Economics 103 – Statistics for Economists

Rossa O'Keeffe-O'Donovan

(heavily indebted to Francis J. DiTraglia)

University of Pennsylvania

Lecture # 12

Sampling Distributions and Estimation – Part I

Remember Lecture 1

1. Descriptive Statistics: summarize data
2. Probability: Population $\rightarrow$ Sample
   - deductive: "safe" argument
     - All ravens are black. Mordecai is a raven, so Mordecai is black.
3. Statistics: Sample $\rightarrow$ Population
   - inductive: "risky" argument
     - I've only every seen black ravens, so all ravens must be black.
     - There is some uncertainty in this argument - today we will begin to quantify that uncertainty

Today: 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?

It will take many lectures to answer these questions. We'll start by "building a bridge" between probability and 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 - each object that we select is a RV

$X_1, X_2, \ldots, X_n \sim$ 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.
  - Think of a bag of n balls, exactly half blue and half red balls
  - Define a Bernoulli RV for the first draw, $X_1 = 1$ if blue, $X_1 = 0$ if red
  - Suppose $n = 10$, the first draw is $X_1 \sim \mathit{Bernoulli}(1/2)$
  - But if the first draw is red, then the second draw is $X_2 \sim \mathit{Bernoulli}(5/9)$ and X_1 and X_2 are not iid
  - However, if $n = 1,000,000$ and our sample is small, then each draw is approximately $\sim \mathit{Bernoulli}(1/2)$

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

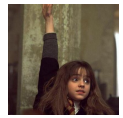
Typically used to estimate an unknown population parameter: e.g. $\bar{x}$ is an estimate of μ.

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 sample statistic is a function of the data in our sample, it is a random variable!

Remember that a function of a random variable is itself 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 sample statistic is a RV. Thus it has a PDF if continuous or a PMF if discrete: this is called its *sampling distribution*.

Sampling Dist. of Sample Mean in Polling Example

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

Some New Terminology

- Under random sampling, a statistic is a RV
- In any real example, however, once we have drawn our dataset it is *fixed* so the sample statistic we calculate is just a *constant number*
- We need some more precise terminology to make this distinction precise: Estimator vs. Estimate
- Use these terms instead of statistic from now on...

Estimator versus Estimate

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* in a particular sample, i.e. the *realizations* of the random variables we use to represent random sampling.

Since it is a function of constants, an estimate is itself a constant.

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?

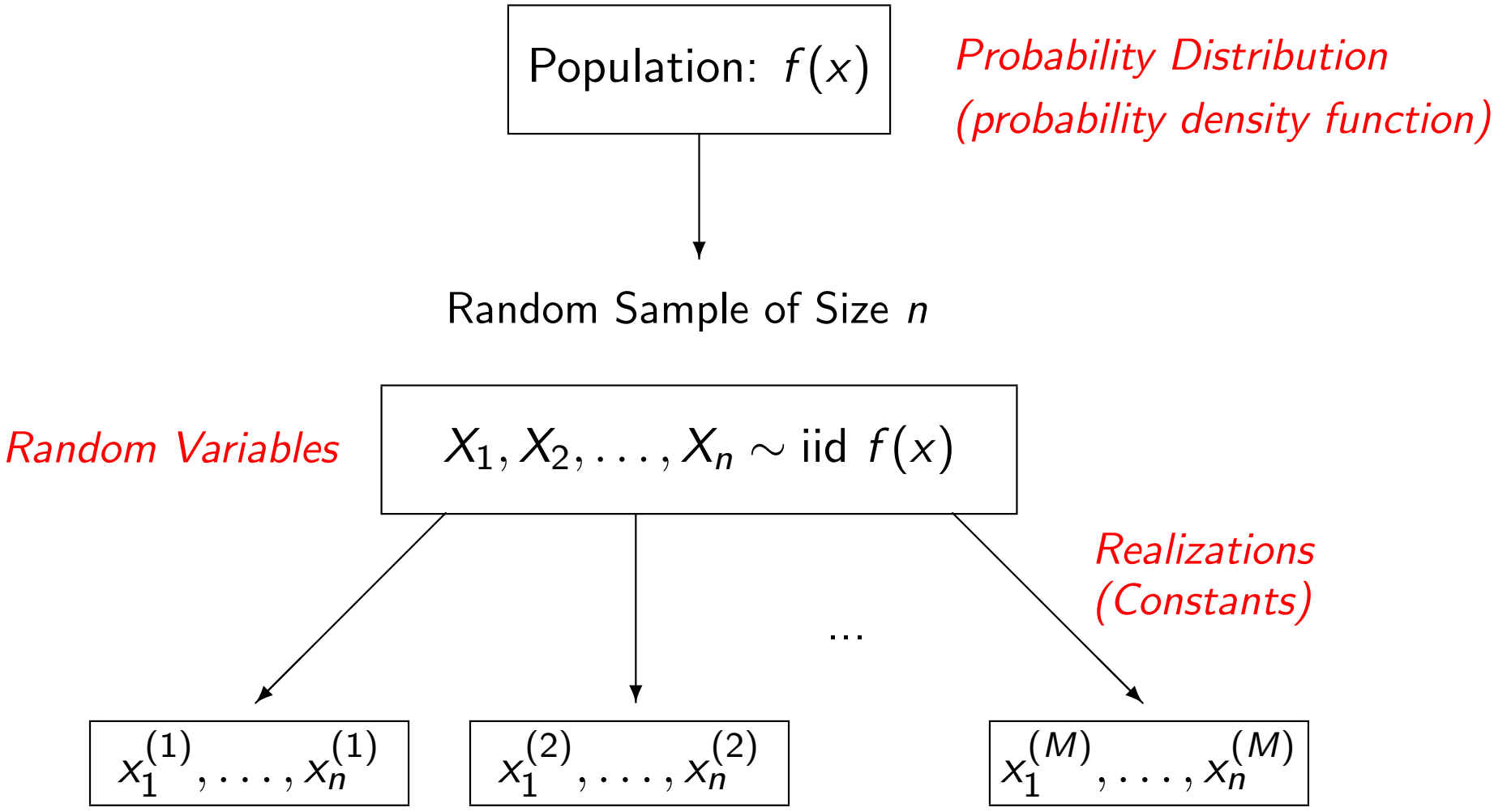
- You may sometimes win, but if you play the lottery many times, on average you will lose money.
- Let X be a random variable representing lottery winnings. I am arguing that $E[X]$ - Cost of Ticket < 0

Procedure = Random Variable

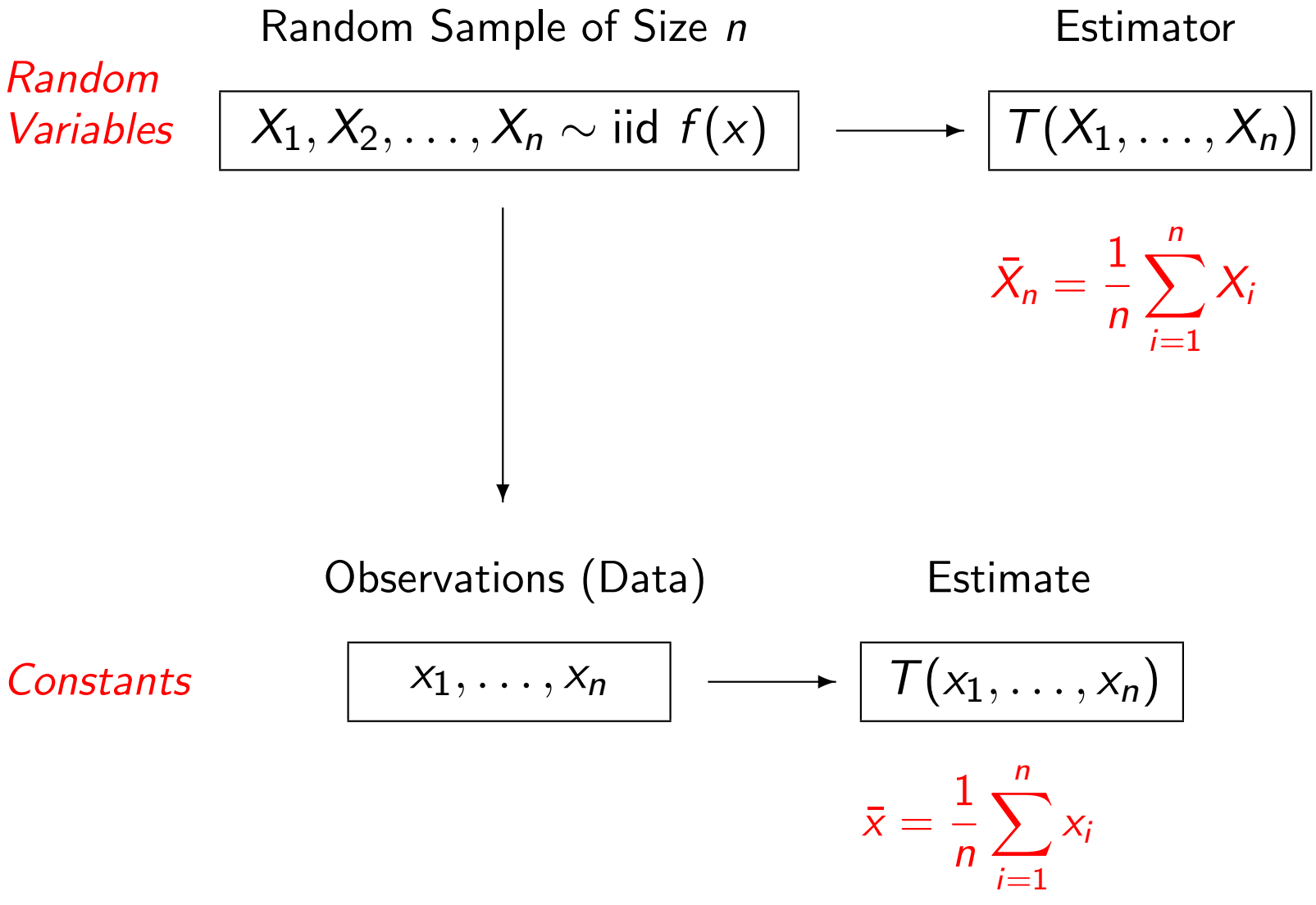
Making a habit of playing the lottery. Expectation is negative.

Result of that Procedure = Constant

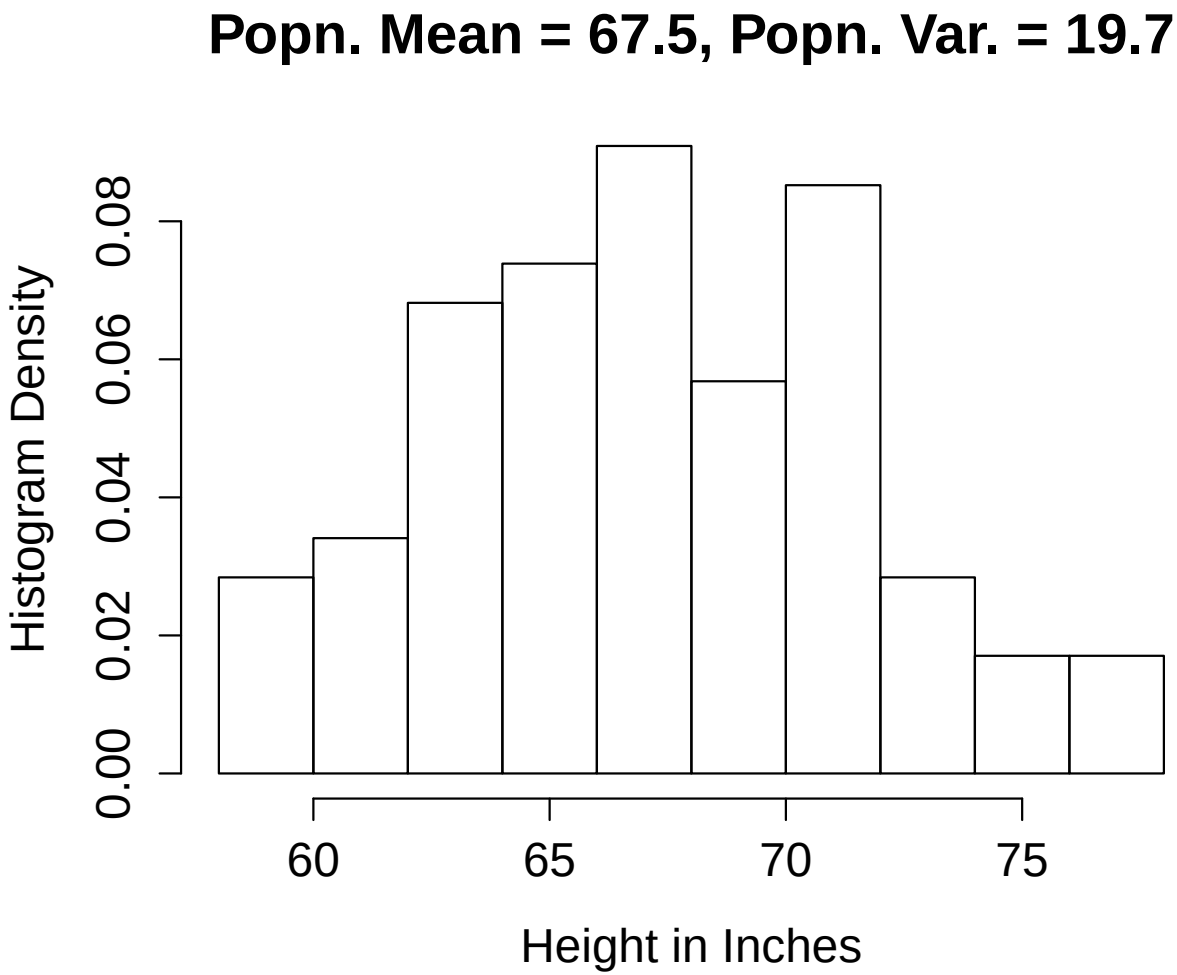
How much you win in a *particular* lottery. Could be greater than or less than cost of ticket in any *individual* instance.



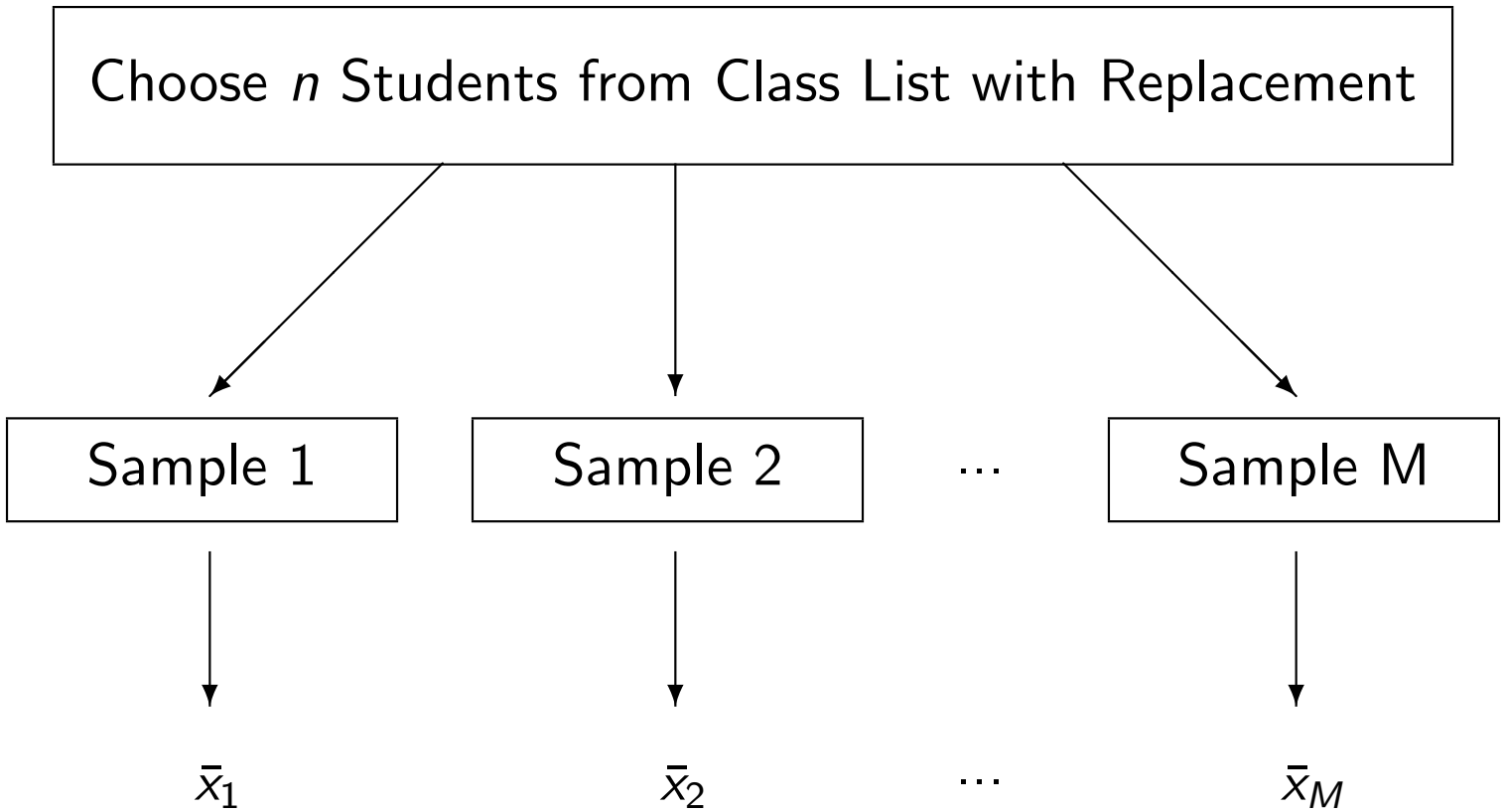
M Replications (samples), each containing n Observations


Random Sample of Size n
Estimator
Random Variables
$X_1, X_2, \ldots, X_n \sim$ iid $f(x)$
$T(X_1, \ldots, X_n)$
$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
Observations (Data)
Estimate
Constants
$x_1, \ldots, x_n$
$T(x_1, \ldots, x_n)$
$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Population: All Students in the Class



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

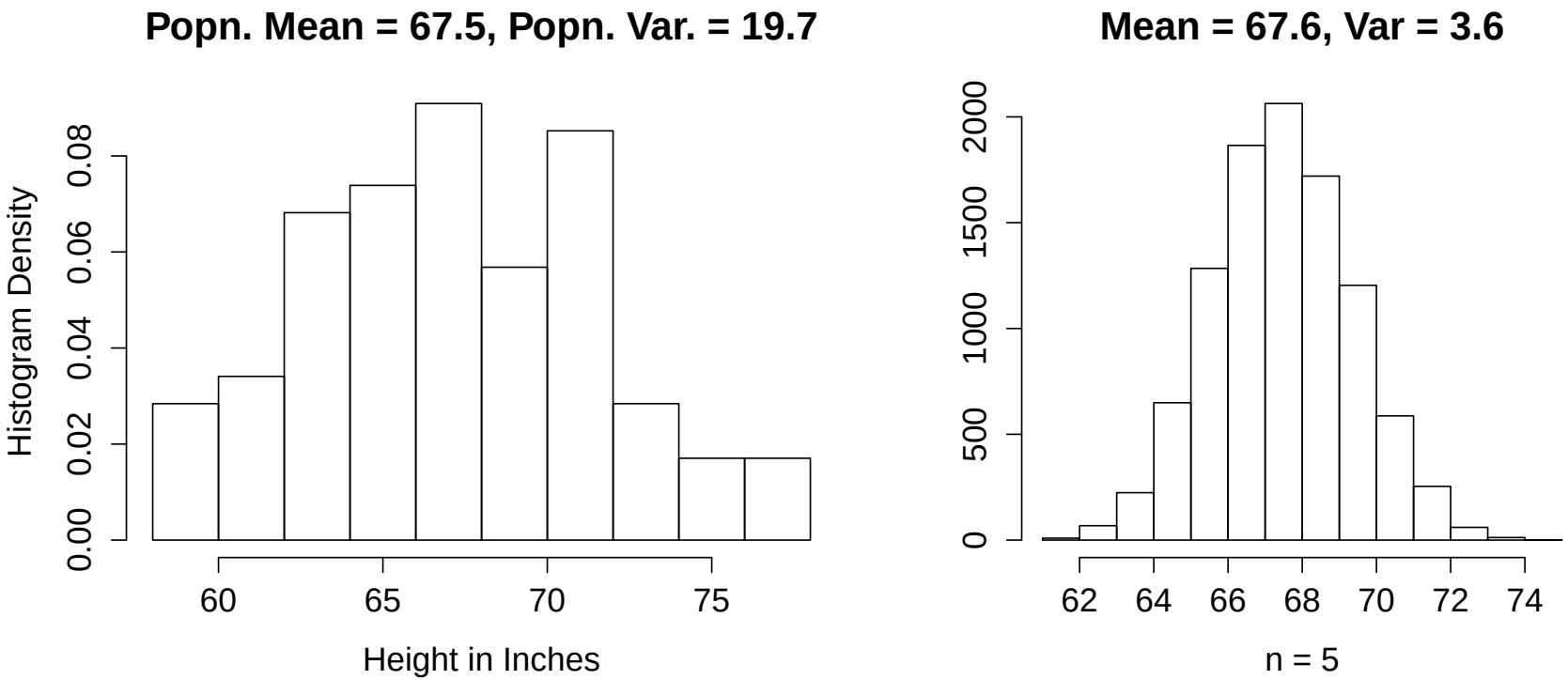
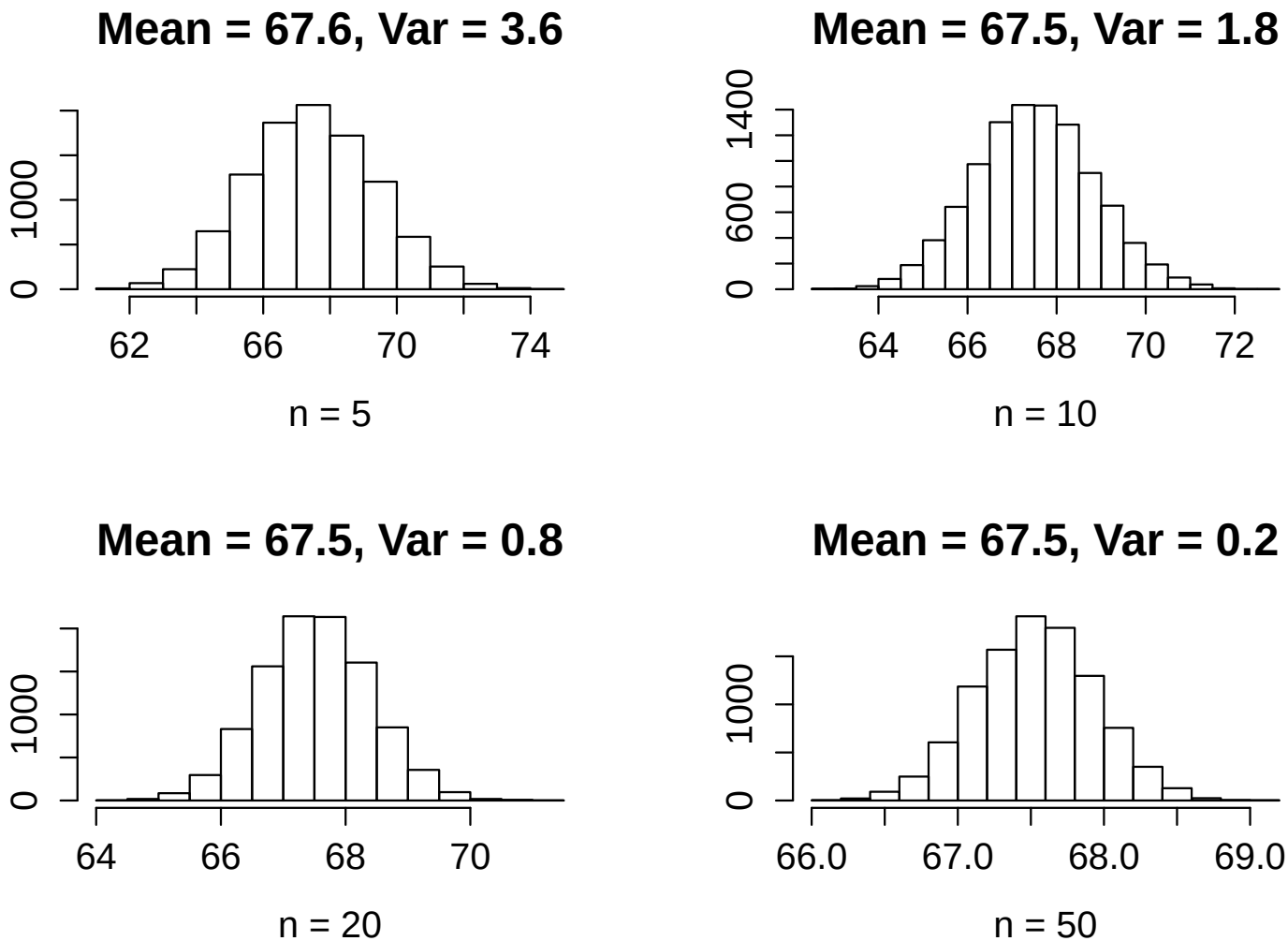


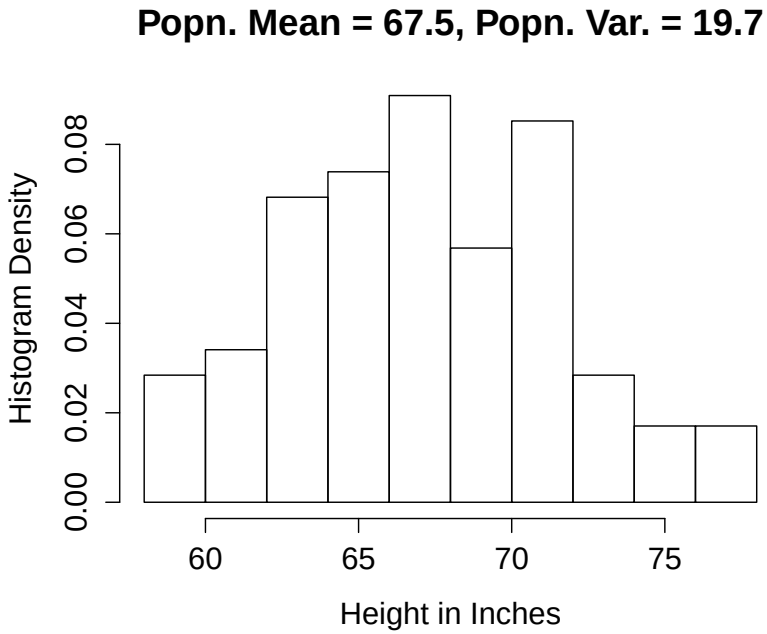
Figure: Left: distribution of the population. Right: sampling distribution of $\bar{X}_5$

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

Random Sampling With Replacement, 10000 Reps. Each



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

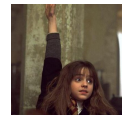


	Sampling Dist. of $\bar{X}_n$	
n	Mean	Variance
5	67.6	3.6
10	67.5	1.8
20	67.5	0.8
50	67.5	0.2

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:

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

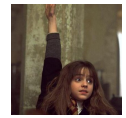
Mean of Sampling Distribution of $\bar{X}_n$

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

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

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

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

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

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