Economics 103 – Statistics for Economists

Francis J. DiTraglia

University of Pennsylvania

Lecture # 14

Sampling Distributions and Estimation – Part I

Weighing a Random Sample

Bag Contains 100 Candies

Estimate total weight of candies by weighing a random sample of size 5 and multiplying the result by 20.

Your Chance to Win

The bag of candies and a digital scale will make their way around the room during the lecture. Each team (2 students) gets a chance to draw 5 candies and weigh them.

Team with closest estimate wins the bag of candy!

Weighing a Random Sample

Procedure

When the bag and scale reach your team, do the following:

1. Fold the top of the bag over and shake to randomize.
2. Randomly draw 5 candies without replacement.
3. Weigh your sample and record the result in grams.
4. Rodrigo will enter your result into his spreadsheet and multiply it by 20 to estimate the weight of the bag.
5. Replace your sample and shake again to re-randomize.
6. Pass bag and scale to next team.

Building a Bridge Between Probability and Statistics

Questions to Answer

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

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
Old Way	**New Way**
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.
- This means our candy experiment (in progress) isn't bogus.

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.

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

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) 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!

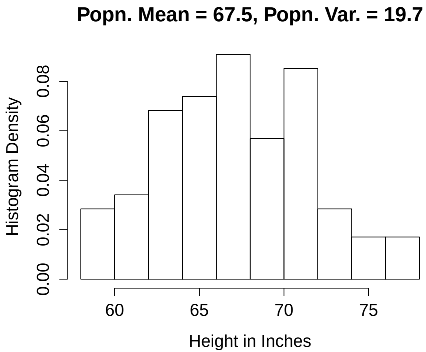
This is *Only* a Thought Experiment

- 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: Sampling from Econ 103 Class List



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

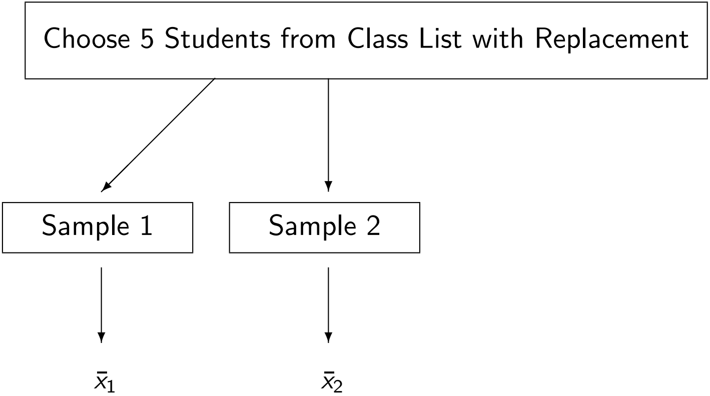
Sampling Distribution of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

Choose 5 Students from Class List with Replacement

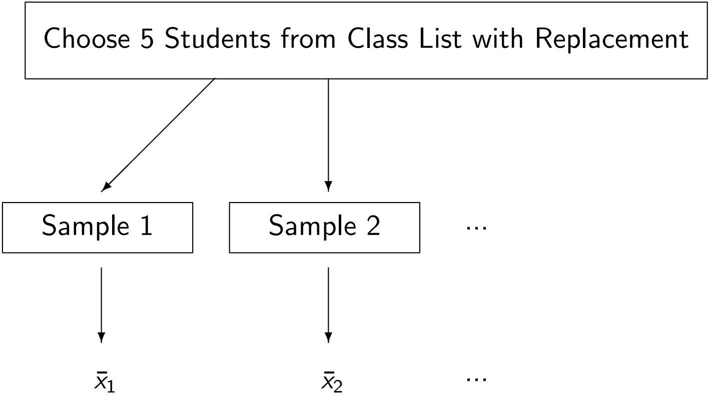
Sample 1

$\bar{x}_1$

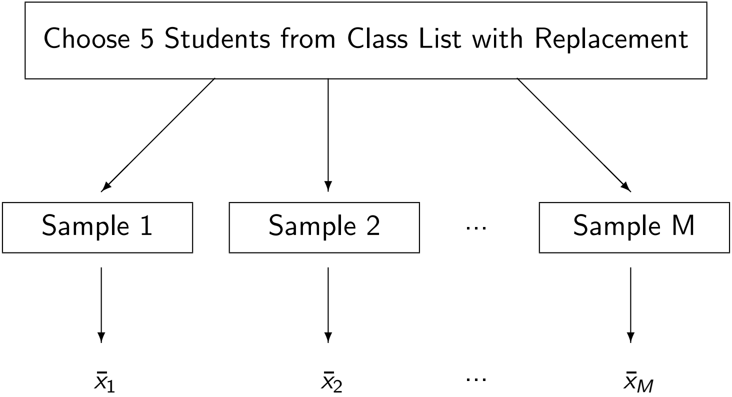
Sampling Distribution of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$



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

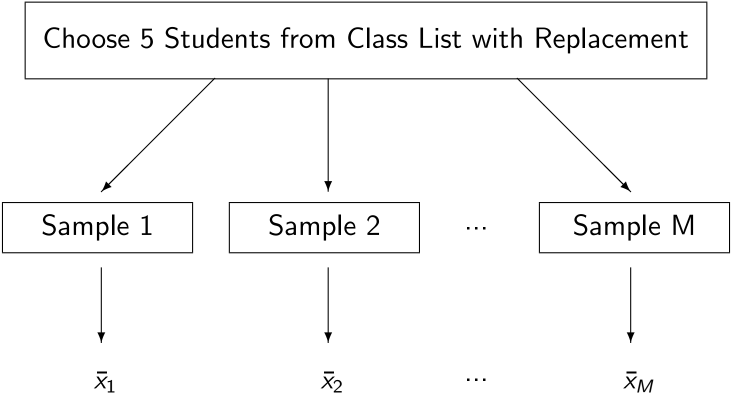


Sampling Distribution of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$



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

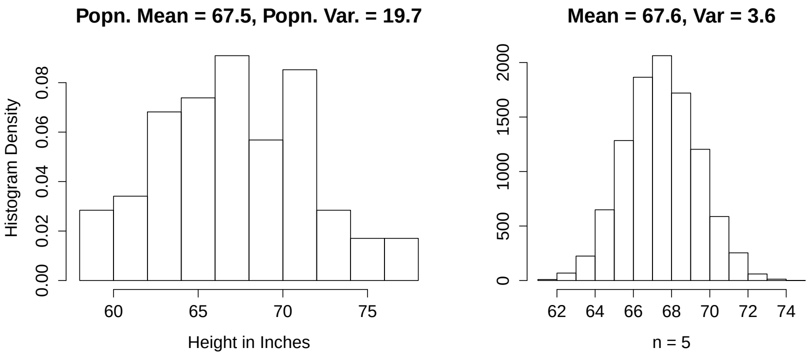
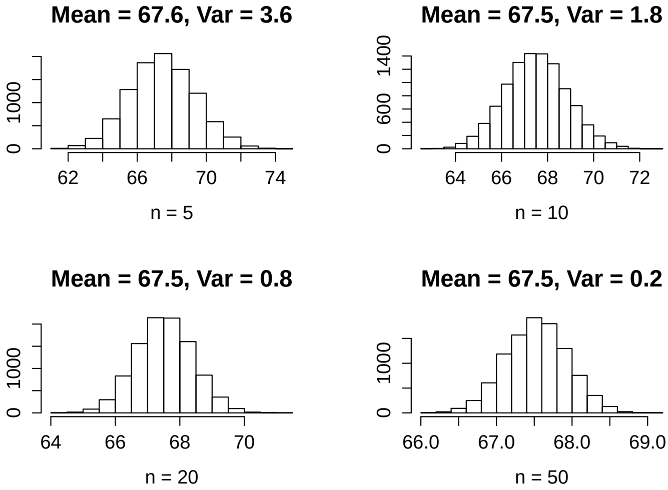


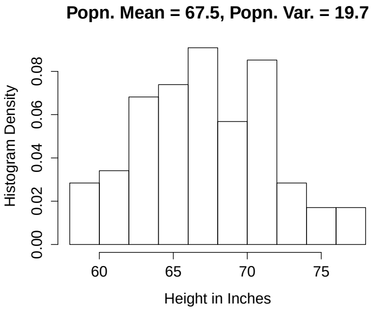
Figure: Left: Population, Right: Sampling distribution of $\bar{X}_5$

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

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

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:

Sample mean isn't the only thing with a sampling distribution!

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 of 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)

$x_1^{(1)}, \ldots, x_n^{(1)}$ $\quad$ $x_1^{(2)}, \ldots, x_n^{(2)}$ $\quad \cdots \quad$ $x_1^{(M)}, \ldots, x_n^{(M)}$

M Replications, each containing n Observations

Random Sample of Size n | Estimator

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

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

$\downarrow$

Observations (Data) | Estimate

Constants $\boxed{x_1, \ldots, x_n} \longrightarrow \boxed{T(x_1, \ldots, x_n)}$

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