

# 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.

# Using a Random Variable to Model a Population

Treat Population as RV rather than list of objects

---

## Old Way

Among 138 million voters, 69 million will vote for Hillary Clinton

## New Way

Bernoulli( $p = 1/2$ ) RV

---

## Old Way

List of heights for 97 million US adult males with mean 69 in and std. dev. 6 in

## New Way

$N(\mu = 69, \sigma^2 = 36)$  RV

---

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

# 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  $\mu$ .



# (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

$X_1, X_2, \dots, X_n \sim \text{iid } f(x)$  if continuous

$X_1, X_2, \dots, X_n \sim \text{iid } p(x)$  if discrete

# Random Sampling

In other words:

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

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.

## Sample with or Without Replacement?

- ▶ Random sampling requires that we draw *with replacement*
- ▶ Without replacement  $\Rightarrow$  dependence between samples
- ▶ Does this mean our candy experiment (in progress) is bogus?

The amount of dependence created by drawing without replacement is negligible if the sample is small relative to the population.

# Estimator versus Estimate

## Estimator

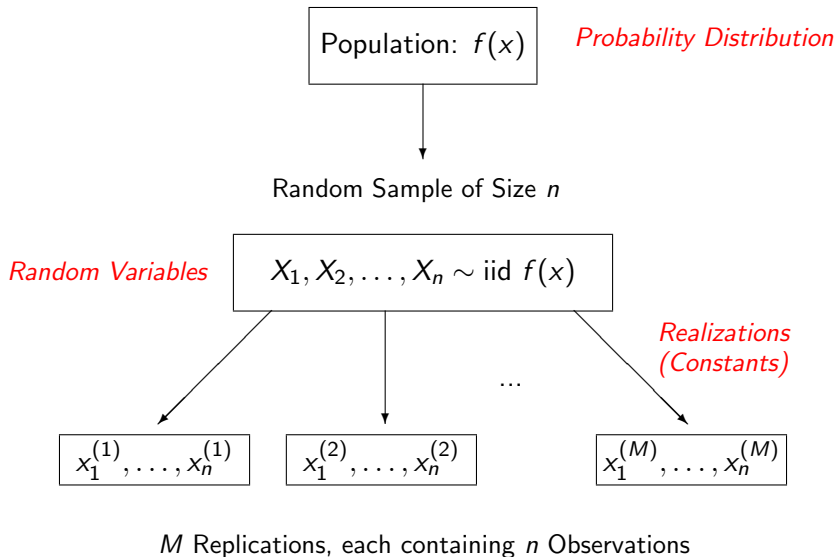
An estimator is a function  $T(X_1, \dots, X_n)$  of the random variables we use to represent the random sampling procedure. Hence, it is a random variable itself.

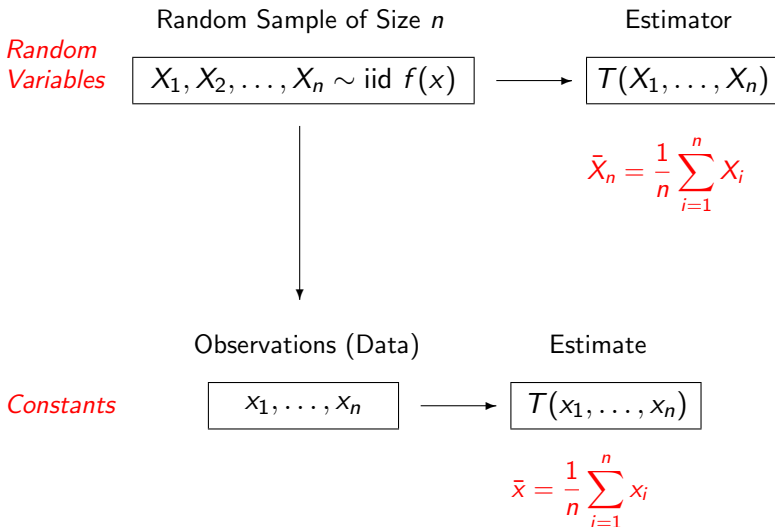
## Sampling Distribution

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

## Estimate

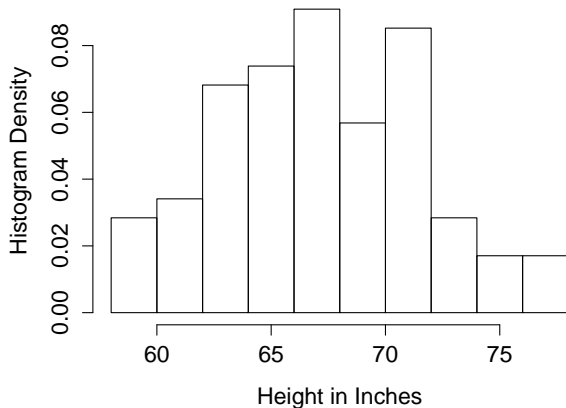
An estimate is a function  $T(x_1, \dots, x_n)$  of the *observed data*, i.e. the *realizations* of the random variables we use to represent random sampling. An estimate is a *constant* since the observed data are *constants*



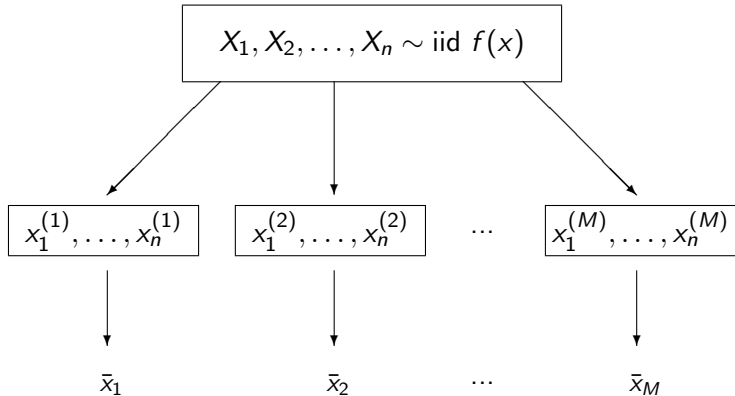


## Population: All Students in the Class

**Popn. Mean = 67.5, Popn. Var. = 19.7**



Random Sample of Size  $n$



$M$  Replications yield  $M$  different estimates

Sampling Distribution: Infinite Replications



# Procedure versus Result of the Procedure

## Procedure = Random Variable

- ▶  $X_1, \dots, X_n$  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

- ▶  $x_1, \dots, x_n$  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 Frequencies?

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  $E[X] - \text{Cost of Ticket} < 0$

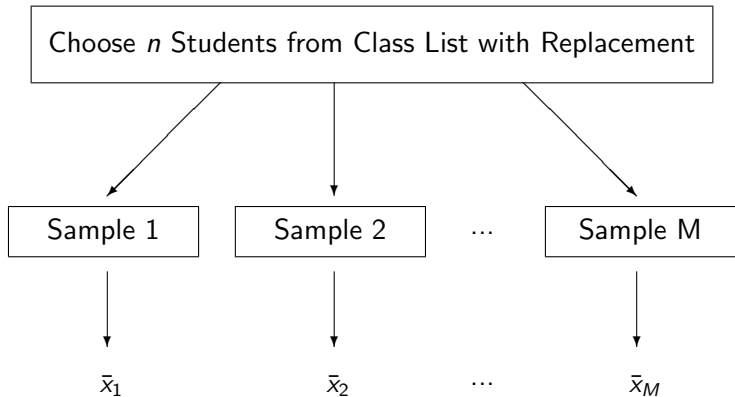
### Procedure = Random Variable

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



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

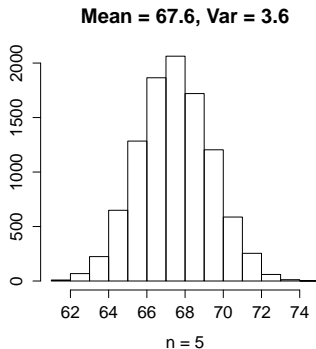
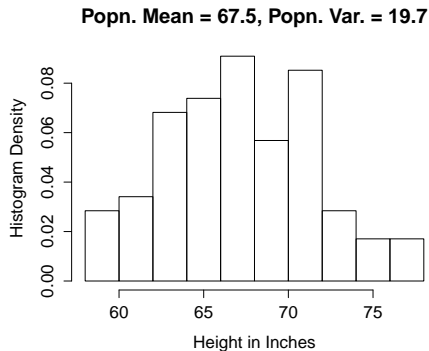
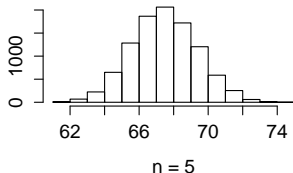


Figure : Left: Population, Right: Sampling distribution of  $\bar{X}_5$

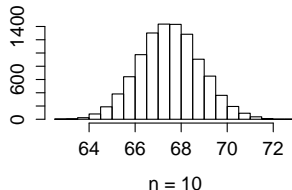
# Histograms of sampling distribution of sample mean $\bar{X}_n$

Random Sampling With Replacement, 10000 Reps. Each

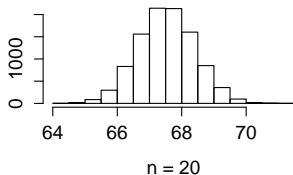
**Mean = 67.6, Var = 3.6**



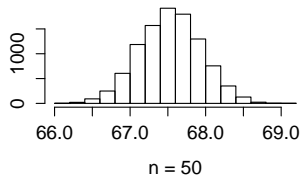
**Mean = 67.5, Var = 1.8**



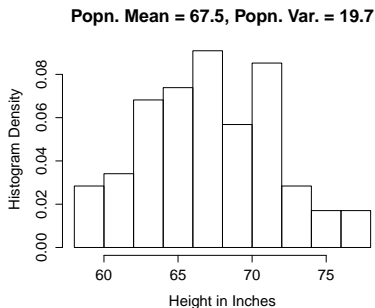
**Mean = 67.5, Var = 0.8**



**Mean = 67.5, Var = 0.2**



# 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  $n$

$X_1, \dots, X_9 \sim \text{iid}$  with  $\mu = 5$ ,  $\sigma^2 = 36$ .



Calculate:

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

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

$X_1, \dots, X_n \sim \text{iid}$  with mean  $\mu$

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

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



$X_1, \dots, X_9 \sim \text{iid}$  with  $\mu = 5$ ,  $\sigma^2 = 36$ .



Calculate:

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

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

$X_1, \dots, X_n \sim \text{iid}$  with mean  $\mu$  and variance  $\sigma^2$

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

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

$X_1, \dots, X_9 \sim \text{iid}$  with  $\mu = 5$ ,  $\sigma^2 = 36$ .



Calculate:

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

# Standard Error

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

## Standard Error of the Sample Mean

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