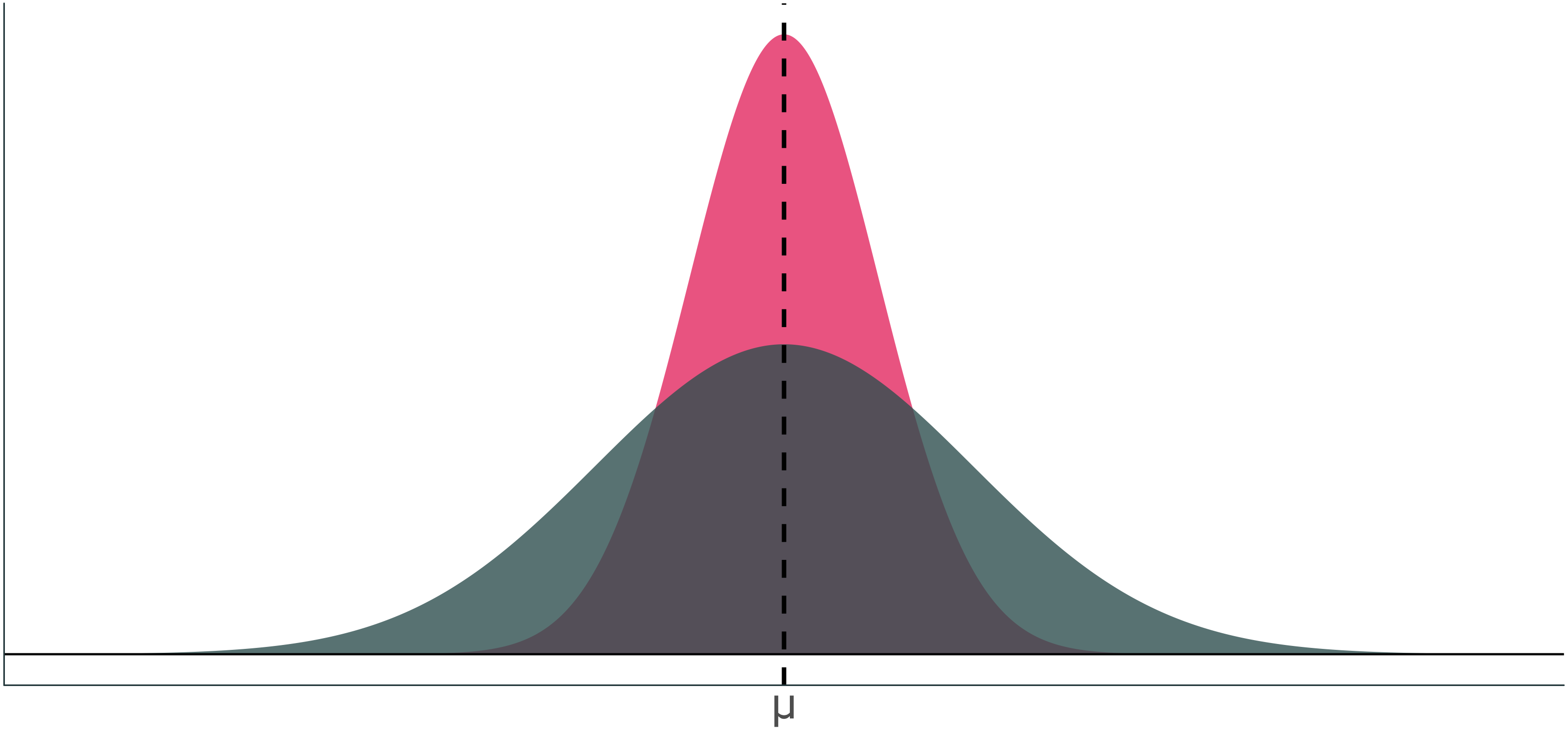
$$\mathbf{Var}(\hat{\mu}) = \mathbb{E}\Big[(\hat{\mu} - \mathbb{E}[\hat{\mu}])^2\Big]$$

Lower variance estimators produce estimates closer to the mean in each sample.

Properties of Estimators

Question: What properties make an estimator reliable?

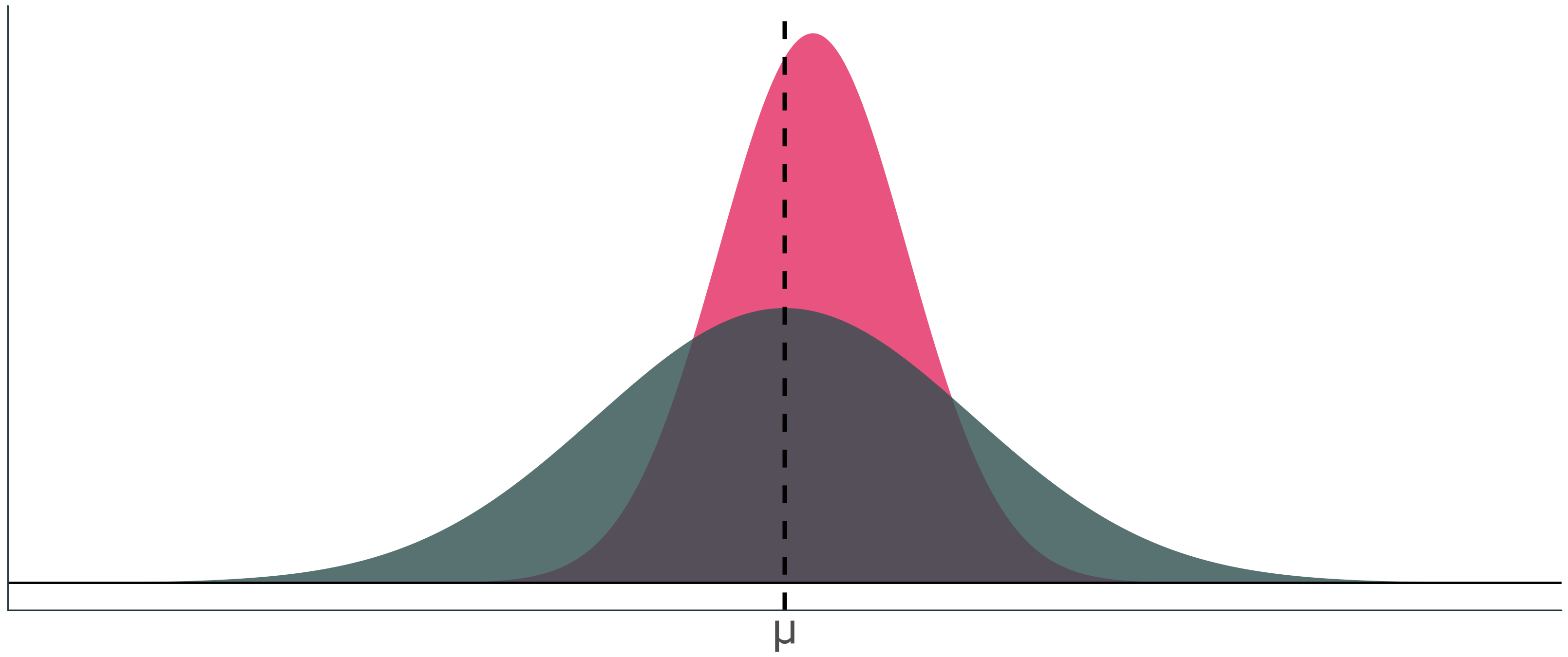
Answer 2: Low Variance (a.k.a. Efficiency).



The Bias-Variance Tradeoff

Should we be willing to take a bit of bias to reduce the variance?

In econometrics, we generally prefer unbiased estimators. Some other disciplines think more about this tradeoff.



Unbiased Estimators

In addition to the sample mean, there are several other unbiased estimators we will use often.

- **Sample variance** to estimate variance σ^2.
- **Sample covariance** to estimate covariance σ_{XY}.
- **Sample correlation** to estimate the population correlation coefficient ρ_{XY}.

Unbiased Estimators

The sample variance S_X^2 is an unbiased estimator of the population variance σ^2:

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

Unbiased Estimators

The sample covariance S_{XY} is an unbiased estimator of the population covariance σ_{XY}:

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}).$$

Unbiased Estimators

The sample correlation r_{XY} is an unbiased estimator of the population correlation coefficient ρ_{XY}:

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_X^2}\sqrt{S_Y^2}}.$$

Hypothesis Testing

Given What do we make of an estimate of the population mean?

- Is it meaningfully different than existing evidence on the population mean?
- Is is *statistically distinguishable* from previously hypothesized values of the population mean?
- Is the estimate extreme enough to update our prior beliefs about the population mean?

We can conduct statistical tests to address these questions.

Hypothesis Testing

Null hypothesis (H_0): $\mu = \mu_0$

Alternative hypothesis (H_1): $\mu \neq \mu_0$

There are four possible outcomes of our test:

1. We **fail to reject** the null hypothesis and the null is true.
2. We **reject** the null hypothesis and the null is false.
3. We **reject** the null hypothesis, but the null is actually true (**Type I error**).
4. We **fail to reject** the null hypothesis, but the null is actually false (**Type II error**).

Hypothesis Testing

We **fail to reject** the null hypothesis and the null is true.

- The defendant was acquitted and he didn't do the crime.

We **reject** the null hypothesis and the null is false.

- The defendant was convicted and he did the crime.

Hypothesis Testing

We **reject** the null hypothesis, but the null is actually true.

- The defendant was convicted, but he didn't do the crime!
- **Type I error** (a.k.a. *false positive*)

We **fail to reject** the null hypothesis, but the null is actually false.

- The defendant was acquitted, but he did the crime!
- **Type II error** (a.k.a. *false negative*)

Hypothesis Testing

$\hat{\mu}$ is random: it could be anything, even if $\mu = \mu_0$ is true.

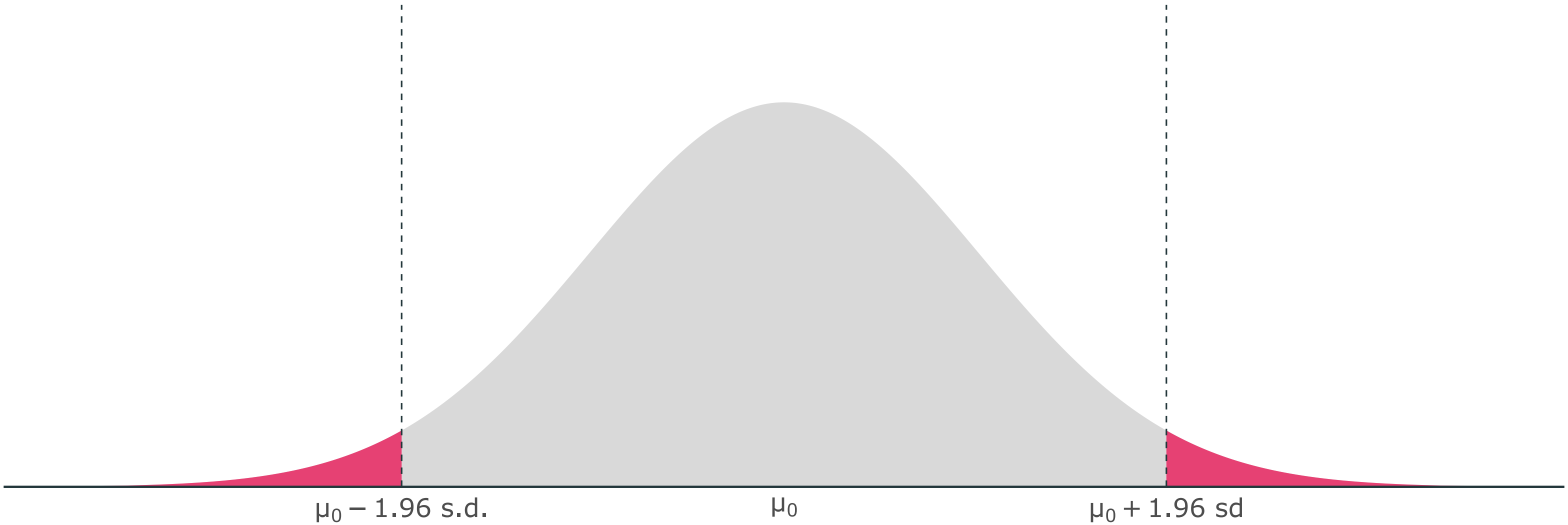
- But if $\mu = 0$ is true, then $\hat{\mu}$ is unlikely to take values far from zero.
- As the variance of $\hat{\mu}$ shrinks, we are even less likely to observe "extreme" values of $\hat{\mu}$ (assuming $\mu = \mu_0$).

Our test should take extreme values of $\hat{\mu}$ as evidence against the null hypothesis, but it should also weight them by what we know about the variance of $\hat{\mu}$.

- For now, we'll assume that the variable of interest X is normally distributed with mean μ and standard deviation σ^2.

Hypothesis Testing

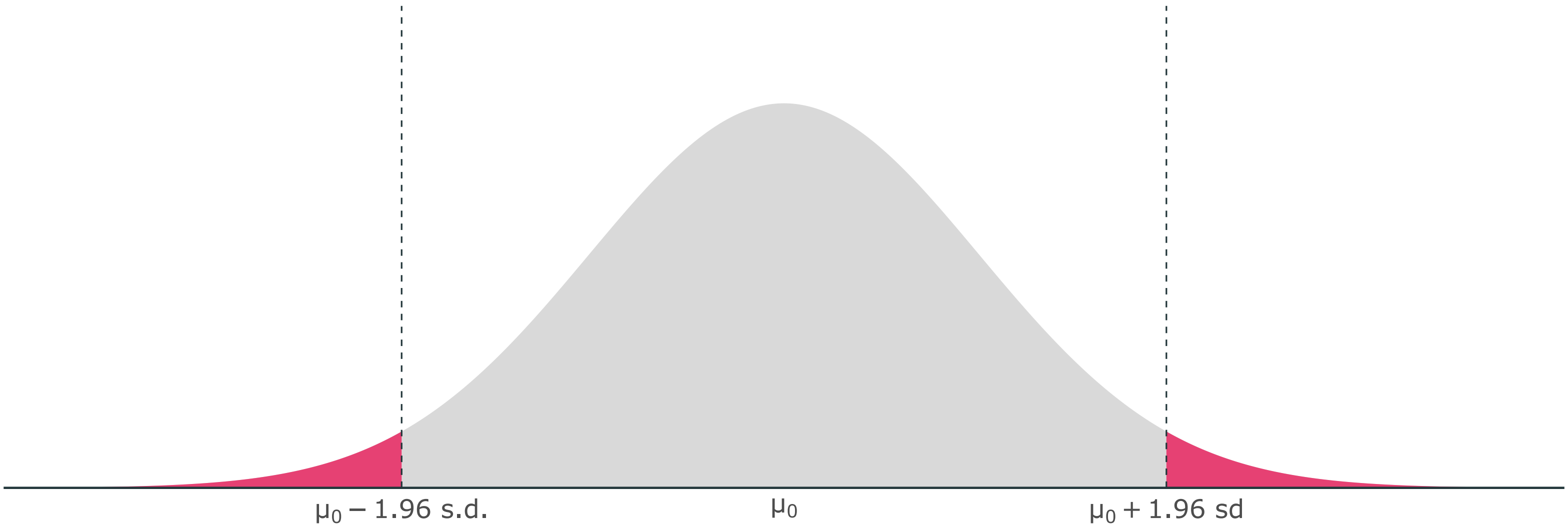
Reject H_0 if $\hat{\mu}$ lies in the **rejection region**.



- The area of the rejection region is defined by the **significance level** of the test.
- In a 5% test, the area is 0.05.
- Significance level = tolerance for Type I error.

Hypothesis Testing

Reject H_0 if $|z| = \left| \frac{\hat{\mu} - \mu_0}{\text{sd}(\hat{\mu})} \right| > 1.96$.



What happens to z as $|\hat{\mu} - \mu_0|$ increases?

What happens to z as $\text{sd}(\hat{\mu})$ increases?

Hypothesis Testing

The formula for the z statistic assumes that we know $\mathrm{sd}(\hat{\mu})$.

- In practice, we don't know $\mathrm{sd}(\hat{\mu})$, so we have to estimate it.

If the variance of X is σ^2, then

$$\sigma_{\hat{\mu}}^2 = \frac{\sigma^2}{n}.$$

- We can estimate σ^2 with the sample variance S_X^2.

The sample variance of the sample mean is

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

Hypothesis Testing

The **standard error** of $\hat{\mu}$ is the square root of $S^2_{\hat{\mu}}$:

$$\mathrm{SE}(\hat{\mu}) = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (X_i - \bar{X})^2}.$$

- Standard error = sample standard deviation of an estimator.

When we use $\mathrm{SE}(\hat{\mu})$ in place of $\mathbf{sd}(\hat{\mu})$, the z statistic becomes a t statistic:

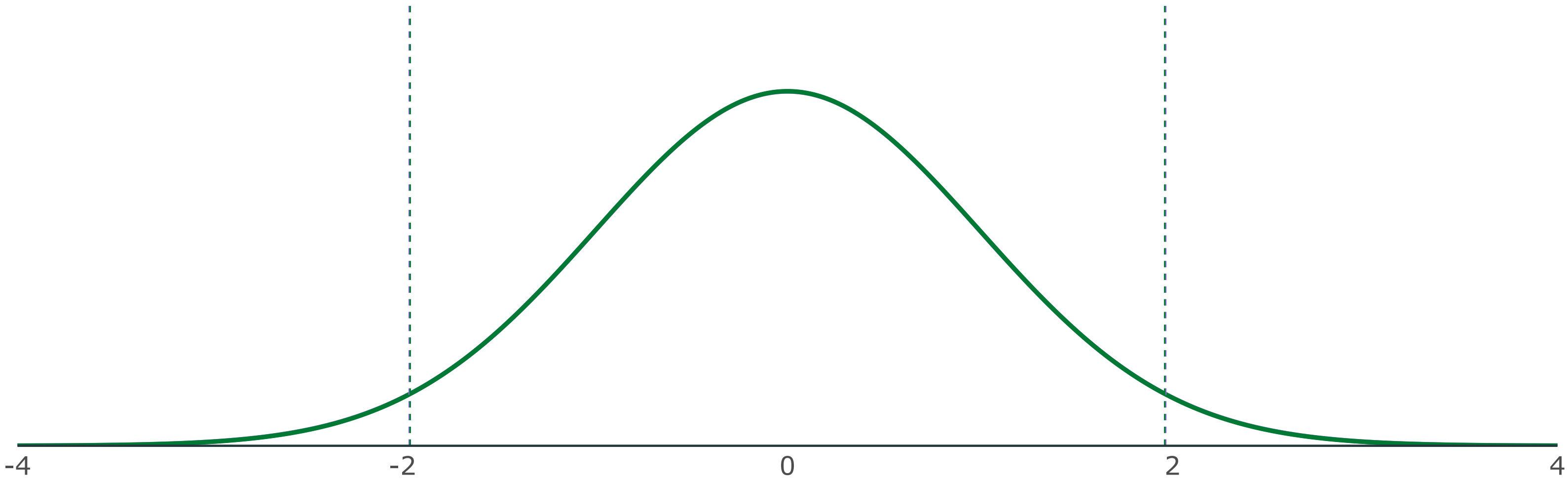
$$t = \frac{\hat{\mu} - \mu_0}{\mathrm{SE}(\hat{\mu})}.$$

- Unlike the standard deviation of $\hat{\mu}$, $\mathrm{SE}(\hat{\mu})$ varies from sample to sample.
- **Consequence:** t statistics do not necessarily have a normal distribution.

Hypothesis Testing

Normal distribution vs. t distribution

- A normal distribution has the same shape for any sample size.
- The shape of the t distribution depends the **degrees of freedom**.



- Degrees of freedom = 500.

Hypothesis Testing

t Tests (two-sided)

To conduct a t test, compare the t statistic to the appropriate **critical value** of the t distribution.

- To find the critical value in a t table, we need the degrees of freedom and the significance level α.

Reject H_0 at the $\alpha \cdot 100$-percent level if

$$|t| = \left| \frac{\hat{\mu} - \mu_0}{\mathrm{SE}(\hat{\mu})} \right| > t_{\text{crit}}.$$

On Your Own

As the term progresses, we will encounter additional flavors of hypothesis testing and other related concepts.

You may find it helpful to review the following topics from Math 243:

- Confidence intervals
- One-sided t tests
- p values

Data and the `tidyverse`

Experimental data

Data generated in controlled, laboratory settings.

Ideal for **causal identification**, but difficult to obtain in the social sciences.

- Intractable logistical problems
- Too expensive
- Morally repugnant

Experiments outside the lab: **randomized control trials** and **A/B testing**.

Data

Observational data

Data generated in non-experimental settings.

- Surveys
- Censuses
- Administrative records
- Environmental data
- Financial and sales transactions
- Social media

Mainstay of economic research, but **poses challenges** to causal identification.

Tidy Data

Search:

	State	Population	Murders
1	Alabama	4779736	135
2	Alaska	710231	19
3	Arizona	6392017	232
4	Arkansas	2915918	93
5	California	37253956	1257
6	Colorado	5029196	65

Showing 1 to 6 of 51 entries

Previous Next

Rows represent **observations**.

Columns represent **variables**.

Each **value** is associated with an **observation** and a **variable**.

Cross Sectional Data

Sample of individuals from a population at a point in time.

Ideally, collected using **random sampling**.

- Random sampling + sufficient sample size = representative sample.
- Random sampling simplifies data analysis, but non-random samples are common (and difficult to work with).

Used extensively in applied microeconomics.*

Main focus of this course.

* Applied microeconomics = Labor, health, education, public finance, development, industrial organization, and urban economics.

Cross Sectional Data

Sample of US workers (Current Population Survey, 1976)

	Wage	Education	Tenure	Female?	Non-white?
1	3.1	11	0	1	0
2	3.24	12	2	1	0
3	3	11	0	0	0
4	6	8	28	0	0
5	5.3	12	2	0	0
6	8.75	16	8	0	0

Showing 1 to 6 of 526 entries

Previous 1 2 3 4 5 … 88 Next

Time Series Data

Observations of variables over time.

- Quarterly US GDP
- Annual US infant mortality rates
- Daily Amazon stock prices

Complication: Observations are not independent draws.

- GDP this quarter highly related to GDP last quarter.

Used extensively in empirical macroeconomics.

Requires more-advanced methods (EC 421 and EC 422).

Time Series Data

Number of US manufacturing strikes per month (Jan. 1968 to Dec. 1976)

	Period	Strikes	Output
1	1	5	0.01517
2	2	4	0.00997
3	3	6	0.0117
4	4	16	0.00473
5	5	5	0.01277
6	6	8	0.01138

Showing 1 to 6 of 108 entries

Previous 1 2 3 4 5 … 18 Next

Pooled Cross Sectional Data

Cross sections from different points in time.

Useful for studying policy changes and relationship that change over time.

Requires more-advanced methods (EC 421 and many 400-level applied micro classes).

Pooled Cross Sectional Data

Sample of US women (General Social Survey, 1972 to 1984)

	Year	Education	Age	Children	Black?
1	72	12	48	4	0
2	72	17	46	3	0
3	72	12	53	2	0
4	72	12	42	2	0
5	72	12	51	2	0
6	72	8	50	4	0

Showing 1 to 6 of 1,129 entries

Previous 1 2 3 4 5 … 189 Next

Panel or Longitudinal Data

Time series for each cross-sectional unit.

- Example: daily attendance data for a sample of students.

Difficult to collect, but useful for causal identification.

- Can control for *unobserved* characteristics.

Requires more-advanced methods (EC 421 and many 400-level applied micro classes).

Panel or Longitudinal Data

Panel of US workers (National Longitudinal Survey of Youth, 1980 to 1987)

	ID	Year	Experience	log(Wage)	Union
1	13	1980	1	1.2	no
2	13	1981	2	1.85	yes
3	13	1982	3	1.34	no
4	13	1983	4	1.43	no
5	13	1984	5	1.57	no
6	13	1985	6	1.7	no

Showing 1 to 6 of 4,360 entries

Previous 1 2 3 4 5 ... 727 Next

Tidy Data?

	worker_id	year	variable	value
1	13	1980	educ	14
2	13	1981	educ	14
3	13	1982	educ	14
4	13	1983	educ	14
5	13	1984	educ	14
6	13	1985	educ	14

Showing 1 to 6 of 21,800 entries

Previous 1 2 3 4 5 … 3,634 Next

Messy Data

Analysis-ready datasets are rare. Most data are "messy."

The focus of this class is data analysis, but **data wrangling** is a non-trivial part of a data scientist/analyst's job.

R has a suite of packages that facilitate data wrangling.

- `readr`, `tidyr`, `dplyr`, `ggplot2` + others.
- Known collectively as the `tidyverse`.

The `tidyverse`: A package of packages

`readr`: Functions to import data.

`tidyr`: Functions to reshape messy data.

`dplyr`: Functions to work with data.

`ggplot2`: Functions to visualize data.

Workflow

Step 1: Load packages with `pacman`

```r
library(pacman)
p_load(tidyverse)
```

If the `tidyverse` hasn't already been installed, `p_load` will install it.

Loading the `tidyverse` automatically loads `readr`, `tidyr`, `dplyr`, `ggplot2`, and a few other packages.

Workflow

Step 2: Import data with `readr`

```
workers <- read_csv("03-example_data.csv")
```

CSV files are a common non-proprietary format for storing tabular data.

The `read_csv` function imports CSV (comma-separated values) files.

- Converts the CSV file to a `tibble`, the `tidyverse` version of a `data.frame`.

Workflow

Step 3: Reshape data with `tidyr`

Variables are stored in rows instead of columns:

```
#> # A tibble: 21,800 × 4
#>    worker_id  year variable value
#>        <dbl> <dbl> <chr>    <dbl>
#>  1        13  1980 educ        14
#>  2        13  1981 educ        14
#>  3        13  1982 educ        14
#>  4        13  1983 educ        14
#>  5        13  1984 educ        14
#>  6        13  1985 educ        14
#>  7        13  1986 educ        14
#>  8        13  1987 educ        14
#>  9        17  1980 educ        13
#> 10        17  1981 educ        13
#> # … with 21,790 more rows
```

Step 3: Reshape data with `tidyr`

Make the data tidy by using the `spread` function:

```
workers <- workers %>%
  spread(key = variable, value = value)
```

Note the use of the **pipe operator**.

- `%>%` = *"and then."*
- Chains multiple commands together without having to define intermediate objects.

Workflow

Step 3: Reshape data with `tidyr`

The result:

```
#> # A tibble: 4,360 × 7
#>    worker_id  year black earnings  educ exper union
#>        <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl> <dbl>
#>  1        13  1980     0    8850.    14     1     0
#>  2        13  1981     0   14800.    14     2     1
#>  3        13  1982     0   11278.    14     3     0
#>  4        13  1983     0   12409.    14     4     0
#>  5        13  1984     0   14734.    14     5     0
#>  6        13  1985     0   15676.    14     6     0
#>  7        13  1986     0    1457.    14     7     0
#>  8        13  1987     0   14013.    14     8     0
#>  9        17  1980     0   13274.    13     4     0
#> 10        17  1981     0   12800.    13     5     0
#> # … with 4,350 more rows
```

Workflow

Step 4: Manipulate data with `dplyr`

Generate new variables with `mutate`:

```
workers <- workers %>%
  mutate(union = ifelse(union == 1, "Yes", "No"))
```

Before, `union` was a binary variable equal to 1 if the worker is in a union or 0 if otherwise.

Now `union` is a character variable.

Workflow

Step 4: Manipulate data with `dplyr`

The result:

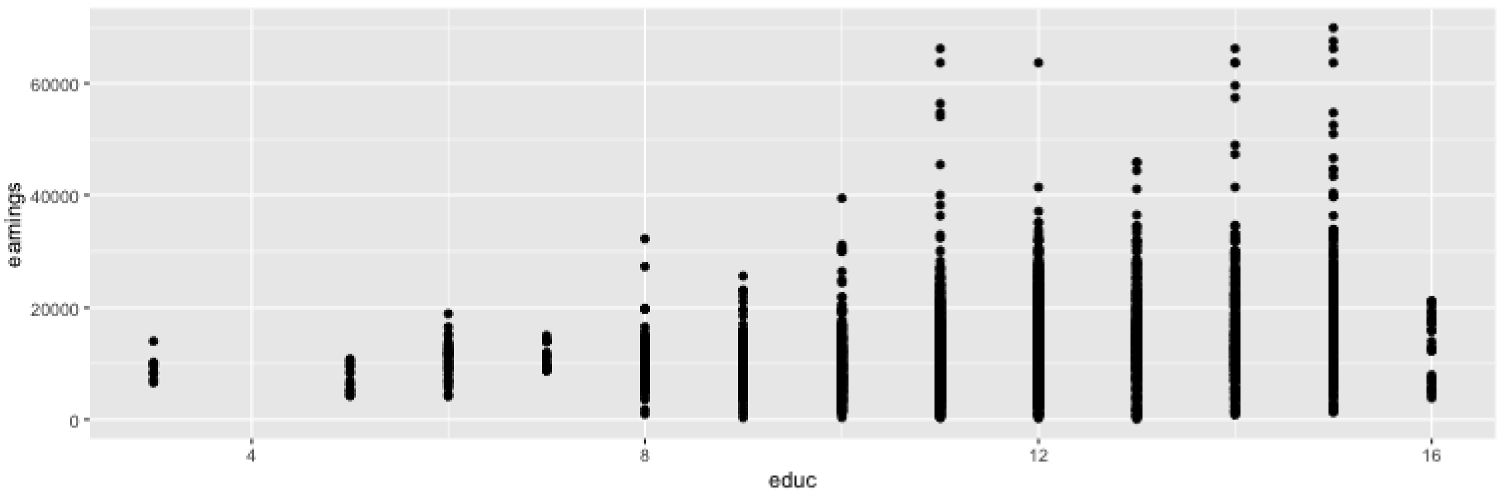
```
#> # A tibble: 4,360 × 7
#>    worker_id  year black earnings  educ exper union
#>        <dbl> <dbl> <dbl>    <dbl> <dbl> <dbl> <chr>
#>  1        13  1980     0    8850.    14     1 No
#>  2        13  1981     0   14800.    14     2 Yes
#>  3        13  1982     0   11278.    14     3 No
#>  4        13  1983     0   12409.    14     4 No
#>  5        13  1984     0   14734.    14     5 No
#>  6        13  1985     0   15676.    14     6 No
#>  7        13  1986     0    1457.    14     7 No
#>  8        13  1987     0   14013.    14     8 No
#>  9        17  1980     0   13274.    13     4 No
#> 10        17  1981     0   12800.    13     5 No
#> # … with 4,350 more rows
```

Workflow

Step 6: Visualize and analyze data with `ggplot2`

How are education and earnings correlated?

```
workers %>%
  ggplot(aes(x = educ, y = earnings)) +
  geom_point()
```



Workflow

Step 6: Visualize and analyze data with `ggplot2`

How are education and earnings correlated?

Can also use the `cor` function from `base` R:

```
cor(workers$educ, workers$earnings)
```

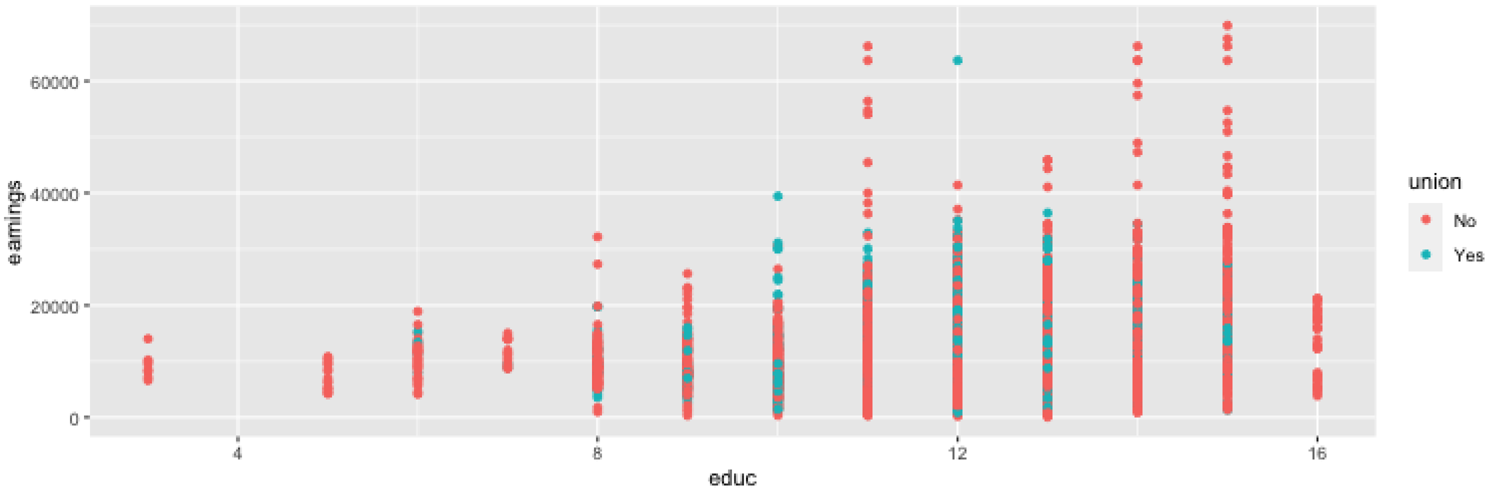
```
#> [1] 0.2685563
```

Workflow

Step 6: Visualize and analyze data with `ggplot2`

How are education and earnings correlated?

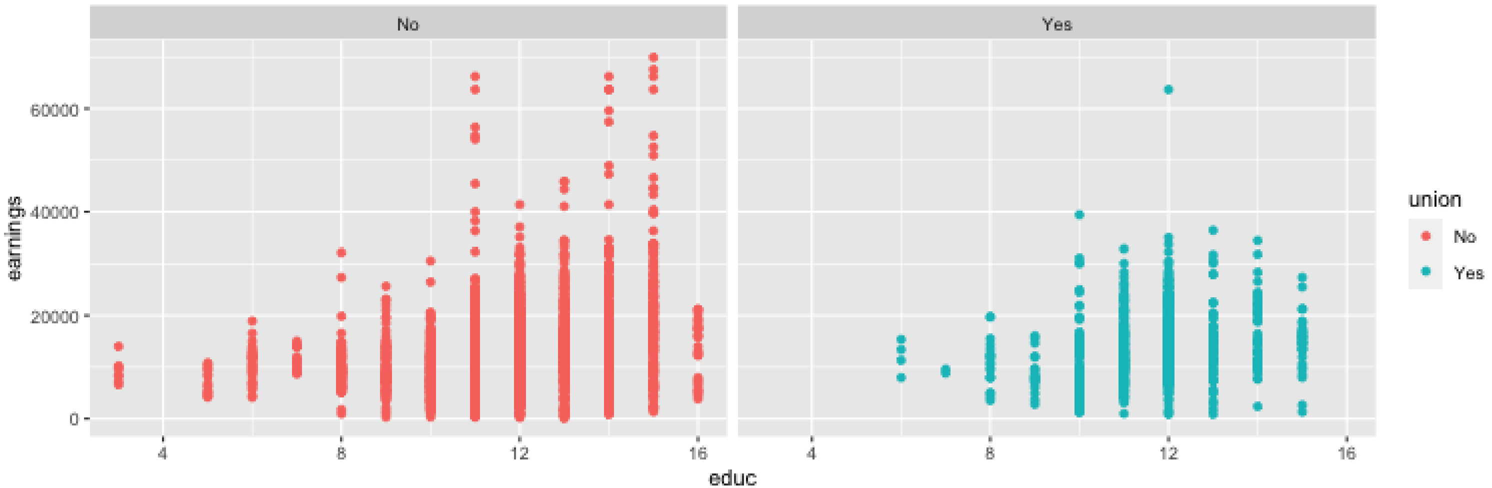
```
workers %>%
  ggplot(aes(x = educ, y = earnings, color = union)) +
  geom_point()
```



Step 6: Visualize and analyze data with `ggplot2`

How are education and earnings correlated?

```
workers %>%
  ggplot(aes(x = educ, y = earnings, color = union)) +
  geom_point() +
  facet_grid(~union)
```



Workflow

Step 6: Visualize and analyze data with `ggplot2`

How are education and earnings correlated?

Can **subset** the data to get group-specific correlations:

```
workers_union <- workers %>%
  filter(union == "Yes")

cor(workers_union$educ, workers_union$earnings)
```

```
#> [1] 0.211482
```

```
workers_nounion <- workers %>%
  filter(union == "No")

cor(workers_nounion$educ, workers_nounion$earnings)
```

```
#> [1] 0.2809786
```

Why Bother?

Q: Why not just use **`MS Excel`** for data wrangling?

A: Reproducibility

- Easier to retrace your steps with R.

A: Portability

- Easy to re-purpose R code for new projects.

A: Scalability

- `Excel` chokes on big datasets.

A: R Saves time (eventually)

- Lower marginal costs in exchange for higher fixed costs.

Further Reading

1. Tidy Data by Hadley Wickham (creator of the `tidyverse`)

2. Cheatsheets