Econ 103 – Statistics for Economists

Chapter 6 and 7: Sampling and Bias

Mallick Hossain

University of Pennsylvania

Motivation

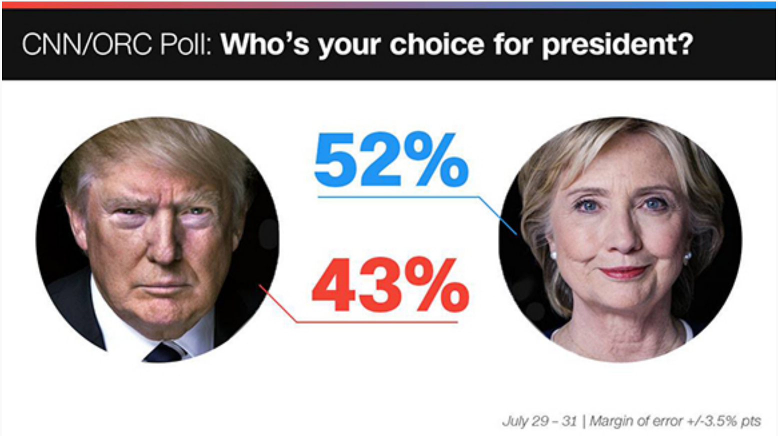
Recall

1. Sampling Error
    - *Random* differences between sample and population
    - Cancel out on average
    - Decreases as sample size grows
2. Nonsampling Error
    - *Systematic* differences between sample and population
    - Does *not* cancel out on average
    - Does *not* decrease as sample size grows

Road Map

1. We are not going to worry about non-sampling error
    - This would be something for a survey design course
2. We will be learning more about sampling errors
    - What do we need to be worried about when sampling from a population? W

Election Polls



Who's Right?!?!

RealClear Politics

Polls Election 2016 Video Writ

RCP Electoral Map | Changes in Electoral Count | Map With No Toss Ups | No Toss Up Changes | Latest Polls

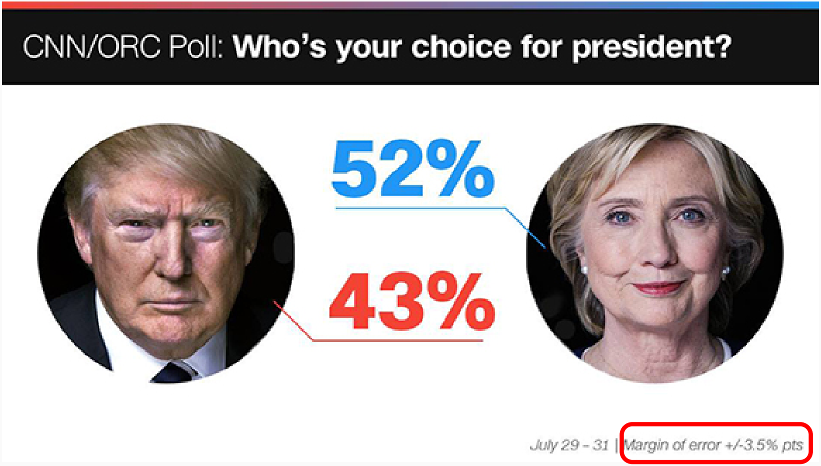
Polling Data						
Poll	**Date**	**Sample**	**MoE**	**Clinton (D)**	**Trump (R)**	**Spread**
RCP Average	**10/3 - 10/14**	**--**	**--**	**48.1**	**41.4**	**Clinton +6.7**
LA Times/USC Tracking	10/8 - 10/14	2870 LV	4.5	44	44	Tie
FOX News	10/10 - 10/12	917 LV	3.0	49	41	Clinton +8
NBC News/Wall St. Jrnl	10/8 - 10/10	806 LV	3.5	50	40	Clinton +10
Reuters/Ipsos	10/6 - 10/10	2363 LV	2.2	44	37	Clinton +7
Economist/YouGov	10/7 - 10/8	971 RV	4.2	48	43	Clinton +5
The Atlantic/PRRI	10/5 - 10/9	886 LV	3.9	49	38	Clinton +11
Quinnipiac	10/5 - 10/6	1064 LV	3.0	50	44	Clinton +6
NBC News/SM	10/3 - 10/9	23329 LV	1.0	51	44	Clinton +7

All General Election: Trump vs. Clinton Polling Data

Questions to Answer

1. How accurately do sample statistics estimate population parameters?
2. How can we quantify the uncertainty in our estimates?

Election Polls



Sampling

Step 1: Population as RV rather than List of Objects

Old Way	New Way
Among 138 million voters, 69 million will vote for Hillary Clinton	Bernoulli($p = 1/2$) RV
List of heights for 97 million US adult males with mean 69 in and std. dev. 6 in	$N(\mu = 69, \sigma^2 = 36)$ RV

Second example assumes distribution of height is bell-shaped.

Random Sample

In Words

Select sample of n objects from population so that:

1. Each member of the population has the same probability of being selected
2. The fact that one individual is selected does not affect the chance that any other individual is selected
3. Each sample of size n is equally likely to be selected

In Math

$X_1, X_2, \ldots, X_n \sim$ iid $f(x)$ if continuous

$X_1, X_2, \ldots, X_n \sim$ iid $p(x)$ if discrete

Random Sample Means *Sample With Replacement*

- Without replacement $\Rightarrow$ dependence between samples
- Sample small relative to popn. $\Rightarrow$ dependence negligible.

Step 2: iid RVs Represent Random Sampling from Popn.

Hillary Clinton Example

Poll random sample of 1000 registered voters:

$$X_1, \ldots, X_{1000} \sim \text{ iid Bernoulli}(p = 1/2)$$

Height Example

Measure the heights of random sample of 50 US males:

$$Y_1, \ldots, Y_{50} \sim \text{ iid } N(\mu = 69, \sigma^2 = 36)$$

Key Question

What do the properties of the population imply about the properties of the sample?

What does the population imply about the sample?

Suppose that exactly half of US voters plan to vote for Hillary Clinton. If you poll a random sample of 4 voters, what is the probability that *exactly half* are Hillary supporters?

What does the population imply about the sample?

Suppose that exactly half of US voters plan to vote for Hillary Clinton. If you poll a random sample of 4 voters, what is the probability that *exactly half* are Hillary supporters?

$$\binom{4}{2}(1/2)^2(1/2)^2 = 3/8 = 0.375$$

The rest of the probabilities...

Suppose that exactly half of US voters plan to vote for Hillary Clinton and we poll a random sample of 4 voters.

$$
\begin{aligned}
P\left(\text{Exactly 0 Hillary Voters in the Sample}\right) &= 0.0625 \\
P\left(\text{Exactly 1 Hillary Voters in the Sample}\right) &= 0.25 \\
P\left(\text{Exactly 2 Hillary Voters in the Sample}\right) &= 0.375 \\
P\left(\text{Exactly 3 Hillary Voters in the Sample}\right) &= 0.25 \\
P\left(\text{Exactly 4 Hillary Voters in the Sample}\right) &= 0.0625
\end{aligned}
$$

You should be able to work these out yourself. If not, review the lecture slides on the Binomial RV.

Population Size is Irrelevant Under Random Sampling

Crucial Point

None of the preceding calculations involved the population size: I didn't even tell you what it was! We'll never talk about population size again in this course.

Why?

Draw with replacement $\implies$ only the sample size and the *proportion* of Hillary supporters in the population matter.

Sample Statistics

(Sample) Statistic

Any function of the data *alone*, e.g. sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.
Typically used to estimate an unknown population parameter:
e.g. $\bar{x}$ is an estimate of μ.

Step 3: Random Sampling $\Rightarrow$ *Sample Statistics* are RVs

This is *the crucial point of the course*: if we draw a random sample, the dataset we get is random. Since a statistic is a function of the data, it is a random variable!

A Sample Statistic in the Polling Example

Suppose that exactly half of voters in the population support Hillary Clinton and we poll a random sample of 4 voters. If we code Hillary supporters as "1" and everyone else as "0" then what are the possible values of the sample mean in our dataset?

(a) $(0, 1)$
(b) $\{0, 0.25, 0.5, 0.75, 1\}$
(c) $\{0, 1, 2, 3, 4\}$
(d) $(-\infty, \infty)$
(e) Not enough information to determine.

Sampling Distribution

Under random sampling, a statistic is a RV so it has a PDF if continuous of PMF if discrete: this is its sampling distribution.

Sampling Dist. of Sample Mean in Polling Example

$$
\begin{aligned}
p(0) &= 0.0625 \\
p(0.25) &= 0.25 \\
p(0.5) &= 0.375 \\
p(0.75) &= 0.25 \\
p(1) &= 0.0625
\end{aligned}
$$

Contradiction? No, but we need better terminology...

- Under random sampling, a statistic is a RV
- Given dataset is *fixed* so statistic is a *constant number*
- Distinguish between: Estimator vs. Estimate

Estimator

Description of a general procedure.

Estimate

Particular result obtained from applying the procedure.

$\bar{X}_n$ is an Estimator = Procedure = Random Variable

1. Take a random sample: $X_1, \ldots, X_n$
2. Average what you get: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

$\bar{X}_n$ is an Estimator = Procedure = Random Variable

1. Take a random sample: $X_1, \ldots, X_n$
2. Average what you get: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

$\bar{x}$ is an Estimate = Result of Procedure = Constant

- Result of taking a random sample was the dataset: $x_1, \ldots, x_n$
- Result of averaging the observed data was $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

$\bar{X}_n$ is an Estimator = Procedure = Random Variable

1. Take a random sample: $X_1, \ldots, X_n$
2. Average what you get: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

$\bar{x}$ is an Estimate = Result of Procedure = Constant

- Result of taking a random sample was the dataset: $x_1, \ldots, x_n$
- Result of averaging the observed data was $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sampling Distribution of $\bar{X}_n$

Thought experiment: suppose I were to repeat the procedure of taking the mean of a random sample over and over forever. What relative frequencies would I get for the sample means?

This is *Only* a Thought Experiment

- Real applications: observe only a single sample:
  - $n =$1,189 voters: 44% Clinton, 43% Trump, 13% Undecided.

This is *Only* a Thought Experiment

- Real applications: observe only a single sample:
  - $n =$1,189 voters: 44% Clinton, 43% Trump, 13% Undecided.
- What does the sample tell us about the population?
  - How close is Trump's *actual* support to 43%?

This is *Only* a Thought Experiment

- Real applications: observe only a single sample:
  - $n =$1,189 voters: 44% Clinton, 43% Trump, 13% Undecided.
- What does the sample tell us about the population?
  - How close is Trump's *actual* support to 43%?
- Can't know for sure without asking *all* voters!
  - Which is impractical and defeats the purpose of the poll!

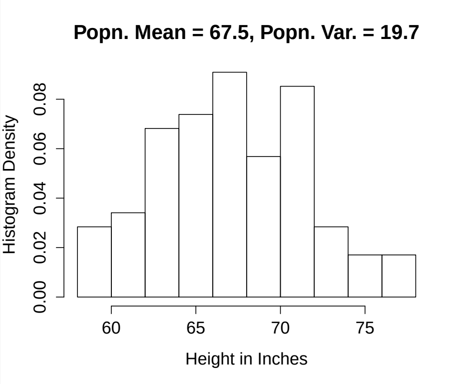
This is *Only* a Thought Experiment

- Real applications: observe only a single sample:
  - $n = 1{,}189$ voters: 44% Clinton, 43% Trump, 13% Undecided.
- What does the sample tell us about the population?
  - How close is Trump's *actual* support to 43%?
- Can't know for sure without asking *all* voters!
  - Which is impractical and defeats the purpose of the poll!
- Since we can't be sure, try to quantify using probability.
  - E.g. what is the prob. that the poll is off by $>$ 2% points?

This is *Only* a Thought Experiment

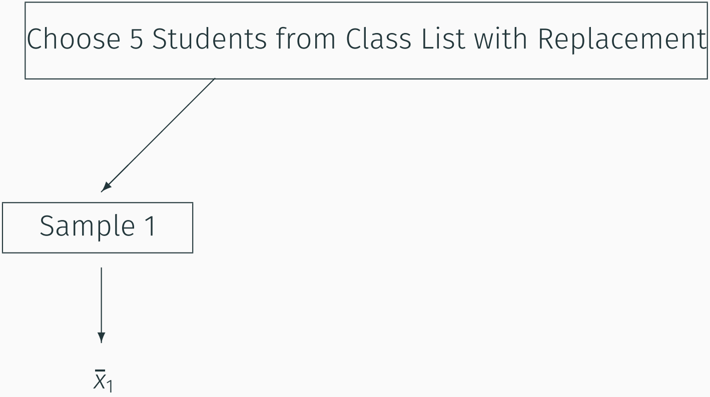
- Real applications: observe only a single sample:
  - n =1,189 voters: 44% Clinton, 43% Trump, 13% Undecided.
- What does the sample tell us about the population?
  - How close is Trump's *actual* support to 43%?
- Can't know for sure without asking *all* voters!
  - Which is impractical and defeats the purpose of the poll!
- Since we can't be sure, try to quantify using probability.
  - E.g. what is the prob. that the poll is off by $>$ 2% points?
- Need to speak in terms of long-run relative frequencies.
  - Remember that is the way we define probability in Econ 103!

Example: Heights of Econ 103 Students

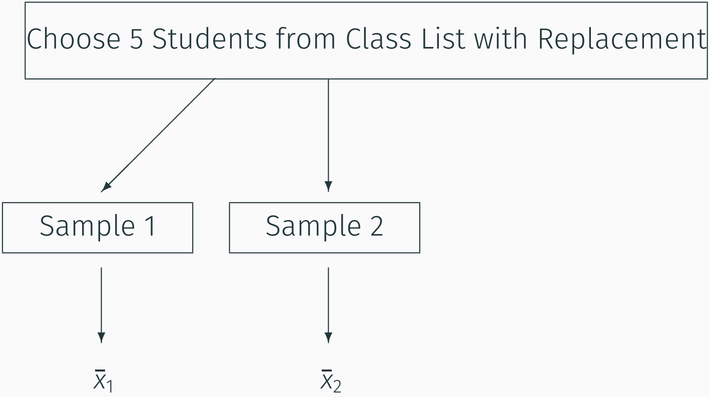


Use R to illustrate an example where we *know* the population. Can't do this in the real applications, but simulate it on the computer...

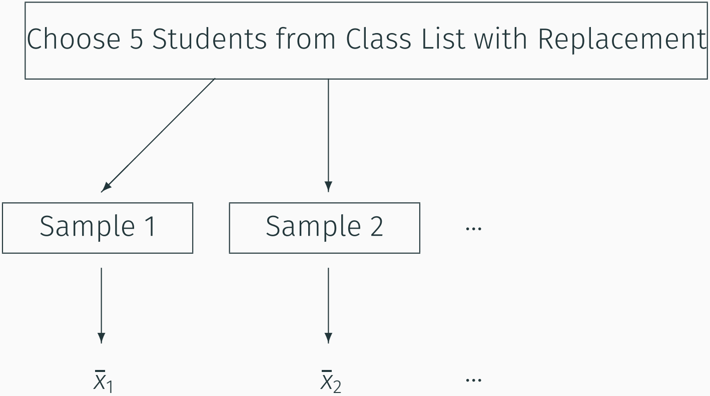
Sampling Distribution of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$



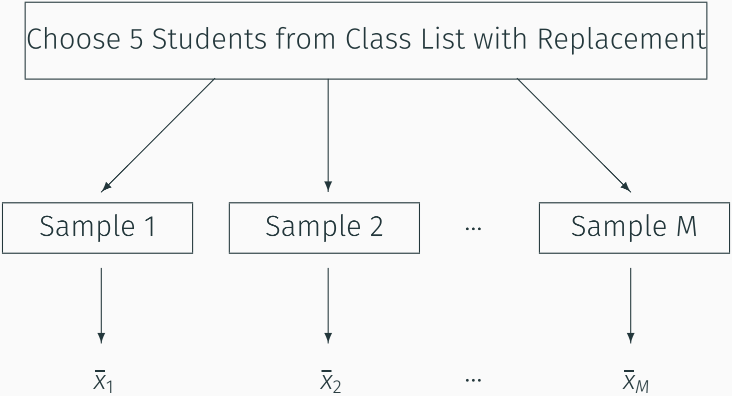
Sampling Distribution of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$



Sampling Distribution of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

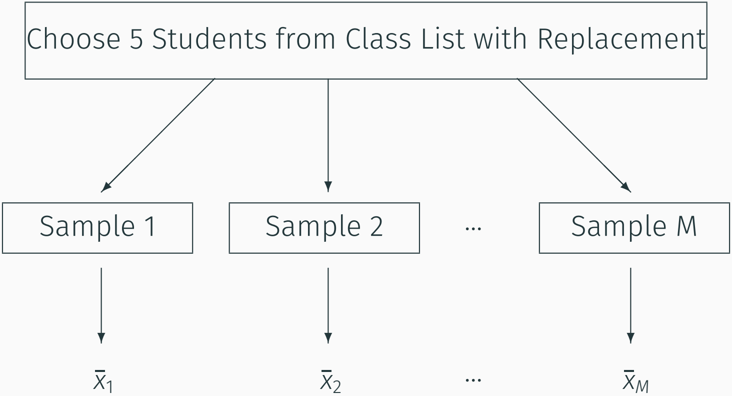


Sampling Distribution of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$



Repeat M times $\rightarrow$ get M different sample means

Sampling Distribution of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$



Repeat M times $\rightarrow$ get M different sample means

Sampling Dist: relative frequencies of the $\bar{x}_i$ when $M = \infty$

Height of Econ 103 Students

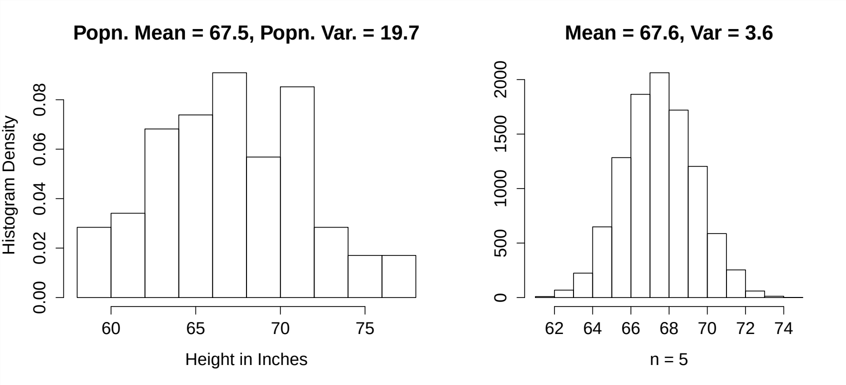
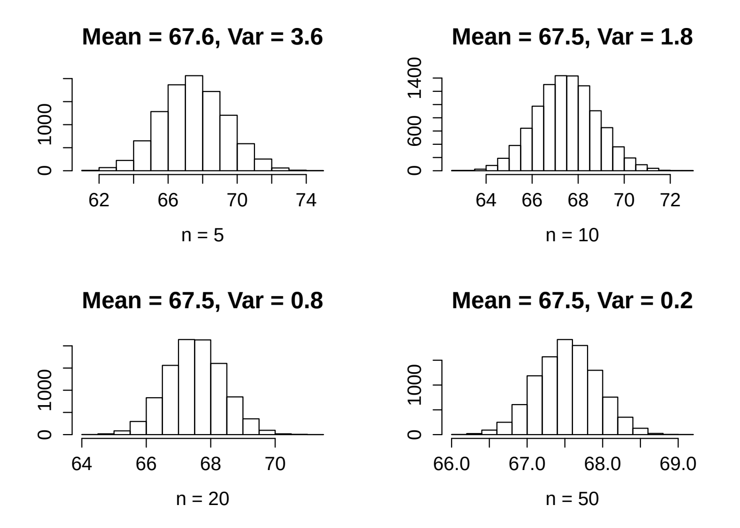


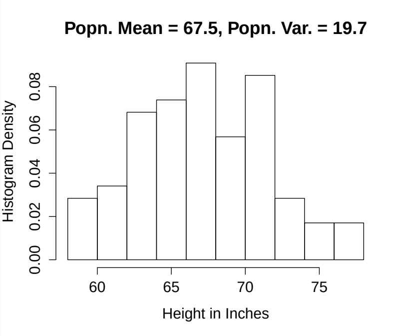
Figure 1: Left: Population, Right: Sampling distribution of $\bar{X}_5$

Histograms of sampling distribution of sample mean $\bar{X}_n$

Random Sampling With Replacement, 10000 Reps. Each



Population Distribution vs. Sampling Distribution of $\bar{X}_n$



	Sampling Dist. of $\bar{X}_n$	
n	Mean	Variance
5	67.6	3.6
10	67.5	1.8
20	67.5	0.8
50	67.5	0.2

Two Things to Notice:

1. Sampling dist. "correct on average"
2. Sampling variability decreases with n

$X_1, \ldots, X_9 \sim$ iid with $\mu = 5$, $\sigma^2 = 36$.

Calculate:

$$E(\bar{X}) = E\left[\frac{1}{9}(X_1 + X_2 + \ldots + X_9)\right]$$

Mean of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$$

Mean of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] =$$

Mean of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n} \mu =$$

Mean of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ

$$E[\bar{X}_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{n\mu}{n} = \mu$$

Hence, sample mean is “correct on average.” The formal term for this is *unbiased*.

$X_1, \ldots, X_9 \sim$ iid with $\mu = 5$, $\sigma^2 = 36$.

Calculate:

$$Var(\bar{X}) = Var\left[\frac{1}{9}(X_1 + X_2 + \ldots + X_9)\right]$$

Variance of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$Var[\bar{X_n}] \quad = \quad Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$$

Variance of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$
\begin{aligned}
Var[\bar{X}_n] &= Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\
&=
\end{aligned}
$$

Variance of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$\begin{aligned} Var[\bar{X}_n] &= Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \end{aligned}$$

Variance of Sampling Distribution of $\bar{X}_n$

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$\begin{aligned} Var[\bar{X}_n] &= Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

Hence the variance of the sample mean *decreases linearly with sample size.*

Standard Error

Std. Dev. of estimator's sampling dist. is called standard error.

Standard Error of the Sample Mean

$$SE(\bar{X}_n) = \sqrt{Var\left(\bar{X}_n\right)} = \sqrt{\sigma^2/n} = \sigma/\sqrt{n}$$

More Generally and More Formally:

Estimator

A function $T(X_1, \ldots, X_n)$ of the RVs that represent the *procedure* of drawing a random sample, hence a RV itself.

Sampling Distribution

The probability distribution (PMF or PDF) of an Estimator.

Estimate

A function $T(x_1, \ldots, x_n)$ of the *observed data*, i.e. the *realizations* of the random variables we use to represent random sampling. Since its a function of constants, an estimate is itself a constant.

Population: $f(x)$

Probability Distribution

Random Sample of Size n

Random Variables

$X_1, X_2, \ldots, X_n \sim \text{iid } f(x)$

Realizations (Constants)

$x_1^{(1)}, \ldots, x_n^{(1)}$

$x_1^{(2)}, \ldots, x_n^{(2)}$

...

$x_1^{(M)}, \ldots, x_n^{(M)}$

M Replications, each containing n Observations

Random Sample of Size n — Estimator

Random Variables

$$X_1, X_2, \ldots, X_n \sim \text{iid } f(x) \longrightarrow T(X_1, \ldots, X_n)$$

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

$$\big\downarrow$$

Observations (Data) — Estimate

Constants

$$x_1, \ldots, x_n \longrightarrow T(x_1, \ldots, x_n)$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Bias

Unbiased means “Right on Average”

Bias of an Estimator

Let $\widehat{\theta}_n$ be a sample estimator of a population parameter θ_0. The *bias* of $\widehat{\theta}_n$ is $E[\widehat{\theta}_n] - \theta_0$.

Unbiased Estimator

A sample estimator $\widehat{\theta}_n$ of a population parameter θ_0 is called *unbiased* if $E[\widehat{\theta}_n] = \theta_0$

Why $(n-1)$ for sample variance?

We will show that having $n-1$ in the denominator ensures:

$$E[S^2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \sigma^2$$

under random sampling.

Why $(n-1)$ for sample variance?

Step # 1 – Tedious but straightforward algebra gives:

$$\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2=\left[\sum_{i=1}^{n}(X_i-\mu)^2\right]-n(\bar{X}-\mu)^2$$

You are not responsible for proving Step #1 on an exam.

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

Step # 2 – Take Expectations of Step # 1:

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

Step # 2 – Take Expectations of Step # 1:

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

Why $(n-1)$ for sample variance?

Step # 2 – Take Expectations of Step # 1:

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

Where we have used the linearity of expectation.

Why $(n-1)$ for sample variance?

Step # 3 – Use assumption of random sampling:

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$E\left[\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right] = \sum_{i=1}^{n} E\left[\left(X_i-\mu\right)^2\right] - n\,E\left[(\bar{X}-\mu)^2\right]$$

$$=$$

Why $(n-1)$ for sample variance?

Step # 3 – Use assumption of random sampling:

$X_1, \dots, X_n \sim$ iid with mean μ and variance σ^2

$$
\begin{aligned}
E\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] &= \sum_{i=1}^{n} E\left[(X_i-\mu)^2\right] - n\, E\left[(\bar{X}-\mu)^2\right] \\
&= \sum_{i=1}^{n} Var(X_i) - n\, E\left[(\bar{X}-E[\bar{X}])^2\right] \\
&=
\end{aligned}
$$

Step # 3 – Use assumption of random sampling:

$X_1, \ldots, X_n \sim$ iid with mean μ and variance σ^2

$$\begin{aligned} E\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] &= \sum_{i=1}^{n} E\left[(X_i - \mu)^2\right] - n\, E\left[(\bar{X} - \mu)^2\right] \\ &= \sum_{i=1}^{n} Var(X_i) - n\, E\left[(\bar{X} - E[\bar{X}])^2\right] \\ &= \sum_{i=1}^{n} Var(X_i) - n\, Var(\bar{X}) = n\sigma^2 - \sigma^2 \\ &= \end{aligned}$$

Why $(n-1)$ for sample variance?

Step # 3 – Use assumption of random sampling:

$X_1, \dots, X_n \sim$ iid with mean μ and variance σ^2

$$\begin{aligned}
E\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right] &= \sum_{i=1}^{n} E\left[(X_i-\mu)^2\right] - n\,E\left[(\bar{X}-\mu)^2\right] \\
&= \sum_{i=1}^{n} Var(X_i) - n\,E\left[(\bar{X}-E[\bar{X}])^2\right] \\
&= \sum_{i=1}^{n} Var(X_i) - n\,Var(\bar{X}) = n\sigma^2 - \sigma^2 \\
&= (n-1)\sigma^2
\end{aligned}$$

Since we showed earlier today that $E[\bar{X}] = \mu$ and $Var(\bar{X}) = \sigma^2/n$ under this random sampling assumption.

Finally – Divide Step # 3 by $(n-1)$:

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

Hence, having $(n-1)$ in the denominator ensures that the sample variance is "correct on average," that is *unbiased*.

A Different Estimator of the Population Variance

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

A Different Estimator of the Population Variance

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

$$E[\widehat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] =$$

A Different Estimator of the Population Variance

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

$$E[\widehat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \frac{(n-1)\sigma^2}{n}$$

Bias of $\widehat{\sigma}^2$

$$E[\widehat{\sigma}^2] - \sigma^2 =$$

A Different Estimator of the Population Variance

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

$$E[\widehat{\sigma}^2] = E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \frac{1}{n}E\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \frac{(n-1)\sigma^2}{n}$$

Bias of $\widehat{\sigma}^2$

$$E[\widehat{\sigma}^2] - \sigma^2 = \frac{(n-1)\sigma^2}{n} - \sigma^2 = \frac{(n-1)\sigma^2}{n} - \frac{n\sigma^2}{n} = -\sigma^2/n$$

How many brothers and sisters are in your family, including yourself?

The average number of children per family was about 2.0 twenty years ago.

What's Going On Here?

What's Going On Here?

Biased Sample!

- Zero children $\Rightarrow$ didn't send any to college
- Sampling by *children* so large families oversampled

Let $X_1, X_2, \ldots X_n \sim iid$ mean μ, variance σ^2 and define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. True or False:

$$\bar{X}_n \textit{ is an unbiased estimator of } \mu$$

(a) True
(b) False

Let $X_1, X_2, \ldots X_n \sim iid$ mean μ, variance σ^2 and define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. True or False:

$\bar{X}_n$ *is an unbiased estimator of* μ

(a) True

(b) False

TRUE!

Let $X_1, X_2, \ldots X_n \sim iid$ mean μ, variance σ^2. True or False:

X_1 is an unbiased estimator of μ

(a) True

(b) False

Let $X_1, X_2, \ldots X_n \sim iid$ mean μ, variance σ^2. True or False:

$$X_1 \textit{ is an unbiased estimator of } \mu$$

(a) True

(b) False

TRUE!

How to choose between two unbiased estimators?

Suppose $X_1, X_2, \ldots X_n \sim iid$ with mean μ and variance σ^2

$$\begin{aligned} E[\bar{X}_n] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu \\ E[X_1] &= \mu \end{aligned}$$

How to choose between two unbiased estimators?

Suppose $X_1, X_2, \ldots X_n \sim iid$ with mean μ and variance σ^2

$$\begin{aligned}
E[\bar{X}_n] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu \\
E[X_1] &= \mu \\
Var(\bar{X}_n) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \sigma^2/n
\end{aligned}$$

How to choose between two unbiased estimators?

Suppose $X_1, X_2, \dots X_n \sim iid$ with mean μ and variance σ^2

$$
\begin{aligned}
E[\bar{X}_n] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \mu \\
E[X_1] &= \mu \\
Var(\bar{X}_n) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \sigma^2/n \\
Var(X_1) &= \sigma^2
\end{aligned}
$$

Efficiency

Efficiency - Compare Unbiased Estimators by Variance

Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be unbiased estimators of θ_0. We say that $\widehat{\theta}_1$ is *more efficient* than $\widehat{\theta}_2$ if $Var(\widehat{\theta}_1) < Var(\widehat{\theta}_2)$.

Mean-Squared Error

Except in very simple situations, unbiased estimators are hard to come by. In fact, in many interesting applications there is a *tradeoff* between bias and variance:

- Low bias estimators often have a high variance
- Low variance estimators often have high bias

Mean-Squared Error

Except in very simple situations, unbiased estimators are hard to come by. In fact, in many interesting applications there is a *tradeoff* between bias and variance:

- Low bias estimators often have a high variance
- Low variance estimators often have high bias

Mean-Squared Error (MSE): Squared Bias plus Variance

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

Root Mean-Squared Error (RMSE): $\sqrt{\text{MSE}}$

Finite Sample versus Asymptotic Properties of Estimators

Finite Sample Properties

For *fixed sample size n* what are the properties of the sampling distribution of $\widehat{\theta}_n$? (E.g. bias and variance.)

Asymptotic Properties

What happens to the sampling distribution of $\widehat{\theta}_n$ *as the sample size n gets larger and larger?* (That is, $n \to \infty$).

Why Asymptotics?

Law of Large Numbers

Make precise what we mean by "bigger samples are better."

Central Limit Theorem

As $n \to \infty$ *pretty much any* sampling distribution is well-approximated by a normal random variable!

Consistency

Consistency

Consistency

If an estimator $\widehat{\theta}_n$ (which is a RV) *converges* to θ_0 (a constant) as $n \to \infty$, we say that *$\widehat{\theta}_n$ is consistent for θ_0*.

What does it mean for a *RV* to converge to a *constant?*

For this course we'll use *MSE Consistency*:

$$\lim_{n \to \infty} \text{MSE}(\widehat{\theta}_n) = 0$$

This makes sense since $\text{MSE}(\widehat{\theta}_n)$ is a *constant*, so this is just an ordinary limit from calculus.

Law of Large Numbers (aka Law of Averages)

Let $X_1, X_2, \ldots X_n \sim iid$ mean μ, variance σ^2. Then the sample mean

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

is consistent for the population mean μ.

Law of Large Numbers (aka Law of Averages)

Let $X_1, X_2, \ldots X_n \sim iid$ mean μ, variance σ^2.

$$
\begin{aligned}
E[\bar{X}_n] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu \\
Var(\bar{X}_n) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \sigma^2/n
\end{aligned}
$$

Law of Large Numbers (aka Law of Averages)

Let $X_1, X_2, \ldots X_n \sim iid$ mean μ, variance σ^2.

$$
\begin{aligned}
E[\bar{X}_n] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu \\
Var(\bar{X}_n) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \sigma^2/n \\
\text{MSE}(\bar{X}_n) &= \text{Bias}(\bar{X}_n)^2 + Var(\bar{X}_n)
\end{aligned}
$$

Law of Large Numbers (aka Law of Averages)

Let $X_1, X_2, \ldots X_n \sim iid$ mean μ, variance σ^2.

$$
\begin{aligned}
E[\bar{X}_n] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu \\
Var(\bar{X}_n) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \sigma^2/n \\
\text{MSE}(\bar{X}_n) &= \text{Bias}(\bar{X}_n)^2 + Var(\bar{X}_n) \\
&= \big(E[\bar{X}_n] - \mu\big)^2 + Var(\bar{X}_n)
\end{aligned}
$$

Law of Large Numbers (aka Law of Averages)

Let $X_1, X_2, \ldots X_n \sim iid$ mean μ, variance σ^2.

$$
\begin{aligned}
E[\bar{X}_n] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu \\
Var(\bar{X}_n) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \sigma^2/n \\
\mathrm{MSE}(\bar{X}_n) &= \mathrm{Bias}(\bar{X}_n)^2 + Var(\bar{X}_n) \\
&= \left(E[\bar{X}_n] - \mu\right)^2 + Var(\bar{X}_n) \\
&= 0 + \sigma^2/n \\
&\to 0
\end{aligned}
$$

Hence $\bar{X}_n$ is consistent for μ

Important!

An estimator *can* be biased but still consistent, as long as the bias disappears as $n \to \infty$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

Bias of $\widehat{\sigma}^2$

$E[\widehat{\sigma}^2] - \sigma^2 =$

Important!

An estimator *can* be biased but still consistent, as long as the bias disappears as $n \to \infty$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

Bias of $\widehat{\sigma}^2$

$$E[\widehat{\sigma}^2] - \sigma^2 = \frac{(n-1)\sigma^2}{n} - \sigma^2 =$$

Important!

An estimator *can* be biased but still consistent, as long as the bias disappears as $n \to \infty$

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

Bias of $\widehat{\sigma}^2$

$$E[\widehat{\sigma}^2] - \sigma^2 = \frac{(n-1)\sigma^2}{n} - \sigma^2 = -\sigma^2/n$$

Important!

An estimator *can* be biased but still consistent, as long as the bias disappears as $n \to \infty$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

Bias of $\widehat{\sigma}^2$

$$E[\widehat{\sigma}^2] - \sigma^2 = \frac{(n-1)\sigma^2}{n} - \sigma^2 = -\sigma^2/n \to 0$$

Suppose $X_1, X_2, \ldots, X_n \sim$ iid $N(\mu, \sigma^2)$. What is the sampling distribution of $\bar{X}_n$?

(a) $N(0, 1)$

(b) $N(\mu, \sigma^2/n)$

(c) $N(\mu, \sigma^2)$

(d) $N(\mu/n, \sigma^2/n)$

(e) $N(n\mu, n\sigma^2)$

But still, how can something random converge to something constant?

Sampling Distribution of $\bar{X}_n$ Collapses to μ

Look at an example where we can directly calculate not only the mean and variance of the sampling distribution of $\bar{X}_n$, but the *sampling distribution itself*:

$$X_1, X_2, \ldots, X_n \sim \text{iid } N(\mu, \sigma^2) \Rightarrow \bar{X}_n \sim N(\mu, \sigma^2/n)$$

Sampling Distribution of $\bar{X}_n$ Collapses to μ

$X_1, X_2, \ldots, X_n \sim \text{iid } N(\mu, \sigma^2 \Rightarrow \bar{X}_n \sim N(\mu, \sigma^2/n).$

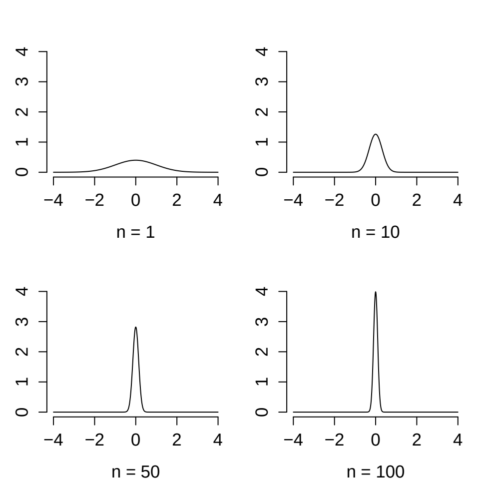
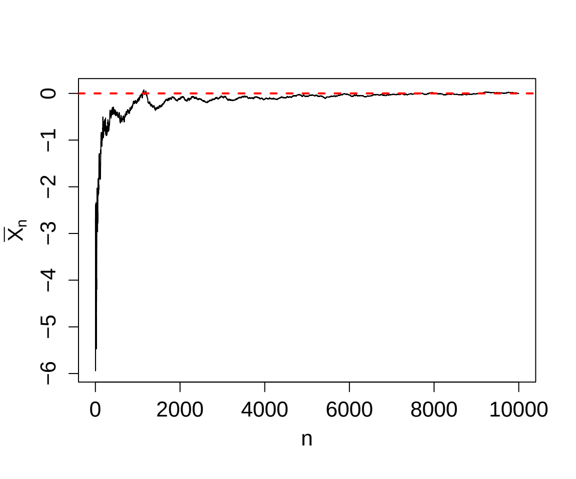


Figure 2: Sampling Distributions for $\bar{X}_n$ where $X_i \sim \text{iid } N(0, 1)$

Another Visualization: Keep Adding Observations



n	$\bar{X}_n$
1	-2.69
2	-3.18
3	-5.94
4	-4.27
5	-2.62
10	-2.89
20	-5.33
50	-2.94
100	-1.58
500	-0.45
1000	-0.13
5000	-0.05
10000	0.00

Figure 3: Running sample means: $X_i \sim$ iid $N(0, 100)$