Statistics Review II

ECON 4651: Principles of Econometrics for Business and Analytics

Jason Cook
Fall 2020

Prologue

Housekeeping

Problem Set 1 available on Canvas. Due 9/3 by 5pm.

- Message Blake if you want a group and need help finding one

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.

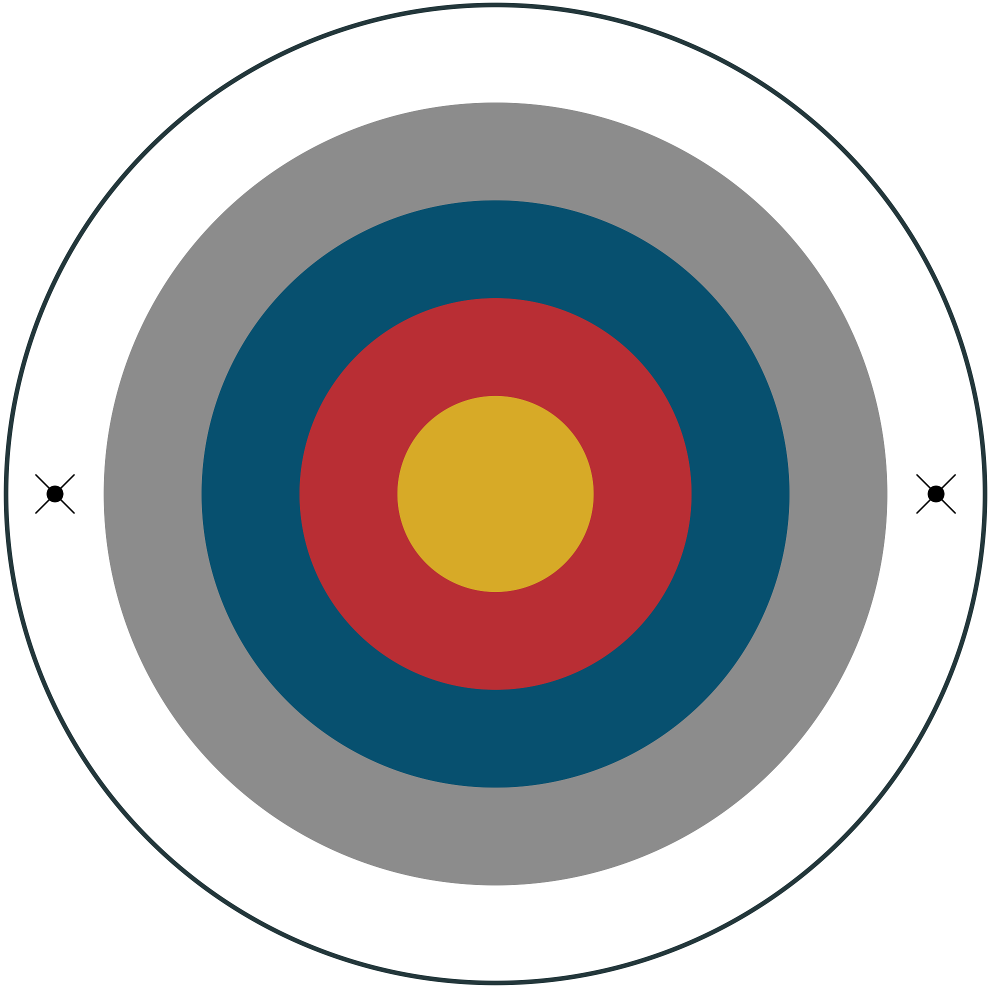
- Each observation in the sample is a random variable.
- An estimator is a combination of random variables $\implies$ it is a random variable.

Example: Sample mean

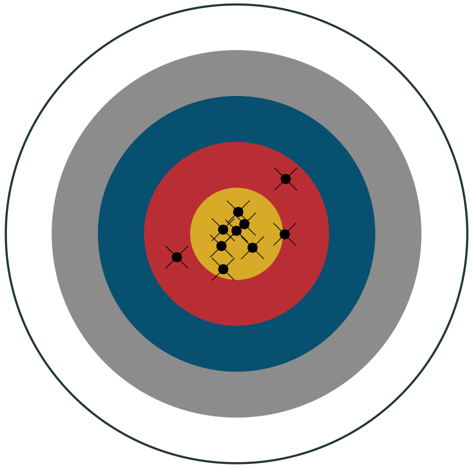
$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

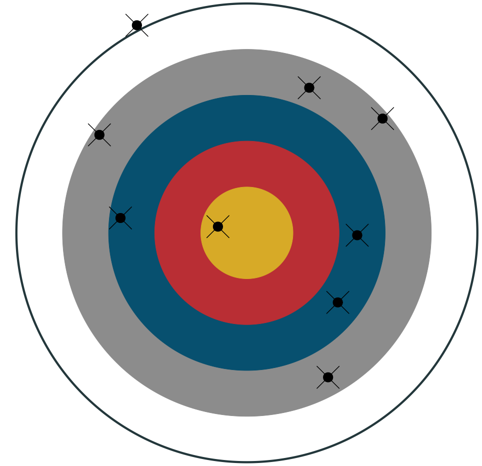
You can think of estimators as trying to hit a bulls-eye at an archery range...



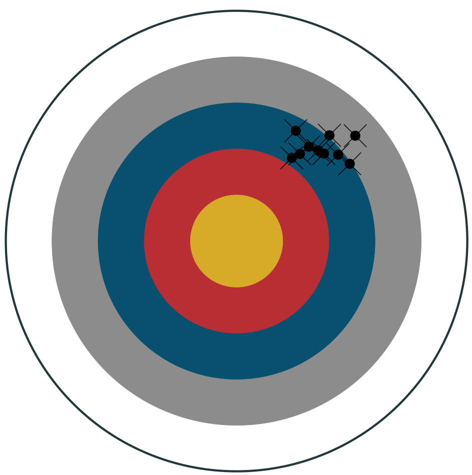
Archer 1



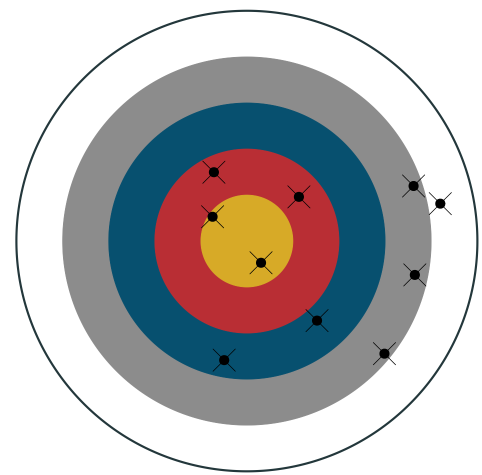
Archer 2



Archer 3

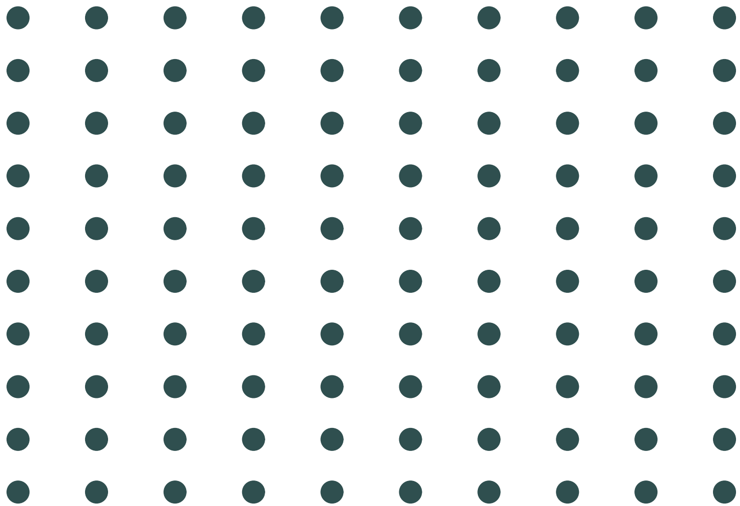


Archer 4

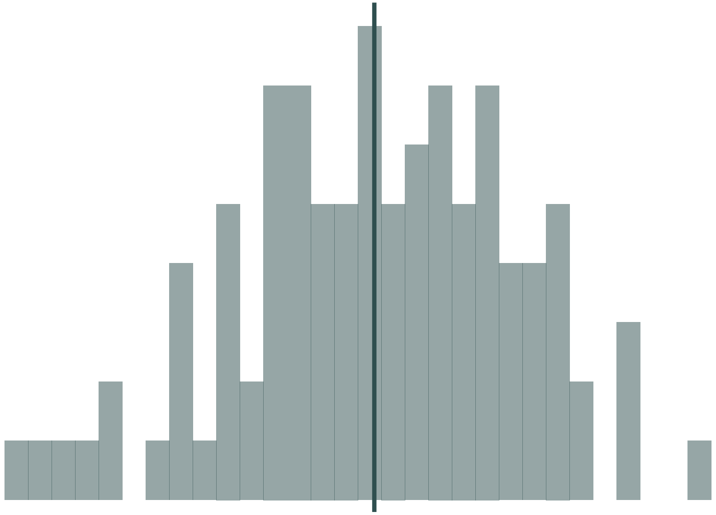


Population *vs.* Sample

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



Population

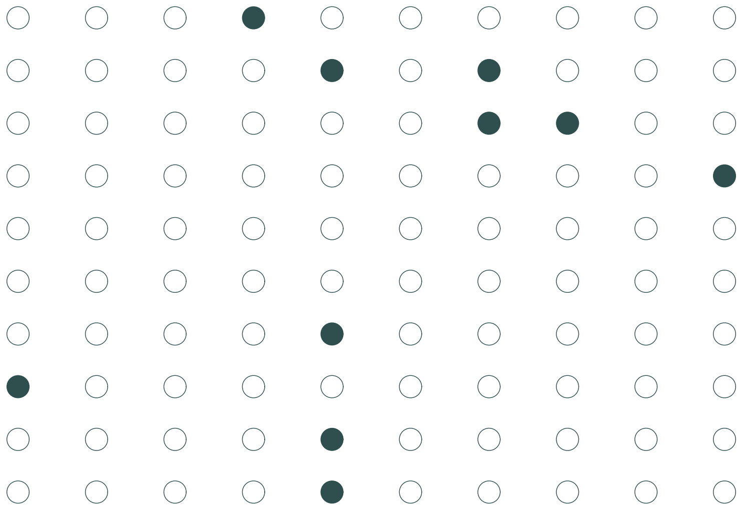


Population relationship

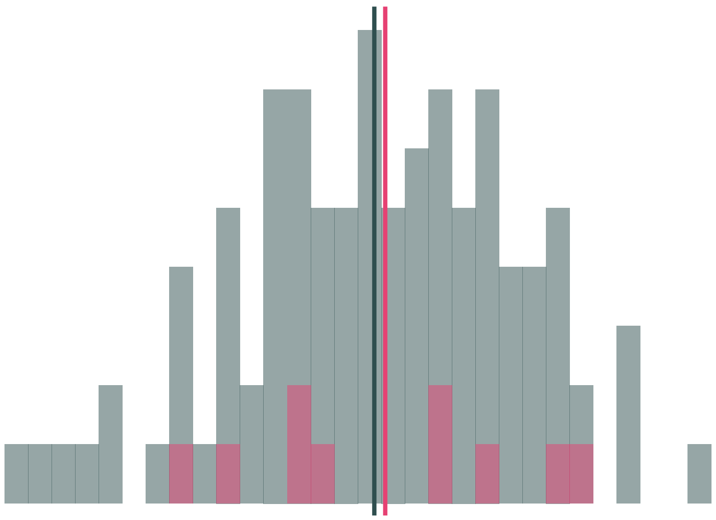
$$\mu = 3.75$$

Population *vs.* Sample

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



Sample 1: 10 random individuals



Population relationship

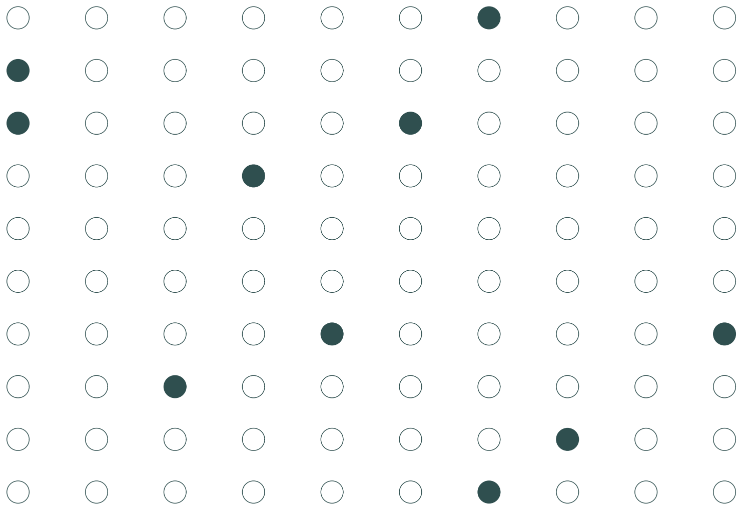
$$\mu = 3.75$$

Sample relationship

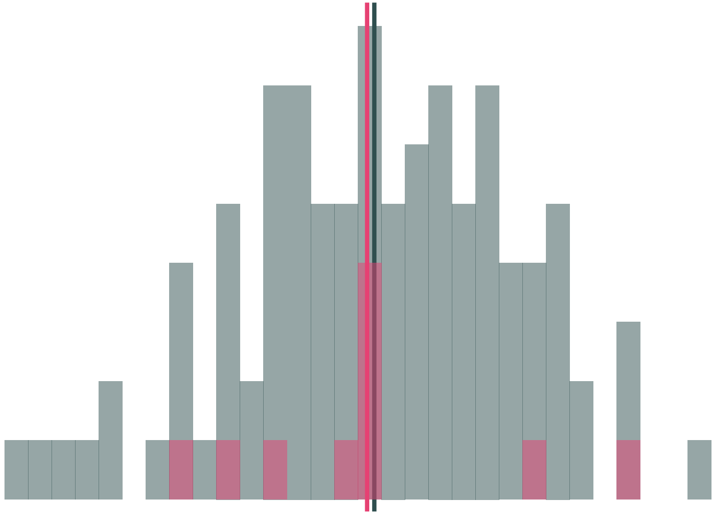
$$\hat{\mu} = 5.19$$

Population *vs.* Sample

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



Sample 2: 10 random individuals



Population relationship

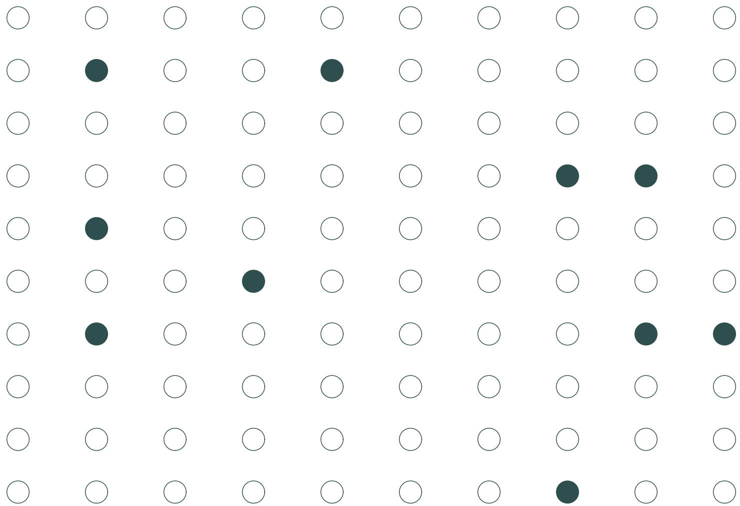
$$\mu = 3.75$$

Sample relationship

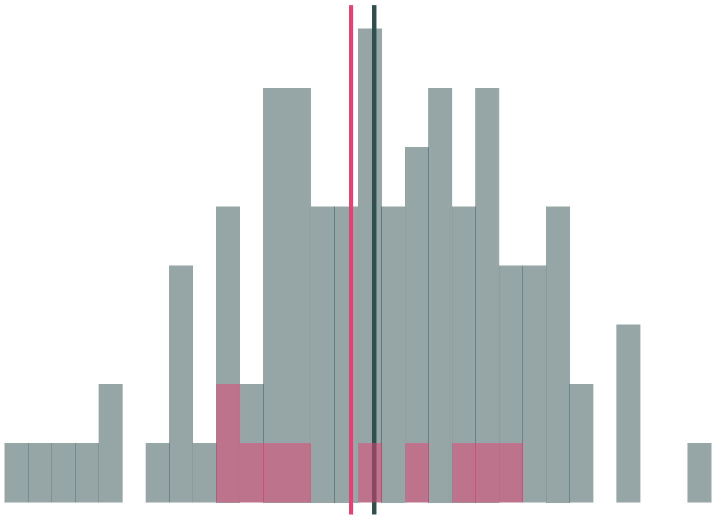
$$\hat{\mu} = 2.79$$

Population *vs.* Sample

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



Sample 3: 10 random individuals



Population relationship

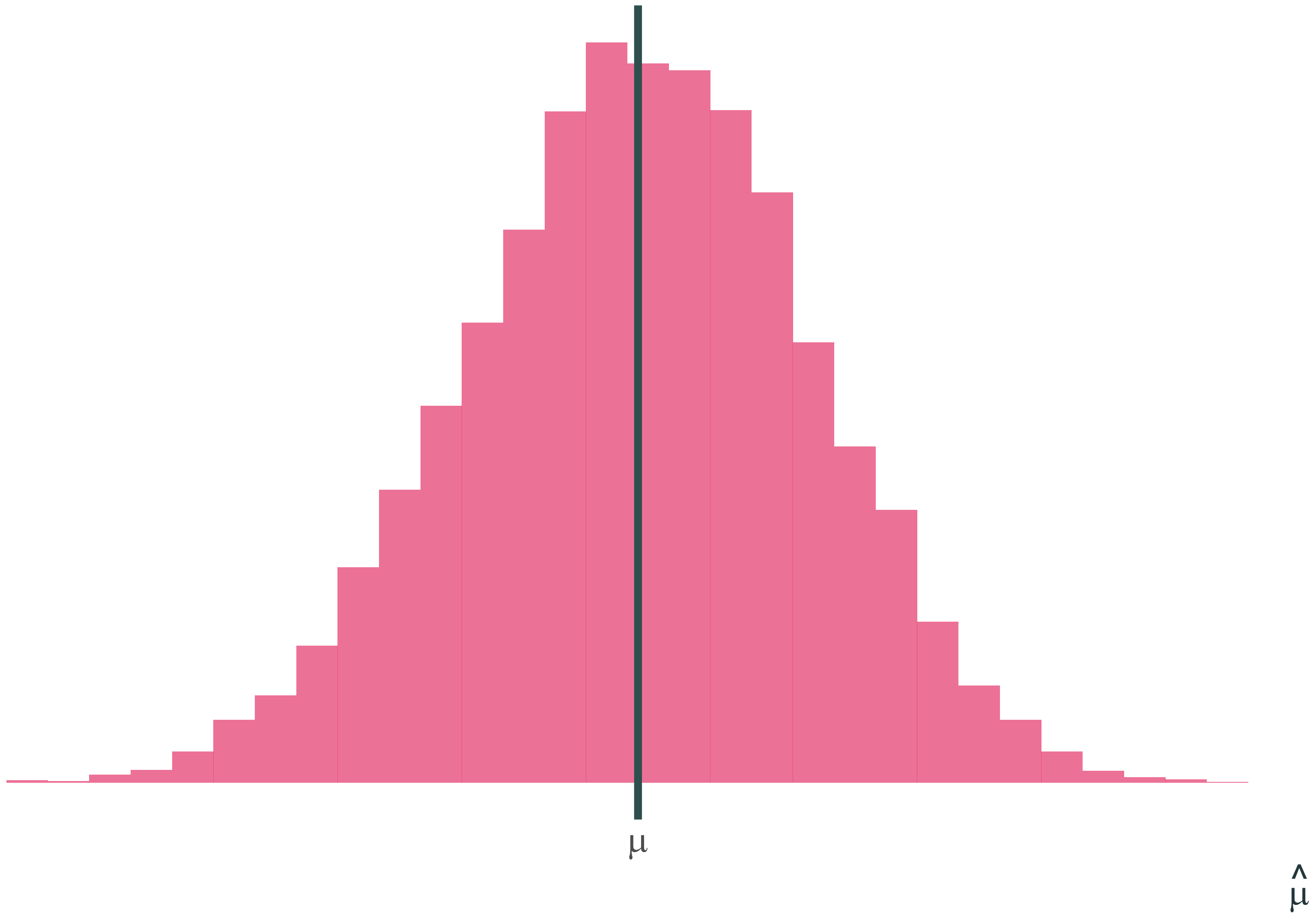
$$\mu = 3.75$$

Sample relationship

$$\hat{\mu} = 0.67$$

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

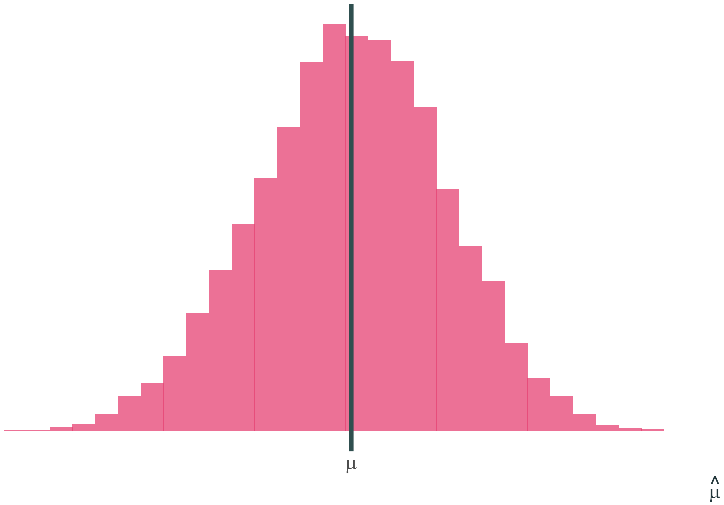
(This exercise is called a Monte Carlo simulation.)



Sampling Distribution

Population *vs.* Sample

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



- 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.
- To do so, we need to discuss **Sampling Distributions**
- But first...

Group Questions

Describe in your own words what the following terms are and how they connect to each other:

- **Population**
- **Sample**
- **Parameter**
- **Estimator**

Sampling Distribution

Sampling Distribution

Recap

- We have a *sample* mean and we are trying to learn about a *population* mean, but we know there will be uncertainty

E.g., Average School Size in CA

- Suppose you want to know the average school size in California
- You can imagine that if we took a different samples of schools, average size would be different
- If you did this many times this would create a distribution, we call this the **sampling distribution**

Sampling Distribution

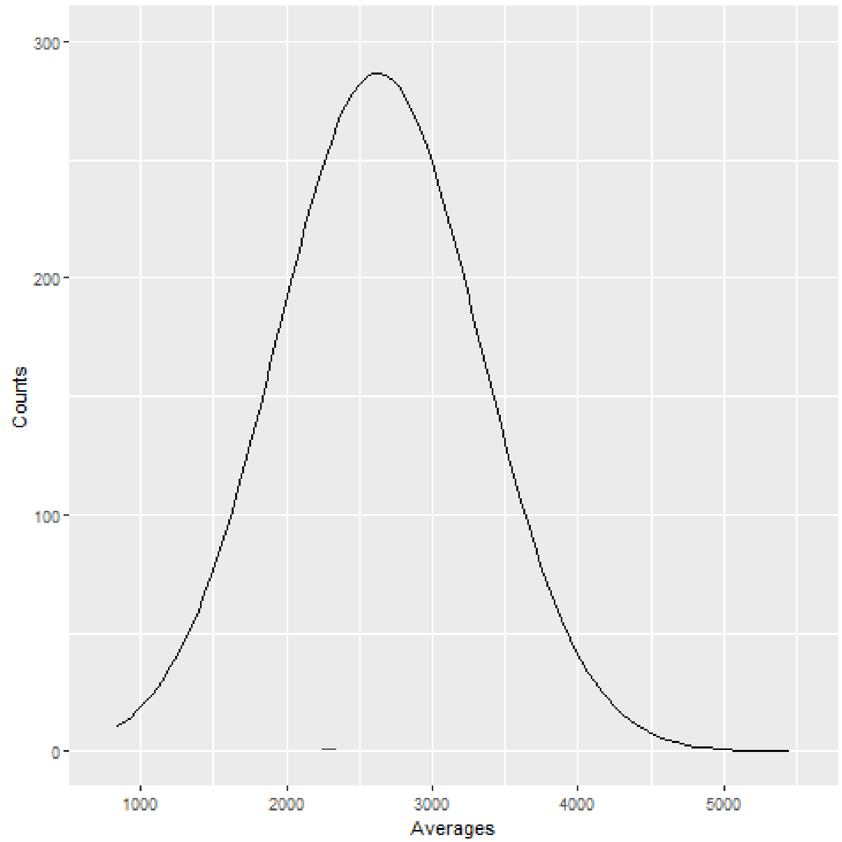
Number of Students in CA Schools

- To illustrate, consider data on the # of students in California schools
- Here is the distribution -- heavily skewed

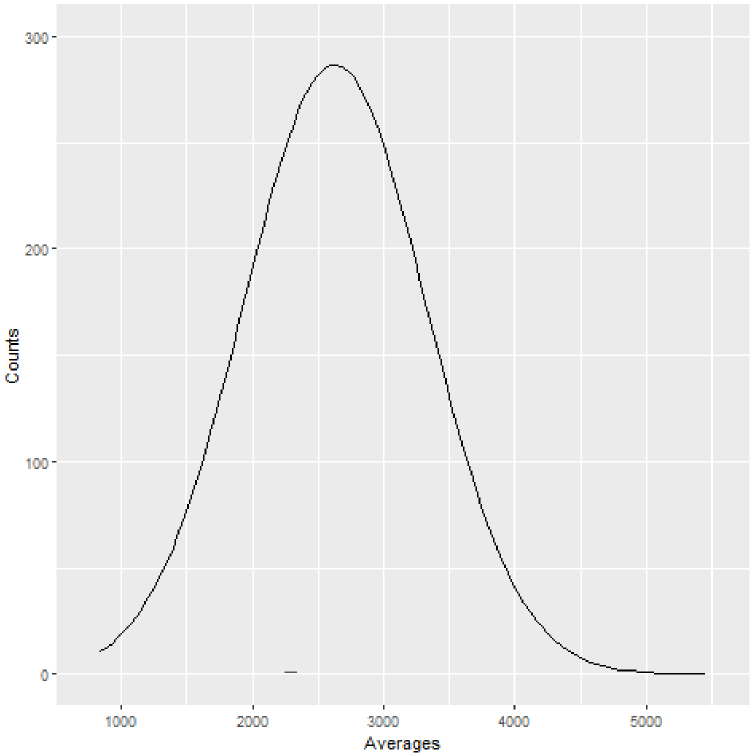
Sampling Distribution

Average # of Students in CA Schools

- Now suppose we take different samples of schools, calculate the average, and kept track

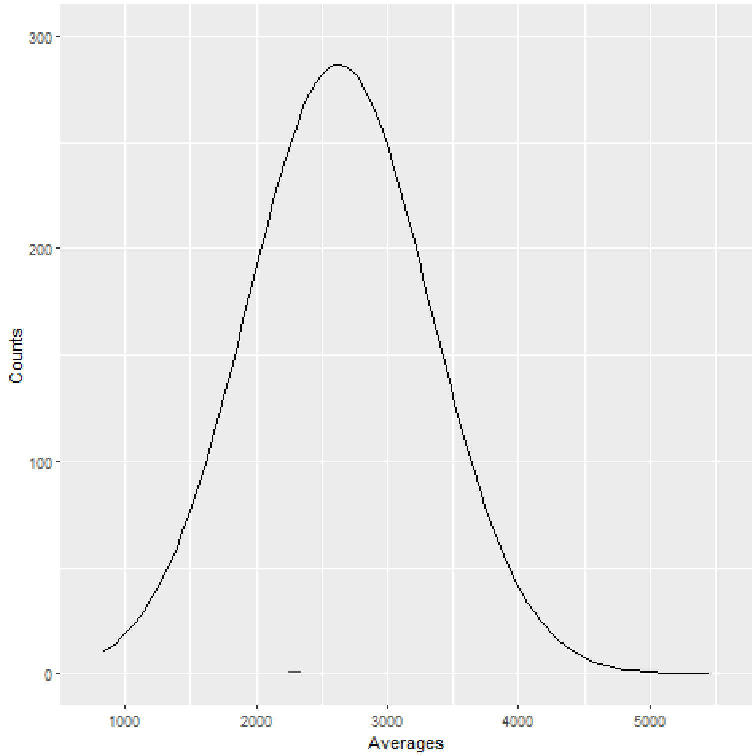


Sampling Distribution



- This is the **sampling distribution** of our estimator (*the sample average*) for the parameter (*the population average*)
    - i.e., distribution of all possible sample averages

Sampling Distribution



Group Question: **What is the difference between the distribution of Y and the sampling distribution of $\bar{Y}$?**

Note:

- Data are skewed, but sampling distribution is **normally distributed** (bell curve)
- Mean of distribution is close to the population average (~2,500)
- Spread of sampling distribution conveys *uncertainty*
  - i.e., more spread means higher chance any given sample is far away from truth

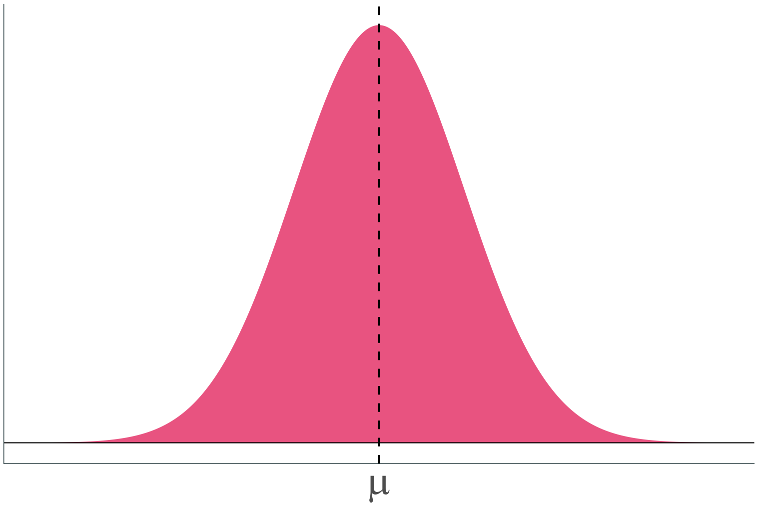
Properties of Estimators

Properties of Estimators

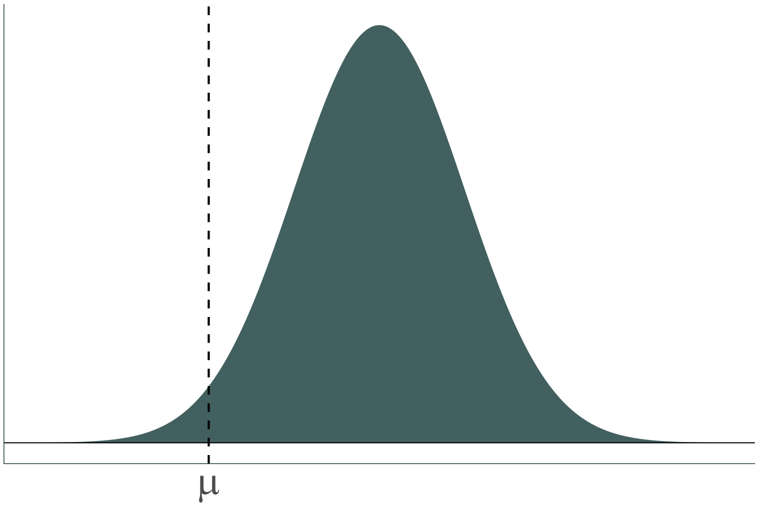
Question: What properties make an estimator reliable?

Answer 1: Unbiasedness.

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



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



- I.e., expected value of sampling distribution = true population parameter

Sample Average is Unbiased

- Sample average turns out to be unbiased estimator of population average
- Recall: $\hat{\mu} \equiv \bar{Y} \equiv \frac{1}{n}\sum_{i=1}^{n} Y_i$ **and** $\mu \equiv \mathbb{E}[Y]$
- **Proof**: WTS $\mathbb{E}[\hat{\mu}] = \mu$

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}n\mu = \mu. \blacksquare$$

  - By simple properties of expectations (first lecture)

Properties of Estimators

Question: What properties make an estimator reliable?

Answer 2: Low Sampling 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, aka, **sampling variance** (variance of *sampling distribution*).

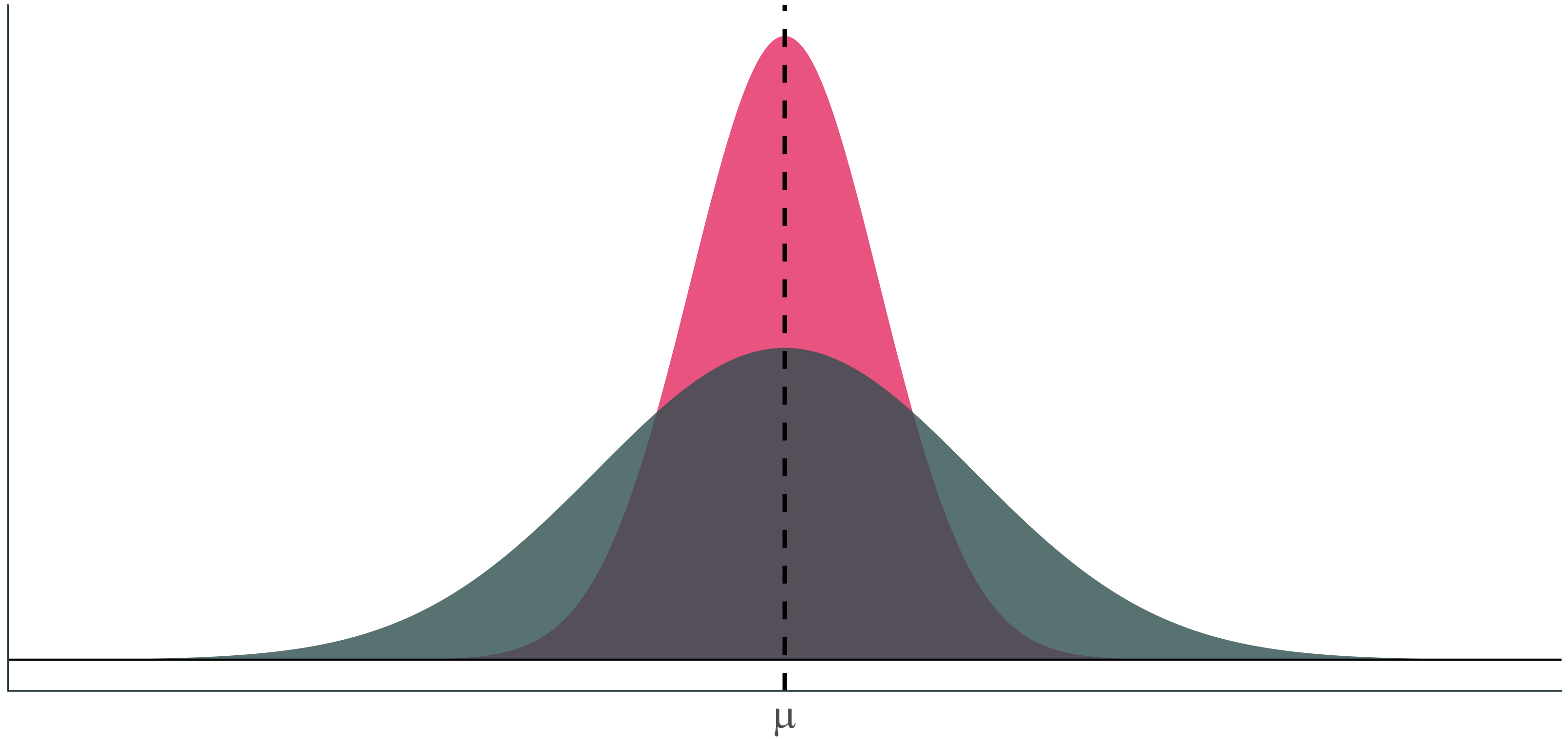
$$\mathbf{Var}(\hat{\mu}) = \mathbb{E}\left[(\hat{\mu} - \mathbb{E}[\hat{\mu}])^2\right]$$

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 Sampling Variance (a.k.a. Efficiency).



Sample Variance

- **Sample Variance:** $S(Y_i)^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$
  - **In Stata:** Summarizing data. `sum Y, detail`
- **Population Variance:** $V(Y_i) = E\left[(Y_i - E[\bar{Y}])^2\right] = \sigma_Y^2$
  - Unknown parameter
- **Standard Deviation:** Square root of the variance $\sigma_Y = \sqrt{\sigma_Y^2}$

Sampling Variance of $\hat{\mu} = \bar{Y}$

We want to characterize the variance of $\bar{Y}$ across repeated samples

- $V(\bar{Y}) = E\left[(\bar{Y} - E[\bar{Y}])^2\right] = E\left[(\bar{Y} - E[Y_i])^2\right]$
- By the unbiasedness property
- $V(\bar{Y})$: variance of sample mean
- $V(Y_i)$ or σ_Y^2: population variance of underlying data

Properties of Estimators

Sampling Variance of $\hat{\mu} = \bar{Y}$

- *Sampling* variance is related to *population* variance

$$V(\bar{Y}) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2 = \frac{n\sigma_Y^2}{n^2} = \frac{\sigma_Y^2}{n}$$

  - Variance of a sum is the sum of variances
  - Constants are squared when pulled out of a variance

- Thus, sampling variance of an average depends on variance of underlying data and number of observations

Properties of Estimators

Standard Errors

- We usually work with standard deviation of sample mean rather than variances

- **Standard error** is the standard deviation of an *estimator*[+]

- $SE(\bar{Y}) = \sqrt{V(\bar{Y})} = \frac{\sigma_Y}{\sqrt{n}}$

- $\widehat{SE}(\bar{Y}) = \frac{S(Y_i)}{\sqrt{n}}$, Estimated Standard Error

- SE summarize variation in estimate from *random sampling*

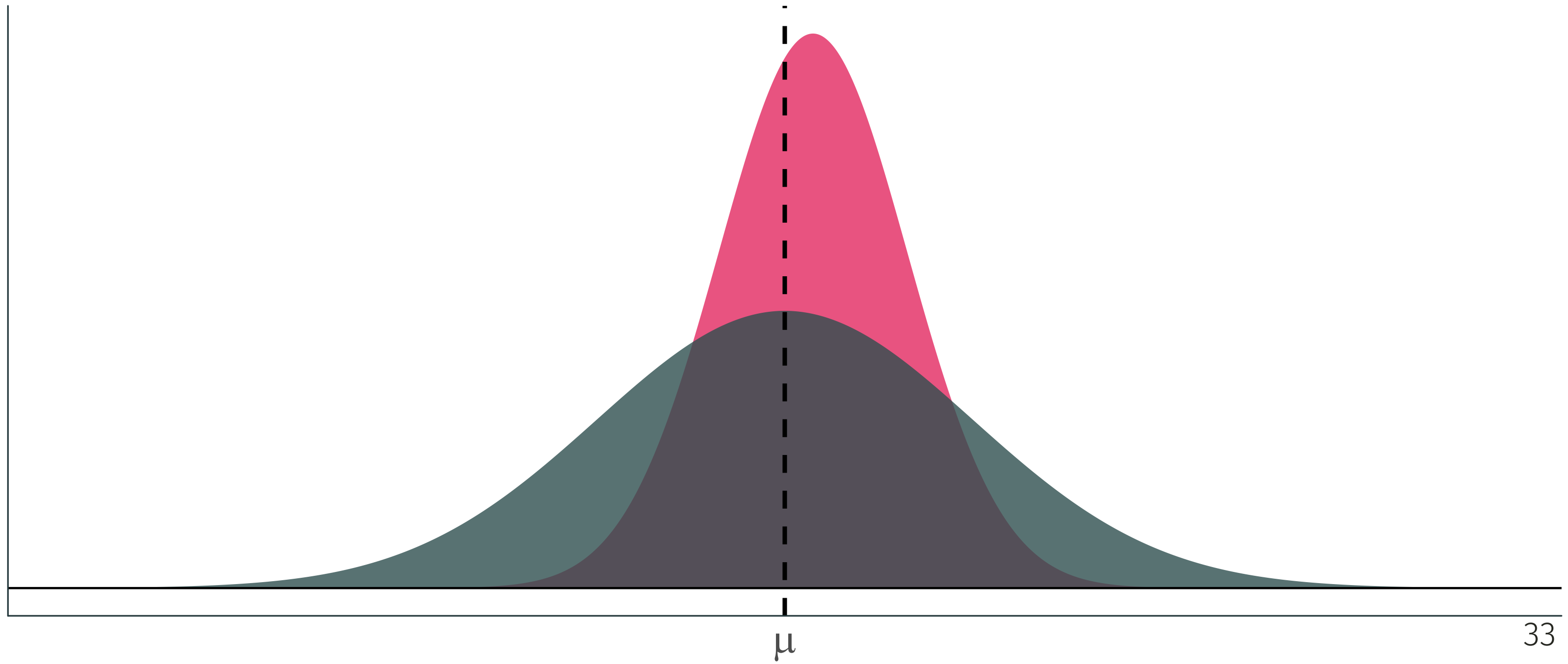
- Again, SE $\neq$ standard deviation of underlying data

[+] The estimator we've considered so far is the sample average. More specifically, the standard error is the standard deviation of the *sampling distribution* of an estimator.

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.



Properties of Estimators

Question: What properties make an estimator reliable?

Answer 3: Consistency.

- We want uncertainty of our estimator to decrease as n grows.
    - I.e., want probability that estimate $\hat{\mu}_Y$ falls within a small interval around parameter μ to get increasingly closer to 1 as n grows.
- **Intuition:** As n grows, our sample size approaches population size $\Rightarrow$ uncertainty should fall
- This is the **Law of Large Numbers (LLN)**

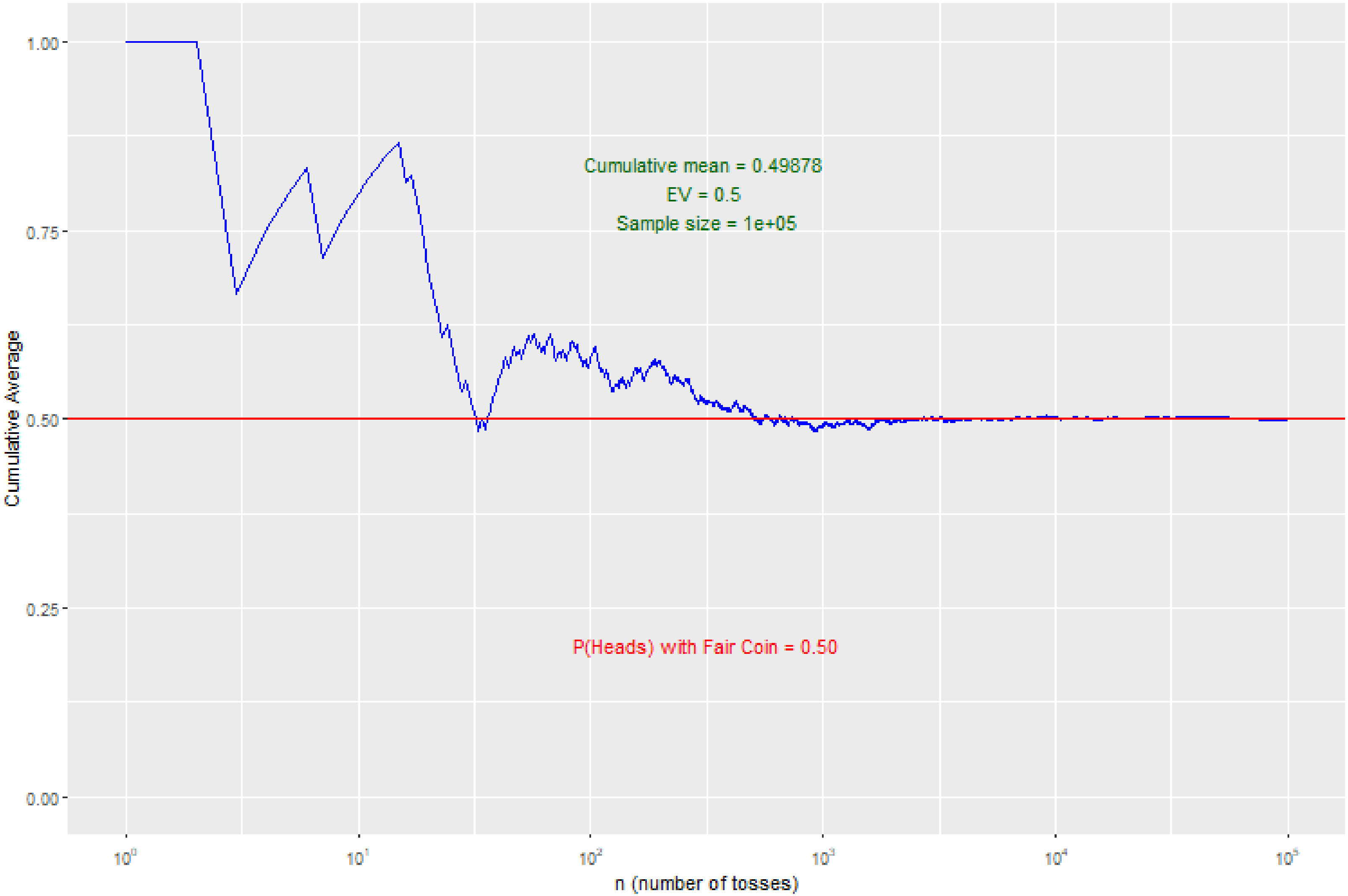
Law of Large Numbers (LLN)

Law of Large Numbers

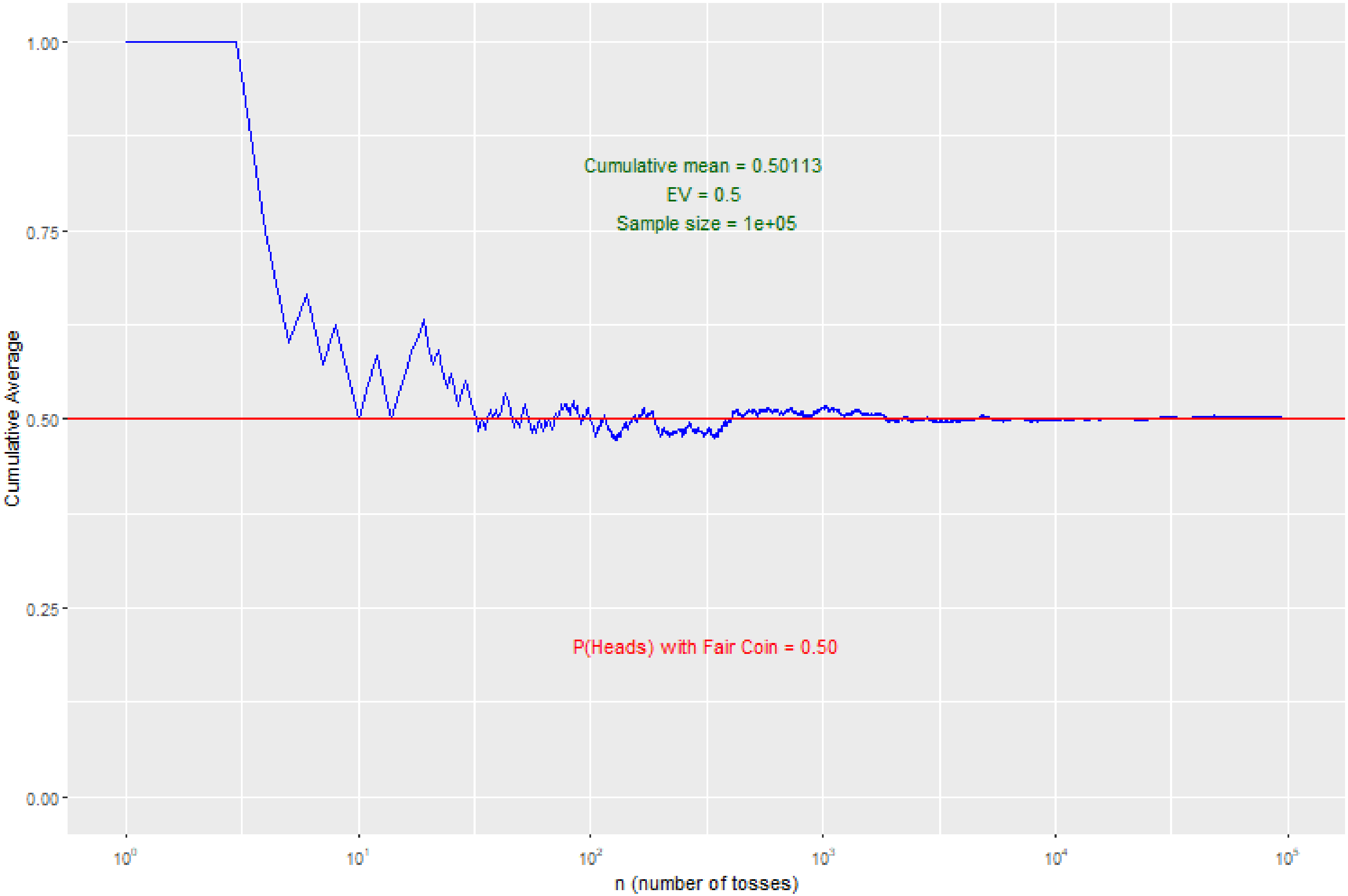
> LLN implies that $\bar{Y}$ will be very close to $E[Y_i]$ as the sample size grows

- Let's empirically test LLN let's flip a fair coin 100,000 times
- Record cumulative average (H=1) (T=0)
- $E[Y_i] = \mathbf{0.5}$

LLN


Cumulative mean = 0.49878
EV = 0.5
Sample size = 1e+05
P(Heads) with Fair Coin = 0.50
Cumulative Average
1.00
0.75
0.50
0.25
0.00
n (number of tosses)
10^0
10^1
10^2
10^3
10^4
10^5

LLN


Cumulative mean = 0.50113
EV = 0.5
Sample size = 1e+05
P(Heads) with Fair Coin = 0.50
Cumulative Average
1.00
0.75
0.50
0.25
0.00
n (number of tosses)
10^0
10^1
10^2
10^3
10^4
10^5

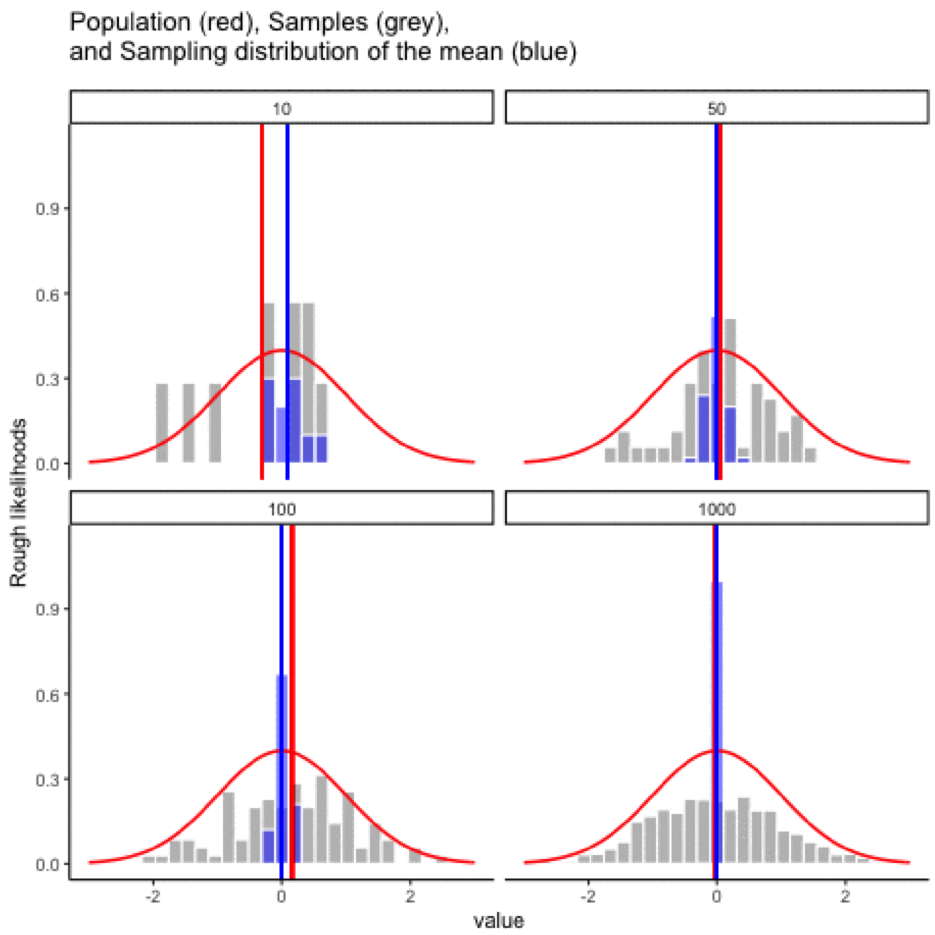
LLN - Analytic Proof

- We've shown that *sampling* variance can be written as

$$V(\bar{Y}) = \frac{\sigma_Y^2}{n}$$

- LLN at work, large n implies little dispersion
  - As $n \to \infty$, $V(\bar{Y}) \to 0$

LLN - One More Visualization



- **Red line:** Mean of Sample. **Blue line:** Mean of Sampling Distribution.
- **True Population Mean:** $\mu = 0$

Unbiased Estimators

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

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

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}:

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

Unbiased Estimators

Poll Questions (1)

Unbiased Estimators

Sorry, lots of questions. It is so easy to lose the forest for the trees with all of these statistical concepts

Group Questions

1. **How does the LLN help us learn about populations using samples?**

2. **What is a standard error and why is it useful?**

Okay, we are ready to actually be researchers and test some hypotheses! Next time.