Economics 103 – Statistics for Economists

Francis J. DiTraglia

University of Pennsylvania

Lecture # 14

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.

Sampling Distributions and Estimation – Part I

Building a Bridge Between Probability and Statistics

Questions to Answer

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

How We'll Proceed

1. Use sequence of iid RVs as a model for random sampling from a population.
2. Parameters of these RVs represent population parameters.
3. Use tools of probability theory to study the behavior of sample statistics.

Step 1: Random Variable as Model for Population

Treat 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

In the second example, our model assumes that the distribution of height is symmetric and bell-shaped.

Recall: (Simple) Random Sample

Definition in Words

Select a sample of n objects from a population in such a way 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

Definition 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
- But 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.

Who Will Vote for Hillary Clinton Example

Poll random sample of 1000 registered voters:

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

Heights of US Males 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 *none of them* 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 *none of them* are Hillary supporters?

$$(1/2)^4 = 1/16 = 0.0625$$

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?

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

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?

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

Population Size is Irrelevant Under Random Sampling

Though we'll see sample size is crucial.

(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.

Random Sampling

In other words:

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

is a Random Sample

Statistics

Sample is drawn randomly, so sample statistics are *also random*.

Use what we know about probability theory to analyze the *distribution* of a statistic under random sampling.

Estimator versus Estimate

Estimator

An estimator is a function $T(X_1, \ldots, X_n)$ of the random variables we use to represent the random sampling procedure. Hence, it is a random variable itself.

Estimator versus Estimate

Estimator

An estimator is a function $T(X_1, \ldots, X_n)$ of the random variables we use to represent the random sampling procedure. Hence, it is a random variable itself.

Sampling Distribution

The probability distribution of an Estimator is called a *sampling distribution*.

Estimator versus Estimate

Estimator

An estimator is a function $T(X_1, \dots, X_n)$ of the random variables we use to represent the random sampling procedure. Hence, it is a random variable itself.

Sampling Distribution

The probability distribution of an Estimator is called a *sampling distribution*.

Estimate

An estimate is a function $T(x_1, \dots, x_n)$ of the *observed data*, i.e. the *realizations* of the random variables we use to represent random sampling. An estimate is a *constant* since the observed data are *constants*

$\boxed{\text{Population: } f(x)}$ *Probability Distribution*

$\boxed{\text{Population: } f(x)}$ *Probability Distribution*

$\downarrow$

Random Sample of Size n

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

$\boxed{\text{Population: } f(x)}$ *Probability Distribution*

$\downarrow$

Random Sample of Size n

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

$\swarrow$

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

Population: $f(x)$

Probability Distribution

Random Sample of Size n

Random Variables

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

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

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

Population: $f(x)$

Probability Distribution

Random Sample of Size n

Random Variables

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

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

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)

$\cdots$

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

M Replications, each containing n Observations

Random Sample of Size n

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

Random Sample of Size n

Estimator

Random Variables

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

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

Random Sample of Size n

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

Estimator

$\boxed{T(X_1, \ldots, X_n)}$

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

Observations (Data)

Constants $\boxed{x_1, \ldots, x_n}$

Random Sample of Size n | Estimator

Random Variables

$$\begin{tikzcd} \boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)} \arrow[r] \arrow[d] & \boxed{T(X_1, \ldots, X_n)} \\ \boxed{x_1, \ldots, x_n} \arrow[r] & \boxed{T(x_1, \ldots, x_n)} \end{tikzcd}$$

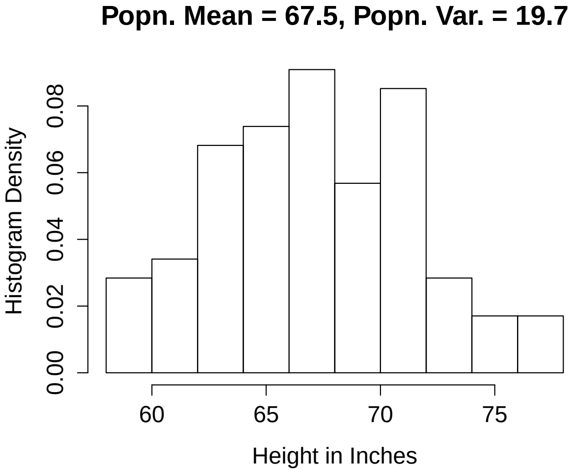
$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Observations (Data) | Estimate

Constants

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

Population: All Students in the Class



Random Sample of Size n

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

Random Sample of Size n

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

$$\downarrow$$

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

Random Sample of Size n

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

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

$$\bar{x}_1$$

Random Sample of Size n

$$
\begin{tikzcd}
 & \boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)} \arrow[dl] \arrow[d] \\
\boxed{x_1^{(1)}, \ldots, x_n^{(1)}} \arrow[d] & \boxed{x_1^{(2)}, \ldots, x_n^{(2)}} \\
\bar{x}_1 &
\end{tikzcd}
$$

Random Sample of Size n

$$
\begin{tikzcd}
 & \boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)} \arrow[dl] \arrow[d] \\
\boxed{x_1^{(1)}, \ldots, x_n^{(1)}} \arrow[d] & \boxed{x_1^{(2)}, \ldots, x_n^{(2)}} \arrow[d] \\
\bar{x}_1 & \bar{x}_2
\end{tikzcd}
$$

Random Sample of Size n

$$
\begin{tikzcd}
 & \boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)} \arrow[dl] \arrow[d] & \\
\boxed{x_1^{(1)}, \ldots, x_n^{(1)}} \arrow[d] & \boxed{x_1^{(2)}, \ldots, x_n^{(2)}} \arrow[d] & \cdots \\
\bar{x}_1 & \bar{x}_2 & \cdots
\end{tikzcd}
$$

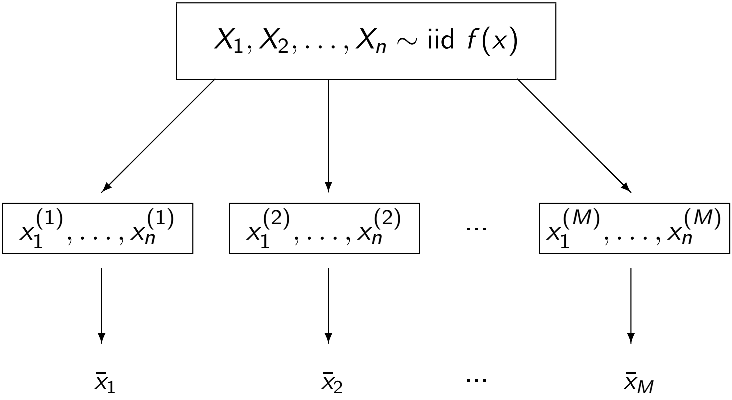
Random Sample of Size n

$$
\begin{tikzcd}
 & \boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)} \arrow[dl] \arrow[d] \arrow[drr] & & \\
\boxed{x_1^{(1)}, \ldots, x_n^{(1)}} \arrow[d] & \boxed{x_1^{(2)}, \ldots, x_n^{(2)}} \arrow[d] & \cdots & \boxed{x_1^{(M)}, \ldots, x_n^{(M)}} \\
\bar{x}_1 & \bar{x}_2 & \cdots &
\end{tikzcd}
$$

Random Sample of Size n

$$\begin{tikzcd}
 & \boxed{X_1, X_2, \ldots, X_n \sim \text{iid } f(x)} \arrow[dl] \arrow[d] \arrow[drr] & & \\
\boxed{x_1^{(1)}, \ldots, x_n^{(1)}} \arrow[d] & \boxed{x_1^{(2)}, \ldots, x_n^{(2)}} \arrow[d] & \cdots & \boxed{x_1^{(M)}, \ldots, x_n^{(M)}} \arrow[d] \\
\bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_M
\end{tikzcd}$$

Random Sample of Size n



M Replications yield M *different estimates*

Random Sample of Size n

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

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

$$\bar{x}_1 \qquad \bar{x}_2 \qquad \cdots \qquad \bar{x}_M$$

M Replications yield M *different estimates*

Sampling Distribution: Infinite Replications

Procedure versus Result of the Procedure

Procedure $=$ Random Variable

- $X_1, \ldots, X_n$ represents procedure of taking a random sample.
- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ represents procedure of taking sample mean

Procedure versus Result of the Procedure

Procedure = Random Variable

- $X_1, \ldots, X_n$ represents procedure of taking a random sample.
- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ represents procedure of taking sample mean

Sampling Dist. = Probabilistic Behavior of Procedure

If I repeat the procedure of taking the mean of a random sample over and over for many samples, what relative frequencies do I get for the sample means?

Procedure versus Result of the Procedure

Procedure = Random Variable

- $X_1, \ldots, X_n$ represents procedure of taking a random sample.
- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ represents procedure of taking sample mean

Sampling Dist. = Probabilistic Behavior of Procedure

If I repeat the procedure of taking the mean of a random sample over and over for many samples, what relative frequencies do I get for the sample means?

Result of Procedure = Constant

- $x_1, \ldots, x_n$ is the result of sampling, the observed data.
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the result of taking sample mean

Procedure? Long-Run Relative Freqencies?

Why would I advise you not to play the lottery?

Procedure? Long-Run Relative Freqencies?

Why would I advise you not to play the lottery?

- You may sometimes win, but if you play the lottery many times, on average you will lose money.

Procedure? Long-Run Relative Freqencies?

Why would I advise you not to play the lottery?

- You may sometimes win, but if you play the lottery many times, on average you will lose money.
- Let X be a random variable representing lottery winnings. I am arguing that $E[X]$ - Cost of Ticket < 0

Procedure? Long-Run Relative Freqencies?

Why would I advise you not to play the lottery?

- You may sometimes win, but if you play the lottery many times, on average you will lose money.
- Let X be a random variable representing lottery winnings. I am arguing that $E[X]$ - Cost of Ticket < 0

Procedure = Random Variable

Making a habit of playing the lottery. Expectation is negative.

Procedure? Long-Run Relative Freqencies?

Why would I advise you not to play the lottery?

- You may sometimes win, but if you play the lottery many times, on average you will lose money.
- Let X be a random variable representing lottery winnings. I am arguing that $E[X]$ - Cost of Ticket < 0

Procedure = Random Variable

Making a habit of playing the lottery. Expectation is negative.

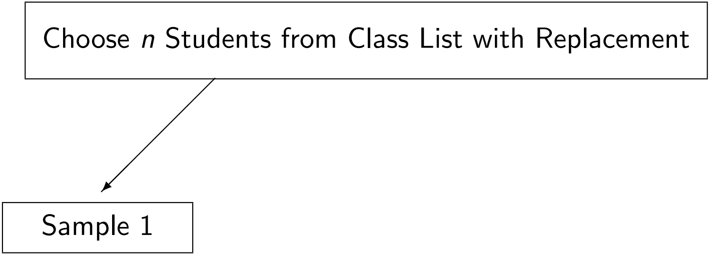
Result of that Procedure = Constant

How much you win in a *particular* lottery. Could be greater than or less than cost of ticket in any *individual* instance.

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

$$\boxed{\text{Choose } n \text{ Students from Class List with Replacement}}$$

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



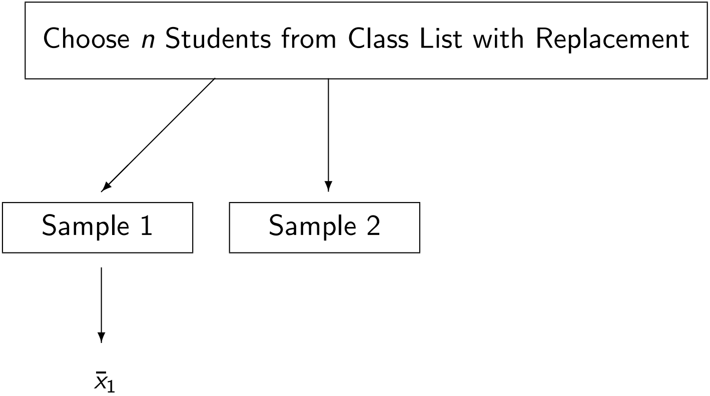
Sampling Distribution of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

Choose n Students from Class List with Replacement

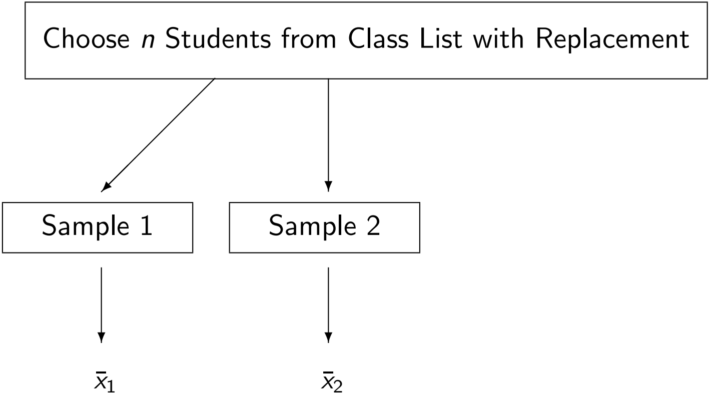
Sample 1

$\bar{x}_1$

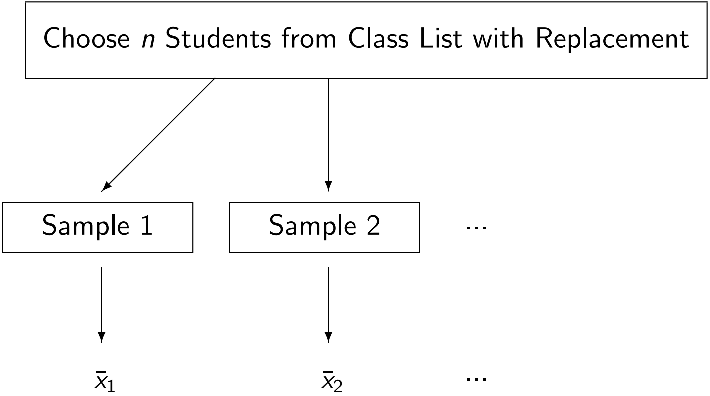
Sampling Distribution of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$



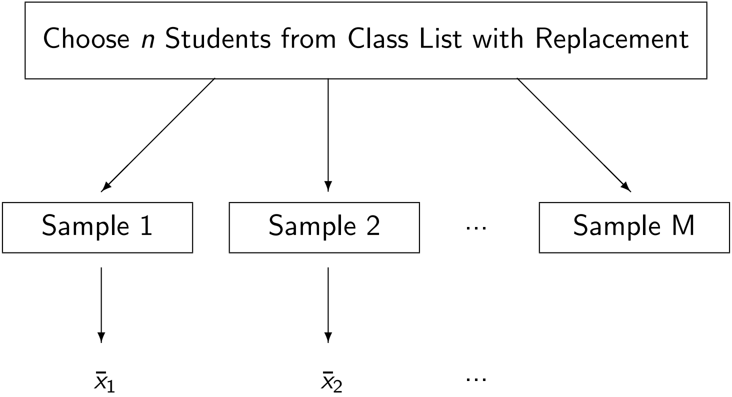
Sampling Distribution of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$



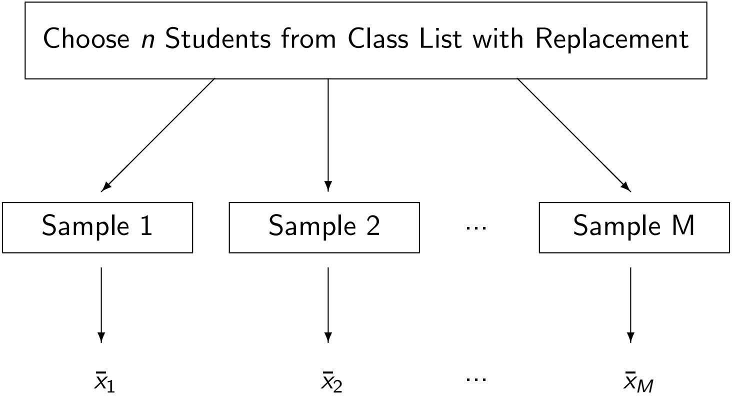
Sampling Distribution of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$



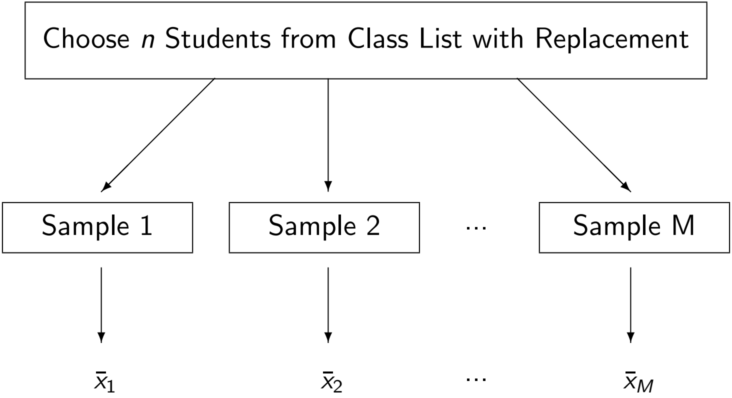
Sampling Distribution of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$



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

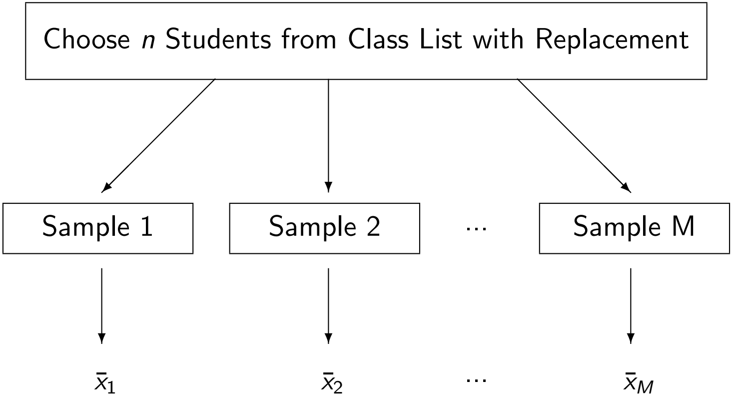


Sampling Distribution of $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$



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: long run relative frequencies of the $\bar{x}_i$

Height of Econ 103 Students

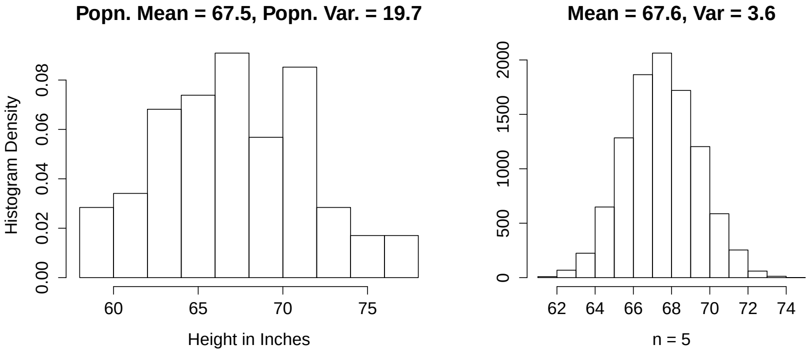
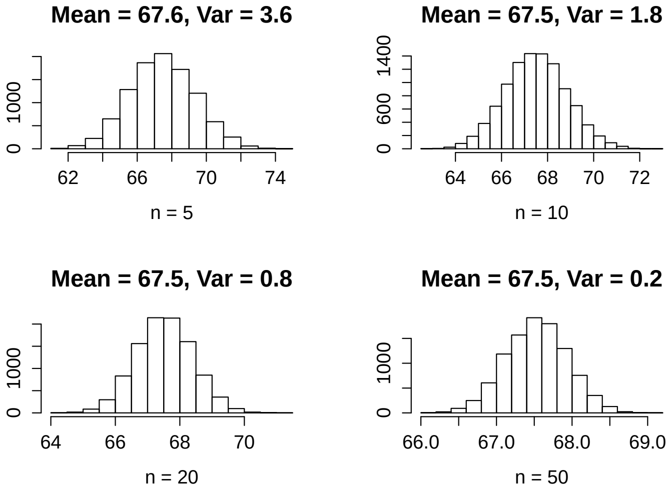


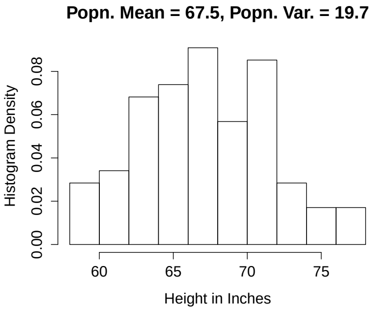
Figure : Left: Population, Right: Sampling distribution of $\bar{X}_5$

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

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

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

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

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

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

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

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

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

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

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.*

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

Calculate:

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

Standard Error

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

Standard Error of the Sample Mean

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