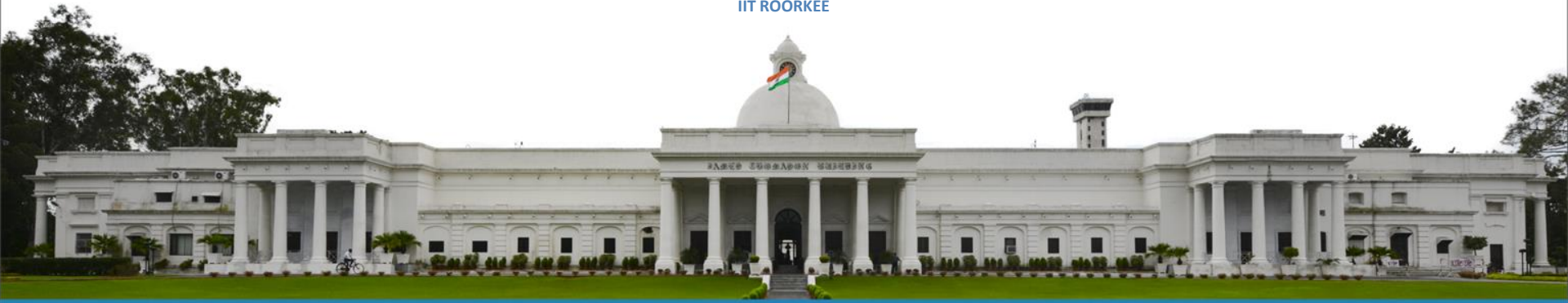
IIT ROORKEE

FREE ONLINE EDUCATION swayam शिक्षित भारत, उन्नत भारत

NPTEL ONLINE CERTIFICATION COURSE

Lecture 11: Sampling and Sampling Distribution

Dr. A. Ramesh
DEPARTMENT OF MANAGEMENT STUDIES
IIT ROORKEE



Lecture Objectives

After completing this lecture, you should be able to:

- Describe a simple random sample and why sampling is important
- Explain the difference between descriptive and inferential statistics
- Define the concept of a sampling distribution
- Determine the mean and standard deviation for the sampling distribution of the sample mean,

Lecture Objectives

- Describe the Central Limit Theorem and its importance
- Determine the mean and standard deviation for the sampling distribution of the sample proportion,
- Describe sampling distributions of sample variances

Descriptive vs Inferential Statistics

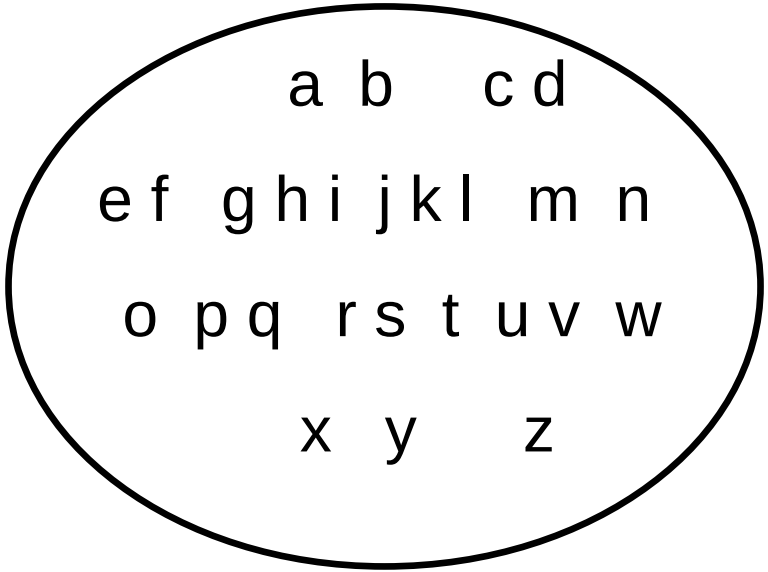
- **Descriptive statistics**
  - Collecting, presenting, and describing data
- **Inferential statistics**
  - Drawing conclusions and/or making decisions concerning a population based only on sample data

Populations and Samples

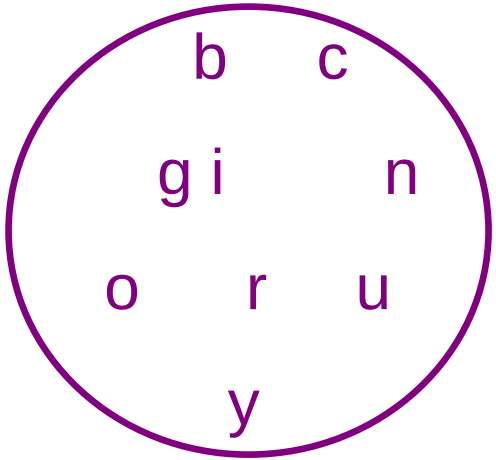
- A **Population** is the set of all items or individuals of interest
  - Examples: All likely voters in the next election
    All parts produced today
    All sales receipts for November
- A **Sample** is a subset of the population
  - Examples: 1000 voters selected at random for interview
    A few parts selected for destructive testing
    Random receipts selected for audit

Population vs. Sample

- **Population**



- **Sample**



Why Sample?

- Less time consuming than a census
- Less costly to administer than a census
- It is possible to obtain statistical results of a sufficiently high precision based on samples.
- Because the research process is sometimes destructive, the sample can save product
- If accessing the population is impossible; sampling is the only option

Reasons for Taking a Census

- Eliminate the possibility that a random sample is not representative of the population
- The person authorizing the study is uncomfortable with sample information

Random Versus Nonrandom Sampling

- **Random sampling**
  - Every unit of the population has the same probability of being included in the sample.
  - A chance mechanism is used in the selection process.
  - Eliminates bias in the selection process
  - Also known as probability sampling
- **Nonrandom Sampling**
  - Every unit of the population does not have the same probability of being included in the sample.
  - Open the selection bias
  - Not appropriate data collection methods for most statistical methods
  - Also known as non-probability sampling

Random Sampling Techniques

- Simple Random Sample
- Stratified Random Sample
  - Proportionate
  - Disproportionate
- Systematic Random Sample
- Cluster (or Area) Sampling

Simple Random Samples

- Every object in the population has an equal chance of being selected
- Objects are selected independently
- Samples can be obtained from a table of random numbers or computer random number generators
- A simple random sample is the ideal against which other sample methods are compared

Simple Random Sample: Numbered Population Frame

01 Andhra Pradesh
02 Himachal Pradesh
03 Gujrath
04 Maharashtra
05 Nagaland
06 Goa
07 West bengal
08 Haryana
09 Punjab
10 Delhi

11 Madhya Pradesh
12 Uttar Pradesh
13 Bihar
14 Rajasthan
15 J & K
16 Tamil Nadu
17 Karantaka
18 Kerala
19 Orissa
20 Manipur

Simple Random Sampling: Random Number Table

9	9	4	3	7	8	7	9	6	1	4	5	7	3	7	3	7	5	5	2	9	7	9	6	9	3	9	0	9	4	3	4	4	7	5	3	1	6	1	8
5	0	6	5	6	0	0	1	2	7	6	8	3	6	7	6	6	8	8	2	0	8	1	5	6	8	0	0	1	6	7	8	2	2	4	5	8	3	2	6
8	0	8	8	0	6	3	1	7	1	4	2	8	7	7	6	6	8	3	5	6	0	5	1	5	7	0	2	9	6	5	0	0	2	6	4	5	5	8	7
8	6	4	2	0	4	0	8	5	3	5	3	7	9	8	8	9	4	5	4	6	8	1	3	0	9	1	2	5	3	8	8	1	0	4	7	4	3	1	9
6	0	0	9	7	8	6	4	3	6	0	1	8	6	9	4	7	7	5	8	8	9	5	3	5	9	9	4	0	0	4	8	2	6	8	3	0	6	0	6
5	2	5	8	7	7	1	9	6	5	8	5	4	5	3	4	6	8	3	4	0	0	9	9	1	9	9	7	2	9	7	6	9	4	8	1	5	9	4	1
8	9	1	5	5	9	0	5	5	3	9	0	6	8	9	4	8	6	3	7	0	7	9	5	5	4	7	0	6	2	7	1	1	8	2	6	4	4	9	3

Simple Random Sample: Sample Members

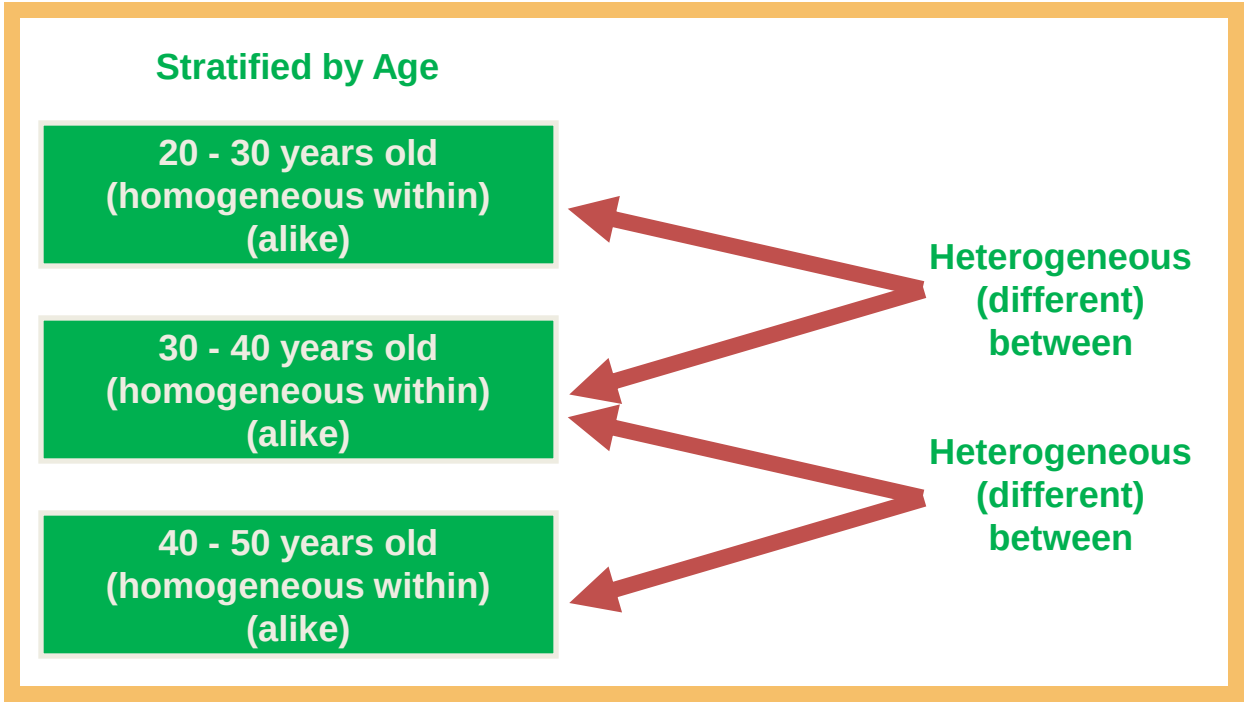
01 Andhra Pradesh	11 Madhya Pradesh
02 Himachal Pradesh	12 Uttar Pradesh
03 Gujrath	13 Bihar
04 Maharashtra	14 Rajasthan
05 Nagaland	15 J & K
06 Goa	16 Tamil Nadu
07 West bengal	17 Karantaka
08 Haryana	18 Kerala
09 Punjab	19 Orissa
10 Delhi	20 Manipur

- $N = 20$
- $n = 4$

Stratified Random Sample

- Population is divided into non-overlapping subpopulations called strata
- A random sample is selected from each stratum
- Potential for reducing sampling error
- Proportionate -- the percentage of these sample taken from each stratum is proportionate to the percentage that each stratum is within the population
- Disproportionate -- proportions of the strata within the sample are different than the proportions of the strata within the population

Stratified Random Sample: Population of FM Radio Listeners



Systematic Sampling

- Convenient and relatively easy to administer
- Population elements are an ordered sequence (at least, conceptually).
- The first sample element is selected randomly from the first *k* population elements.
- Thereafter, sample elements are selected at a constant interval, *k*, from the ordered sequence frame.

$$k = \frac{N}{n},$$

where:

n = sample size

N = population size

k = size of selection interval

Systematic Sampling: Example

- Purchase orders for the previous fiscal year are serialized 1 to 10,000 (N = 10,000).
- A sample of fifty (n = 50) purchases orders is needed for an audit.
- k = 10,000/50 = 200
- First sample element randomly selected from the first 200 purchase orders. Assume the 45th purchase order was selected.
- **Subsequent sample elements: 245, 445, 645, . . .**

Cluster Sampling

- Population is divided into non-overlapping clusters or areas
- Each cluster is a miniature of the population.
- A subset of the clusters is selected randomly for the sample.
- If the number of elements in the subset of clusters is larger than the desired value of *n*, these clusters may be subdivided to form a new set of clusters and subjected to a random selection process.

Cluster Sampling

- **Advantages**
  - More convenient for geographically dispersed populations
  - Reduced travel costs to contact sample elements
  - Simplified administration of the survey
  - Unavailability of sampling frame prohibits using other random sampling methods
- **Disadvantages**
  - Statistically less efficient when the cluster elements are similar
  - Costs and problems of statistical analysis are greater than for simple random sampling

Nonrandom Sampling

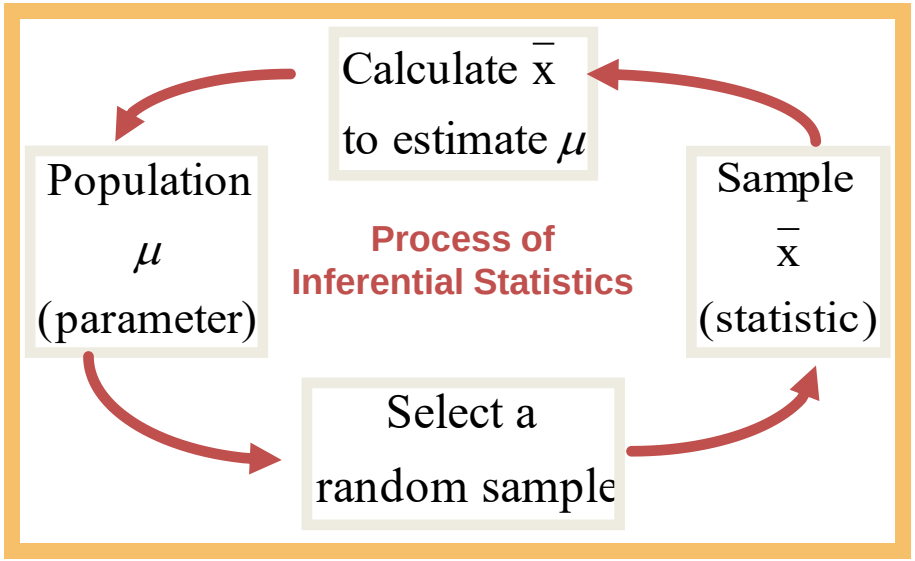
- **Convenience Sampling:** Sample elements are selected for the convenience of the researcher
- **Judgment Sampling:** Sample elements are selected by the judgment of the researcher
- **Quota Sampling:** Sample elements are selected until the quota controls are satisfied
- **Snowball Sampling:** Survey subjects are selected based on referral from other survey respondents

Errors

- Data from nonrandom samples are not appropriate for analysis by inferential statistical methods.
- Sampling Error occurs when the sample is not representative of the population
- Non-sampling Errors
  - Missing Data, Recording, Data Entry, and Analysis Errors
  - Poorly conceived concepts , unclear definitions, and defective questionnaires
  - Response errors occur when people so not know, will not say, or overstate in their answers

Sampling Distribution of $\bar{x}$

Proper analysis and interpretation of a sample statistic requires knowledge of its distribution.

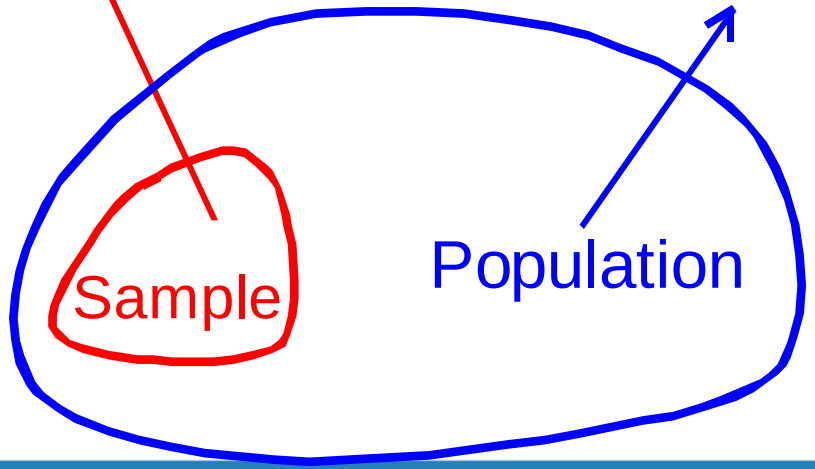


Inferential Statistics

- Making statements about a population by examining sample results

Sample statistics (known) → Population parameters (unknown, but can be estimated from sample evidence)

Inference



Inferential Statistics

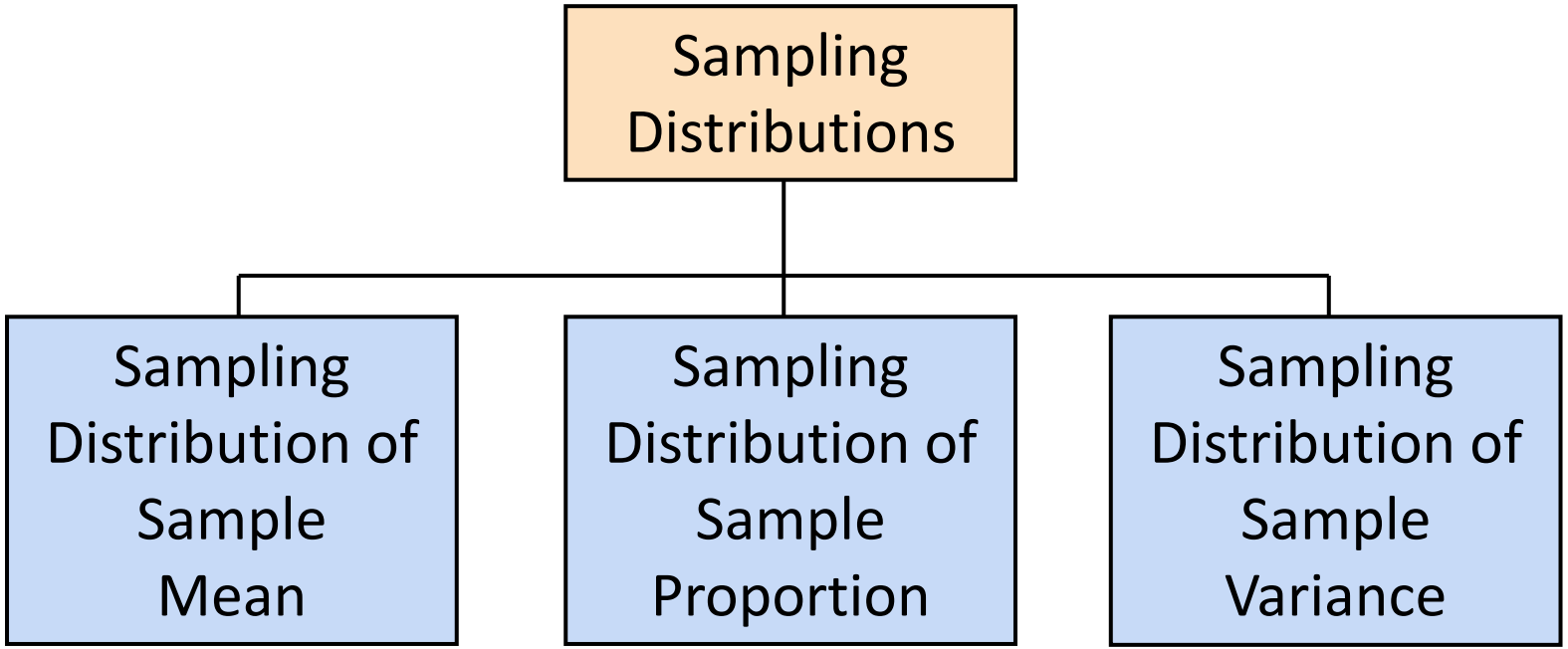
Drawing conclusions and/or making decisions concerning a population based on sample results.

- **Estimation**
  - e.g., Estimate the population mean weight using the sample mean weight
- **Hypothesis Testing**
  - e.g., Use sample evidence to test the claim that the population mean weight is 120 pounds

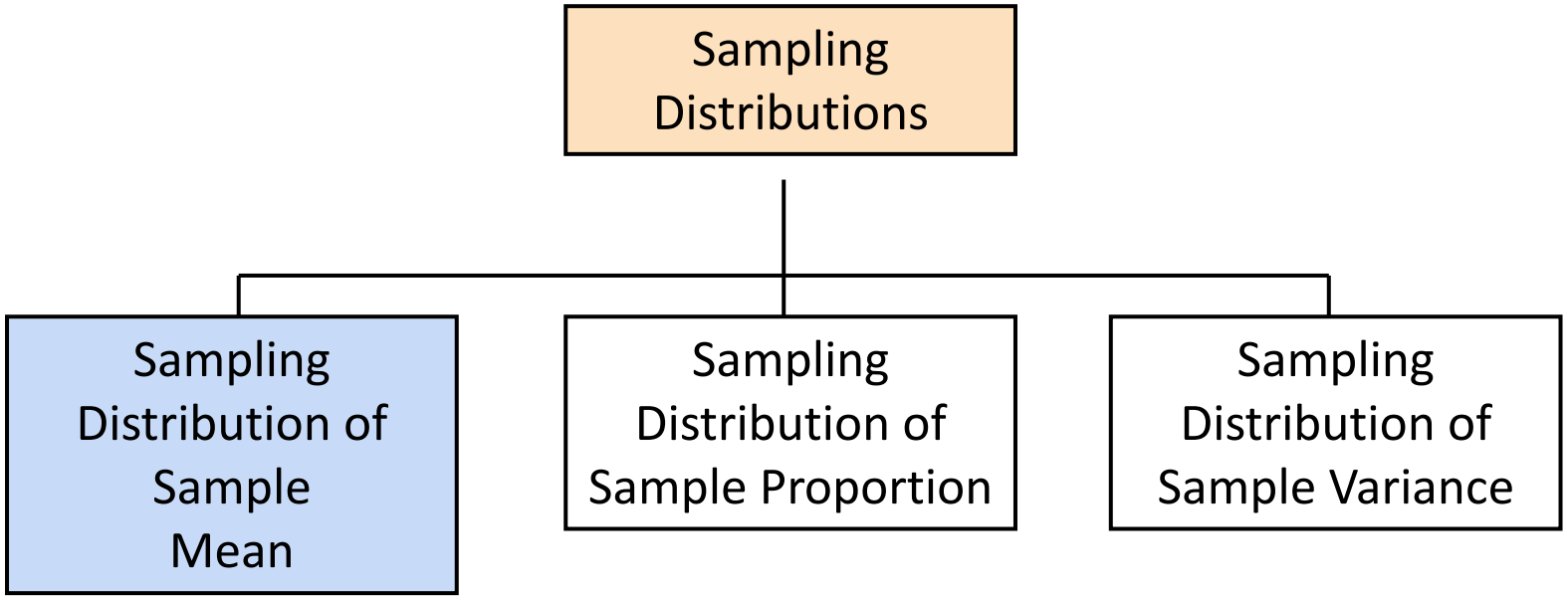
Sampling Distributions

- A sampling distribution is a distribution of all of the possible values of a statistic for a given size sample selected from a population

Types of sampling distributions



Sampling Distributions of Sample Means



Developing a Sampling Distribution

- **Assume there is a population …**
- Population size N=4
- Random variable, X,
  is age of individuals
- Values of X:

  18, 20, 22, 24 (years)

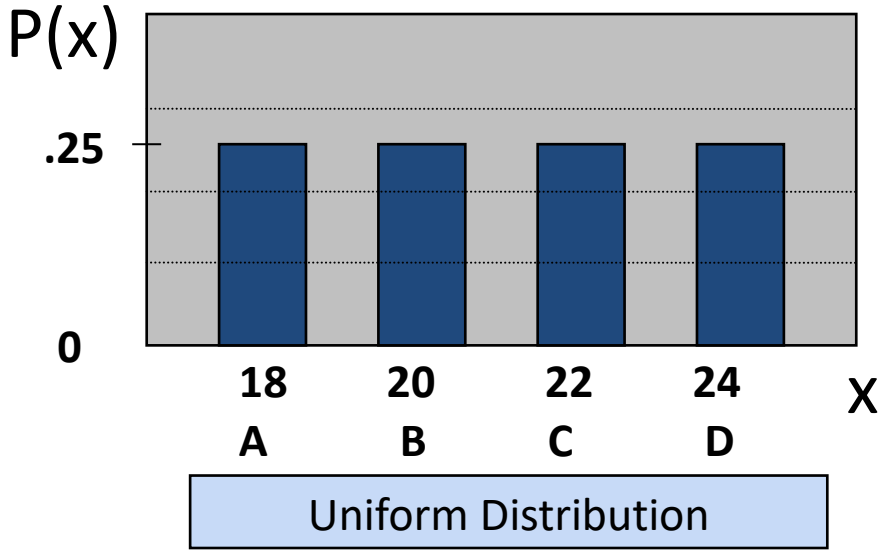
Developing a Sampling Distribution *(continued)*

Summary Measures for the Population Distribution:

$$\mu = \frac{\sum X_i}{N}$$

$$= \frac{18 + 20 + 22 + 24}{4} = 21$$

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}} = 2.236$$



Developing a Sampling Distribution *(continued)*

Now consider all possible samples of size n = 2

1st Obs	2nd Observation			
	18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

16 possible samples (sampling with replacement)

16 Sample Means

1st Obs	2nd Observation			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

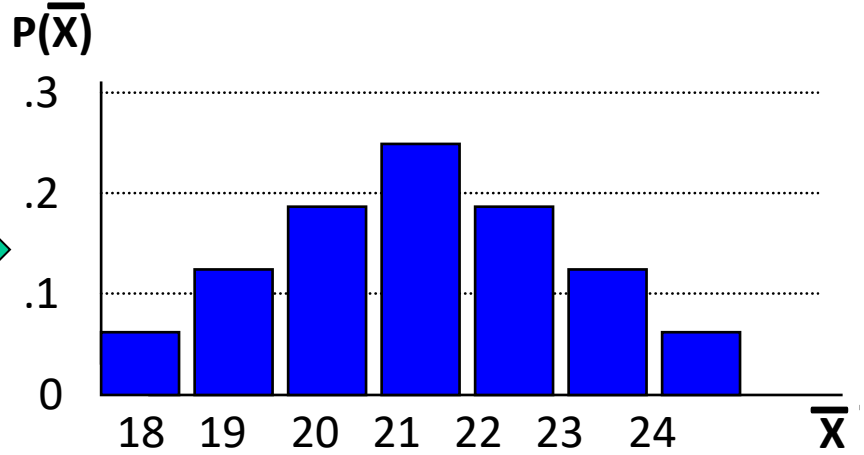
Developing a Sampling Distribution *(continued)*

- Sampling Distribution of All Sample Means

16 Sample Means

1st Obs	2nd Observation 18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Sample Means Distribution



(no longer uniform)

Developing a Sampling Distribution *(continued)*

- Summary Measures of this Sampling Distribution:

$$E(\bar{X}) = \frac{\sum \bar{X}_i}{N} = \frac{18+19+21+\cdots+24}{16} = 21 = \mu$$

$$\sigma_{\bar{X}} = \sqrt{\frac{\sum(\bar{X}_i - \mu)^2}{N}}$$

$$= \sqrt{\frac{(18-21)^2+(19-21)^2+\cdots+(24-21)^2}{16}} = 1.58$$

Comparing the Population with its Sampling Distribution

Population

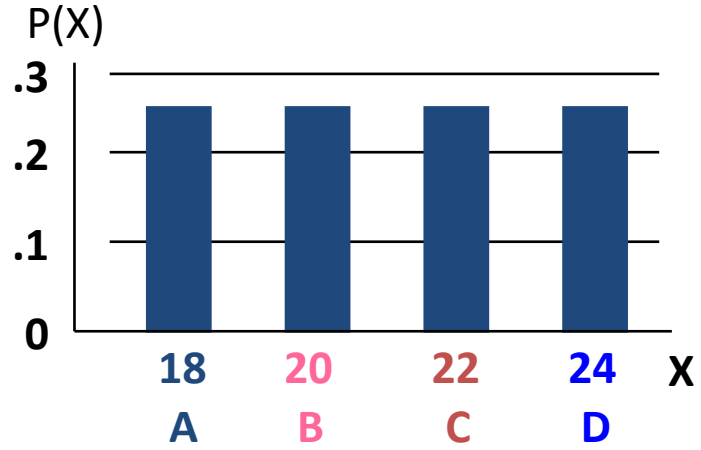
N = 4

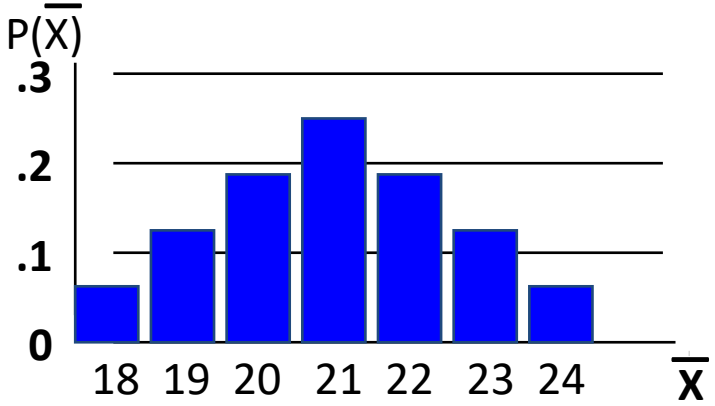
$\mu = 21 \quad \sigma = 2.236$

Sample Means Distribution

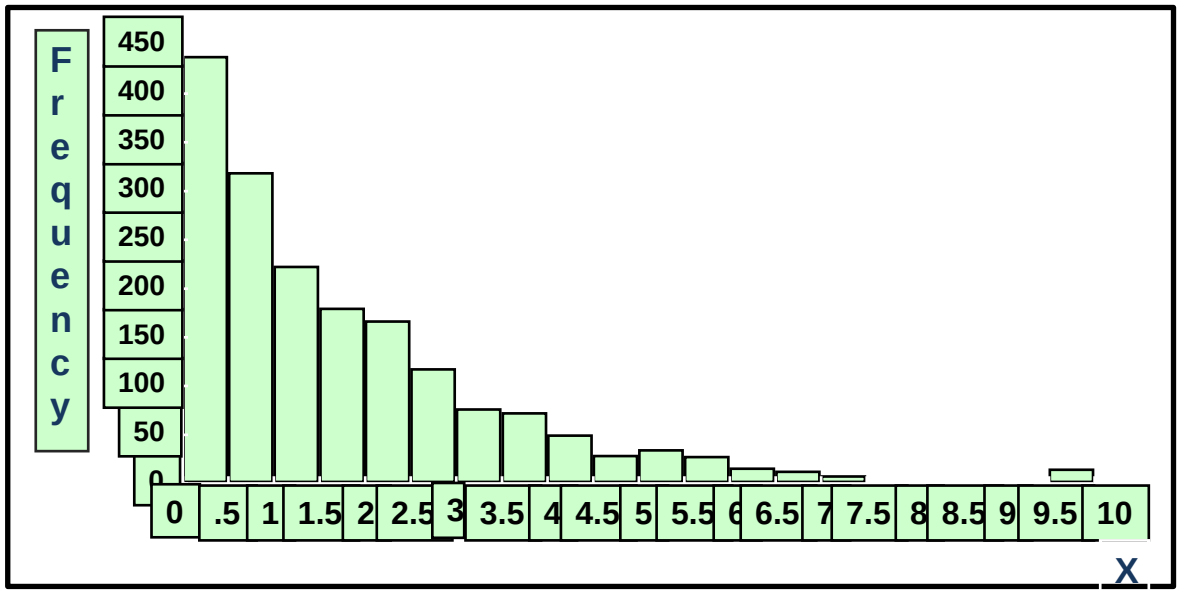
n = 2

$\mu_{\bar{X}} = 21 \quad \sigma_{\bar{X}} = 1.58$

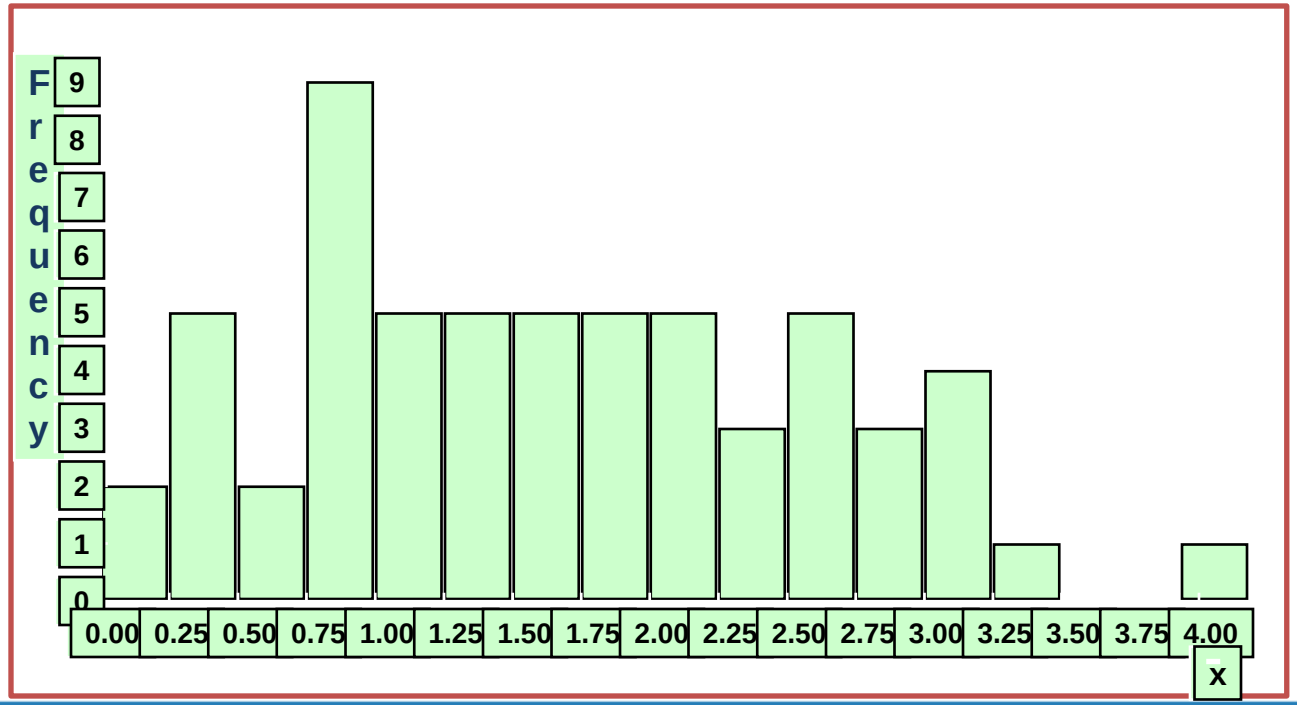




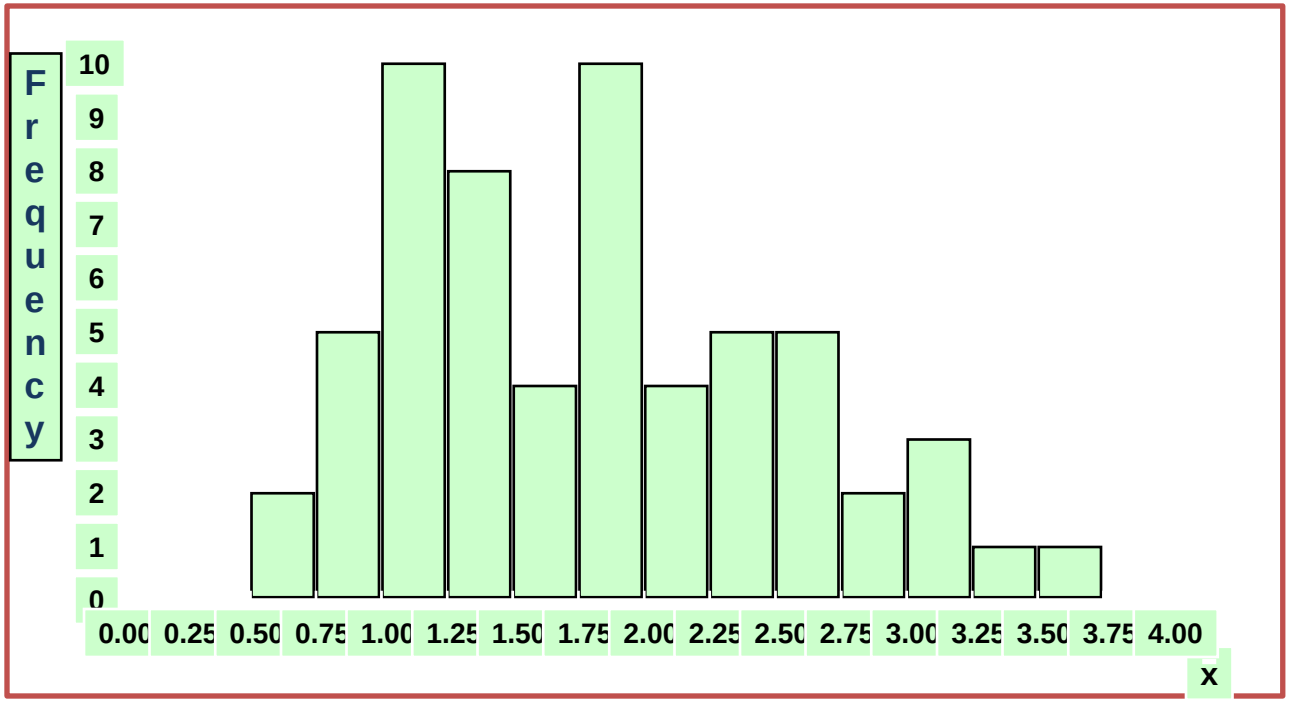
1,800 Randomly Selected Values from an Exponential Distribution



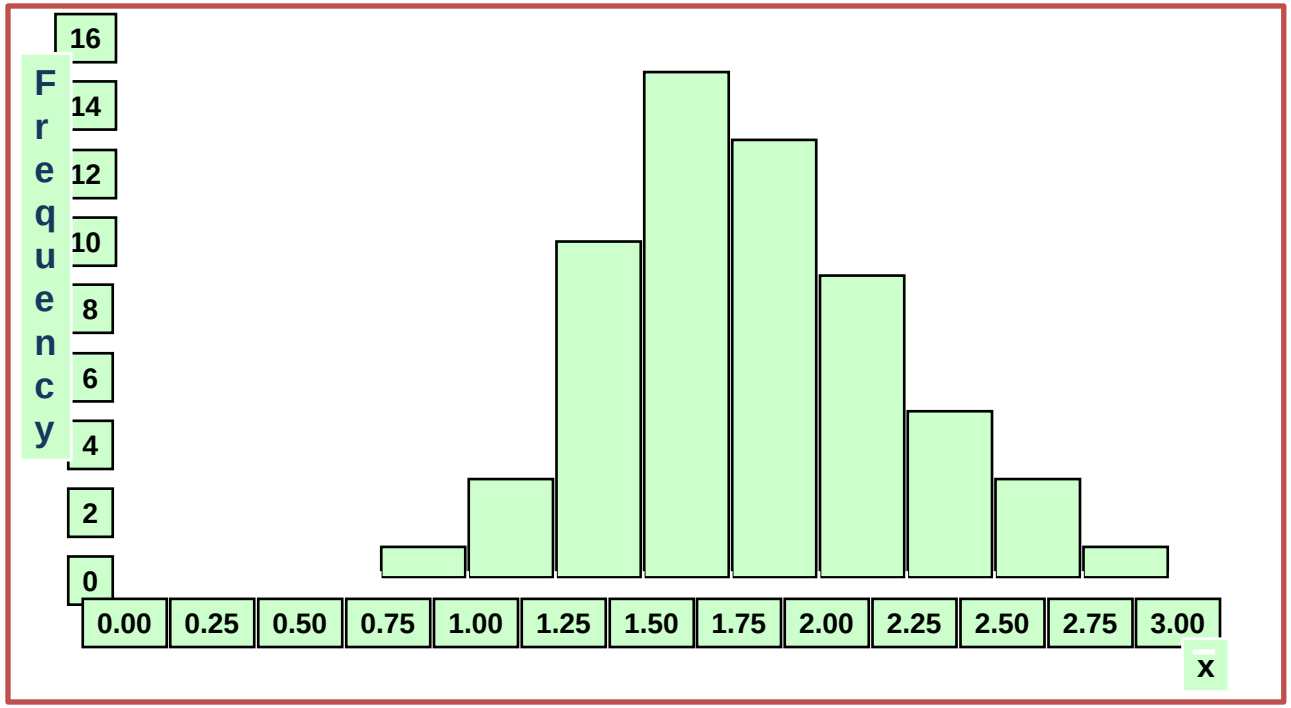
Means of 60 Samples (n = 2) from an Exponential Distribution



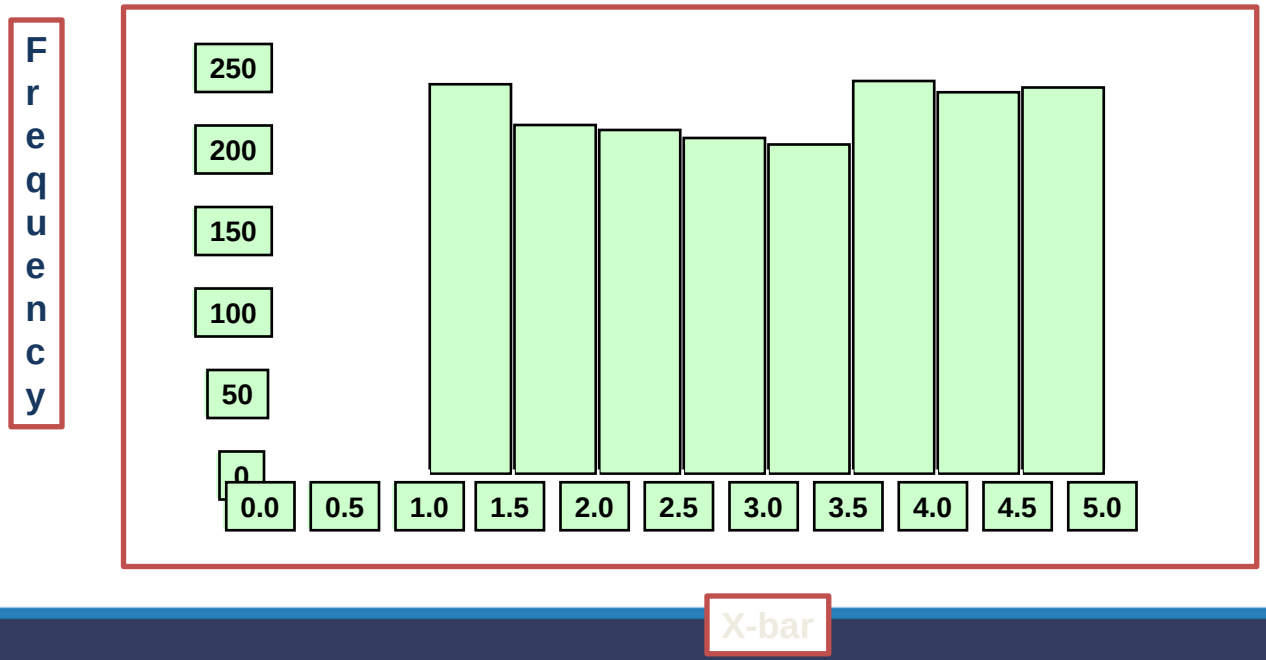
Means of 60 Samples (n = 5) from an Exponential Distribution



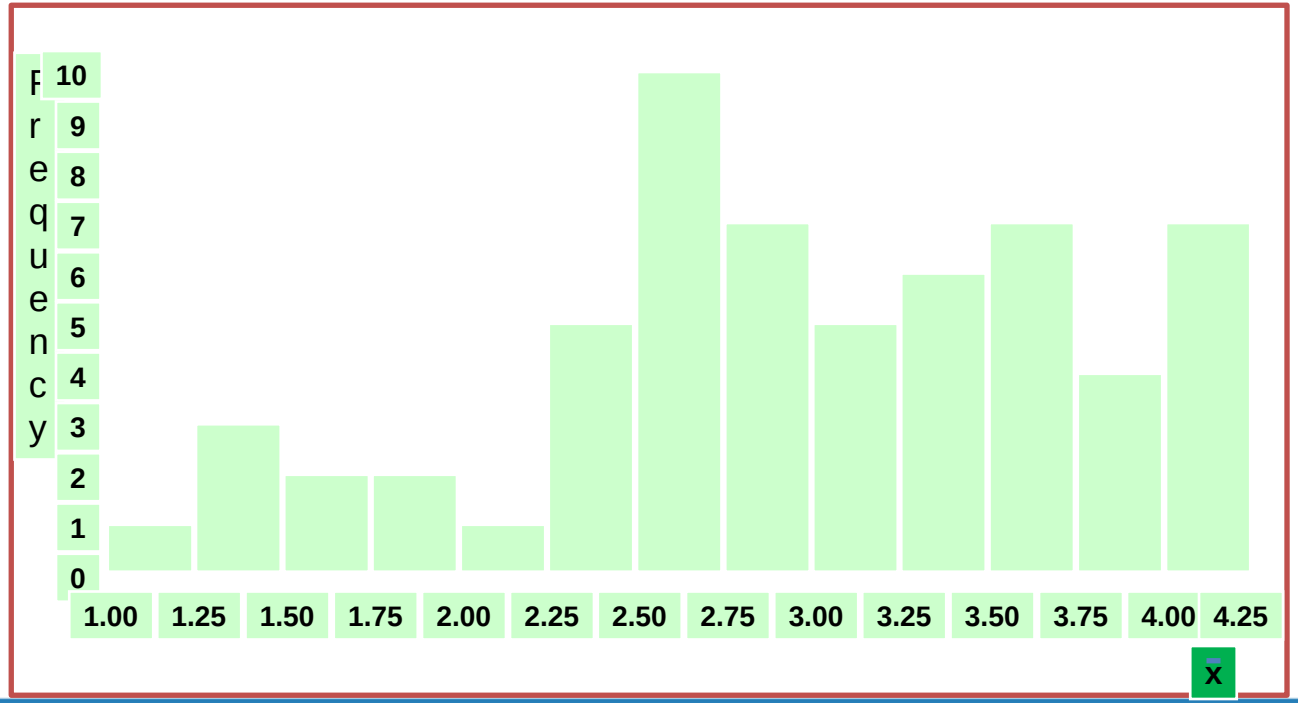
Means of 60 Samples (n = 30) from an Exponential Distribution



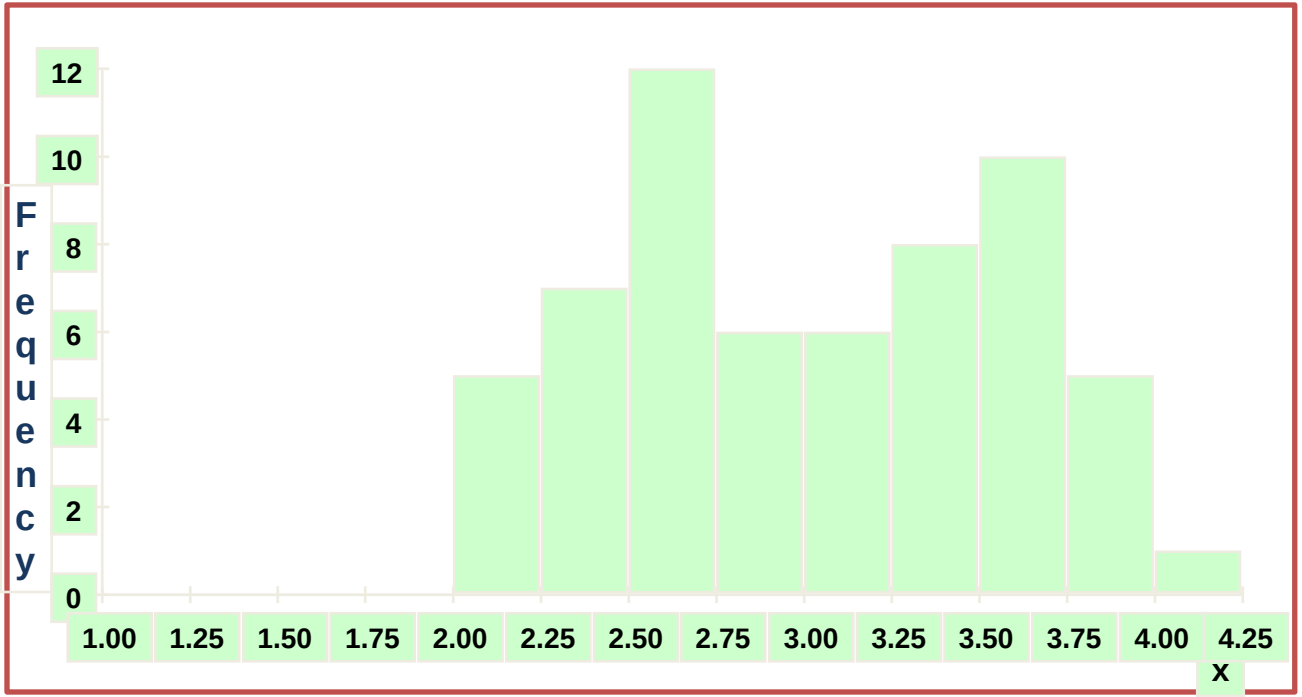
1,800 Randomly Selected Values from a Uniform Distribution



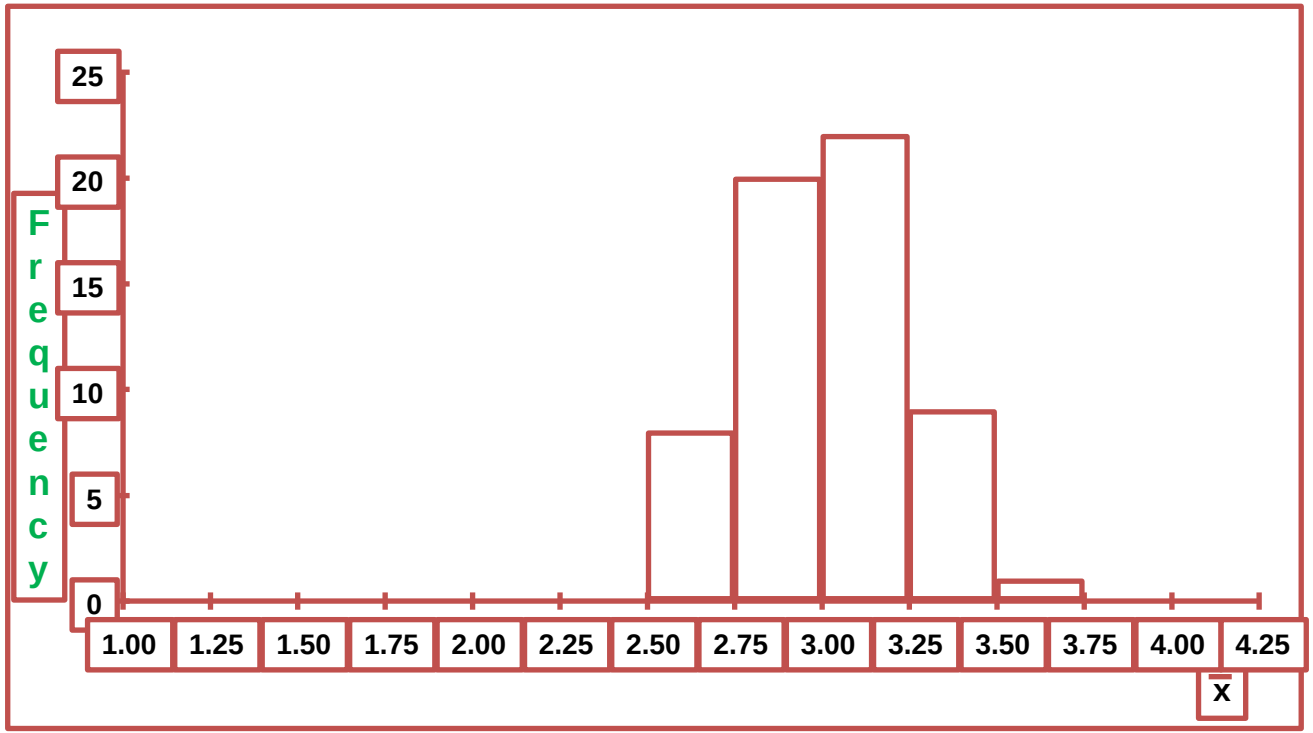
Means of 60 Samples (n = 2) from a Uniform Distribution



Means of 60 Samples (n = 5) from a Uniform Distribution



Means of 60 Samples (n = 30) from a Uniform Distribution



Expected Value of Sample Mean

- Let $X_1, X_2, \ldots X_n$ represent a random sample from a population
- The sample mean value of these observations is defined as

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

Standard Error of the Mean

- Different samples of the same size from the same population will yield different sample means
- A measure of the variability in the mean from sample to sample is given by the Standard Error of the Mean:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- Note that the standard error of the mean decreases as the sample size increases

If sample values are not independent *(continued)*

- If the sample size n is not a small fraction of the population size N, then individual sample members are not distributed independently of one another
- Thus, observations are not selected independently
- A correction is made to account for this:

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1}$$

or

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

If the Population is Normal

- If a population is normal with mean μ and standard deviation σ, the sampling distribution of $\bar{X}$ is also normally distributed with

$$\mu_{\bar{X}} = \mu \quad \text{and} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

- If the sample size n is not large relative to the population size N, then

$$\mu_{\bar{X}} = \mu \quad \text{and} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

Z-value for Sampling Distribution of the Mean

- Z-value for the sampling distribution of :

$$Z = \frac{(\bar{X} - \mu)}{\sigma_{\bar{X}}}$$

where:

$\bar{X}$ = sample mean

μ = population mean

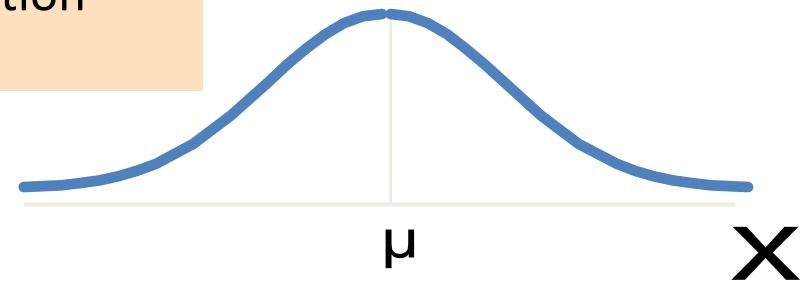
$\sigma_{\bar{X}}$ = standard error of the mean

Sampling Distribution Properties

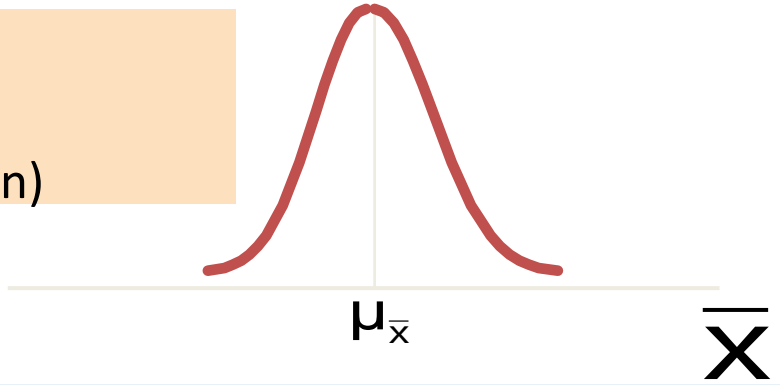
$$\mu_{\bar{x}} = \mu$$

(i.e. $\bar{x}$ is unbiased)

Normal Population Distribution



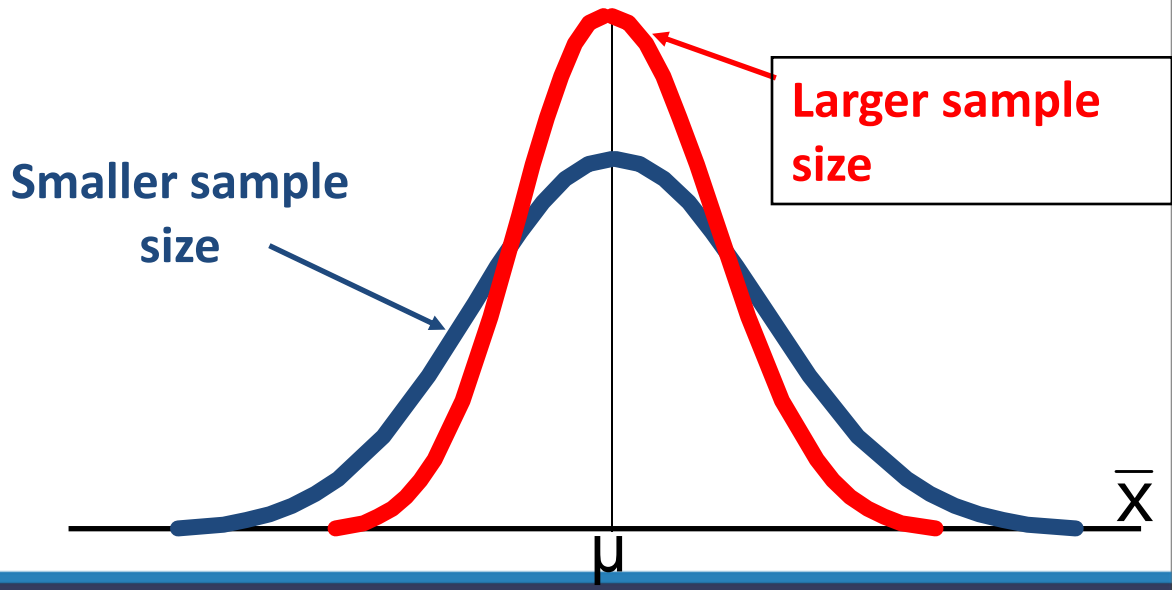
Normal Sampling Distribution (has the same mean)



Sampling Distribution Properties

- For sampling with replacement:

As n increases,
$\sigma_{\bar{X}}$ decreases



If the Population is not Normal- Central Limit Theorem

We can apply the Central Limit Theorem:

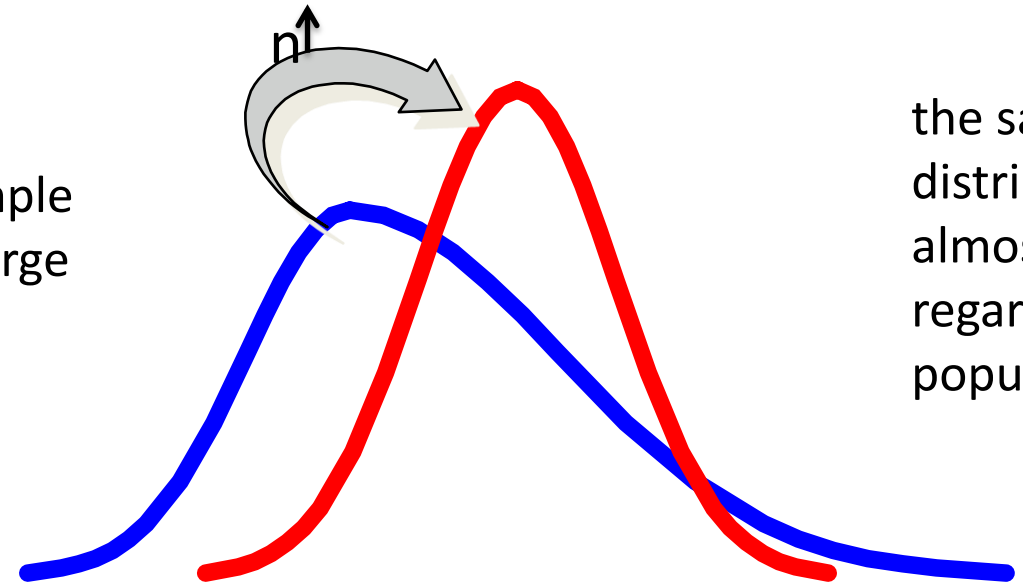
- Even if the population is not normal,
- sample means from the population will be approximately normal as long as the sample size is large enough.

Properties of the sampling distribution:

$$\mu_{\bar{x}} = \mu$$ And $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Central Limit Theorem

As the sample size gets large enough…



the sampling distribution becomes almost normal regardless of shape of population

If the Population is not Normal

(continued)

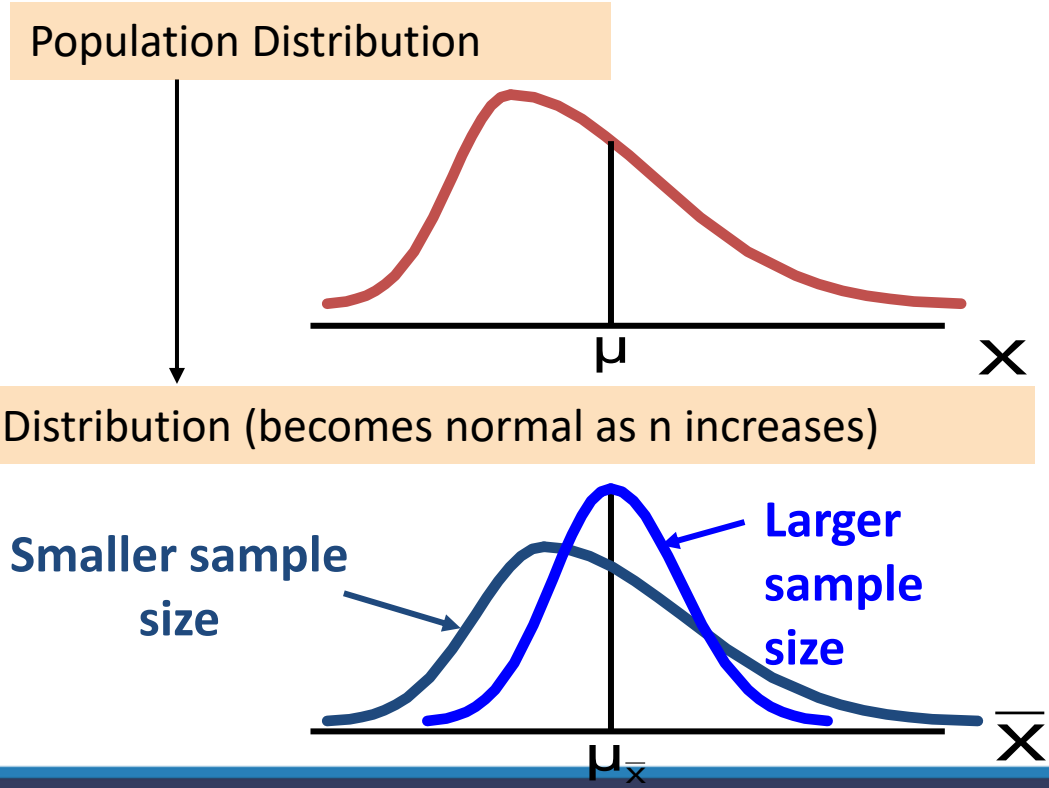
Sampling distribution properties:

Central Tendency

$$\mu_{\bar{x}} = \mu$$

Variation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$



How Large is Large Enough?

- For most distributions, $n > 25$ will give a sampling distribution that is nearly normal
- For normal population distributions, the sampling distribution of the mean is always normally distributed

Example

- Suppose a large population has mean $\mu = 8$ and standard deviation $\sigma = 3$. Suppose a random sample of size $n = 36$ is selected.

- What is the probability that the sample mean is between 7.8 and 8.2?

Example

Solution:

- Even if the population is not normally distributed, the central limit theorem can be used (n > 25)
- … so the sampling distribution of $\overline{X}$ is approximately normal
- … with mean $\mu_{\bar{x}} = 8$
- …and standard deviation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{3}{\sqrt{36}} = 0.5$$

Example *(continued)*

Solution (continued)

$$P(7.8 < \mu_{\bar{x}} < 8.2) = P\left(\frac{7.8 - 8}{3/\sqrt{36}} < \frac{\mu_{\bar{x}} - \mu}{\sigma/\sqrt{n}} < \frac{8.2 - 8}{3/\sqrt{36}}\right)$$

$$= P(-0.5 < Z < 0.5) = 0.3830$$