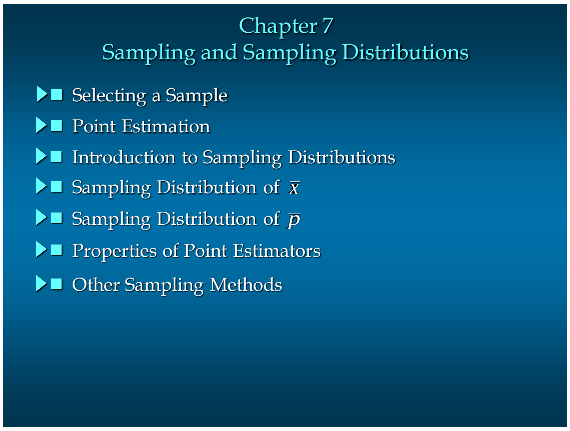
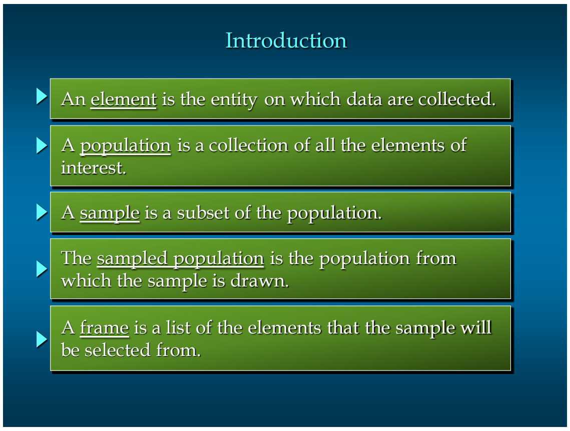


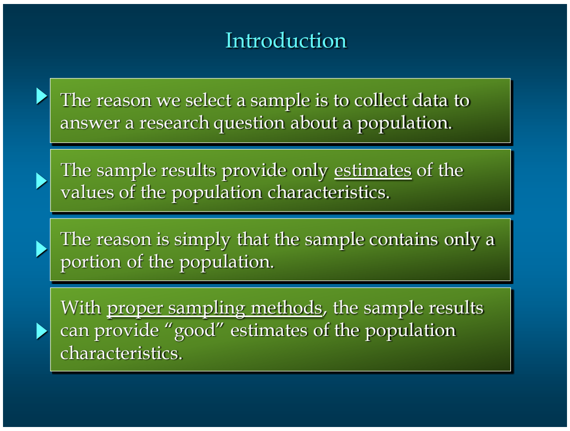
Chapter 7
Sampling and Sampling Distributions

- Selecting a Sample
- Point Estimation
- Introduction to Sampling Distributions
- Sampling Distribution of $\bar{x}$
- Sampling Distribution of $\bar{p}$
- Properties of Point Estimators
- Other Sampling Methods



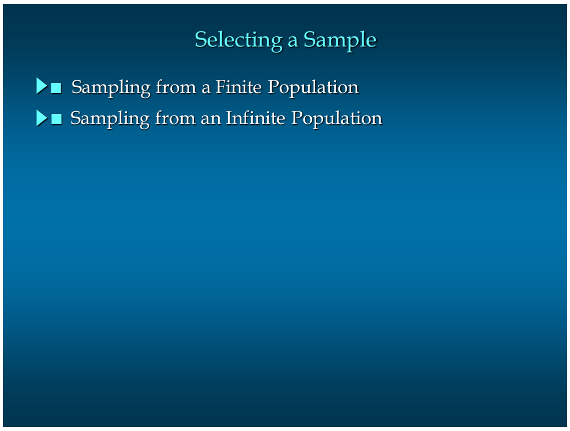
Introduction

- An element is the entity on which data are collected.
- A population is a collection of all the elements of interest.
- A sample is a subset of the population.
- The sampled population is the population from which the sample is drawn.
- A frame is a list of the elements that the sample will be selected from.



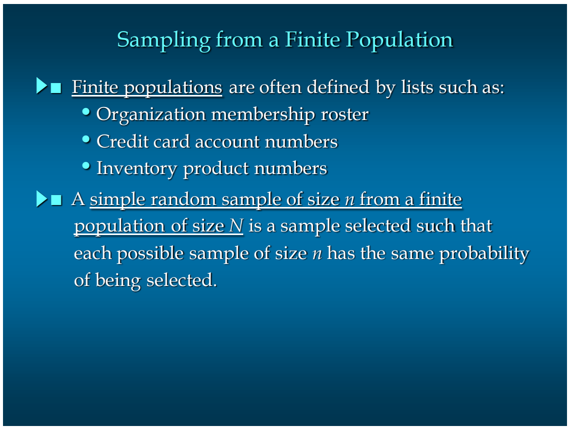
Introduction

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



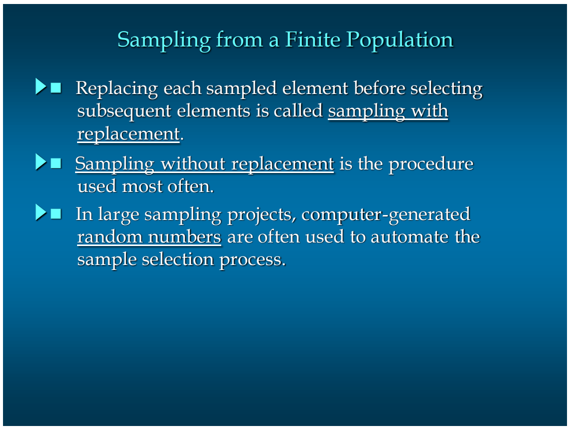
Selecting a Sample

- Sampling from a Finite Population
- Sampling from an Infinite Population



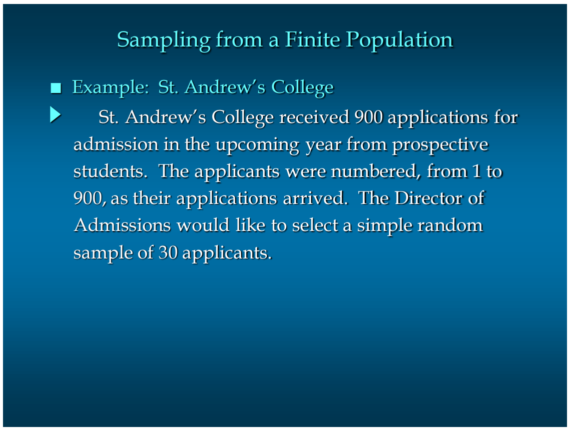
Sampling from a Finite Population

- Finite populations are often defined by lists such as:
  - Organization membership roster
  - Credit card account numbers
  - Inventory product numbers
- A simple random sample of size n from a finite population of size N is a sample selected such that each possible sample of size n has the same probability of being selected.



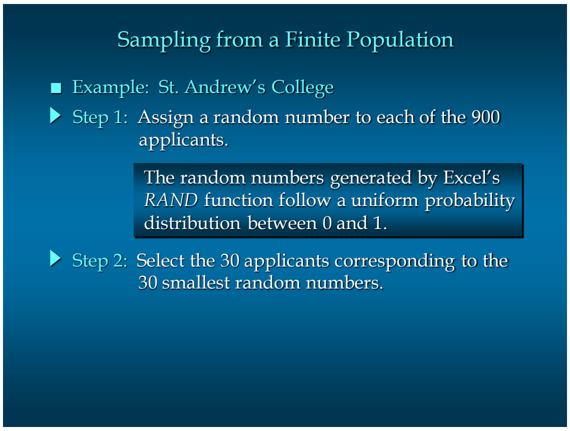
Sampling from a Finite Population

- Replacing each sampled element before selecting subsequent elements is called sampling with replacement.
- Sampling without replacement is the procedure used most often.
- In large sampling projects, computer-generated random numbers are often used to automate the sample selection process.



Sampling from a Finite Population

- Example: St. Andrew's College
- St. Andrew's College received 900 applications for admission in the upcoming year from prospective students. The applicants were numbered, from 1 to 900, as their applications arrived. The Director of Admissions would like to select a simple random sample of 30 applicants.

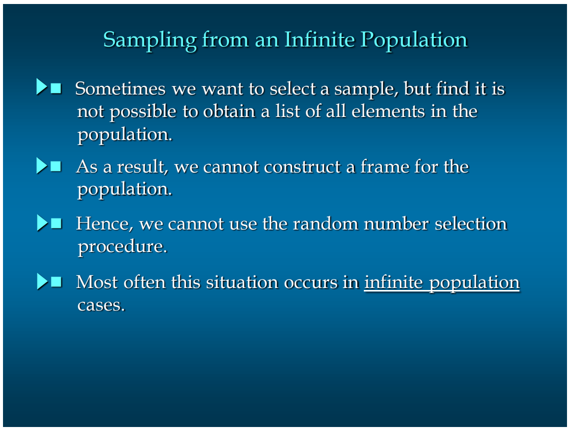


Sampling from a Finite Population

- Example: St. Andrew's College
- Step 1: Assign a random number to each of the 900 applicants.

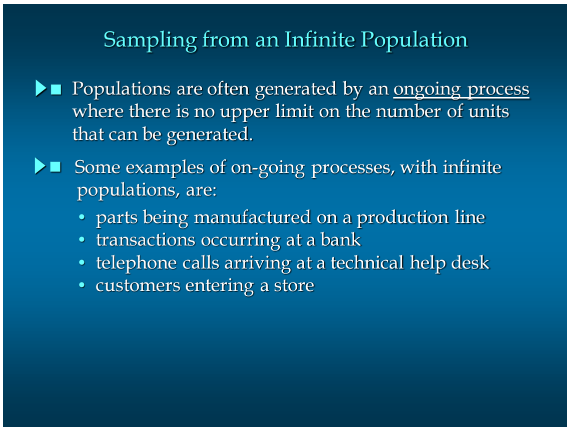
The random numbers generated by Excel's *RAND* function follow a uniform probability distribution between 0 and 1.

- Step 2: Select the 30 applicants corresponding to the 30 smallest random numbers.



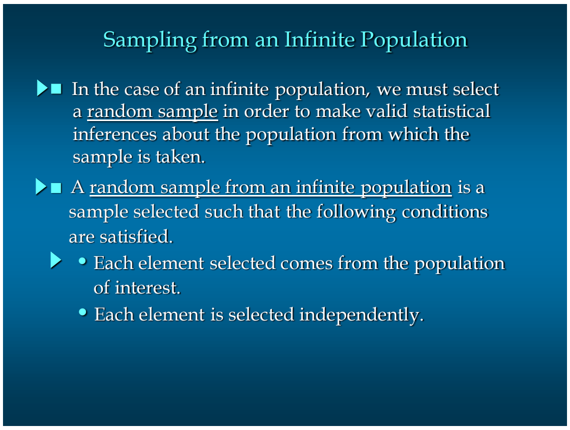
Sampling from an Infinite Population

- Sometimes we want to select a sample, but find it is not possible to obtain a list of all elements in the population.
- As a result, we cannot construct a frame for the population.
- Hence, we cannot use the random number selection procedure.
- Most often this situation occurs in infinite population cases.



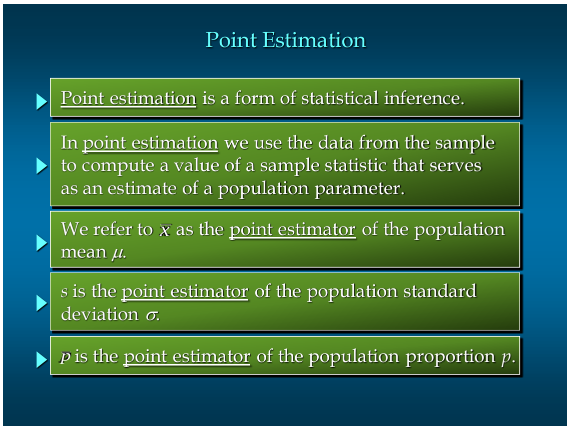
Sampling from an Infinite Population

- Populations are often generated by an ongoing process where there is no upper limit on the number of units that can be generated.
- Some examples of on-going processes, with infinite populations, are:
  - parts being manufactured on a production line
  - transactions occurring at a bank
  - telephone calls arriving at a technical help desk
  - customers entering a store



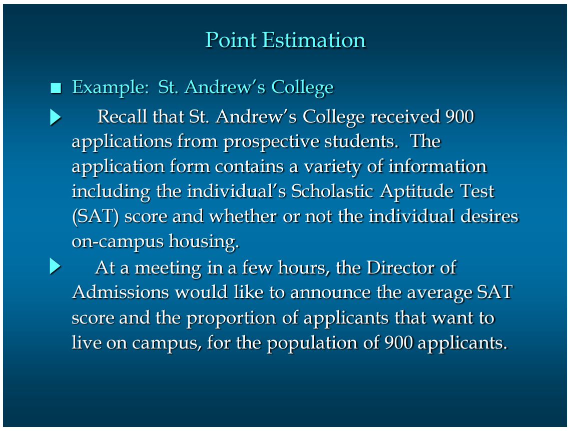
Sampling from an Infinite Population

- In the case of an infinite population, we must select a random sample in order to make valid statistical inferences about the population from which the sample is taken.
- A random sample from an infinite population is a sample selected such that the following conditions are satisfied.
  - Each element selected comes from the population of interest.
  - Each element is selected independently.



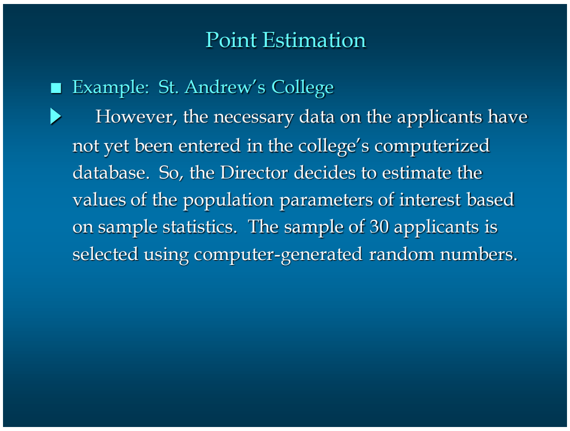
Point Estimation

- Point estimation is a form of statistical inference.
- In point estimation we use the data from the sample to compute a value of a sample statistic that serves as an estimate of a population parameter.
- We refer to $\bar{x}$ as the point estimator of the population mean μ.
- s is the point estimator of the population standard deviation σ.
- $\bar{p}$ is the point estimator of the population proportion p.



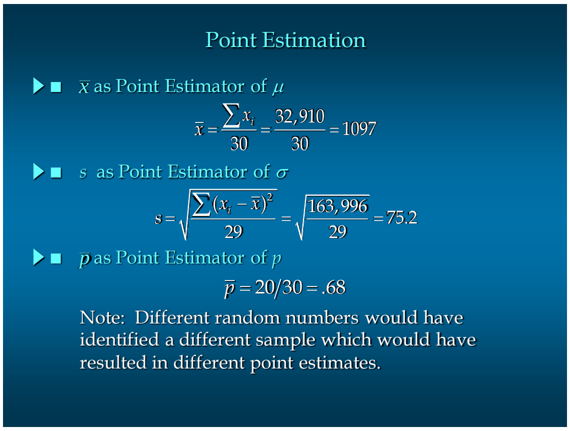
Point Estimation

- Example: St. Andrew's College
  - Recall that St. Andrew's College received 900 applications from prospective students. The application form contains a variety of information including the individual's Scholastic Aptitude Test (SAT) score and whether or not the individual desires on-campus housing.
  - At a meeting in a few hours, the Director of Admissions would like to announce the average SAT score and the proportion of applicants that want to live on campus, for the population of 900 applicants.



Point Estimation

- Example: St. Andrew's College
  - However, the necessary data on the applicants have not yet been entered in the college's computerized database. So, the Director decides to estimate the values of the population parameters of interest based on sample statistics. The sample of 30 applicants is selected using computer-generated random numbers.



Point Estimation

- $\bar{x}$ as Point Estimator of μ

$$\bar{x} = \frac{\sum x_i}{30} = \frac{32,910}{30} = 1097$$

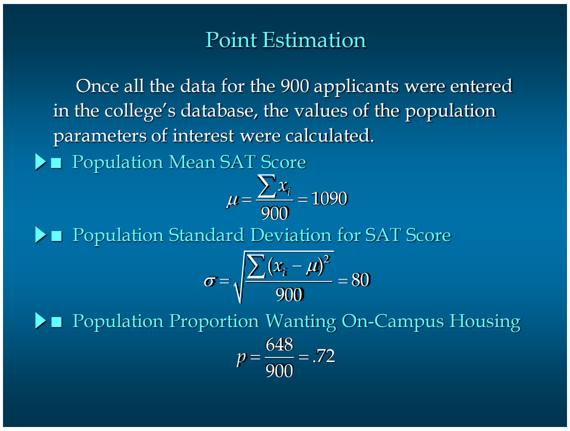
- s as Point Estimator of σ

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{29}} = \sqrt{\frac{163,996}{29}} = 75.2$$

- $\bar{p}$ as Point Estimator of p

$$\bar{p} = 20/30 = .68$$

Note: Different random numbers would have identified a different sample which would have resulted in different point estimates.



Point Estimation

Once all the data for the 900 applicants were entered in the college's database, the values of the population parameters of interest were calculated.

- Population Mean SAT Score

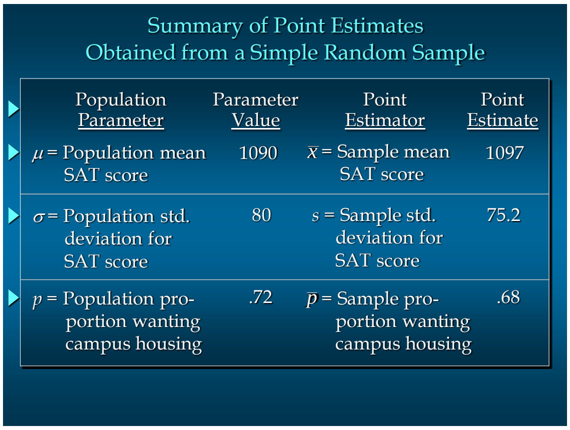
$$\mu = \frac{\sum x_i}{900} = 1090$$

- Population Standard Deviation for SAT Score

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{900}} = 80$$

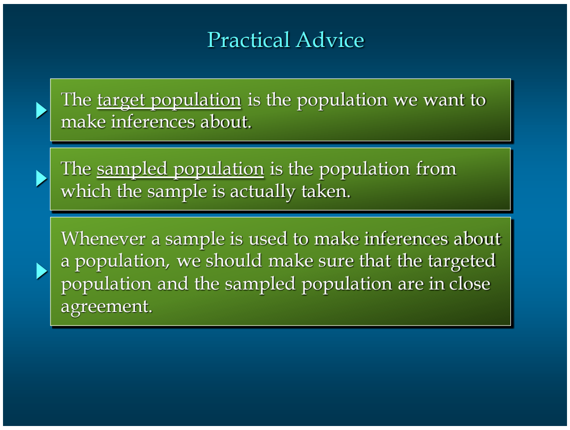
- Population Proportion Wanting On-Campus Housing

$$p = \frac{648}{900} = .72$$



Summary of Point Estimates Obtained from a Simple Random Sample

Population Parameter	Parameter Value	Point Estimator	Point Estimate
μ = Population mean SAT score	1090	$\bar{x}$ = Sample mean SAT score	1097
σ = Population std. deviation for SAT score	80	s = Sample std. deviation for SAT score	75.2
p = Population proportion wanting campus housing	.72	$\bar{p}$ = Sample proportion wanting campus housing	.68

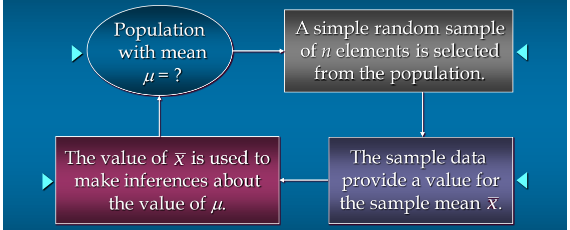


Practical Advice

- The target population is the population we want to make inferences about.
- The sampled population is the population from which the sample is actually taken.
- Whenever a sample is used to make inferences about a population, we should make sure that the targeted population and the sampled population are in close agreement.

Sampling Distribution of $\bar{x}$

- Process of Statistical Inference



Population with mean μ = ?

A simple random sample of n elements is selected from the population.

The sample data provide a value for the sample mean $\bar{x}$.

The value of $\bar{x}$ is used to make inferences about the value of μ.

Sampling Distribution of $\bar{x}$

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

- Expected Value of $\bar{x}$

$$E(\bar{x}) = \mu$$

where: μ = the population mean

When the expected value of the point estimator equals the population parameter, we say the point estimator is unbiased.

Sampling Distribution of $\bar{x}$

- Standard Deviation of $\bar{x}$

We will use the following notation to define the standard deviation of the sampling distribution of $\bar{x}$.

$\sigma_{\bar{x}}$ = the standard deviation of $\bar{x}$

σ = the standard deviation of the population

n = the sample size

N = the population size

Sampling Distribution of $\bar{x}$

- Standard Deviation of $\bar{x}$

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

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

Sampling Distribution of $\bar{x}$

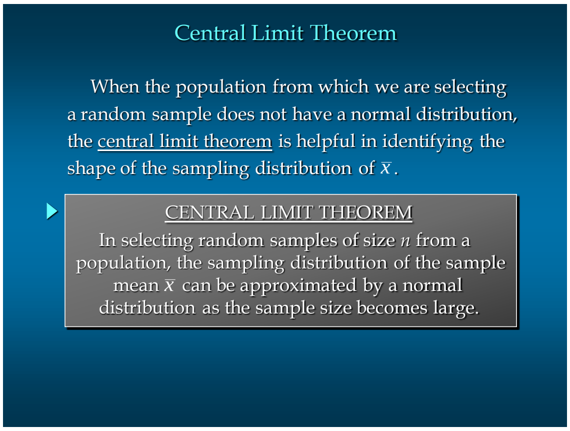
When the population has a normal distribution, the sampling distribution of $\bar{x}$ is normally distributed for any sample size.

In most applications, the sampling distribution of $\bar{x}$ can be approximated by a normal distribution whenever the sample is size 30 or more.

In cases where the population is highly skewed or outliers are present, samples of size 50 may be needed.

Sampling Distribution of $\bar{x}$

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.

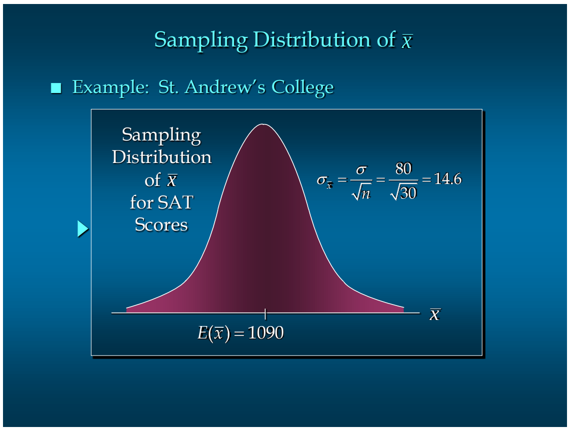


Central Limit Theorem

When the population from which we are selecting a random sample does not have a normal distribution, the central limit theorem is helpful in identifying the shape of the sampling distribution of $\bar{x}$.

CENTRAL LIMIT THEOREM

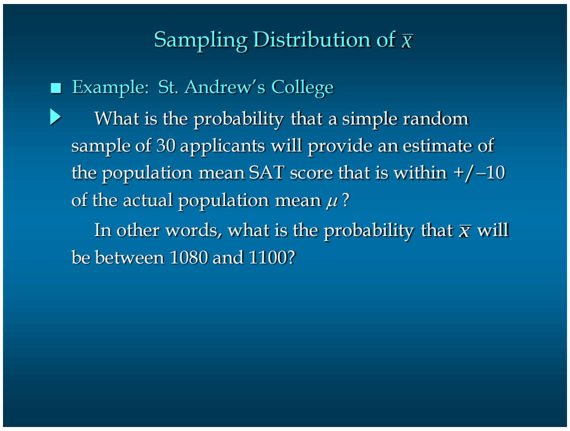
In selecting random samples of size n from a population, the sampling distribution of the sample mean $\bar{x}$ can be approximated by a normal distribution as the sample size becomes large.



Sampling Distribution of $\bar{x}$

- Example: St. Andrew's College

Sampling Distribution of $\bar{x}$ for SAT Scores
$\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{80}{\sqrt{30}} = 14.6$
$\bar{x}$
$E(\bar{x}) = 1090$

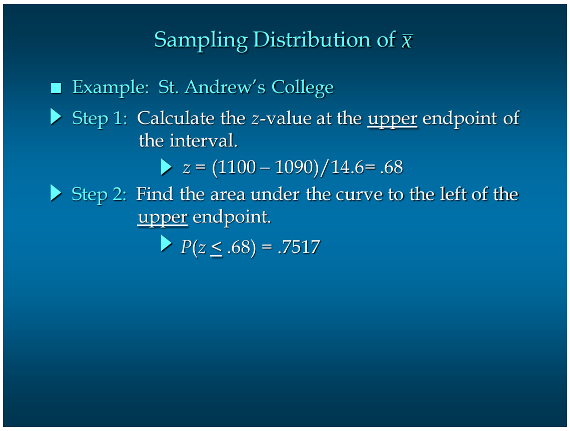


Sampling Distribution of $\bar{x}$

- Example: St. Andrew's College

What is the probability that a simple random sample of 30 applicants will provide an estimate of the population mean SAT score that is within +/−10 of the actual population mean μ?

In other words, what is the probability that $\bar{x}$ will be between 1080 and 1100?



Sampling Distribution of $\bar{x}$

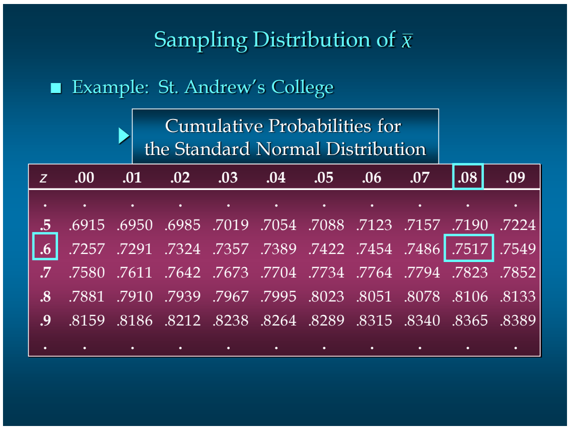
- Example: St. Andrew's College

Step 1: Calculate the z-value at the upper endpoint of the interval.

$z = (1100 - 1090)/14.6 = .68$

Step 2: Find the area under the curve to the left of the upper endpoint.

$P(z \leq .68) = .7517$

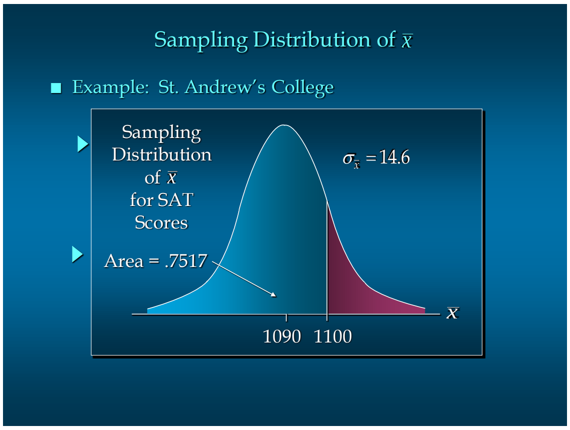


Sampling Distribution of $\bar{x}$

- Example: St. Andrew's College

Cumulative Probabilities for the Standard Normal Distribution

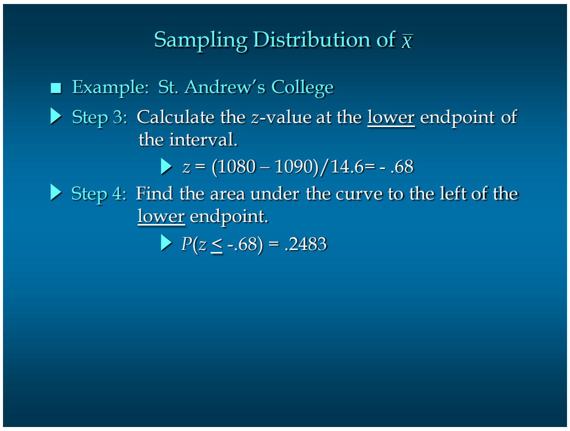
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.	.	.	.	.	.	.	.	.	.	.
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
.	.	.	.	.	.	.	.	.	.	.



Sampling Distribution of $\bar{x}$

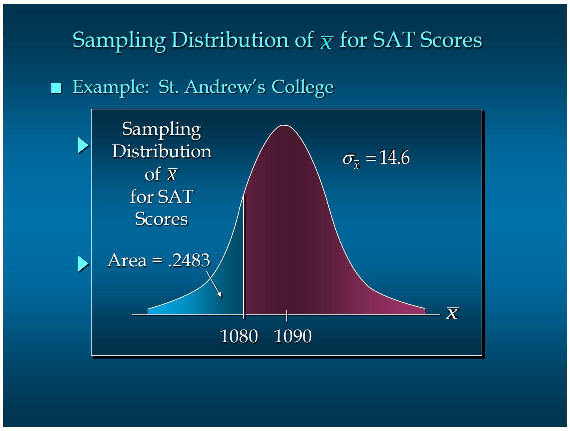
- Example: St. Andrew's College

Sampling Distribution of $\bar{x}$ for SAT Scores
$\sigma_{\bar{x}} = 14.6$
Area = .7517
$\bar{x}$
1090
1100



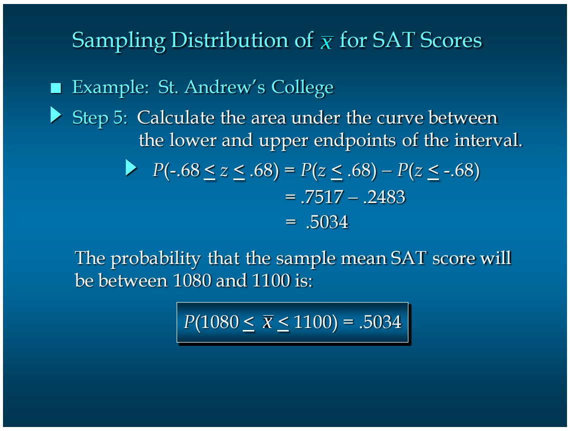
Sampling Distribution of $\bar{x}$

- Example: St. Andrew's College
- Step 3: Calculate the *z*-value at the lower endpoint of the interval.
  - $z = (1080 – 1090)/14.6= - .68$
- Step 4: Find the area under the curve to the left of the lower endpoint.
  - $P(z \le -.68) = .2483$



Sampling Distribution of $\bar{x}$ for SAT Scores

- Example: St. Andrew's College

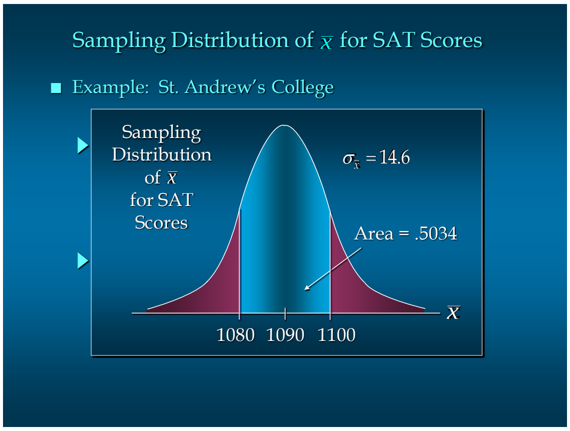


Sampling Distribution of $\bar{x}$ for SAT Scores

- Example: St. Andrew's College
- Step 5: Calculate the area under the curve between the lower and upper endpoints of the interval.
  - $P(-.68 \le z \le .68) = P(z \le .68) – P(z \le -.68)$
    $= .7517 – .2483$
    $= .5034$

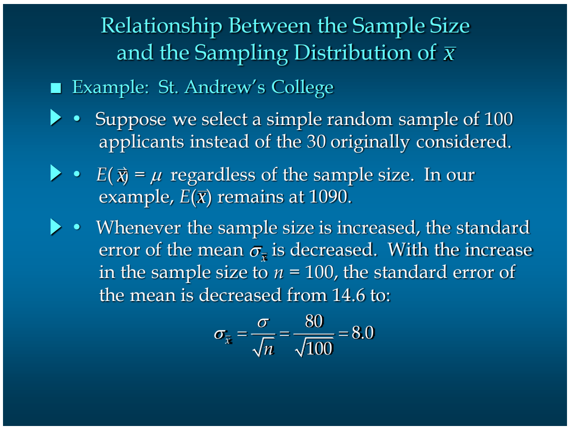
The probability that the sample mean SAT score will be between 1080 and 1100 is:

$$\boxed{P(1080 \le \bar{x} \le 1100) = .5034}$$



Sampling Distribution of $\bar{x}$ for SAT Scores

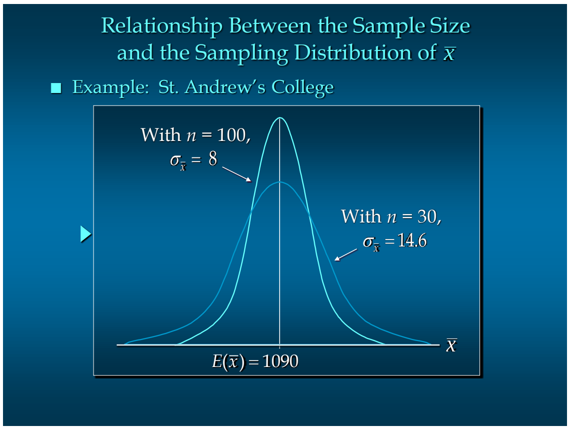
- Example: St. Andrew's College



Relationship Between the Sample Size and the Sampling Distribution of $\bar{x}$

- Example: St. Andrew's College
- Suppose we select a simple random sample of 100 applicants instead of the 30 originally considered.
- $E(\bar{x}) = \mu$ regardless of the sample size. In our example, $E(\bar{x})$ remains at 1090.
- Whenever the sample size is increased, the standard error of the mean $\sigma_{\bar{x}}$ is decreased. With the increase in the sample size to n = 100, the standard error of the mean is decreased from 14.6 to:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{80}{\sqrt{100}} = 8.0$$



Relationship Between the Sample Size and the Sampling Distribution of $\bar{x}$

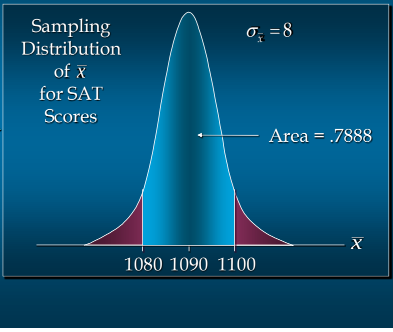
- Example: St. Andrew's College

Relationship Between the Sample Size and the Sampling Distribution of $\bar{x}$

- Example: St. Andrew's College
  - Recall that when $n = 30$, $P(1080 \le \bar{x} \le 1100) = .5034$.
  - We follow the same steps to solve for $P(1080 \le \bar{x} \le 1100)$ when $n = 100$ as we showed earlier when $n = 30$.
  - Now, with $n = 100$, $P(1080 \le \bar{x} \le 1100) = .7888$.
  - Because the sampling distribution with $n = 100$ 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 $n = 30$.

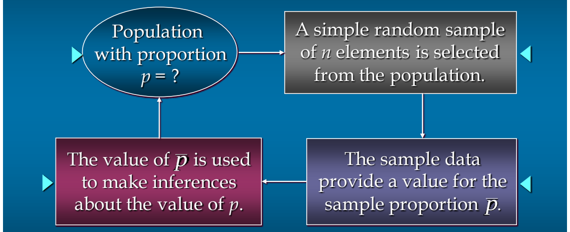
Relationship Between the Sample Size and the Sampling Distribution of $\bar{x}$

- Example: St. Andrew's College



Sampling Distribution of $\bar{p}$

- Making Inferences about a Population Proportion



Population with proportion p = ?

A simple random sample of n elements is selected from the population.

The sample data provide a value for the sample proportion $\bar{p}$.

The value of $\bar{p}$ is used to make inferences about the value of p.

Sampling Distribution of $\bar{p}$

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

- Expected Value of $\bar{p}$

$$E(\bar{p}) = p$$

where:

p = the population proportion

Sampling Distribution of $\bar{p}$

- Standard Deviation of $\bar{p}$

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

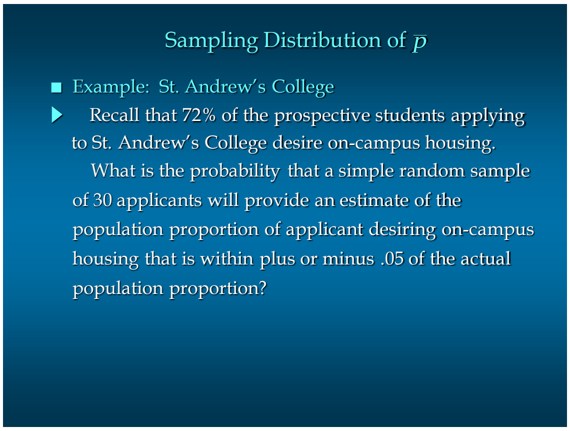
- σ_p is referred to as the standard error of the proportion.
- $\sqrt{(N-n)/(N-1)}$ is the finite population correction factor.

Form of the Sampling Distribution of $\bar{p}$

The sampling distribution of $\bar{p}$ can be approximated by a normal distribution whenever the sample size is large enough to satisfy the two conditions:

$np \ge 5$ and $n(1 - p) \ge 5$

. . . because when these conditions are satisfied, the probability distribution of x in the sample proportion, $\bar{p} = x/n$, can be approximated by normal distribution (and because n is a constant).

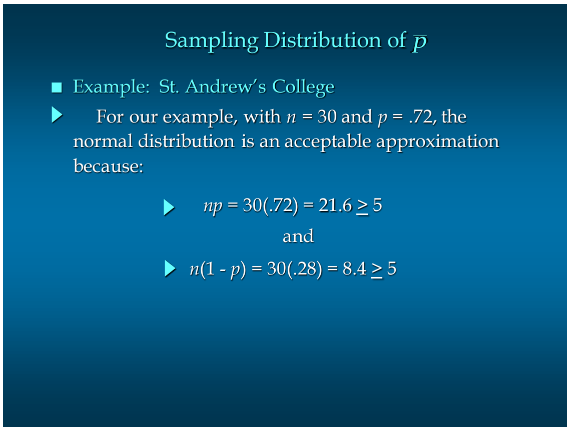


Sampling Distribution of $\bar{p}$

- Example: St. Andrew's College

Recall 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 an estimate of the population proportion of applicant desiring on-campus housing that is within plus or minus .05 of the actual population proportion?



Sampling Distribution of $\bar{p}$

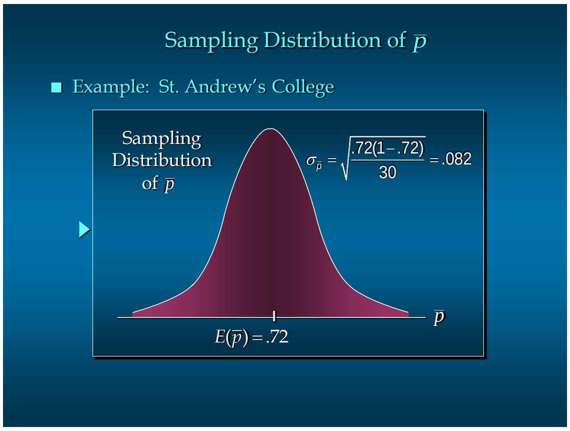
- Example: St. Andrew's College

For our example, with $n = 30$ and $p = .72$, the normal distribution is an acceptable approximation because:

$np = 30(.72) = 21.6 \geq 5$

and

$n(1 - p) = 30(.28) = 8.4 \geq 5$



Sampling Distribution of $\bar{p}$

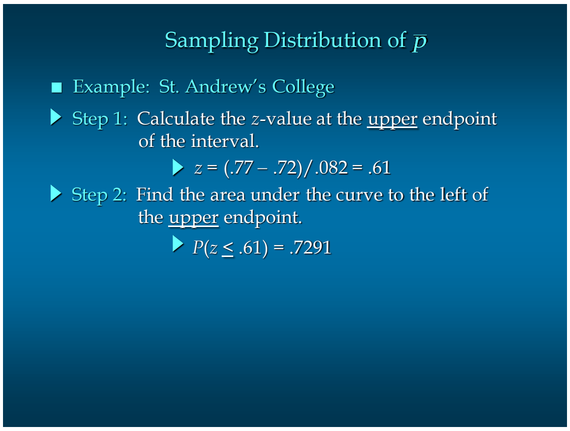
- Example: St. Andrew's College

Sampling Distribution of $\bar{p}$

$\sigma_{\bar{p}} = \sqrt{\dfrac{.72(1-.72)}{30}} = .082$

$\bar{p}$

$E(\bar{p}) = .72$



Sampling Distribution of $\bar{p}$

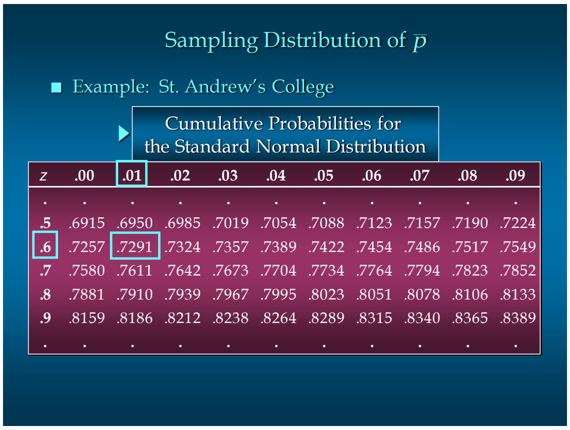
- Example: St. Andrew's College

Step 1: Calculate the z-value at the upper endpoint of the interval.

$z = (.77 - .72)/.082 = .61$

Step 2: Find the area under the curve to the left of the upper endpoint.

$P(z \leq .61) = .7291$

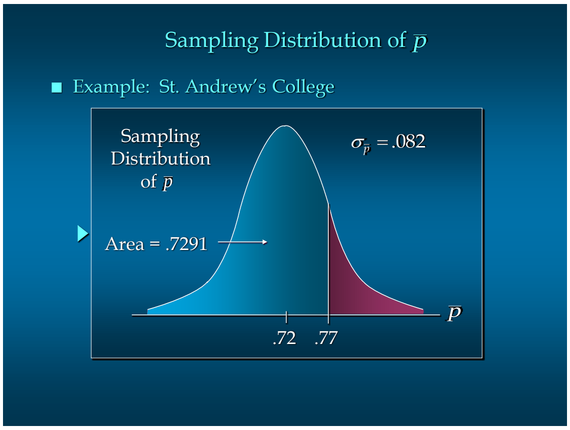


Sampling Distribution of $\bar{p}$

- Example: St. Andrew's College

Cumulative Probabilities for the Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.	.	.	.	.	.	.	.	.	.	.
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
.	.	.	.	.	.	.	.	.	.	.



Sampling Distribution of $\bar{p}$

- Example: St. Andrew's College

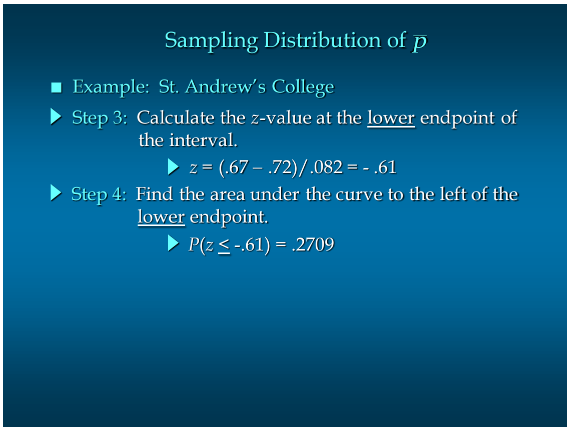
Sampling Distribution of $\bar{p}$

$\sigma_{\bar{p}} = .082$

Area = .7291

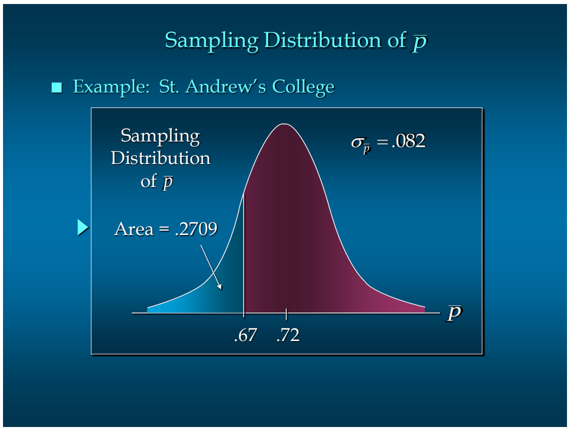
$\bar{p}$

.72 .77



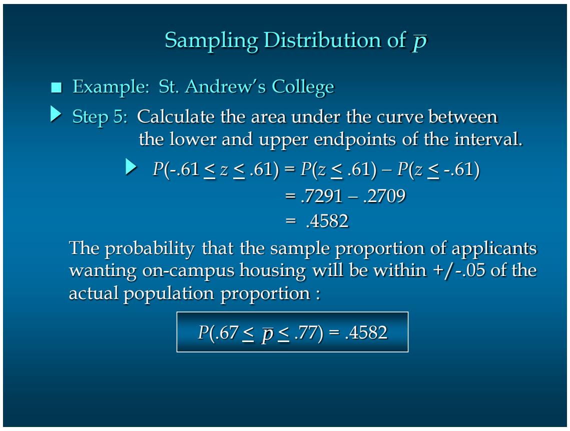
Sampling Distribution of $\bar{p}$

- Example: St. Andrew's College
- Step 3: Calculate the *z*-value at the lower endpoint of the interval.
  - $z = (.67 – .72)/.082 = - .61$
- Step 4: Find the area under the curve to the left of the lower endpoint.
  - $P(z \leq -.61) = .2709$



Sampling Distribution of $\bar{p}$

- Example: St. Andrew's College

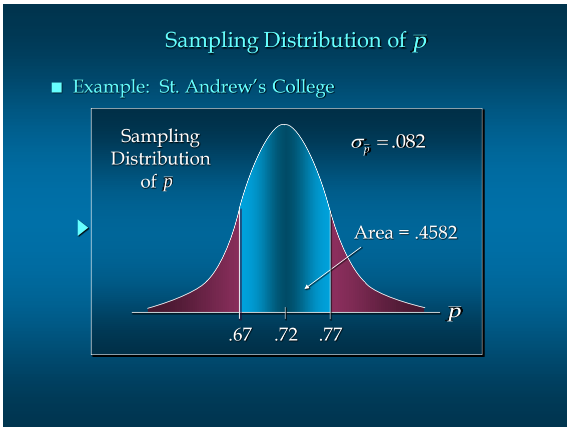


Sampling Distribution of $\bar{p}$

- Example: St. Andrew's College
- Step 5: Calculate the area under the curve between the lower and upper endpoints of the interval.
  - $P(-.61 \leq z \leq .61) = P(z \leq .61) – P(z \leq -.61)$
    $= .7291 – .2709$
    $= .4582$

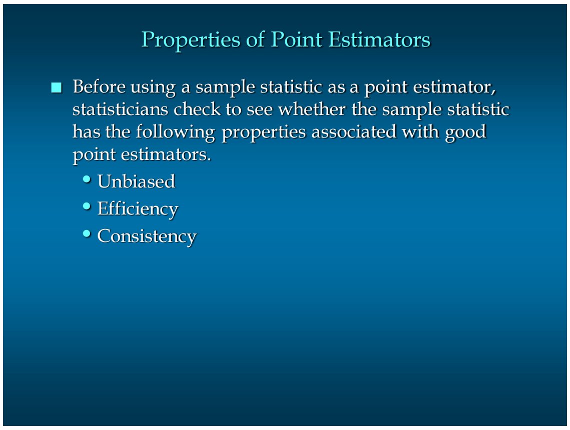
The probability that the sample proportion of applicants wanting on-campus housing will be within +/-.05 of the actual population proportion :

$$\boxed{P(.67 \leq \bar{p} \leq .77) = .4582}$$



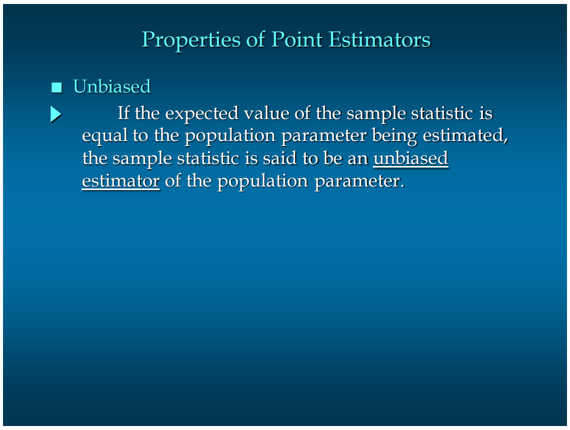
Sampling Distribution of $\bar{p}$

- Example: St. Andrew's College



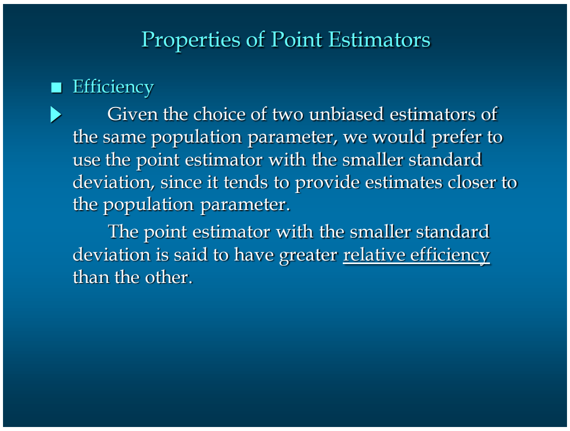
Properties of Point Estimators

- Before using a sample statistic as a point estimator, statisticians check to see whether the sample statistic has the following properties associated with good point estimators.
  - Unbiased
  - Efficiency
  - Consistency



Properties of Point Estimators

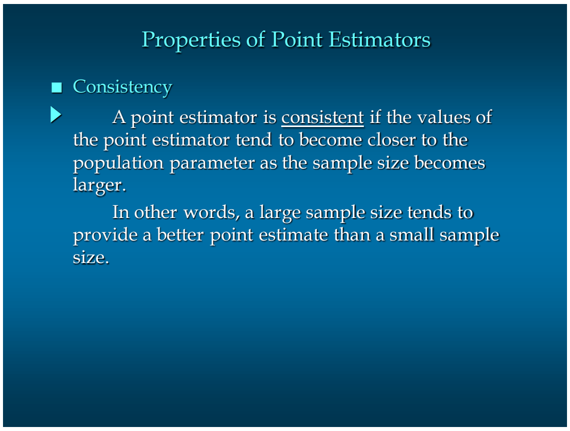
- Unbiased
- If the expected value of the sample statistic is equal to the population parameter being estimated, the sample statistic is said to be an unbiased estimator of the population parameter.



Properties of Point Estimators

- Efficiency
  - Given the choice of two unbiased estimators of the same population parameter, we would prefer to use the point estimator with the smaller standard deviation, since it tends to provide estimates closer to the population parameter.

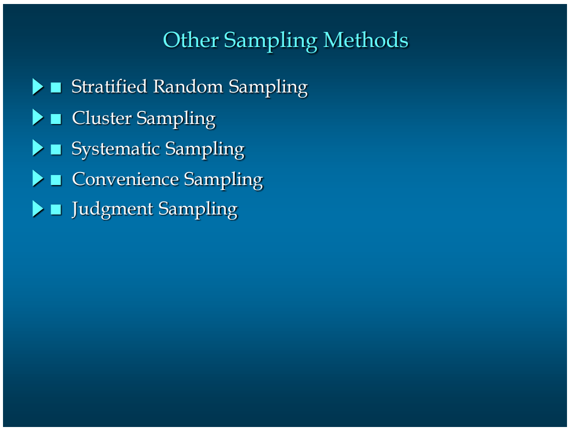
    The point estimator with the smaller standard deviation is said to have greater relative efficiency than the other.



Properties of Point Estimators

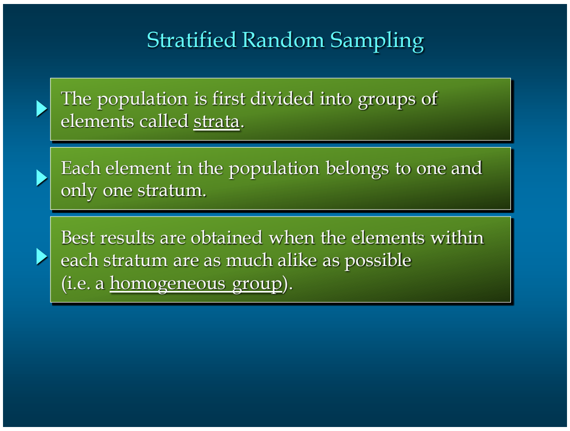
- Consistency
  - A point estimator is consistent if the values of the point estimator tend to become closer to the population parameter as the sample size becomes larger.

    In other words, a large sample size tends to provide a better point estimate than a small sample size.



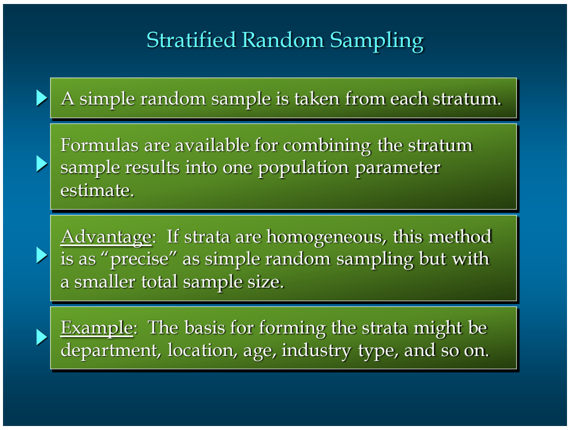
Other Sampling Methods

- Stratified Random Sampling
- Cluster Sampling
- Systematic Sampling
- Convenience Sampling
- Judgment Sampling



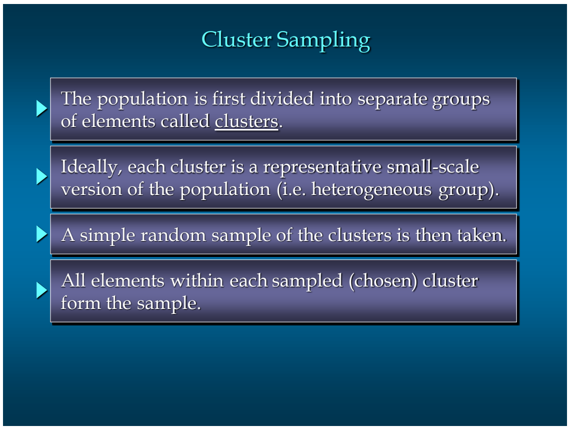
Stratified Random Sampling

- The population is first divided into groups of elements called strata.
- Each element in the population belongs to one and only one stratum.
- Best results are obtained when the elements within each stratum are as much alike as possible (i.e. a homogeneous group).



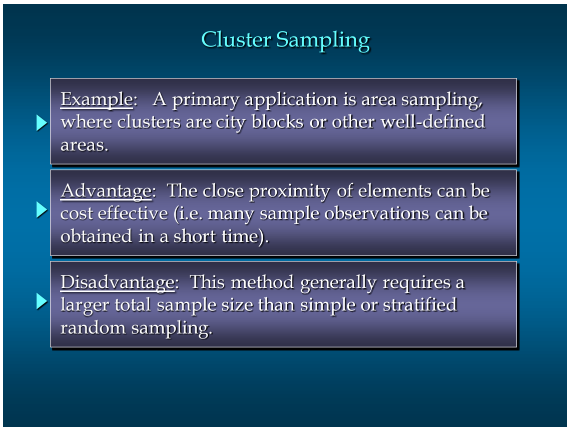
Stratified Random Sampling

- A simple random sample is taken from each stratum.
- Formulas are available for combining the stratum sample results into one population parameter estimate.
- Advantage: If strata are homogeneous, this method is as "precise" as simple random sampling but with a smaller total sample size.
- Example: The basis for forming the strata might be department, location, age, industry type, and so on.



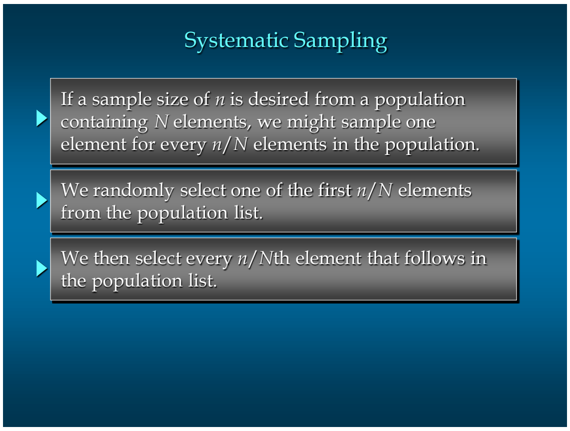
Cluster Sampling

- The population is first divided into separate groups of elements called clusters.
- Ideally, each cluster is a representative small-scale version of the population (i.e. heterogeneous group).
- A simple random sample of the clusters is then taken.
- All elements within each sampled (chosen) cluster form the sample.



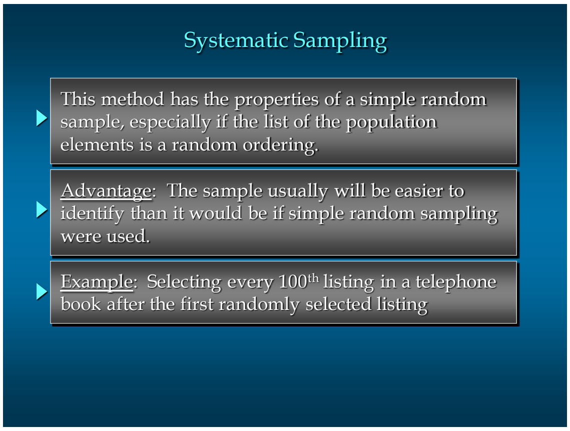
Cluster Sampling

- Example: A primary application is area sampling, where clusters are city blocks or other well-defined areas.
- Advantage: The close proximity of elements can be cost effective (i.e. many sample observations can be obtained in a short time).
- Disadvantage: This method generally requires a larger total sample size than simple or stratified random sampling.



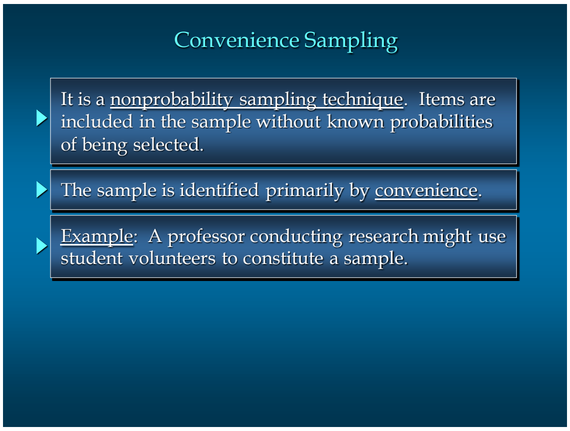
Systematic Sampling

- If a sample size of n is desired from a population containing N elements, we might sample one element for every n/N elements in the population.
- We randomly select one of the first n/N elements from the population list.
- We then select every n/Nth element that follows in the population list.



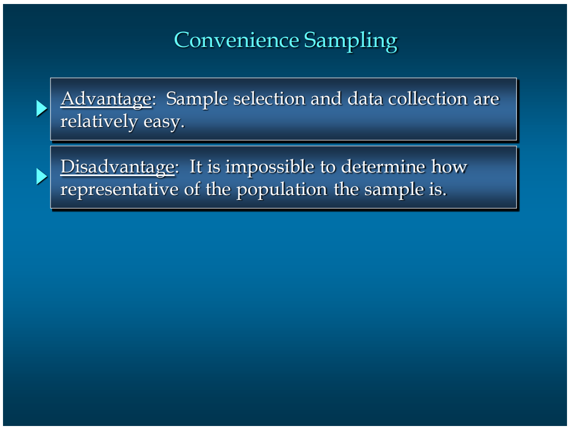
Systematic Sampling

- This method has the properties of a simple random sample, especially if the list of the population elements is a random ordering.
- Advantage: The sample usually will be easier to identify than it would be if simple random sampling were used.
- Example: Selecting every 100th listing in a telephone book after the first randomly selected listing



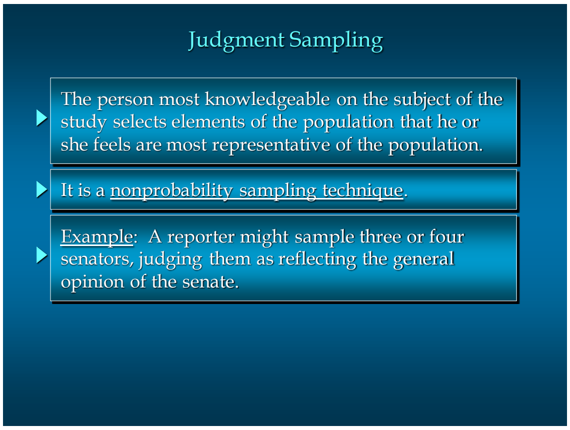
Convenience Sampling

- It is a nonprobability sampling technique. Items are included in the sample without known probabilities of being selected.
- The sample is identified primarily by convenience.
- Example: A professor conducting research might use student volunteers to constitute a sample.



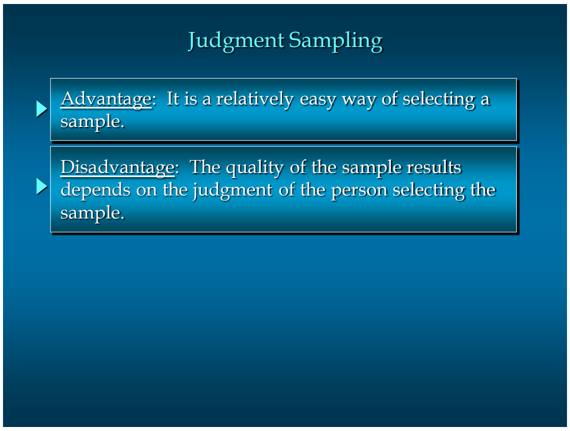
Convenience Sampling

- Advantage: Sample selection and data collection are relatively easy.
- Disadvantage: It is impossible to determine how representative of the population the sample is.



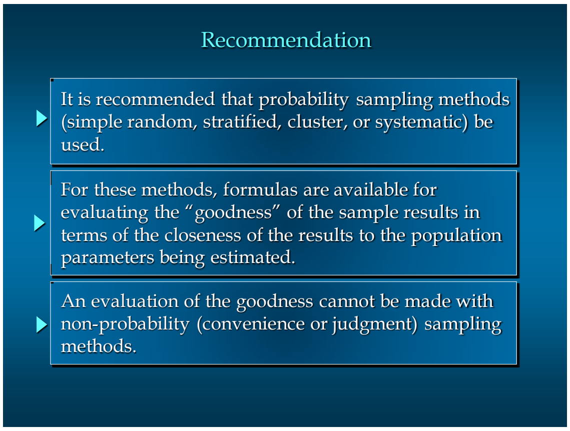
Judgment Sampling

- The person most knowledgeable on the subject of the study selects elements of the population that he or she feels are most representative of the population.
- It is a nonprobability sampling technique.
- Example: A reporter might sample three or four senators, judging them as reflecting the general opinion of the senate.



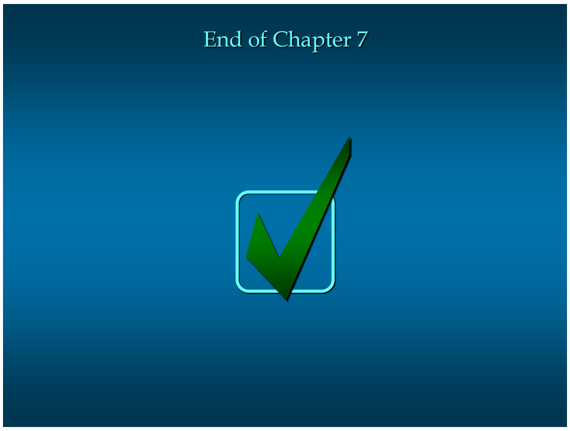
Judgment Sampling

- Advantage: It is a relatively easy way of selecting a sample.
- Disadvantage: The quality of the sample results depends on the judgment of the person selecting the sample.



Recommendation

- It is recommended that probability sampling methods (simple random, stratified, cluster, or systematic) be used.
- For these methods, formulas are available for evaluating the "goodness" of the sample results in terms of the closeness of the results to the population parameters being estimated.
- An evaluation of the goodness cannot be made with non-probability (convenience or judgment) sampling methods.



End of Chapter 7