BAN 602: Quantitative Fundamentals

Spring, 2020 Lecture Slides – Week 3

CAL STATE
EAST BAY

Agenda

- Sampling Distribution
  - Selecting a Sample
  - Point Estimation
  - Sampling Distribution of $\bar{x}$
  - Sampling Distribution of $\bar{p}$
- Interval Estimation
  - Interval Estimate of a Population Mean: σ Known
  - Interval Estimate of a Population Mean: σ Unknown
  - Sample Size for an Interval Estimate of a Population Mean
  - Interval Estimate of a Population Proportion
  - Sample Size for an Interval Estimate of a Population Proportion

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.

- 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.

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.

- Replacing each sampled element before selecting subsequent elements is called sampling with replacement. An element can appear in the sample more than once.

- 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 an Infinite Population

- Sometimes we want to select a sample, but find that 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 the case of 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

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.

The data on the applicants have not yet been entered in the college's database. So the Director decides to estimate the values of the population parameters of interest based on sample statistics. A sample of 30 applicants is selected using computer-generated random numbers.

Point Estimation

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

- $\bar{x}$ is the point estimator of the population mean, μ

$$\bar{x} = \frac{\sum x_i}{n} = \frac{50{,}520}{30} = 1684$$

- s is the point estimator of the population standard deviation, σ

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{2470.8}{29}} = 85.2$$

- $\bar{p}$ is the point estimator of the population proportion, p.

$$\bar{p} = {}^{20}/_{30} = .67$$

Point Estimation

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

- Population Mean SAT Score: $\mu = \frac{\sum x_i}{900} = 1697$
- Population Standard Deviation for SAT Score: $\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{900}} = 87.4$
- Population proportion wanting On-Campus Housing: $p = 648/900 = 0.72$

Summary of Point Estimates Obtained from a Simple Random Sample

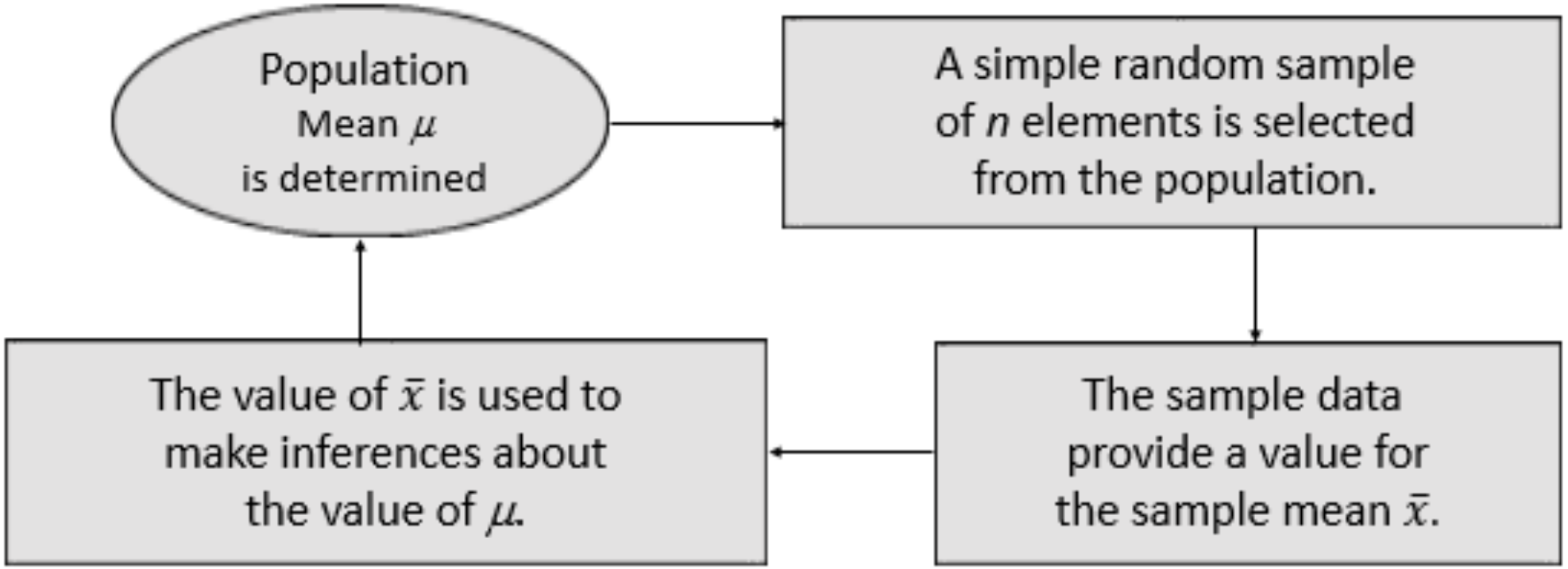
Population Parameter	Parameter Value	Point Estimator	Point Estimate
μ = Population mean SAT score	1697	$\bar{x}$ = Sample mean SAT score	1684
σ = Population std. deviation for SAT score	87.4	s = Sample std. deviation for SAT score	85.2
p = Population proportion wanting campus housing	0.72	$\bar{p}$ = Sample proportion wanting on campus housing	0.67

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



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 $\bar{x}$ is $E(\bar{x}) = \mu$, where μ = the population mean.
- When the expected value of the point estimator equals the population parameter, we say the point estimator is unbiased.
- 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 $\bar{x}$

- The standard deviation of $\bar{x}$, for a finite population is $\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right)$.
- The standard deviation of $\bar{x}$, for an infinite population $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.
- A finite population is treated as being infinite if $n/N \leq 0.05$.
- $\sqrt{\frac{N-n}{N-1}}$ 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.
- 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

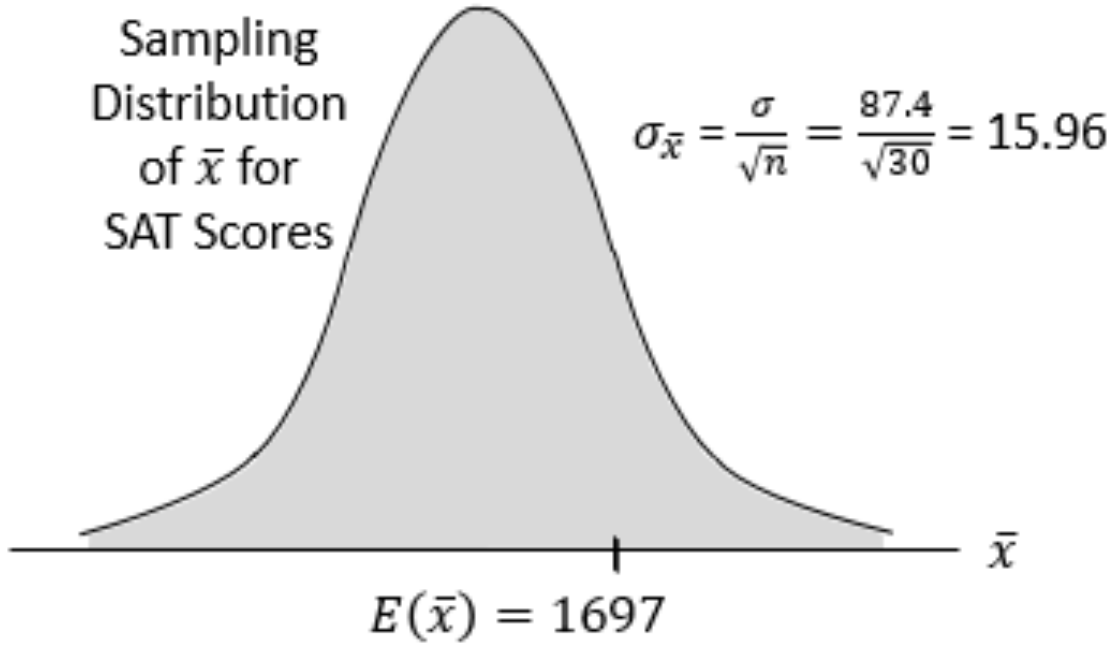
When the population from which we are selecting a random sample does not have a normal distribution, the central limit theorem is helpful in identifying the shape of the sampling distribution of $\bar{x}$.

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



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 1687 and 1707?

Sampling Distribution of $\bar{x}$

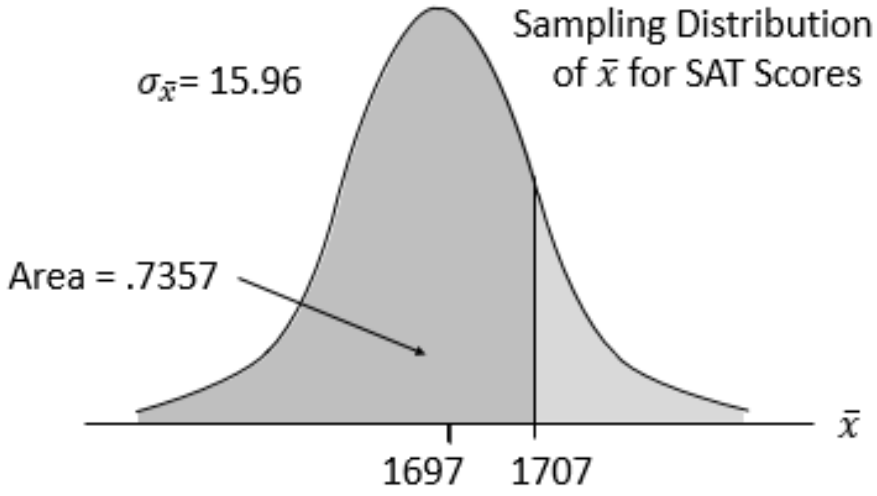
Example: St. Andrew's College

Step 1: Calculate the *z*-value at the upper endpoint of the interval. $z = \dfrac{(1707 - 1697)}{15.96} = 0.63$

Step 2: Find the area under the curve to the left of the upper endpoint. $P(z \leq 0.63) = 0.7357$

Cumulative Probabilities for the Standard Normal Distribution

z	.00	.01	.02	.03	.04
.	.	.	.	.	.
.5	.6915	.6950	.6985	.7019	.7054
.6	.7257	.7291	.7324	.737	.7389
.7	.7580	.7611	.7642	.7673	.7704
.8	.7881	.7910	.7939	.7967	.7995
.9	.8159	.8186	.8212	.8238	.8264

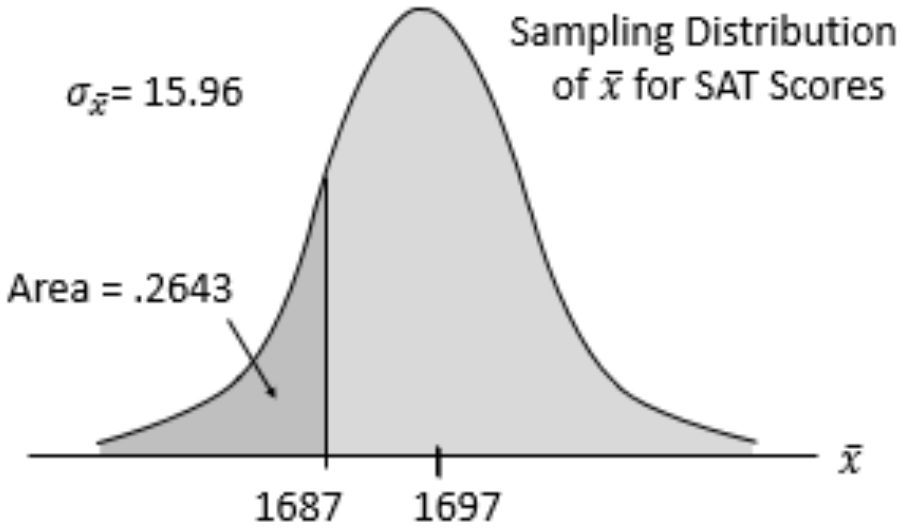


Sampling Distribution of $\bar{x}$

Example: St. Andrew's College

Step 3: Calculate the *z*-value at the lower endpoint of the interval. $z = \dfrac{(1687 - 1697)}{15.96} = -0.63$

Step 4: Find the area under the curve to the left of the lower endpoint. $P(z \leq -0.63) = 0.2643$



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

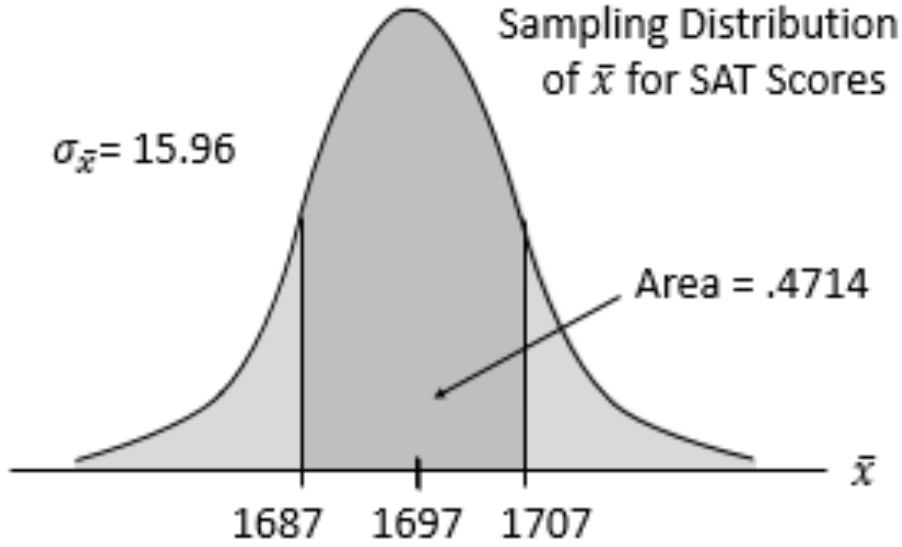
Example: St. Andrew's College

Step 5: Calculate the area under the curve between the lower and upper endpoints of the interval.

$$P(-0.63 <= z <= 0.63) = P(z <= 0.63) - P(z <= -0.63)$$
$$= 0.7357 - 0.2643$$
$$= 0.4714$$

The probability that the estimate of population mean SAT score will be between 1687 and 1707 is:

$$P(1687 \leq \bar{x} \leq 1707) = 0.4714$$

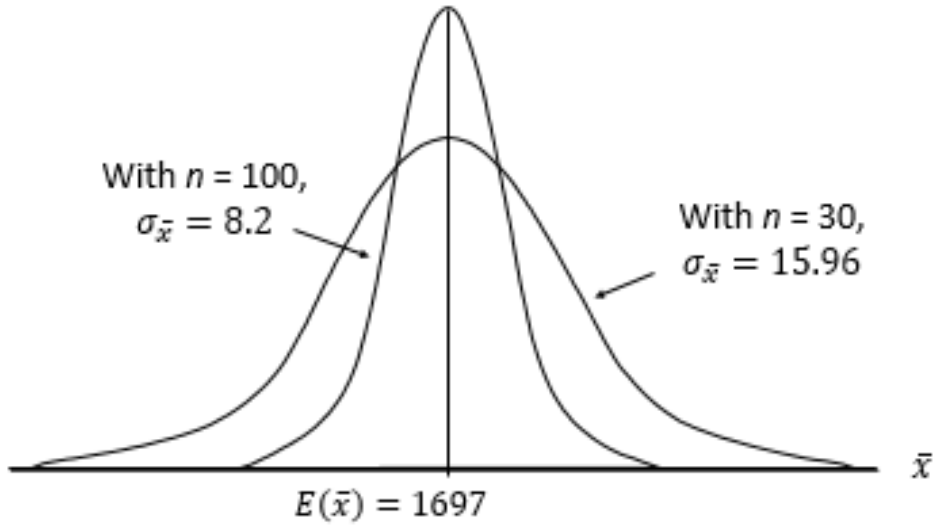


Relationship Between Sample Size Sampling Distribution of $\bar{x}$

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 1697.
- 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 15.96 to:

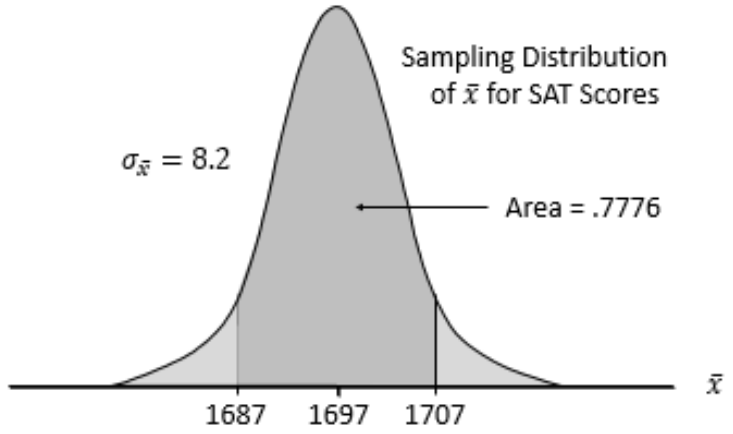
$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\left(\frac{\sigma}{\sqrt{n}}\right) = \sqrt{\frac{900-100}{900-1}}\left(\frac{87.4}{\sqrt{100}}\right) = 0.9433(8.74) = 8.2$$



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

Example: St. Andrew's College

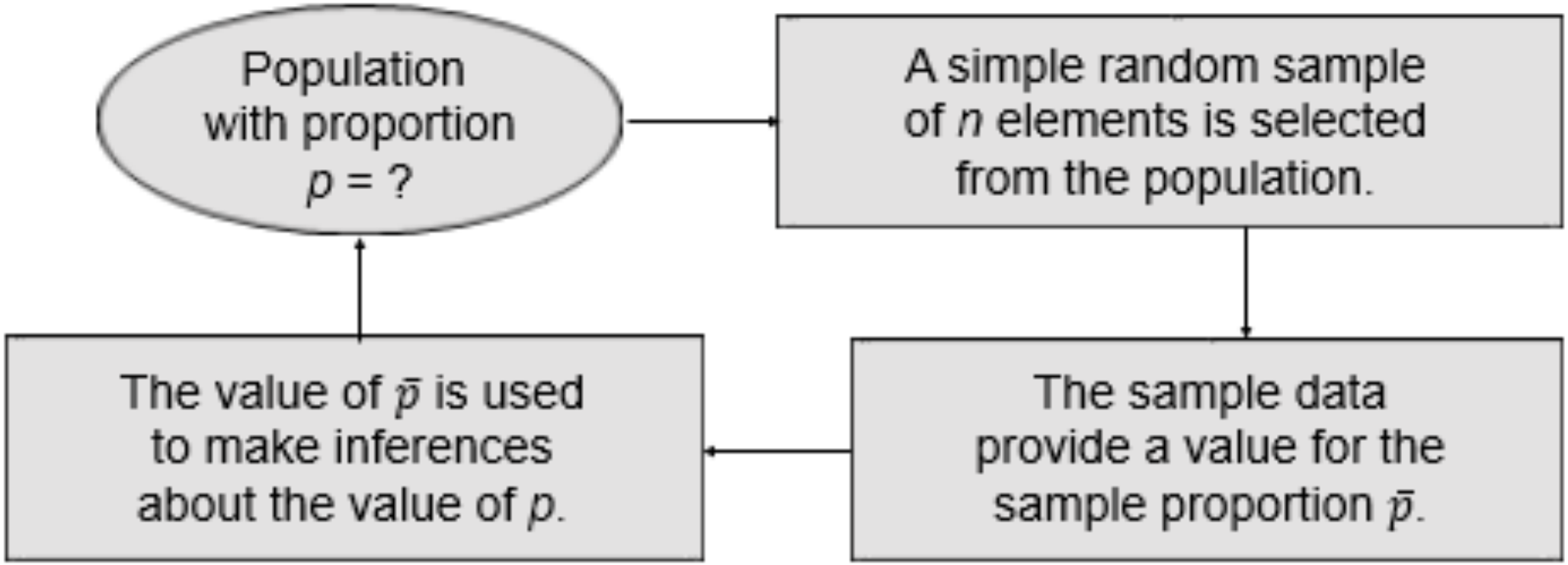
- Recall that when $n = 30$, $P(1687 \leq \bar{x} \leq 1707) = .4714$.
- We follow the same steps to solve for $P(1687 \leq \bar{x} \leq 1707)$ when $n = 100$ as we showed earlier when $n = 30$.
- Now, with $n = 100$, $P(1687 \leq \bar{x} \leq 1707) = .7776$.
- Because the sampling distribution with $n = 100$ 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 $n = 30$.



CAL STATE EAST BAY

Sampling Distribution of $\bar{p}$

Making Inferences about a Population Proportion



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

- Standard Deviation of $\bar{p}$

Finite Population

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

Infinite Population

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

- $\sigma_{\bar{p}}$ is referred to as the standard error of the proportion.
- $\sqrt{(N-n)/(N-1)}$ is the finite population correction factor.
- 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: $np \geq 5$ and $n(1-p) \geq 5$

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

Sampling Distribution of $\bar{p}$

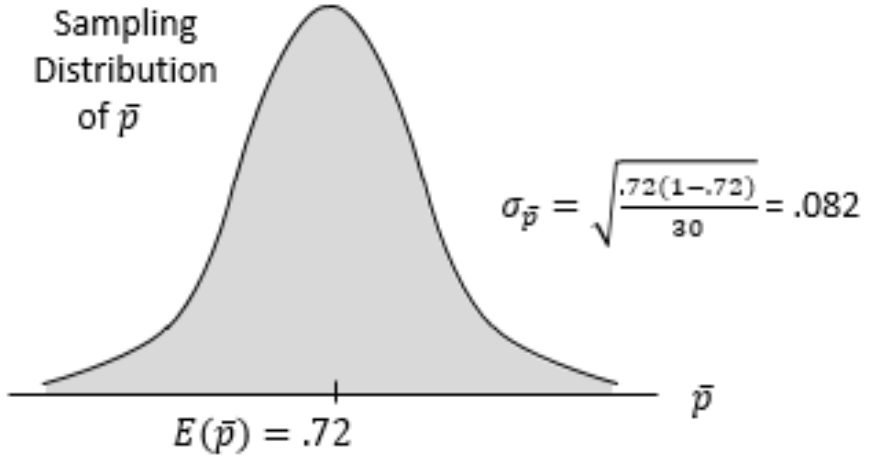
Example: St. Andrew's College

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?

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

$$np = 30(0.72) = 21.6 \geq 5 \text{ and } n(1-p) = 30(0.28) = 8.4 \geq 5.$$



Sampling Distribution of $\bar{p}$

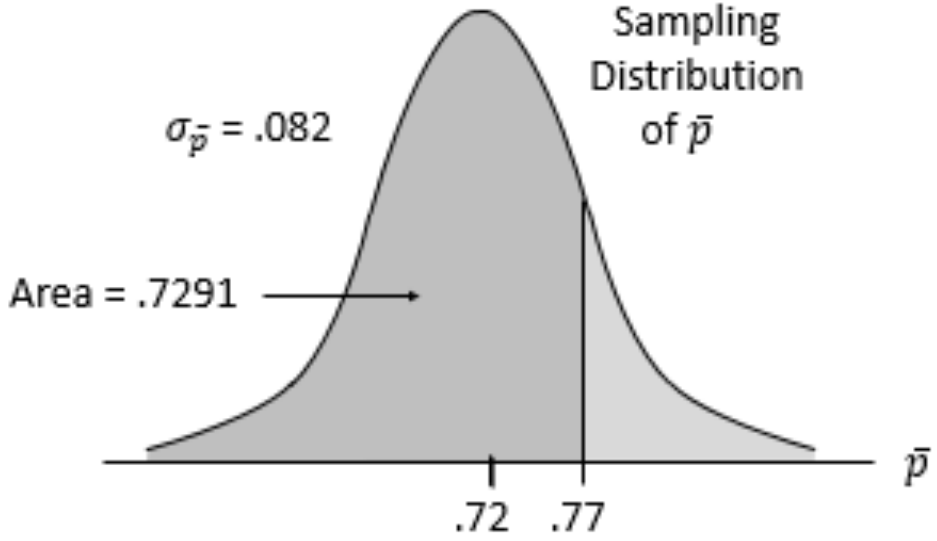
Example: St. Andrew's College

Step 1: Calculate the z-value at the upper endpoint of the interval. $z = \dfrac{(0.77 - 0.72)}{0.082} = 0.61$

Step 2: Find the area under the curve to the left of the upper endpoint. $P(z \le 0.61) = 0.7291$

Cumulative Probabilities for the Standard Normal Distribution

z	.00	.01	.02	.03	.04
.	.	.	.	.	.
.5	.6915	.6950	.6985	.7019	.7054
.6	.7257	.7291	.7324	.7387	.7389
.7	.7580	.7611	.7642	.7673	.7704
.8	.7881	.7910	.7939	.7967	.7995
.9	.8159	.8186	.8212	.8238	.8264
.	.	.	.	.	.



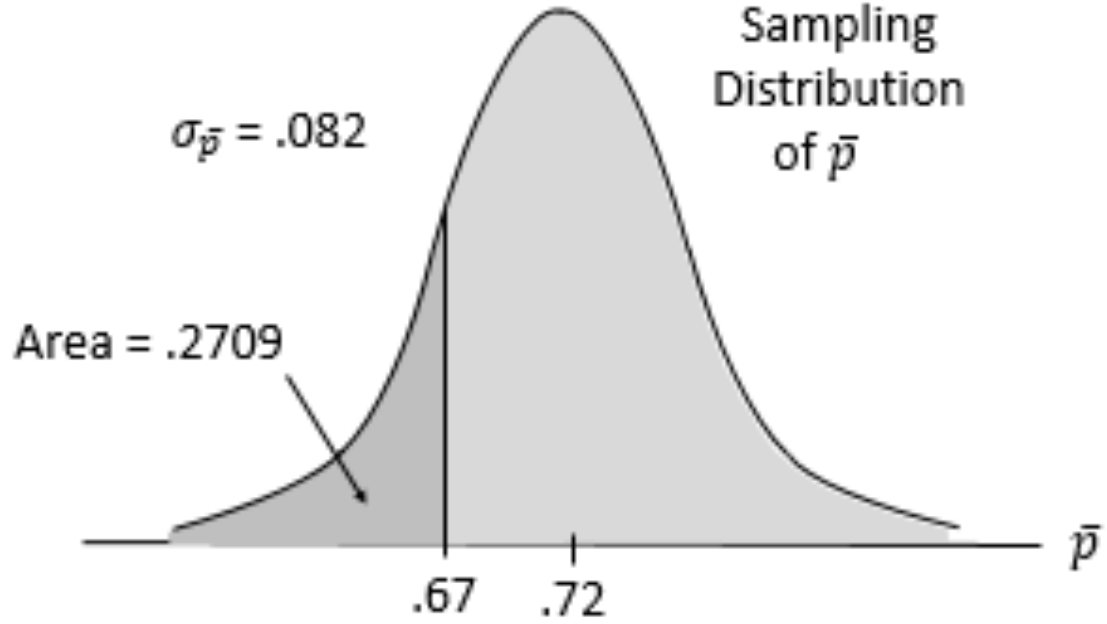
Sampling Distribution of $\bar{p}$

Example: St. Andrew's College

Step 3: Calculate the z-value at the lower endpoint of the interval. $z = \frac{(0.67-0.72)}{0.082} = -0.61$

Step 4: Find the area under the curve to the left of the lower endpoint.

$P(z \leq -0.61) = 0.2709$

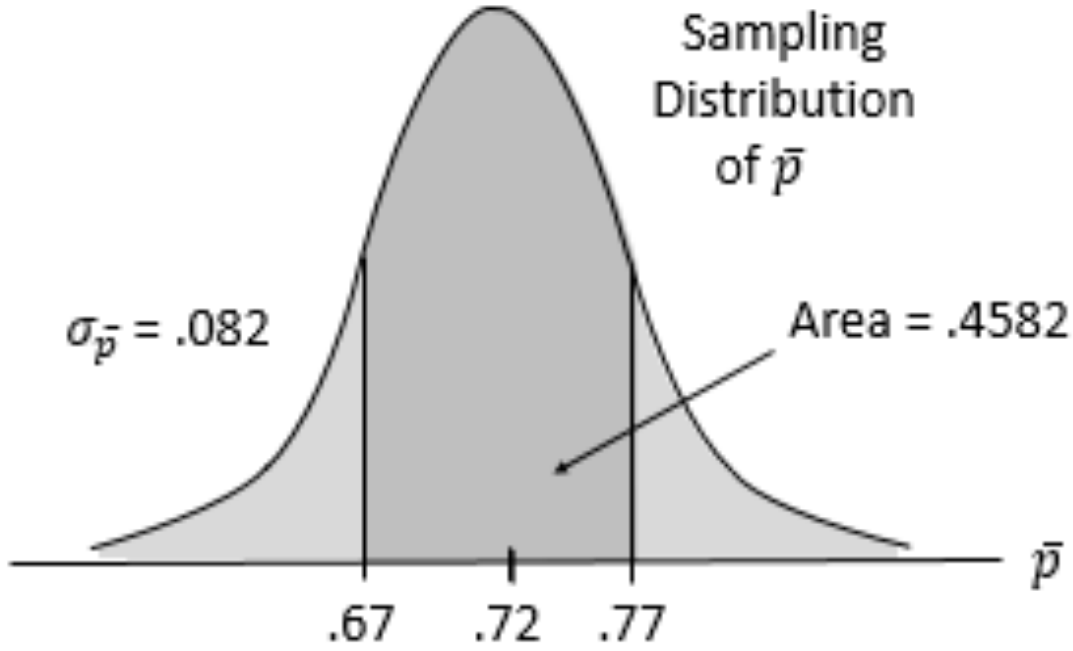


Sampling Distribution of $\bar{p}$

The probability that the estimate of the population proportion of applicants desiring on-campus housing is within ± 0.05 of the actual population proportion:

Example: St. Andrew's College

Step 5: Calculate the area under the curve between the lower and upper endpoints of the interval.



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).
- 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 provides results that 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.
- 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.
- 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.
- 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.
- 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.

Margin of error and the Interval Estimate

- A point estimator cannot be expected to provide the exact value of the population parameter.
- An interval estimate can be computed by adding and subtracting a margin of error to the point estimate.

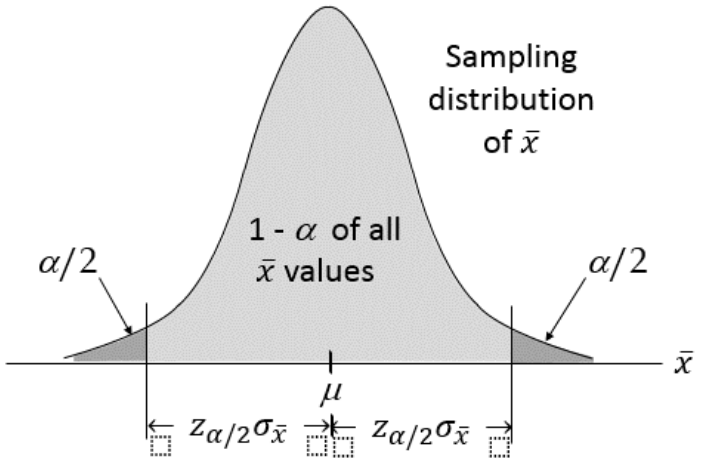
  Point estimate $\pm$ Margin of error

- The purpose of an interval estimate is to provide information about how close the point estimate is to the value of the parameter.
- The general form of an interval estimate of a population mean is $\bar{x} \pm$ Margin of error
- In order to develop an interval estimate of a population mean, the margin of error must be computed using either:
  - the population standard deviation σ, or
  - the sample standard deviation s
- σ is rarely known exactly. But often a good estimate can be obtained based on historical data or other information.
- We refer to such cases as the σ known case.

Interval Estimate of a Population Mean: σ Known

There is a $1 - \alpha$ probability that the value of a sample mean will provide a margin of error of $z_{\alpha/2}\sigma_{\bar{x}}$ or less.

Interval Estimate of μ: $\bar{x} \pm z_{\alpha/2}\dfrac{\sigma}{\sqrt{n}}$



Where:

$\bar{x}$ = the sample mean

$1 - \alpha$ = the confidence coefficient.

$z_{\alpha/2}$= the z-value providing an area of $\alpha/2$ in the upper tail of the standard normal distribution.

σ = the population standard deviation

n = the sample size

Values of $z_{\alpha/2}$ for the most commonly used confidence levels.

Confidence Level	α	$\alpha/2$	Table Look-up Area	$z_{\alpha/2}$
90%	.10	.05	.9500	1.645
95%	.05	.025	.9750	1.960
99%	.01	.005	.9950	2.576

Meaning of C% Confidence

- Because 90% of all the intervals constructed using $\bar{x} \pm 1.645\sigma_{\bar{x}}$ will contain the population mean, we say that we are 90% confident that the interval $\bar{x} \pm 1.645\sigma_{\bar{x}}$ includes the population mean, μ.
- We say that this interval has been established at the 90% confidence level.
- The value 0.90 is referred to as the confidence coefficient.

Interval Estimate of a Population Mean: σ Known

Example: Discount Sounds

Discount Sounds has 260 retail outlets throughout the United States. The firm is evaluating a potential location for a new outlet, based in part, on the mean annual income of the individuals in the marketing area of the new location.

A sample of size $n = 36$ was taken; the sample mean income is \$41,100. The population is not believed to be highly skewed. The population standard deviation is estimated to be \$4,500, and the confidence coefficient to be used in the interval estimate is 0.95.

95% of the sample means that can be observed are within $\pm 1.96\sigma_{\bar{x}}$ of the population mean, μ. Therefore the margin of error is

$$z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) = 1.96\left(\frac{4{,}500}{\sqrt{36}}\right) = 1{,}470$$

Thus at 95% confidence, the margin of error is \$1,470.

Interval Estimate of a Population Mean: σ Known

Example: Discount Sounds

The interval estimate of μ is

$$\$41{,}100 \pm \$1{,}470$$

or

$$\$39{,}630 \text{ to } \$42{,}570$$

We are 95% confident that the interval contains the population mean.

Confidence level	Margin of error	Interval estimate
90%	1,234	39,866 to 42,334
95%	1,470	39,630 to 42,570
99%	1,932	39,168 to 43,032

In order to have a higher degree of confidence, the margin of error and thus the width of the confidence interval must be larger.

Interval Estimate of a Population Mean: σ Known

Adequate Sample Size

- In most applications, a sample size of $n \geq 30$ is adequate.
- If the population distribution is highly skewed or contains outliers, a sample size of 50 or more is recommended.
- If the population is not normally distributed but is roughly symmetric, a sample size as small as 15 will suffice.
- If the population is believed to be at least approximately normal, a sample size of less than 15 can be used.

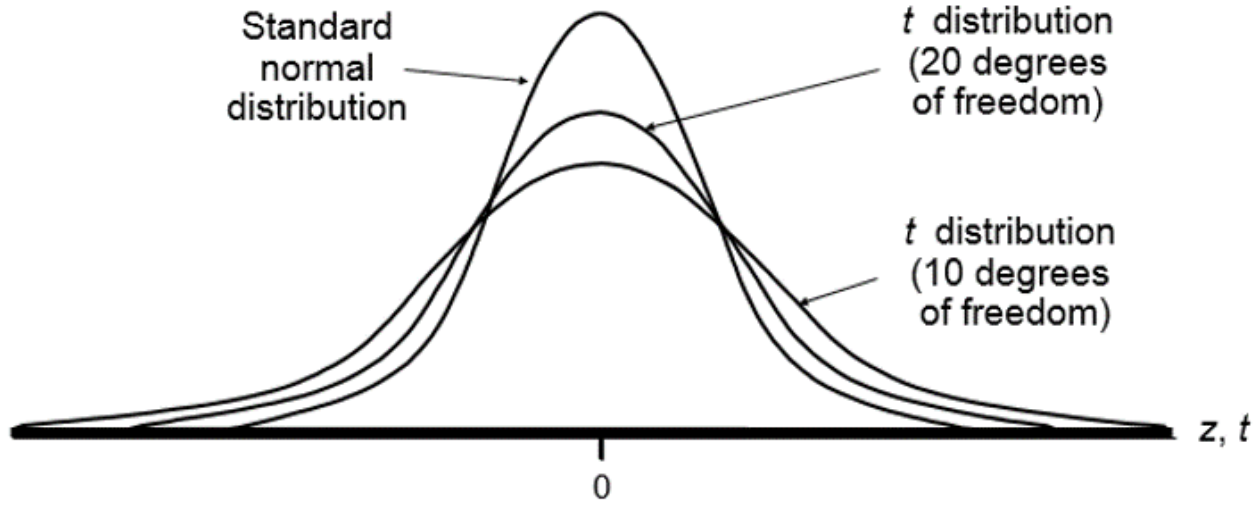
Interval Estimate of a Population Mean: σ Unknown

- If an estimate of the population standard deviation, σ, cannot be developed prior to sampling, we use the sample standard deviation, s, to estimate σ.
- This is the σ unknown case.
- In this case, the interval estimate for μ is based on the t distribution.
- We'll assume for now that the population is normally distributed.

t Distribution

- William Gosset, writing under the name "Student", is the founder of the *t* distribution.
- Gosset was an Oxford graduate in mathematics and worked for the Guinness Brewery in Dublin.
- He developed the *t* distribution while working on small-scale materials and temperature experiments.
- The *t* distribution is a family of similar probability distributions.
- A specific *t* distribution depends on a parameter known as the degrees of freedom.
- Degrees of freedom refer to the number of independent pieces of information that go into the computation of *s*.
- A *t* distribution with more degrees of freedom has less dispersion.
- As the degrees of freedom increases, the difference between the *t* distribution and the standard normal probability distribution becomes smaller and smaller.
- For more than 100 degrees of freedom, the standard normal *z* value provides a good approximation to the *t* value.
- The standard normal *z* values can be found in the infinite degrees (∞) row of the *t* distribution table.

t Distribution



Degrees of Freedom	Area in Upper Tail					
	.20	.10	.05	.025	.01	.005
.	.	.	.	.	.	.
50	.849	1.299	1.676	2.009	2.403	2.678
60	.848	1.296	1.671	2.000	2.390	2.660
80	.846	1.292	1.664	1.990	2.374	2.639
100	.845	1.290	1.660	1.984	2.364	2.626
∞	.842	1.282	1.645	1.960	2.326	2.576

(bottom row is standard normal z values)

Interval Estimate of a Population Mean: σ Unknown

Interval Estimate of μ:

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Where:

$\bar{x}$ = the sample mean

$1 - \alpha$ = the confidence coefficient.

$t_{\alpha/2}$ = the t-value providing an area of $\alpha/2$ in the upper tail of the t distribution with $n - 1$ degrees of freedom.

s = the sample standard deviation

n = the sample size

Interval Estimate of a Population Mean: σ Unknown

Example: Apartment Rents

A reporter for a student newspaper is writing an article on the cost of off-campus housing. A sample of 16 one-bedroom apartments within a half-mile of campus resulted in a sample mean of $750 per month and a sample standard deviation of $55.

Let us provide a 95% confidence interval estimate of the mean rent per month for the population of one-bedroom apartments within a half-mile of campus. We will assume this population to be normally distributed.

Degrees	Area in Upper Tail					
of Freedom	.20	.10	.05	.025	.01	.005
15	.866	1.341	1.753	2.131	2.602	2.947
16	.865	1.337	1.746	2.120	2.583	2.921
17	.863	1.333	1.740	2.110	2.567	2.898
18	.862	1.330	1.734	2.101	2.520	2.878
19	.861	1.328	1.729	2.093	2.539	2.861
.	.	.	.	.	.	.

Interval Estimate of a Population Mean: σ Unknown

Interval Estimate $$\bar{x} \pm t_{.025}\frac{s}{\sqrt{n}}$$

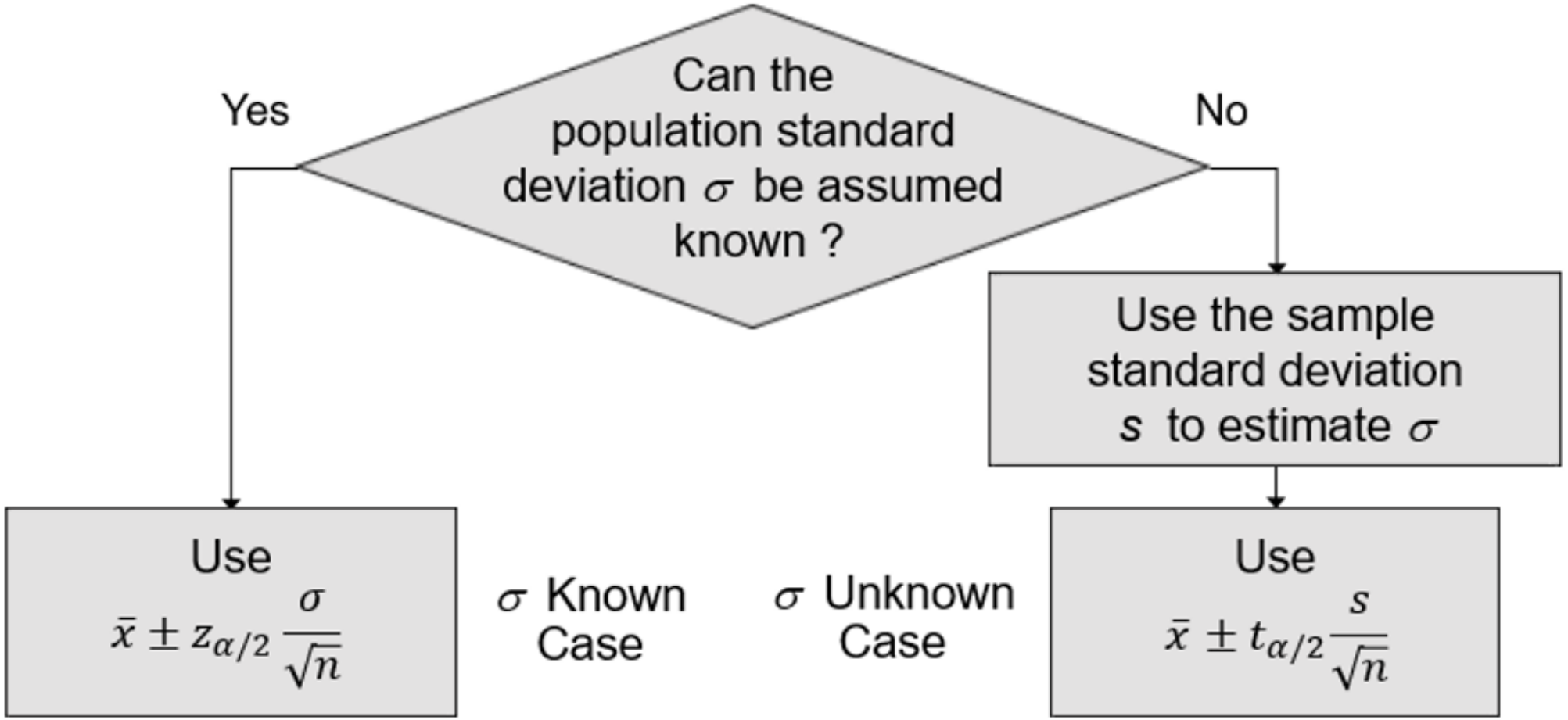
$$750 \pm 2.131\frac{55}{\sqrt{16}} = 750 \pm 29.30$$

We are 95% confident that the mean rent per month for the population of one-bedroom apartments within a half-mile of campus is between $720.70 and $779.30.

Adequate Sample Size:

- Usually, a sample size of at least 30 is adequate when using a t interval to estimate a population mean.
- If the population distribution is highly skewed or contains outliers, a sample size of 50 or more is recommended.
- If the population is not normally distributed but is roughly symmetric, a sample size as small as 15 will suffice.
- If the population is believed to be at least approximately normal, a sample size of less than 15 can be used.

Summary of Interval Estimation Procedures for a Population Mean



Sample Size for an Interval Estimate of a Population Mean

- Let E = the desired margin of error.
- E is the amount added to and subtracted from the point estimate to obtain an interval estimate.
- If a desired margin of error is selected prior to sampling, the sample size necessary to satisfy the margin of error can be determined.
- Margin of error $E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
- Necessary sample size $n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2}$

The Necessary Sample Size equation requires a value for the population standard deviation, σ.
If σ is unknown, a preliminary or planning value for σ can be used in the equation.

1. Use the estimate of the population standard deviation computed in a previous study.
2. Use a pilot study to select a preliminary sample and use the sample standard deviation from the study.
3. Use judgment or a "best guess" for the value of s.

Sample Size for an Interval Estimate of a Population Mean

Example: Discount Sounds

Recall that Discount Sounds is evaluating a potential location for a new retail outlet, based in part, on the mean annual income of the individuals in the marketing area of the new location.

Suppose that Discount Sounds' management team wants an estimate of the population mean such that there is a 0.95 probability that the sampling error is $500 or less.

How large a sample size is needed to meet the required precision?

$$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$E = 500, \sigma = 4{,}500$, at 95% confidence $z_{0.025} = 1.96$

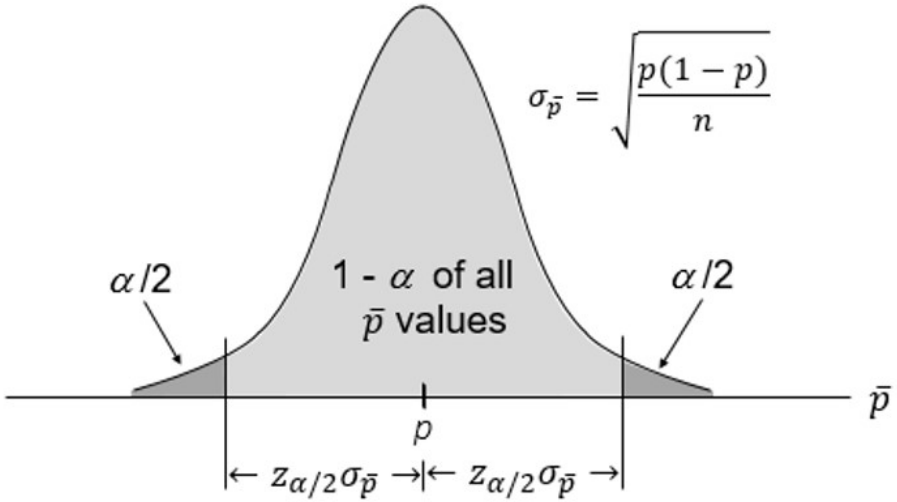
$$n = \frac{(z_{\alpha/2})^2\sigma^2}{E^2} = \frac{(1.96)^2(4500)^2}{500^2} = 311.17 \approx 312$$

A sample size of 312 is needed to reach the desired precision of ± 500 at 95% confidence.

Interval Estimate of a Population Proportion

- The general form of an interval estimate of a population proportion is: $\bar{p} \pm$ Margin of error
- The sampling distribution of $\bar{p}$ plays a key role in computing the margin of error for this interval estimate.
- The sampling distribution of $\bar{p}$ can be approximated by a normal distribution whenever $np \geq 5$ and $n(1-p) \geq 5$.

Normal Approximation of the Sampling Distribution of $\bar{p}$.



Interval Estimate of a Population Proportion

Interval Estimate of p:

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Where:

$\bar{p}$ = the sample proportion

$1-\alpha$ = the confidence coefficient

$z_{\alpha/2}$= the z-value providing an area of $\alpha/2$ in the upper tail of the standard normal distribution.

n = the sample size

Interval Estimate of a Population Proportion

Example: Political Science, Inc.

Political Science Inc. (PSI) specializes in voter polls and surveys designed to keep political office seekers informed of their position in a race. Using telephone surveys, PSI interviewers ask registered voters who they would vote for if the election were held that day. In a current election campaign, PSI has just found that 220 registered voters, out of 500 contacted, favor a particular candidate. PSI wants to develop a 95% confidence interval estimate for the proportion of the population of registered voters that favor the candidate.

$$\bar{p} \pm z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$

Where n = 500, $\bar{p}$ = 220/500 = 0.44, $z_{\alpha/2}$ =1.96

$$0.44 \pm 1.96\sqrt{\frac{0.44\,(1-0.44)}{500}}$$

$$0.44 \pm 0.0435$$

PSI is 95% confident that the proportion of all voters that favor the candidate is between 0.3965 and 0.4835.

Sample Size for an Interval Estimate of a Population Proportion

- Margin of error $E = z_{\alpha/2}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$
- Solving for n, the necessary sample size is $n = \frac{(z_{\alpha/2})^2 \bar{p}(1-\bar{p})}{E^2}$

- However, $\bar{p}$ will not be known until after we have selected the sample. Therefore, we will use the planning value p^* for $\bar{p}$.

Necessary Sample Size $n = \frac{(z_{\alpha/2})^2 p^*(1-p^*)}{E^2}$

The planning value p^* can be chosen by:

1. Using the sample proportion from a previous sample of the same or similar size.
2. Selecting a preliminary sample and using the sample proportion from that sample.
3. Using judgment or a "best guess" for the p^* value.
4. Otherwise, use $p^* = 0.5$.

Sample Size for an Interval Estimate of a Population Proportion

Example: Political Science, Inc.

Suppose that PSI would like a 0.99 probability that the sample proportion is within $\pm$ 0.03 of the population proportion. How large a sample size is needed to meet the required precision? (A previous sample of similar units yielded 0.44 for the sample proportion.)

E =0.03, p^* =0.44, and at 99% confidence, $z_{0.005}$ =2.576.

$$n = \frac{(z_{\alpha/2})^2 p^*(1-p^*)}{E^2}$$

$$n = \frac{(2.576)^2 (0.44)(0.56)}{(0.03)^2} = 1817$$

A sample size of 1817 is needed to reach the desired precision of ±0.03 at 99% confidence.

Note: We used 0.44 as the best estimate of p. If no information is available about p, then 0.5 is often used because it provides the greatest possible sample size. If we had used $p^* = 0.5$, the recommended n would have been 1843.