Sampling and Sampling Distributions

- Why Sample?
- Types of Sampling
- The Sampling Distribution of the Mean
- The Central Limit Theorem
- The Sampling Distribution of the Proportion

- Reading: Chapter 7

Why Sample?

The reason we select a sample is to collect data to answer a research question about a population

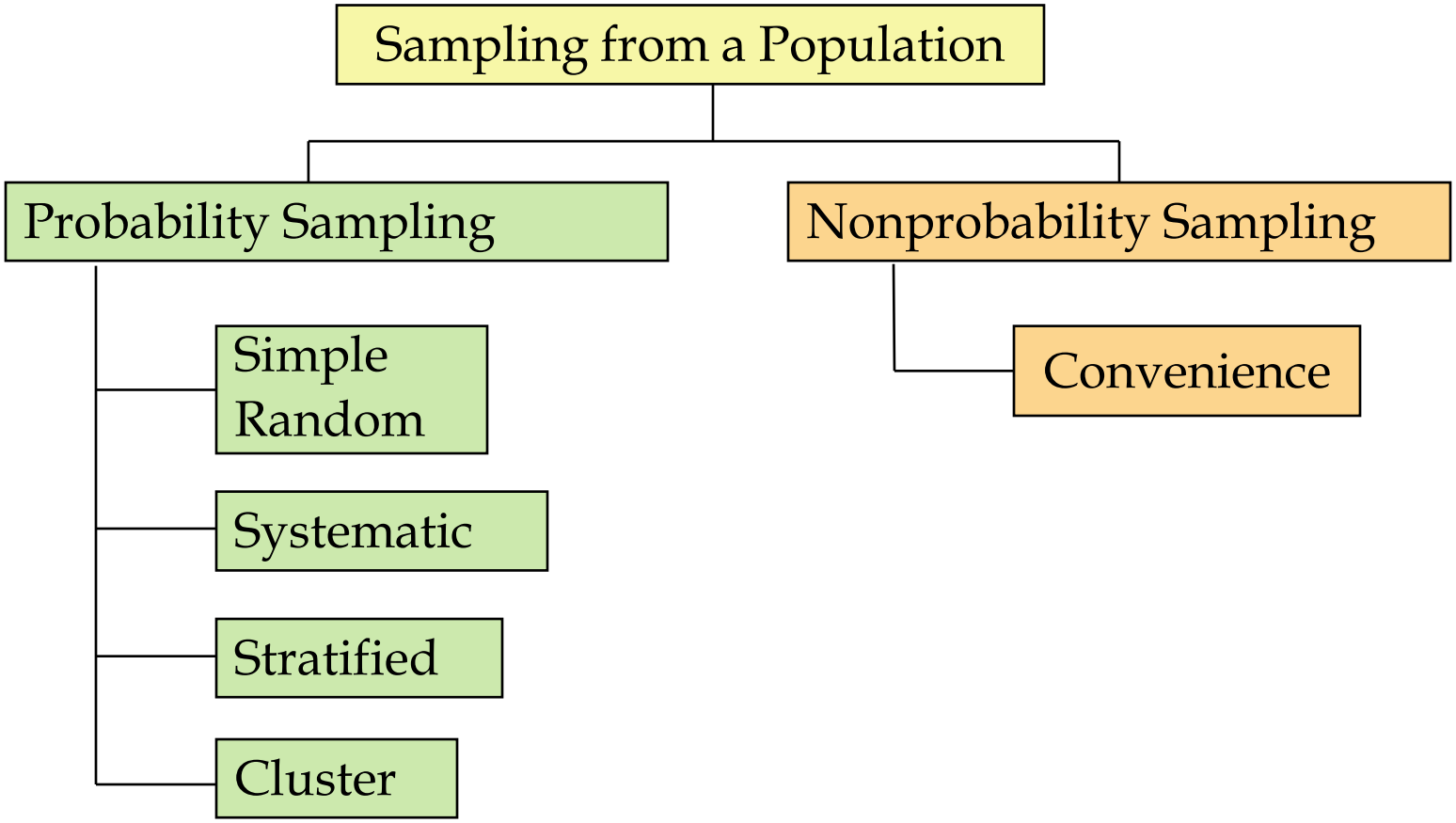
- Examining the entire population is expensive and time consuming
- Just infeasible to examine the entire population

The sample results provide only estimates of the values of the population characteristics

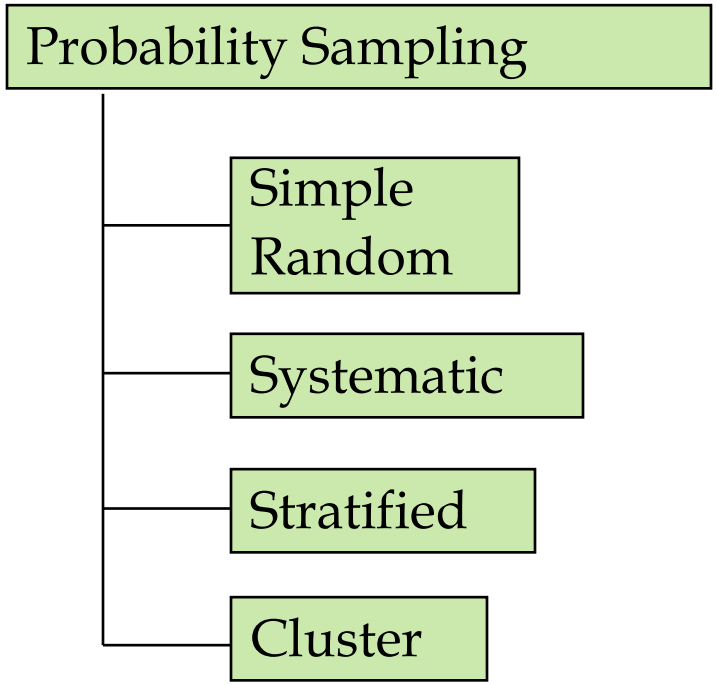
- The reason is simply that the sample contains only a portion of the population

With proper sampling methods, the sample results can provide "good" estimates of the population characteristics

Types of Sampling

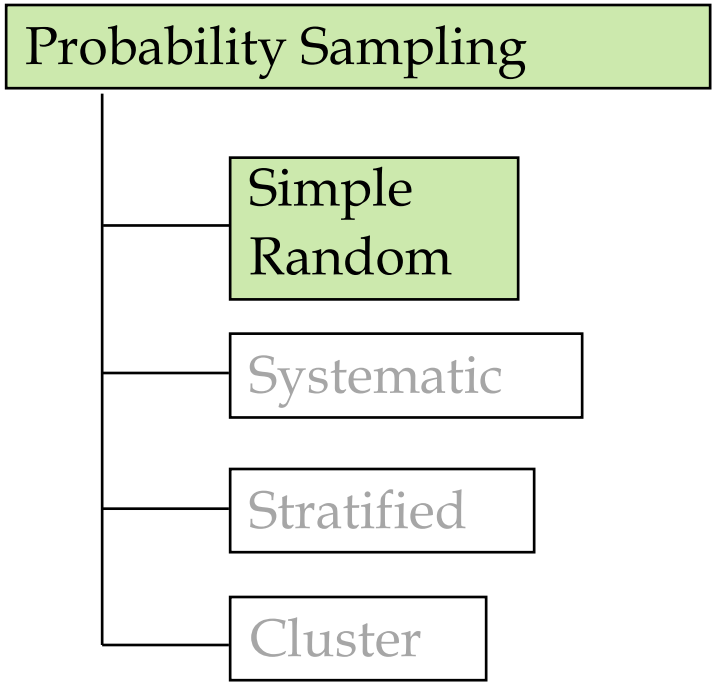


Types of Sampling



- A **probability sample** is a sample in which each member of the population has a known, nonzero, chance of being selected for the sample
- **Advantage:** we can perform inferential statistical tests to draw reliable conclusions about the population

Simple Random Sampling



- A **simple random sample** is a sample in which every member of the population has an **equal** chance of being chosen
- Many statistical methods rely simple random samples

Simple Random Sample

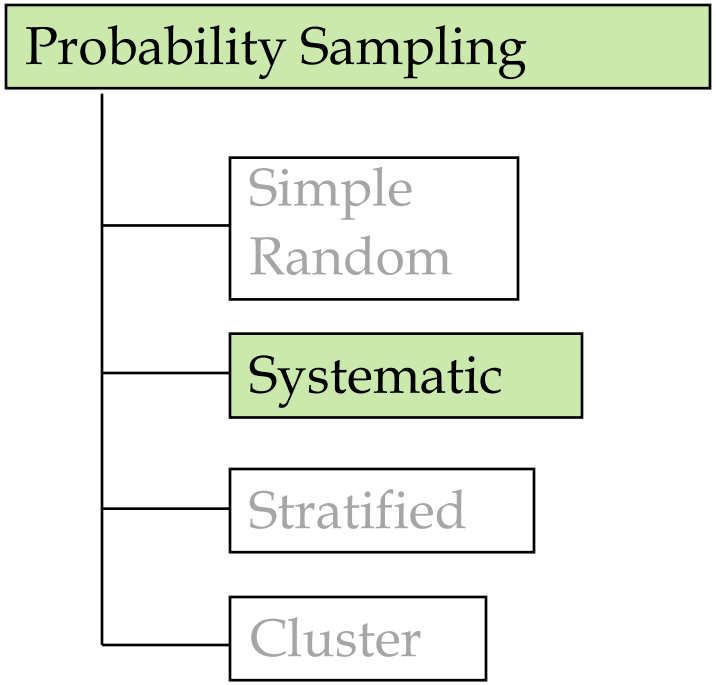
Sampling **with replacement**

- After a value from the population has been selected for the sample, the value is *returned* back into the population and *can be chosen again* for the same sample

Sampling **without replacement**

- After a value from the population is selected for the sample, it is *not returned* to the population so that value *cannot be chosen again* for the same sample

Systematic Sampling



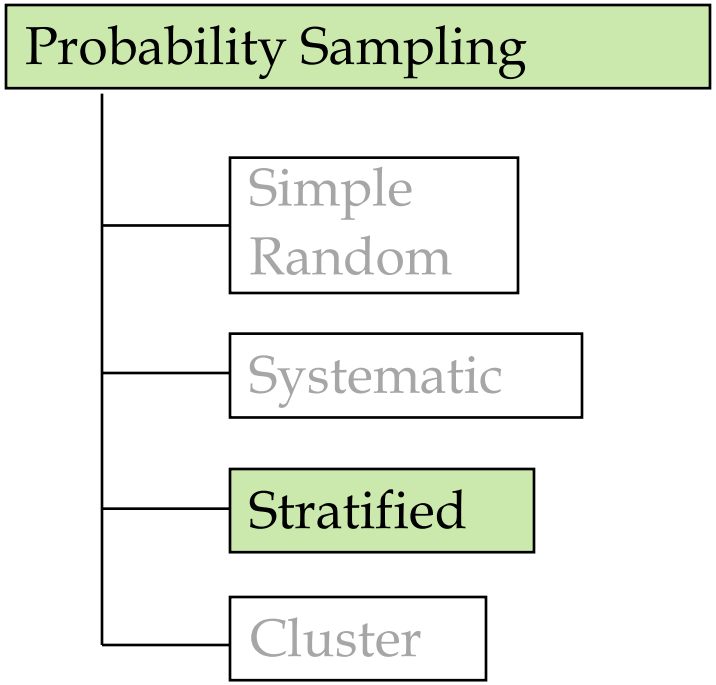
- In **systematic sampling**, every k^{th} member of the population is chosen for the sample. The value of k is determined by dividing the size of the population (N) by the size of the sample (n).

Example: Select a systematic sample of size $n = 5$ from a population of $N = 30$ students in this classroom?

$$k = N/n = 30/5 = 6$$

$\Rightarrow$ Choose every 6th student in the classroom for the sample

Stratified Sampling



- **Stratified sampling** divides the population into mutually exclusive and collectively exhaustive groups, or **strata**
- Strata are based on one or more classification criteria
  - Usually, some important variables which can affect research results
- A random sample is selected from each stratum
  - Number of observation per stratum is proportional to the stratum's size in the population

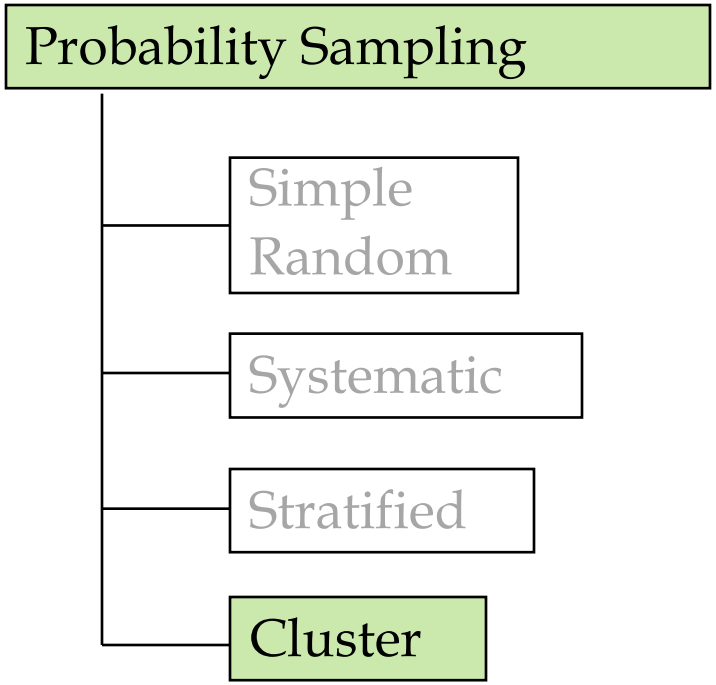
Stratified Sampling

Examples of strata:

- For an undergraduate population, strata could be class standing: Freshman, Sophomore, Junior, and Senior
- For factory production, strata could be 1st shift, 2nd shift, and 3rd shift
- For a population of workers, strata might be different age categories of workers

Advantages: Using stratified sampling helps ensure that all population subdivisions (classes, shifts, ages) are represented in the sample

Cluster Sampling



- **Cluster sampling** divides the population into mutually exclusive and collectively exhaustive groups, or **clusters**
- Each cluster is representative of the population (mini-version of the population)
- Final sample includes observations from randomly selected clusters:
  - We randomly select clusters
  - Sample all observations in those randomly selected clusters

Cluster Sampling

Advantages: This method is useful when clusters occur naturally in the population (city blocks, schools, and other geographic areas)

- That is why clusters are often selected based on geography to help simplify the sampling process

Examples of clusters:

- Individual cities where a new product is introduced
- Customer account balances arranged in clusters by first letter of last name

Stratified vs. Cluster Sampling

Stratified Sampling:

- Strata are homogeneous collections: each strata has a certain characteristic of interest
  - All values within a strata have some characteristic in common (e.g. each student is a freshman)
- The final sample consists of randomly selected elements from each strata

Cluster Sampling:

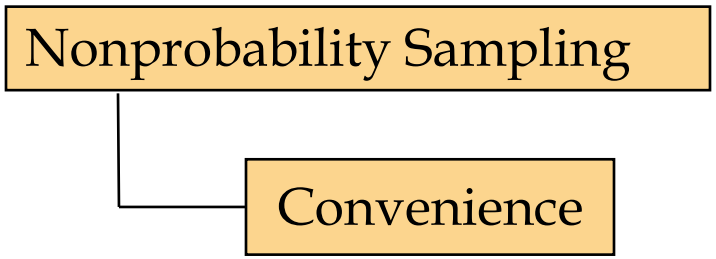
- Each cluster is representative of the entire population
  - Each cluster has a combination of various characteristics
- The final sample consists of all elements from the randomly selected clusters

Stratified vs. Cluster Sampling

Stratified Sampling

Cluster Sampling

Nonprobability Sampling

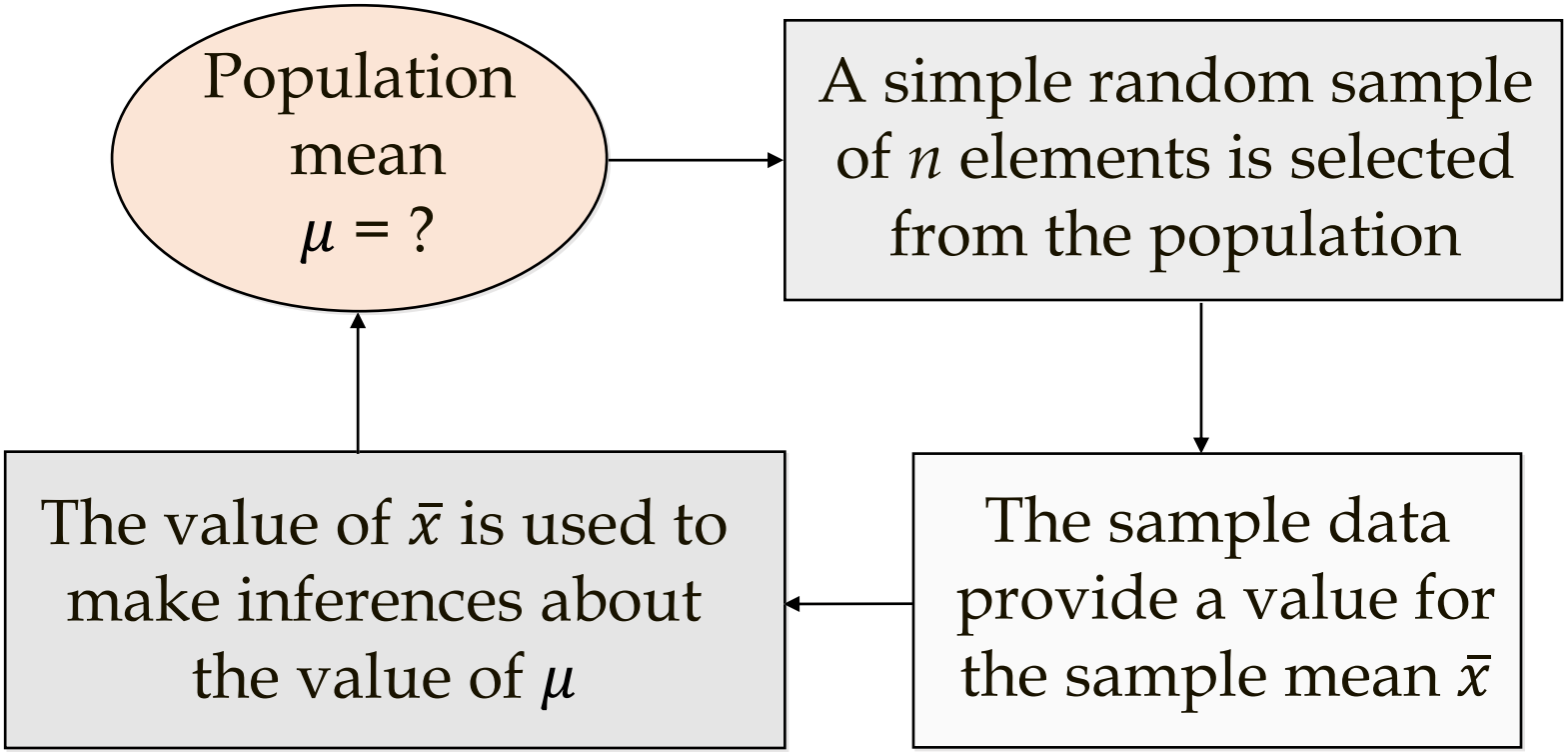


A **nonprobability sample** is a sample in which the probability of a population member being selected for the sample is **not known**

- Advantages:
  - Quick and easy to get sample data
  - Provides general information about the population
- Disadvantages:
  - May not be representative of the population

Sampling Distribution of $\bar{x}$

Process of Statistical Inference



Sampling Distribution of $\bar{x}$

The sampling distribution of $\bar{x}$ is the probability distribution of all possible values of the sample mean $\bar{x}$

- Each random sample of size n drawn from the population provides the sample mean $\bar{x}$
- Drawing many samples of size n results in many different sample means, one for each sample
- The sampling distribution of the mean is the frequency or probability distribution of sample means $\bar{x}$ derived from all possible samples of a given size n

Sampling Distribution of $\bar{x}$

We will use the following notation to describe the sampling distribution of $\bar{x}$:

μ = the population mean
$\mu_{\bar{x}}$ = the mean of the sampling distribution of $\bar{x}$
σ = the standard deviation of the population
$\sigma_{\bar{x}}$ = the standard deviation of the sampling distribution of $\bar{x}$
N = the population size
n = the sample size

Sampling Distribution of $\bar{x}$

If we know the sampling distribution of $\bar{x}$,

i.e. we know possible values of $\bar{x}$ and probabilities that $\bar{x}$ takes these values or the probability density function for $\bar{x}$,

then we can use our general formulas to find the mean and the standard deviation of the sampling distribution

For example, if $\bar{x}$ takes a finite number of values and the sampling distribution of $\bar{x}$ is known:

$$\mu_{\bar{x}} = \sum_{i=1}^{K} \bar{x}_i P(\bar{x}_i) \qquad \sigma_{\bar{x}} = \sqrt{\sum_{i=1}^{K} (\bar{x}_i - \mu_{\bar{x}})^2 P(\bar{x}_i)}$$

where $\bar{x}_i$ = values that $\bar{x}$ can take, from 1 to K

$P(\bar{x}_i)$ = probability that $\bar{x}$ takes value $\bar{x}_i$

Sampling Distribution of $\bar{x}$

Question: can we relate $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ to μ and σ - to the parameters of the population distribution? YES!

The mean of the sampling distribution of $\bar{x}$ or the expected value of $\bar{x}$

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

The standard deviation of the sampling distribution of $\bar{x}$ or the standard deviation of $\bar{x}$

- Is called the *standard error of the mean*
- We will distinguish between the standard error of the mean for the finite and infinite populations

Sampling Distribution of $\bar{x}$

The standard deviation of the sampling distribution of $\bar{x}$ or the standard deviation of $\bar{x}$

Finite Population

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

Infinite Population

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

- $\sqrt{(N-n)/(N-1)}$ is the *finite population correction factor*
- A finite population is treated as infinite if $n/N \leq .05$
- $\sigma_{\bar{x}}$ is referred to as the **standard error** of the mean

Sampling Distribution of $\bar{x}$

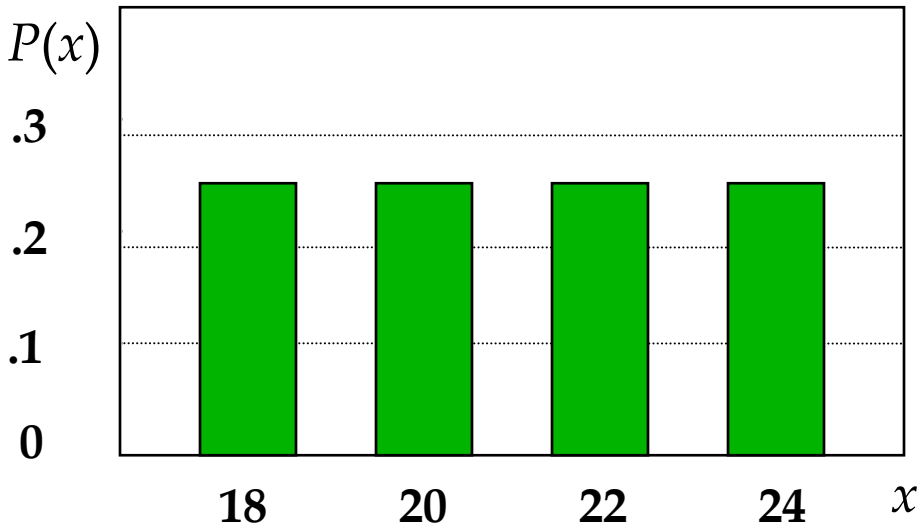
Example: Assume there is a population …

- Population size $N = 4$
- Random variable, x, is age of individuals
- Values of x: 18, 20, 22, 24 (years)

Summary Measures for the Population Distribution:

$$\mu = \frac{\sum x_i}{N} = 21$$

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



Sampling Distribution of $\bar{x}$

Now consider all possible samples of size $n = 2$

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

12 possible samples (sampling *w/o replacement*)

12 Sample Means

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

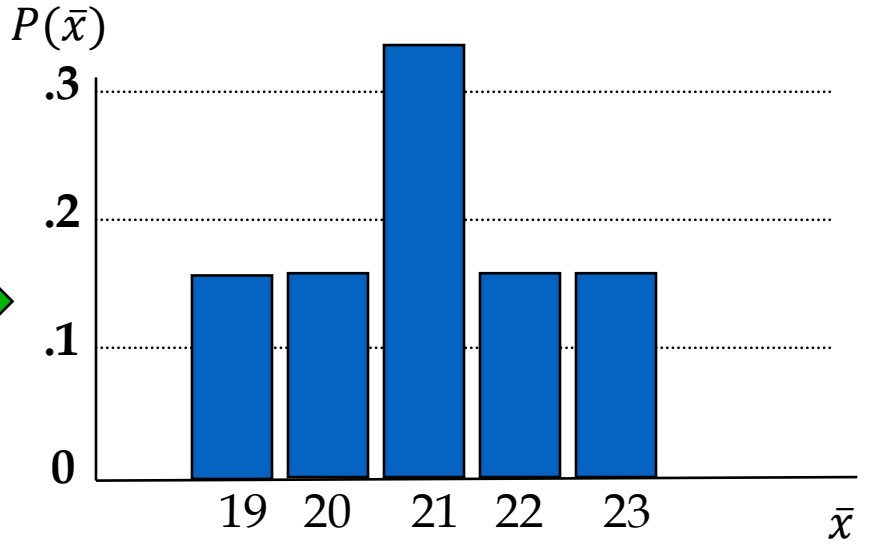
Sampling Distribution of $\bar{x}$

The **sampling distribution of the mean** describes all possible sample means and their frequencies when samples are randomly drawn from a population

12 Sample Means

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

Sampling Distribution of the Mean, $n = 2$



(no longer uniform)

Sampling Distribution of $\bar{x}$

Summary measures of this **sampling distribution**:

$$\mu_{\bar{x}} = \frac{\sum \bar{x}_i}{N_{\bar{x}}} = \frac{19 + 19 + 20 + 20 + 21 \ldots + 23}{12} = 21$$

$$\sigma_{\bar{x}} = \sqrt{\frac{\sum(\bar{x}_i - \mu_{\bar{x}})^2}{N_{\bar{x}}}} =$$

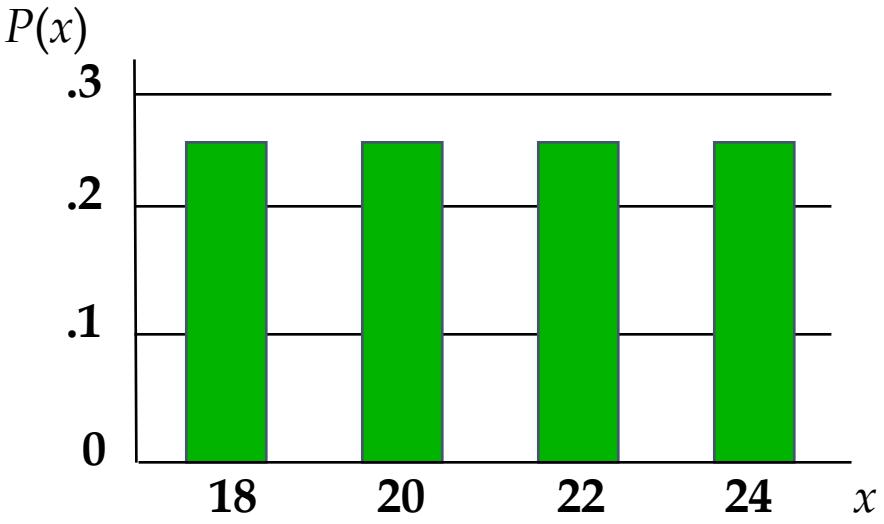
$$= \sqrt{\frac{(19-21)^2+(19-21)^2+(20-21)^2 \ldots+(23-21)^2}{12}} = 1.29$$

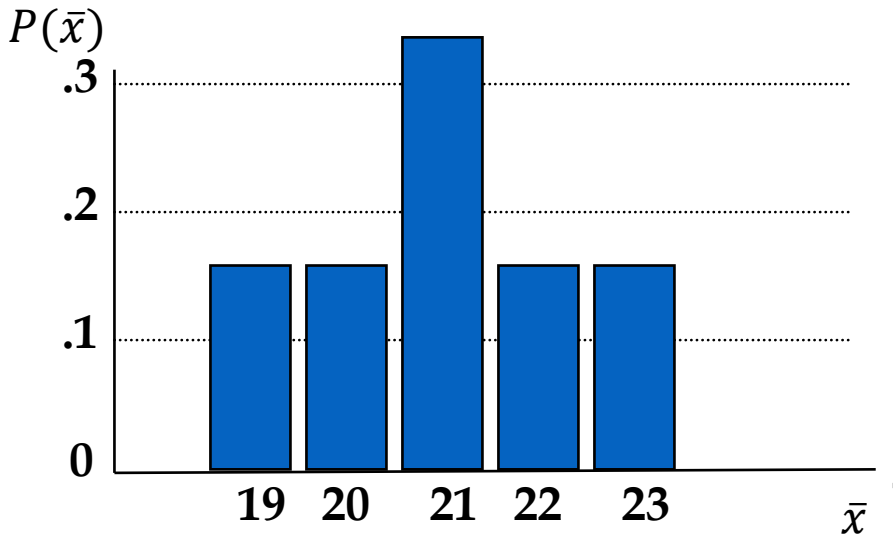
We consider the population of $\bar{x}$'s ⇒ divide by the number of members in the population of $\bar{x}$'s ($N_{\bar{x}}$ here is used to distinguish it from N, the number of members of the original population)

Sampling Distribution of $\bar{x}$

Population		Sampling Distribution of the Mean, $n = 2$
$N = 4$		
$\mu = 21$	$\mu_{\bar{x}} = \mu$	$\mu_{\bar{x}} = 21$
$\sigma = 2.236$	$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$	$\sigma_{\bar{x}} = 1.29$

Population is finite ($N = 4$)
Additionally, $n/N > 0.05$





Sampling Distribution of $\bar{x}$

Question: what about the shape of the sampling distribution of $\bar{x}$?

- **When the population has a normal distribution**, the sampling distribution of $\bar{x}$ is normally distributed for **any** sample size
- When the population from which we are selecting a random sample **does not have a normal distribution**, the Central Limit Theorem (CLT) is helpful in identifying the shape of the sampling distribution of $\bar{x}$

Central Limit Theorem

CENTRAL LIMIT THEOREM

In selecting random samples of size n from a population, the sampling distribution of the sample mean $\bar{x}$ can be approximated by a normal distribution ***as the sample size becomes large***

Sampling Distribution of $\bar{x}$: Notes

According to the **Central Limit Theorem**, sample means from samples of **sufficient size**, drawn from **any** population, will be **normally distributed**

- In most applications, the sampling distribution of $\bar{x}$ can be approximated by a normal distribution whenever the sample is size 30 or more, *regardless of the shape of the population distribution*
  - In cases where the population is highly skewed or outliers are present, samples of size 50 may be needed

- If the population follows the normal probability distribution, the sample means will also be normally distributed, regardless of the size of the sample

Sampling Distribution of $\bar{x}$: Example

The sampling distribution of $\bar{x}$ can be used to provide probability information about how close the sample mean $\bar{x}$ is to the population mean μ

Example: St. Andrew's College

St. Andrew's College received 900 applications for admission from prospective students. Once all SAT scores for the 900 applicants were entered in the college's database, the values of the population parameters were calculated: $\mu = 1090$, $\sigma = 80$.

What is the probability that a simple random sample of 30 applicants will provide a **sample mean** SAT score within +/– 10 of the actual population mean (i.e. between 1080 and 1100)?

Sampling Distribution of $\bar{x}$: Example

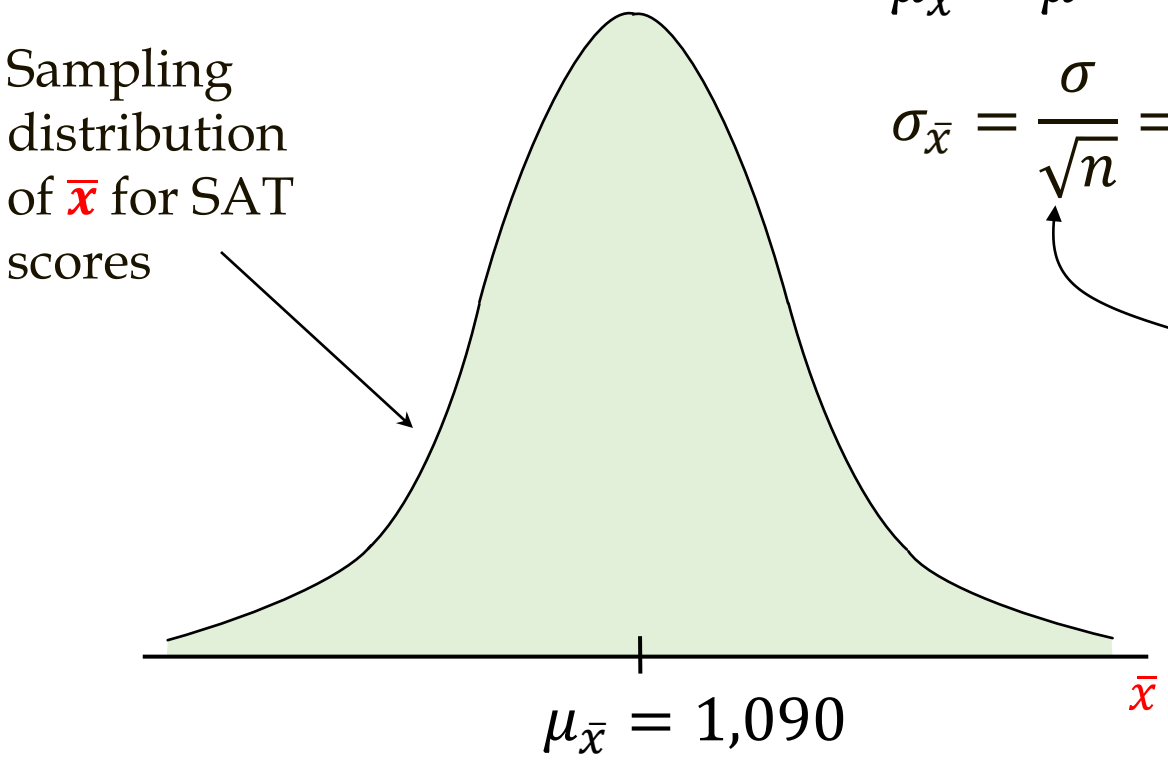
Example: St. Andrew's College

Question is to find $P(1080 \leq \bar{x} \leq 1100)$? ⇒ Need a sampling distribution of $\bar{x}$

$$\mu_{\bar{x}} = \mu = 1090$$

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{80}{\sqrt{30}} = 14.6059$$

Population is finite (N = 900) But $n/N > 0.05$



Sampling Distribution of $\bar{x}$: Example

Example: St. Andrew's College

- Calculate the z-score for the upper and lower endpoints of the interval

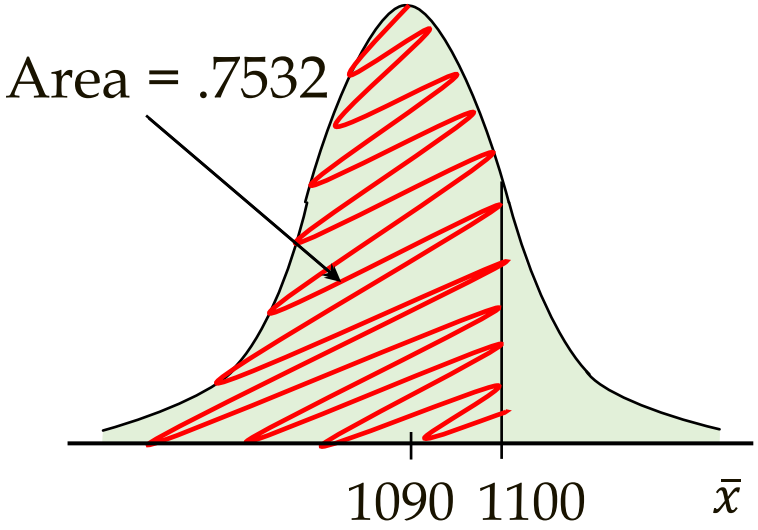
$$z_{1100} = \frac{1100 - 1090}{14.6059} = .6847 \qquad z_{1080} = \frac{1080 - 1090}{14.6059} = -.6847$$

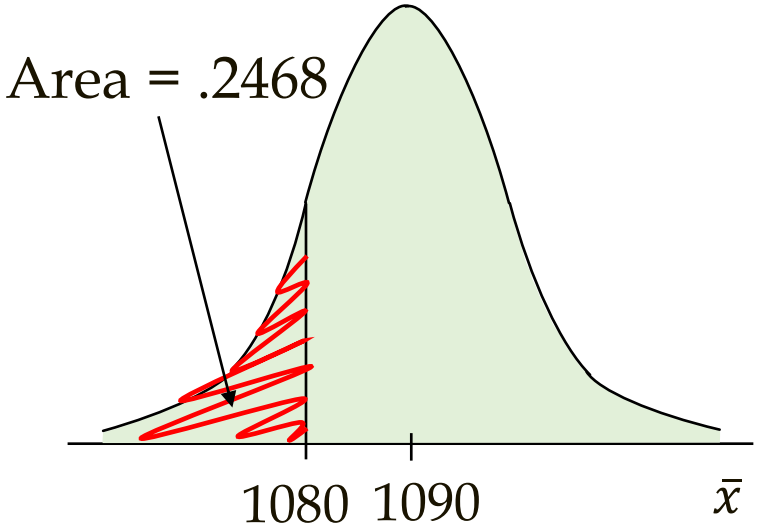
- Find cumulative probabilities for the z-scores of the endpoints (the areas under the probability density function to the left of the z-score for each endpoint):

$P(z \leq .6847) = .7532$ $\qquad$ $P(z \leq -.6847) = .2468$

Sampling Distribution of $\bar{x}$: Example

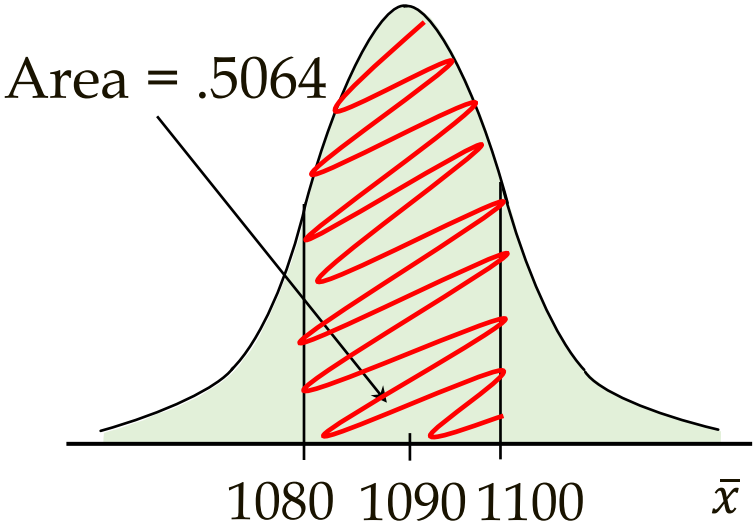
Example: St. Andrew's College





Sampling Distribution of $\bar{x}$: Example

Example: St. Andrew's College



- Calculate the area under the curve between the lower and upper endpoints of the interval:

  $P(-.6847 \le z \le .6847) = P(z \le .6847) - P(z \le -.6847) = .7532 - .2468 = .5064$

- The probability that the **sample mean** SAT score will be between 1080 and 1100 is $P(1080 \le \bar{x} \le 1100) = .5064$

The Effect of the Sample Size on the Sampling Distribution

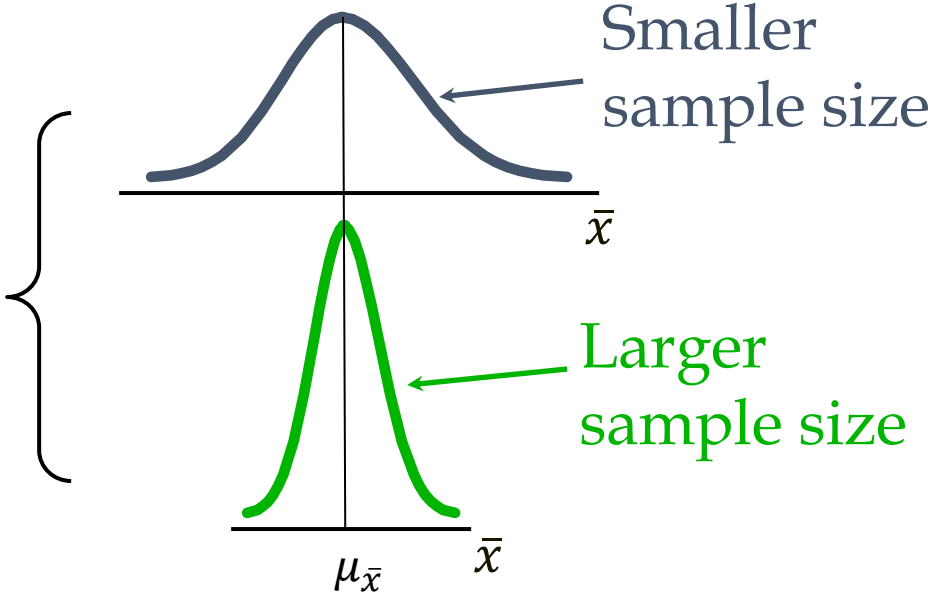
As the sample size **increases**

- The standard error of the mean becomes **smaller**

$\Rightarrow$ The sampling distribution becomes taller

As n increases, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ decreases

I.e. $n \uparrow \Rightarrow \sigma_{\bar{x}} \downarrow \Rightarrow$ The **sampling** distribution is skinnier



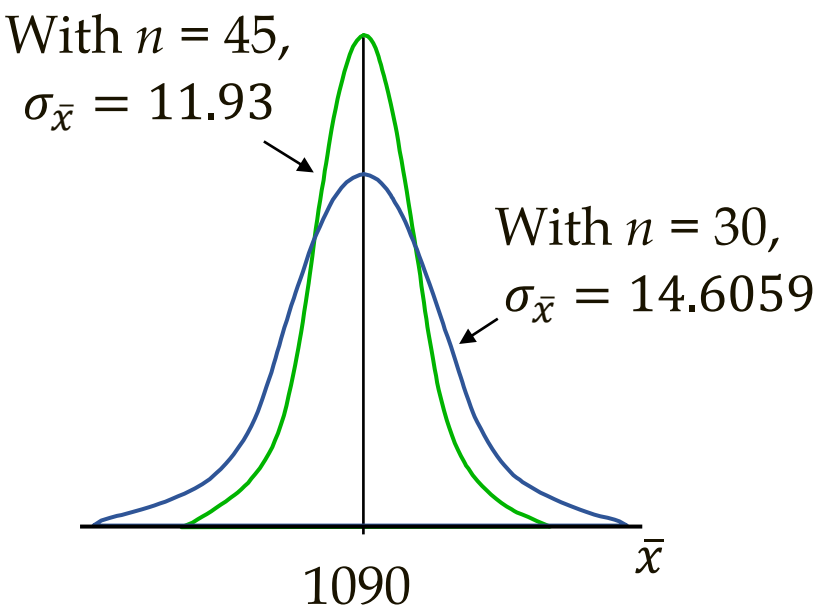
The Effect of the Sample Size on the Sampling Distribution: Example

Example: St. Andrew's College

Suppose we select a simple random sample of 45 applicants instead of the 30 originally considered

- $\mu_{\bar{x}} = \mu$ regardless of the sample size $\Rightarrow \mu_{\bar{x}} = 1090$
- As the sample size ↑, the standard error of the mean $\sigma_{\bar{x}}$ ↓
  - With n ↑ to 45, the standard error of the mean ↓ from 14.6 to:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{80}{\sqrt{45}} \approx 11.93$$



The Effect of the Sample Size on the Sampling Distribution: Example

Example: St. Andrew's College

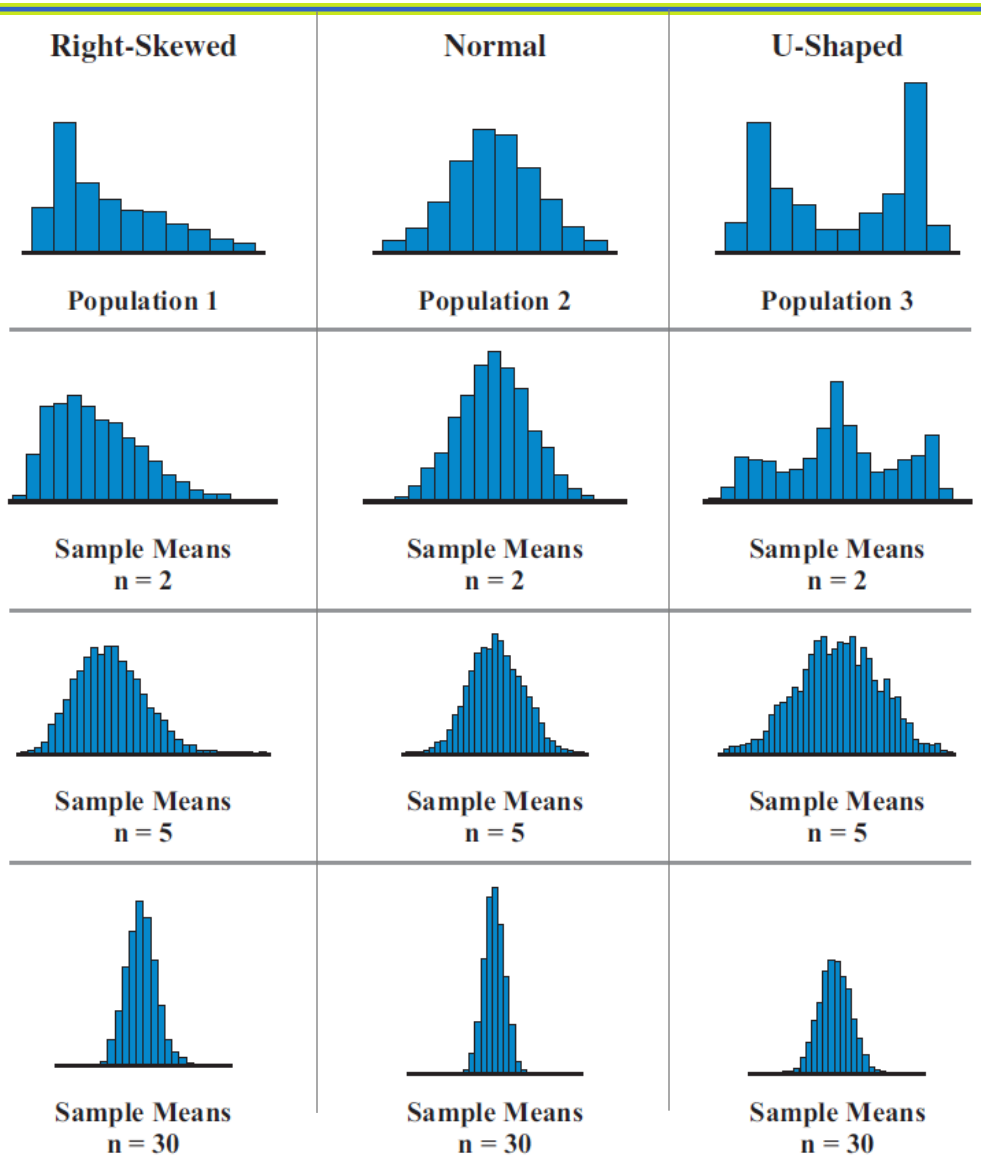
- Recall that when $n = 30$, $P(1080 \leq \bar{x} \leq 1100) = .5064$
- Follow the same steps to solve for $P(1080 \leq \bar{x} \leq 1100)$ when $n = 45$ as when $n = 30$
- Now, with $n = 45$, $P(1080 \leq \bar{x} \leq 1100) = .5983$
- Because the sampling distribution with $n = 45$ has a smaller standard error, the values of $\bar{x}$ have less variability and tend to be closer to the population mean than the values of $\bar{x}$ with $n = 30$

Excel Exercise >>

The Effect of the Sample Size on the Sampling Distribution

Will the shape of the population distribution affect the shape of the sampling distribution of $\bar{x}$?

- **Recall CLT:** No if the sample size if large enough
- Yes if the sample size is small

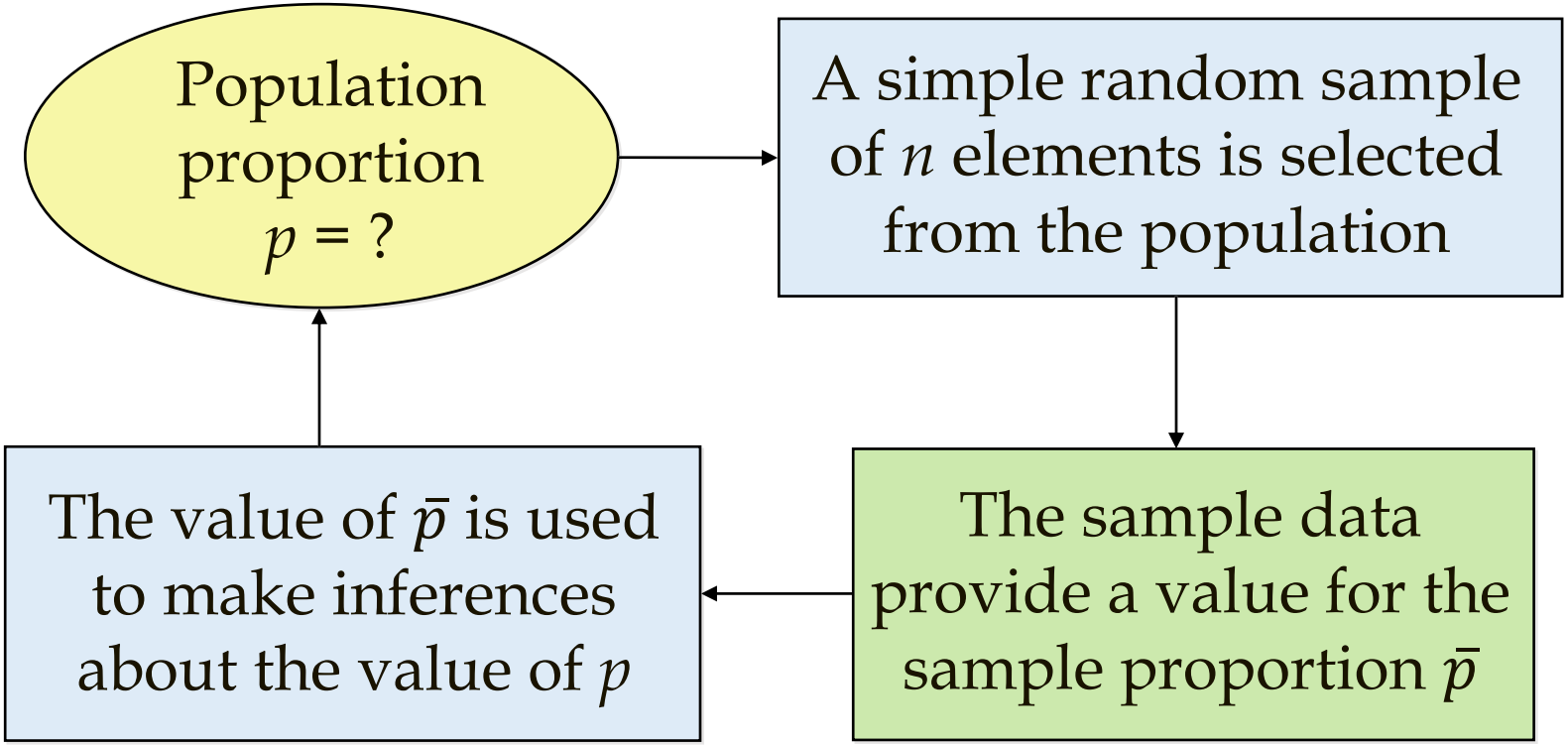


Approaching Assignment Problems

Notice that we need to pay attention to several things (or a combination of them) to describe the sampling distribution

1. Original population: finite or infinite
2. For finite populations, the population size relative to the sample size (n/N)
   - Affects which formula for the standard error we use
3. Population distribution: normal or other shape
   - Normally distributed $x \Rightarrow$ sampling distribution of $\bar{x}$ is **always** normal
   - Other distribution $\Rightarrow$ think if n is sufficiently big to use CLT
4. Sample size
   - If n is big, the CLT states that the sampling distribution of $\bar{x}$ is normal for any population distribution

Inferences about a Population Proportion



Sampling Distribution of the Proportion

So far ⇒ sampling distribution of a **sample mean**

Now ⇒ sampling distribution of a **sample proportion**

Sample Proportion

$$\bar{p} = \frac{x}{n}$$

where x = The number of observations of interest in the sample (successes)

n = Sample size (trials)

Sampling Distribution of $\bar{p}$

The sampling distribution of $\bar{p}$ is the probability distribution of all possible values of the sample proportion

Our notation:

p = the population proportion
$\bar{p}$ = the sample proportion
$\mu_{\bar{p}}$ = the mean of the sampling distribution of $\bar{p}$
$\sigma_{\bar{p}}$ = the standard deviation of the sampling distribution of $\bar{p}$
N = the population size
n = the sample size

Sampling Distribution of $\bar{p}$

The mean of the sampling distribution of $\bar{p}$ or the expected value of $\bar{p}$

$$\mu_{\bar{p}} = p$$

The standard deviation of the sampling distribution of $\bar{p}$ or the standard deviation of $\bar{p}$

- Is called the *standard error of the proportion*
- We will distinguish the standard error of the proportion for the finite and infinite populations

Sampling Distribution of $\bar{p}$

The standard deviation of the sampling distribution of $\bar{p}$ or the standard deviation of $\bar{p}$

Finite Population

$$\sigma_{\bar{p}} = \sqrt{\frac{N-n}{N-1}}\sqrt{\frac{p(1-p)}{n}}$$

Infinite Population

$$\sigma_{\bar{p}} = \sqrt{\frac{p(1-p)}{n}}$$

- $\sqrt{(N-n)/(N-1)}$ is the *finite population correction factor*
- A finite population is treated as infinite if $n/N \leq .05$
- $\sigma_{\bar{p}}$ is referred to as the **standard error** of the proportion

Sampling Distribution of $\bar{p}$

Question: Can we describe the shape of the sampling distribution of $\bar{p}$?

- The underlying distribution of $\bar{p}$ is the binomial distribution
  - We examine the number of successes (x) in n trials
  - Because n is constant, the proportion x/n resembles the distribution of x, i.e. binomial
- Binomial distribution can be approximated by the normal:

$$np \geq 5 \qquad \text{and} \qquad n(1 - p) \geq 5$$

⇒ The sampling distribution of $\bar{p}$ can be approximated by a normal distribution whenever these two condition are satisfied

Sampling Distribution of $\bar{p}$

The z-score for the Sample Proportion

- to calculate normal probabilities for the sample proportion with the z-transformation

$$z_{\bar{p}} = \frac{\bar{p} - p}{\sigma_{\bar{p}}}$$

p = the population proportion

$\bar{p}$ = the sample proportion

$\sigma_{\bar{p}}$= the standard error of $\bar{p}$

Sampling Distribution of $\bar{p}$: Example

Example: St. Andrew's College

Suppose that 72% of the prospective students applying to St. Andrew's College desire on-campus housing.

What is the probability that a simple random sample of 30 applicants will provide a **sample proportion** of applicants desiring on-campus housing that is within plus or minus .05 of the actual population proportion?

Sampling Distribution of $\bar{p}$: Example

Example: St. Andrew's College

Question is to find $P(0.67 \leq \bar{p} \leq 0.77)$? ⇒ Need a sampling distribution of $\bar{p}$

For our example,

1. $n = 30$ and $p = .72$
2. The normal distribution is an acceptable approximation:

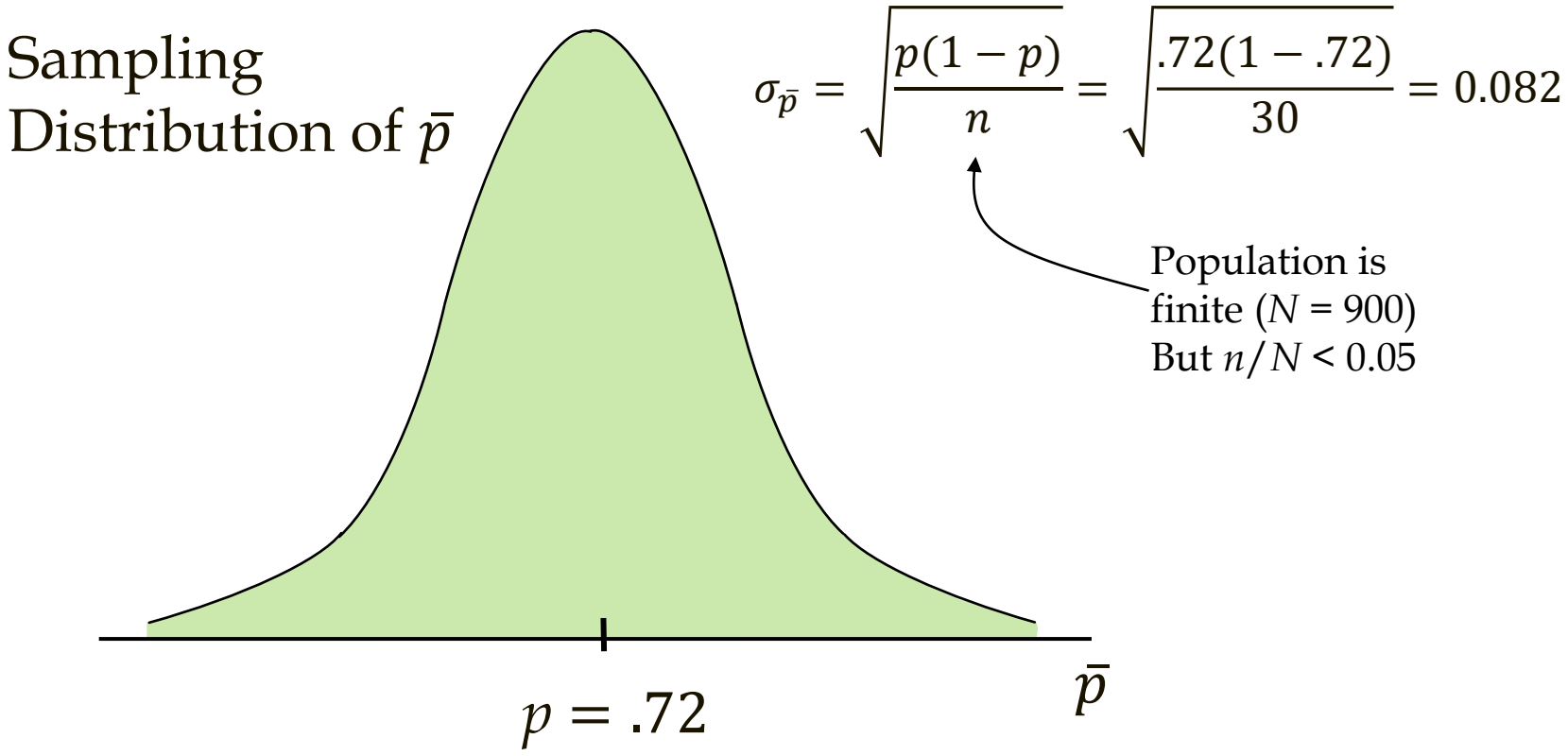
$$np = 30 \times 0.72 = 21.6 \geq 5$$

and

$$n(1 - p) = 30 \times (1 - 0.72) = 8.4 \geq 5$$

Sampling Distribution of $\bar{p}$: Example

Example: St. Andrew's College



Sampling Distribution of $\bar{p}$: Example

Example: St. Andrew's College

- Calculate the z-scores at the upper and lower endpoints of the interval

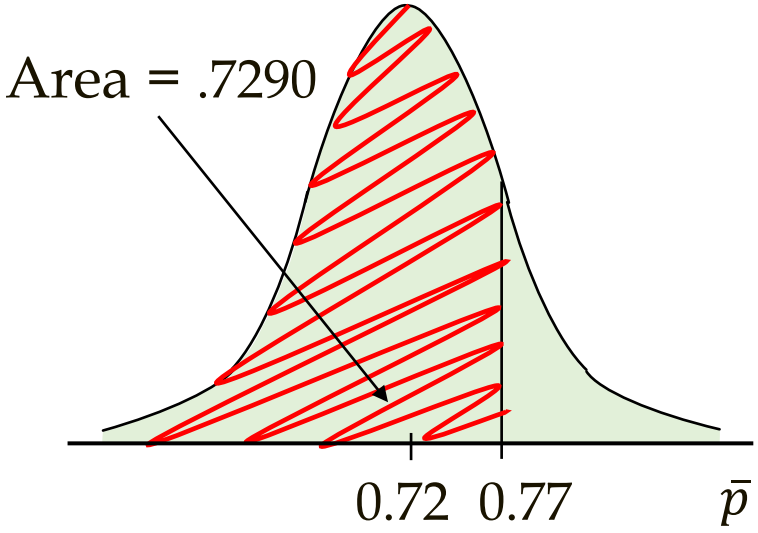
$$z_{0.77} = \frac{0.77 - 0.72}{0.082} = .6098 \quad z_{0.67} = \frac{0.67 - 0.72}{0.082} = -.6098$$

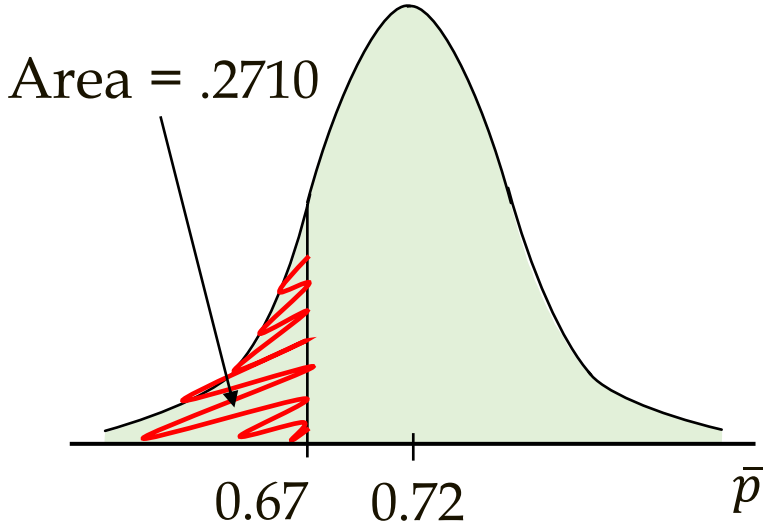
- Find cumulative probabilities for the z-scores of the endpoints (the areas under the probability density function to the left of each endpoint):

$$P(z \leq .6098) = .7290 \qquad P(z \leq -.6098) = .2710$$

Sampling Distribution of $\bar{p}$: Example

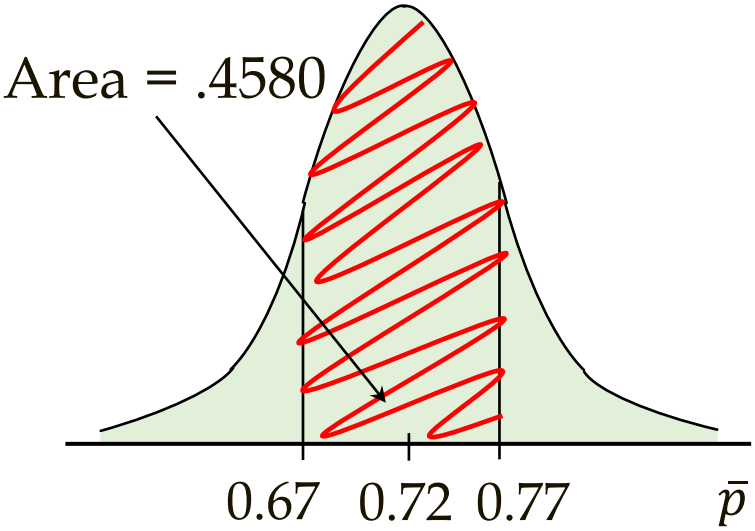
Example: St. Andrew's College





Sampling Distribution of $\bar{p}$: Example

Example: St. Andrew's College



- Calculate the area under the curve between the lower and the upper endpoints of the interval:
  $P(-.6098 \le z \le .6098) = P(z \le .6098) - P(z \le -.6098) = .7290 - .2710 = .4580$
- The probability that the sample proportion of applicants wanting on-campus housing will be within +/-.05 of the population proportion is $P(0.67 \le \bar{p} \le 0.77) = .4580$

Excel Exercise >>

Excel Time: Exercise 1 (not in the textbook)

The mean preparation fee H&R Block charged retail customers in 2011 was $183. Use this price as the population mean and assume that the population standard deviation of preparation fees is $50.

1. What is the probability that the mean price **for a sample** of 30 H&R Block retail customers is within $8 of the population mean?
2. What is the probability that the mean price **for a sample** of 50 H&R Block retail customers is within $8 of the population mean?
3. Compare the results. What can you conclude regarding the probability as the sample size increases? Why this conclusion applies (relate it to the shape of the sampling distribution)?

Back <<

Excel Time: Exercise 2 (not in the textbook)

The Grocery Manufacturers of America reported that 76% of consumers read the ingredients listed on a product's label. Assume that the population proportion 0.76 and a sample of 400 customers is selected from the population.

1. Show the sampling distribution of the **sample proportion** of customers who read the ingredients, i.e. find the mean and the standard deviation of the sampling distribution.
2. What is the probability that the sample proportion will be within ±0.03 of the population proportion?
3. Answer previous question using a sample of 750 consumers.

Back <<

Excel Time: Exercise 7.38 (Extra Practice)

According to Bureau of Labor Statistics, the average salary for a federal employee in 2017 was $76,250. Assume the population standard deviation is $28,025. A random sample of 34 federal employees is selected.

a. What is the probability that the sample mean will be less than $70,000?
b. What is the probability that the sample mean will be more than $74,000?
c. What is the probability that the sample mean will be between $76,000 and $81,000?

Excel Time: Exercise 7.39 (Extra Practice)

According to the Bureau of Labor Statistics, the unemployment rate for workers aged 20 to 24 in April 2018 was 6.7%. Consider a random sample of 110 workers from this age group.

a. What is the probability that the sample proportion of unemployed will be below 9%?
b. What is the probability that the sample proportion of unemployed will be below 4.5%?
c. What is the probability that the sample proportion of unemployed will be between 4.5% and 13.6%?