USC
School of Business

Data Driven Decision Making: Sampling Distributions & Estimation

GSBA 545, Fall 2021

Professor Dawn Porter

Sampling Distributions

- Sampling
  - Expected Value & Standard Error of Sample Means
- Application - Probability of Averages
- Confidence Intervals
  - Large Sample, Means
  - Small Sample, Means
  - Population Proportions
- Sample Size Determination
  - Means
  - Proportions

Notation

Estimating Parameters

Parameter: a characteristic of the population

Statistic: an observed characteristic of a sample

Name	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Standard deviation	s	σ
Correlation	r	ρ
Proportion	$\hat{p}$	p
Regression intercept	b_0	β_0
Regression slope	b_i	β_i

Sampling Variation:

- The variability in the value of a statistic from sample to sample.
- The price we pay for working with a sample rather than the population.

Sampling Distributions

Assume two separate samples are taken to estimate the population average income in LA:

$n_1 = 10$, sample average = \$45,000, $n_2 = 1000$, sample average = \$45,000

- Which result would you trust more? Why?
- If the experiment were repeated, which result would vary more? Why?

Sampling Distributions

If a random sample of size *n* is taken from a (normal) population with mean μ and standard deviation σ, then the sampling distribution of the sample mean $\bar{x}$ has

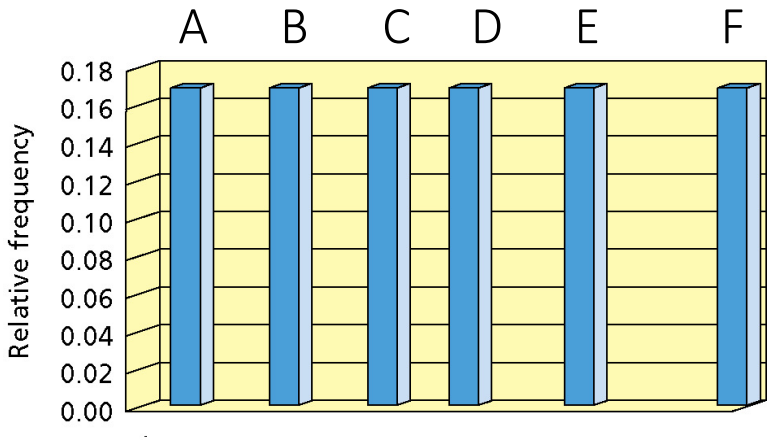
mean: $E(\bar{x}) = \mu_{\bar{x}} = \mu$

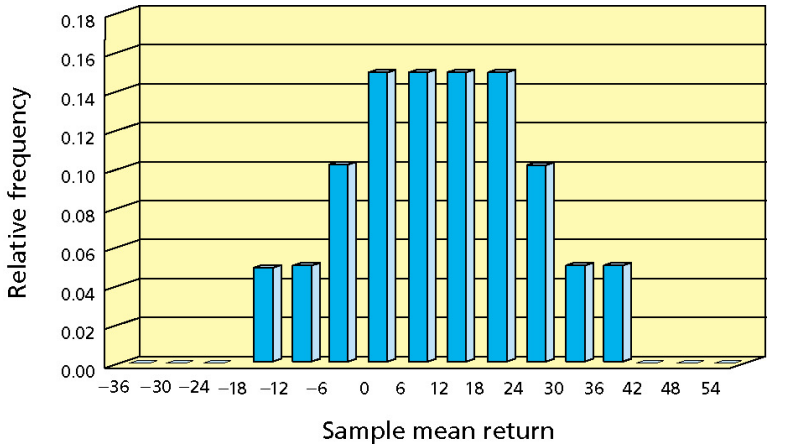
standard error: $SE(\bar{x}) = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

1. The best guess of μ is $\bar{x}$ and the reverse.
2. As *n* increases, estimates become more reliable.
3. Regardless of underlying population distribution, possible sample averages should be normally distributed (if $n > 30$).

Sample Means

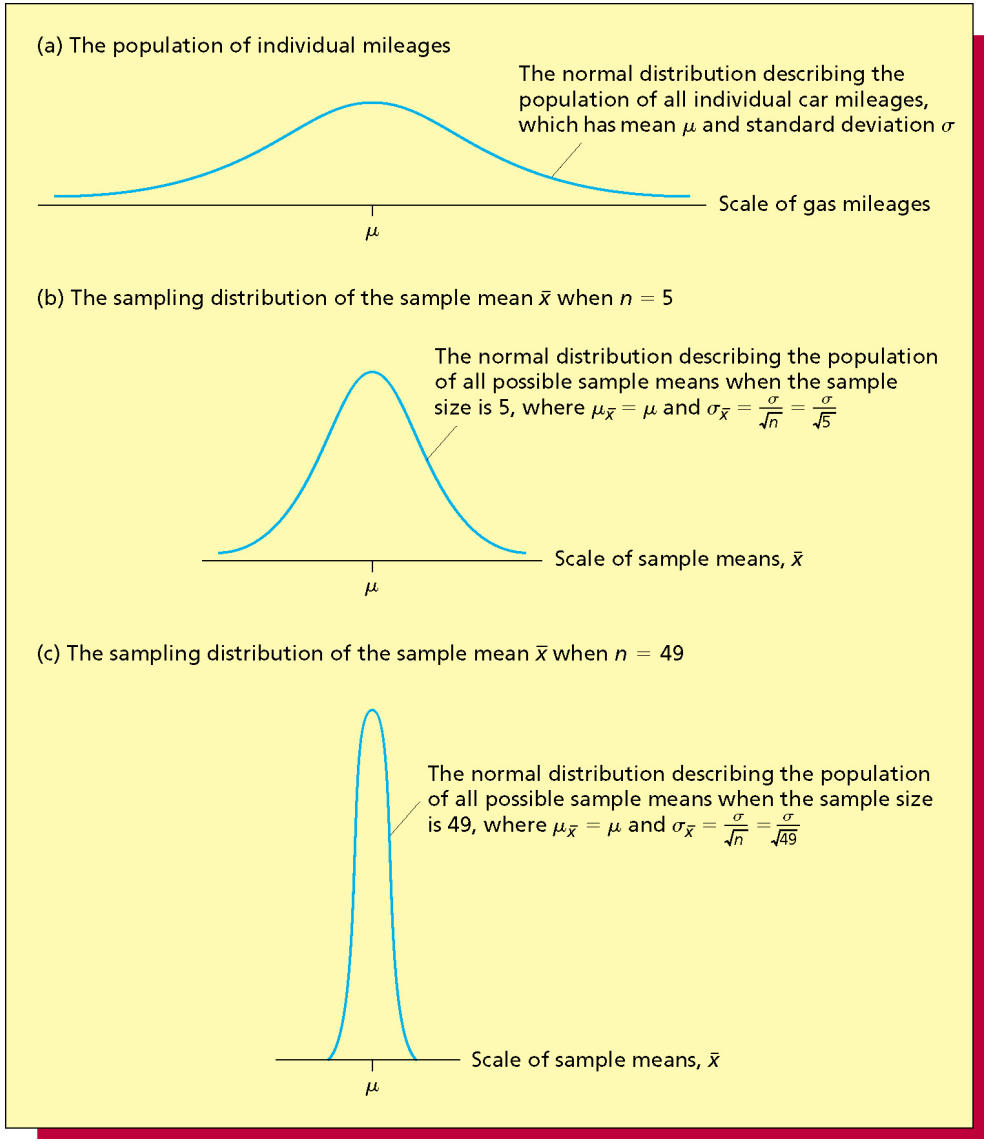
	n = 3 stocks in each sample				
	Sample	Stock Returns			Mean
1	A,B,C	-36	-15	3	-16





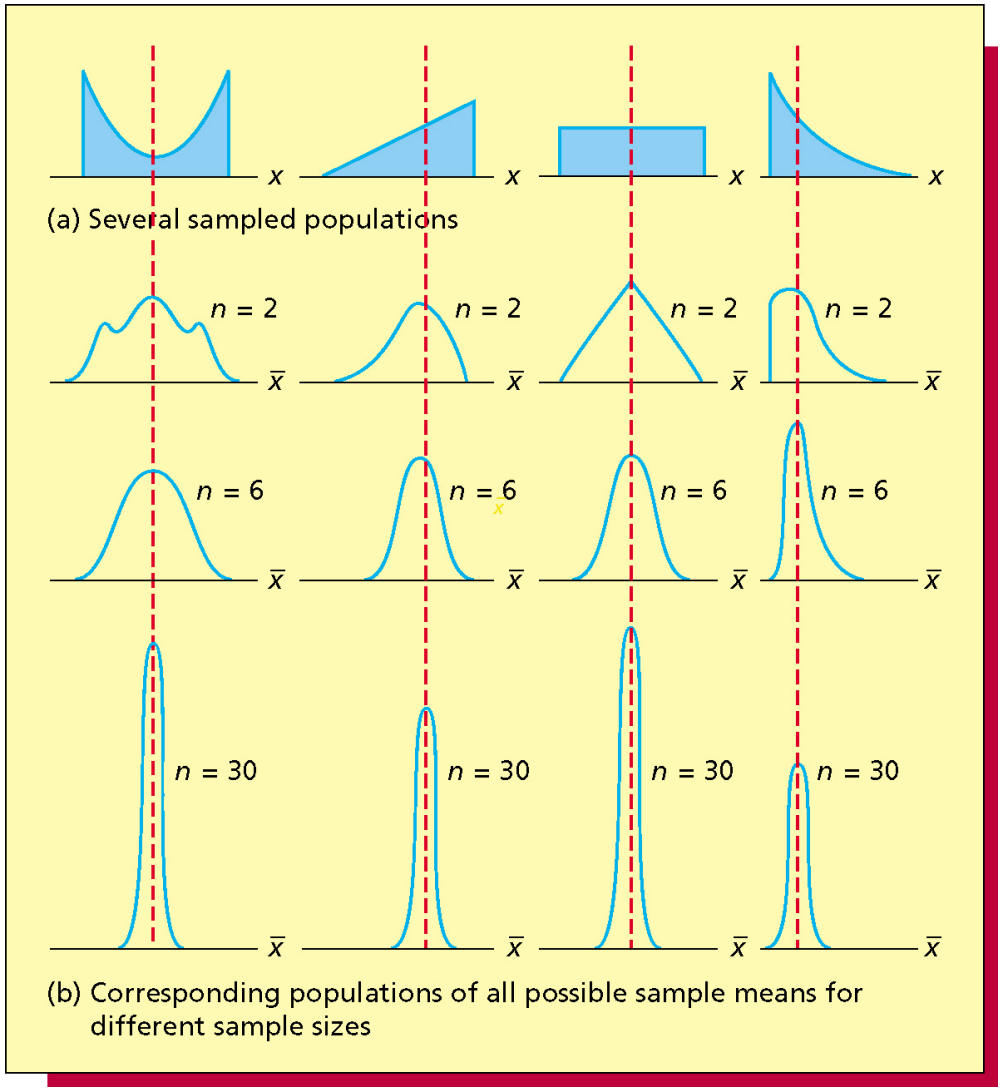
Tetris for statisticians ☺

Sample Size Effect



Sample Size Effect

The larger the sample size, the more nearly normally distributed is the population of all possible sample means, even if the original distribution was far from normal.



Sampling Distributions

Why is this concept important and where is it used?

- If population parameters are known, we can calculate the probability of certain sample averages occurring.
- In Statistical Inference, we use sample results to test hypotheses.
- In Regression Analysis, sample results (including standard errors!) are needed to draw intelligent conclusions.

Probability of Averages

Assume the prices of meals at a restaurant are normally distributed with an average of $35 and a standard deviation of $5. I have $33 in my pocket.

1. What is the probability I have enough money for my meal?

2. I'm out to dinner with friends and we're splitting the check evenly by 10.
   a) What is the expected price per meal now?
   b) What is the probability I have enough money now?
   c) Why did the probability change?

Probability of Averages

Assume the prices of meals at a restaurant are normally distributed with an average of $35 and a standard deviation of $5. I have $33 in my pocket.

1. What is the probability I have enough money for my meal?

   Soln: $P(X \leq 33) = P\left(Z \leq \frac{33-35}{5}\right) = P(Z \leq -0.4) = 0.3446$

2. I'm out to dinner with friends and we're splitting the check evenly by 10.

   a) What is the expected price per meal now?

   Soln: $E(\bar{X}) = \mu = \$35$

   b) What is the probability I have enough money now?

   Soln: $P(\bar{X} \leq 33) = P\left(Z \leq \frac{33-35}{5/\sqrt{10}}\right) = P(Z \leq -1.26) = 0.1038$

   c) Why did the probability change?

   Soln: Since we're *averaging* 10 values, the std error is less and values converge toward $35.

Margin of Error

Sample estimates are generally point estimates, but these are approximations.

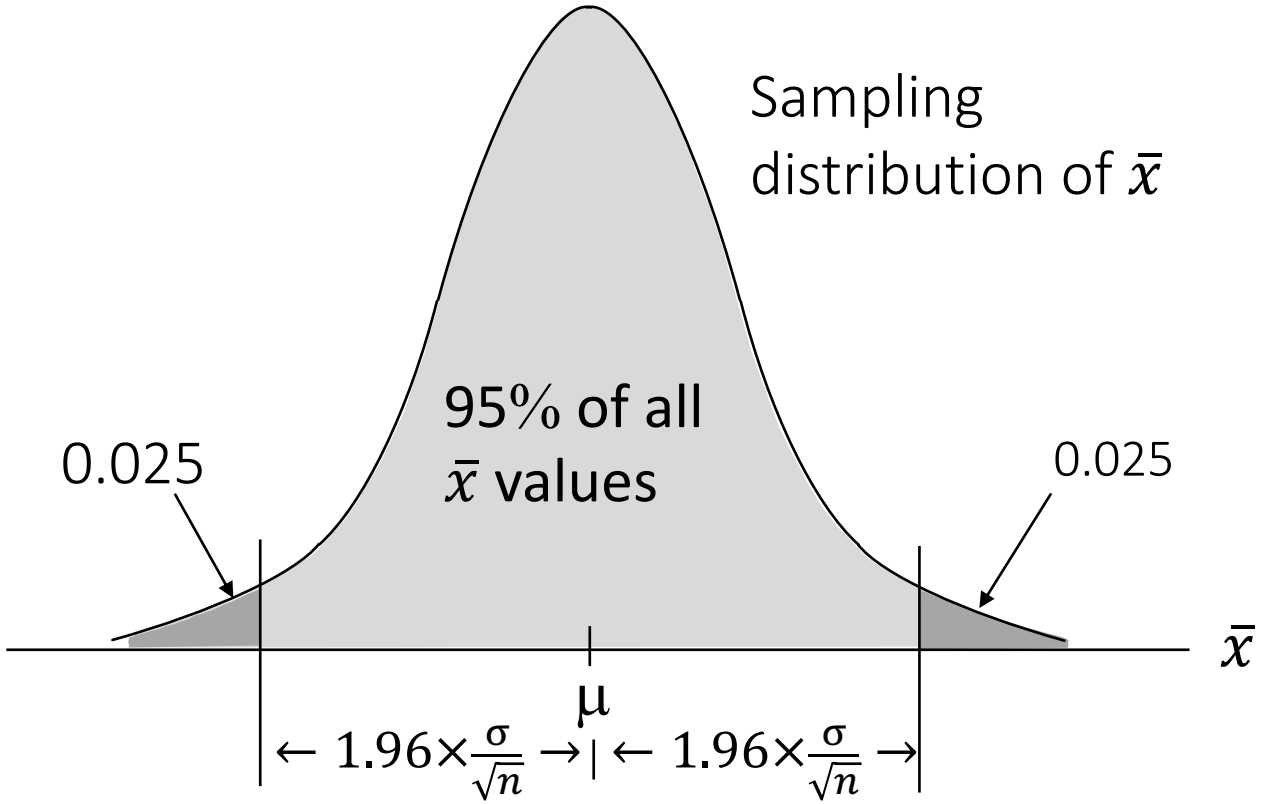
Therefore, we need to have an understanding of the data to create a **margin of error** and create an interval around a point estimate.

The margin of error is 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; in that case, we can assume *s* is known

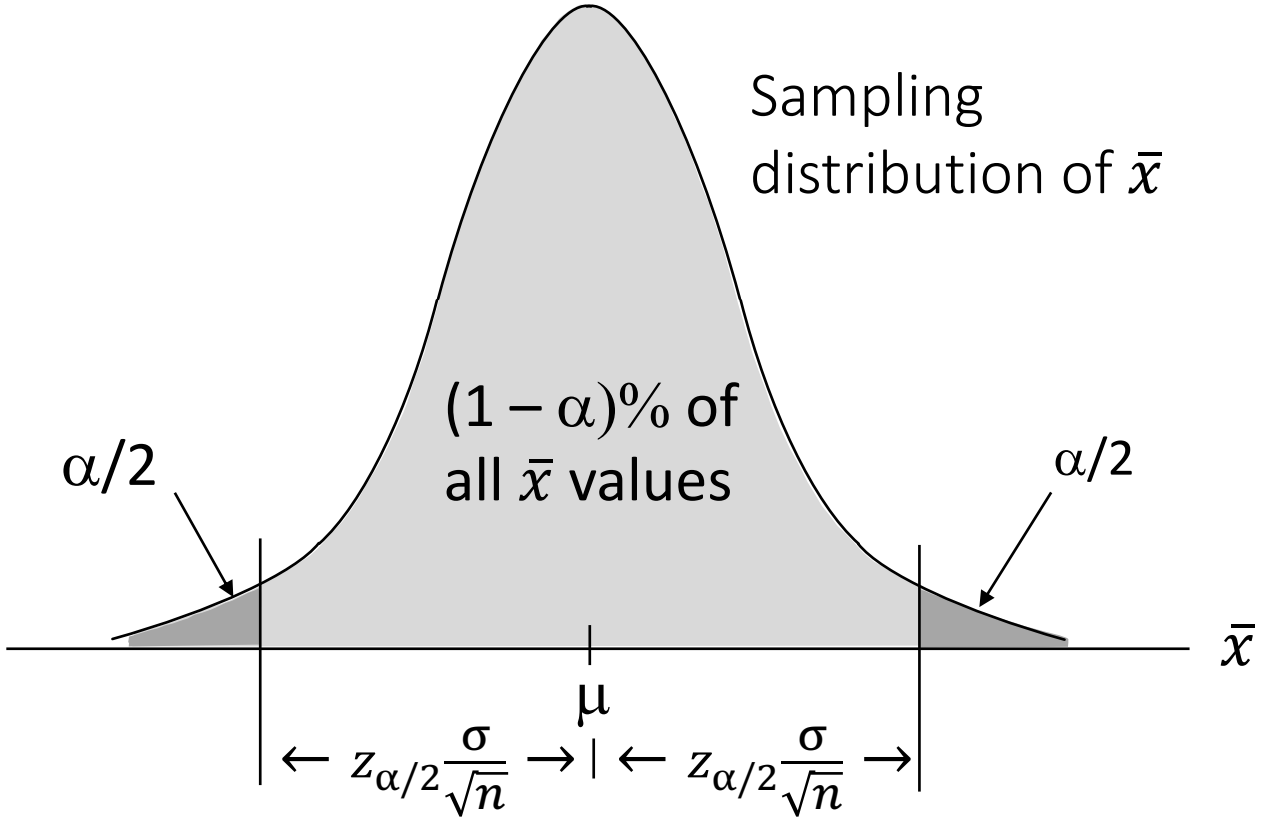
Margin of Error

There is a 0.95 probability that the value of a sample mean will provide a margin of error of $1.96\times\frac{\sigma}{\sqrt{n}}$ or less.



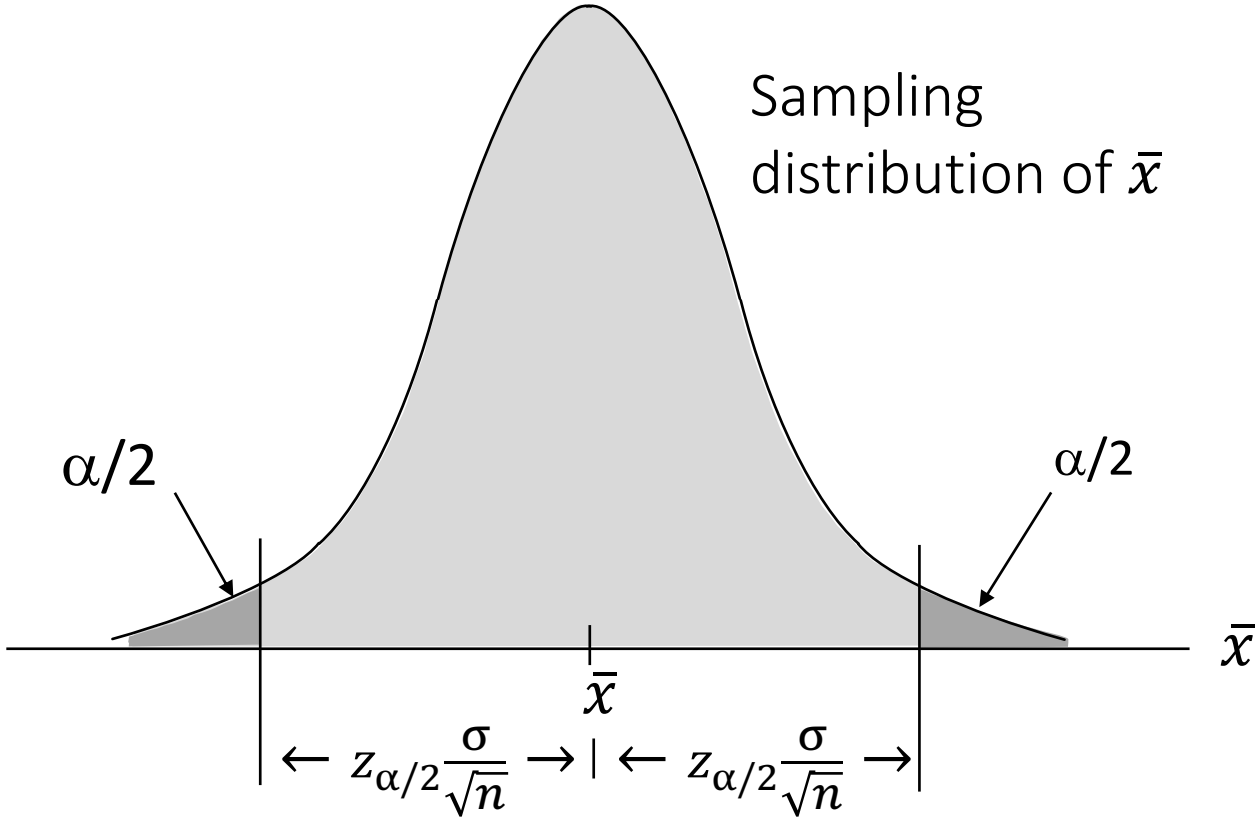
Margin of Error

There is a $(1 - \alpha)$ probability that the value of a sample mean will provide a margin of error of $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ or less.



Margin of Error

There is a $(1 - \alpha)$ probability μ will be within a margin of error of $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ or less from a sample mean.



Large Sample CIs: Mean

If σ is known, or if $n \geq 30$, we can be $100(1-\alpha)\%$ confident that μ falls between:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

where:

$\bar{x}$	is the sample mean
$(1-\alpha)$	is the confidence coefficient
$z_{\alpha/2}$	is z value for an area of $\alpha/2$ in the upper tail
σ (s)	is the population (sample) standard deviation
n	is the sample size

if $n \geq 30$, σ (or) s are similar

Large Sample CI: Retail Store

Example: Discount Sounds

Discount Sounds has several retail outlets throughout the United States and 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 individuals was taken; the sample mean income is $41,100. The population is not believed to be highly skewed. The sample standard deviation is calculated to be $4,500, and the confidence level to be used is 95%.

Create a 95% Confidence Interval for the *population mean income.*

Large Sample CI: Retail Store

Example: Discount Sounds

95% Confidence Interval

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} \rightarrow 41{,}100 \pm 1.96\frac{4500}{\sqrt{36}}$$

4500 → std dev

$\frac{4500}{\sqrt{36}}$ → std error

$$\rightarrow 41{,}000 \pm 1470$$

$$\rightarrow [\$39{,}630,\ \$42{,}570]$$

"We are 95% sure that the true population average income for that location is somewhere between $39,639 and $42,570."

Large Sample CI: Retail Store

Example: Discount Sounds

Various confidence level intervals

Confidence level	Margin of Error	Interval estimate
90%	1234	[39,866, 42,334]
95%	1470	[39,630, 42,570]
99%	1932	[39,168, 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.

Small Sample CIs: Mean

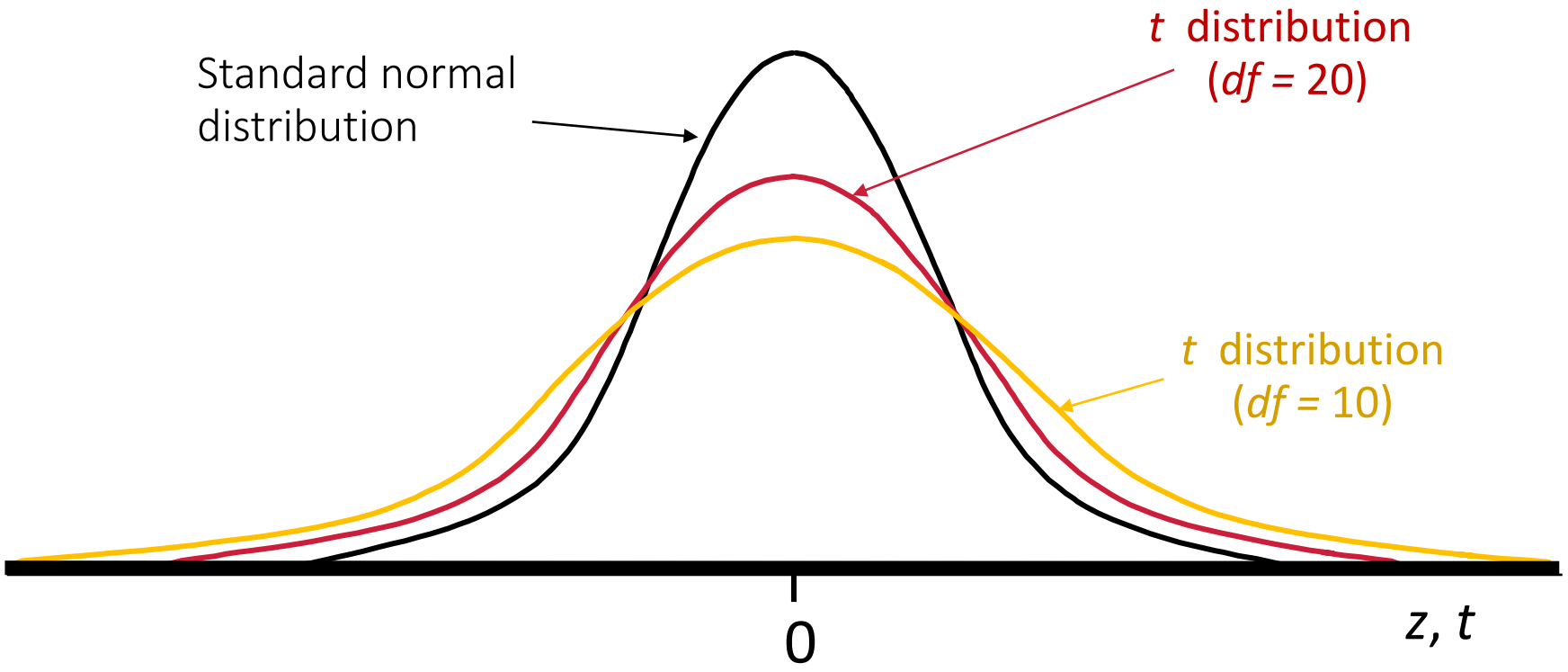
- If $n < 30$, an adjustment needs to be made to account for a higher level of uncertainty.
- Instead of using the **Normal Distribution** and *Z* values, we rely on the ***t*-Distribution**.
- Using the ***t*-Distribution** is more conservative because its shape is more spread out than the Normal; as the sample size increases, though, the two curves will converge.
- To use this distribution, we need the degrees of freedom, or *df.* In these interval calculations, $df = n - 1$. → changes according to method you're using.

if $\bar{x}$ is given, & $n=5$ one no. remains fixed & other 4 can vary.

$\therefore df = n-1 = 4$

if $df = 4$ & 95% CI required $\Rightarrow t_{0.025,4} = 2.776$.

Effect of df on t-distribution



When *df* > 100, the *z* values are good approximations to the *t* value.

Standard normal Z values can be found on the infinite degrees (∞) row of the *t*-distribution table.

Small Sample CIs: Mean

If $n < 30$, an adjustment needs to be made to account for a higher level of uncertainty. The new interval for μ, at a $100(1-\alpha)\%$ confidence level, is:

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

where:

$\bar{x}$	is the sample mean
$(1-\alpha)$	is the confidence coefficient
$t_{\alpha/2}$	is t value for an area of $\alpha/2$ in the upper tail
s	is the sample standard deviation
n	is the sample size

Small Sample CIs: Apartments

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.

Create 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.

$df = n-1 = 15, \alpha/2 = 0.025 \quad t = 2.131$

$\Rightarrow \quad CI \Rightarrow 750 \pm 2.131 \frac{55}{\sqrt{16}}$

Small Sample CIs: Apartments

Example: Apartment Rents

$df = n - 1 = 15$ and $\alpha/2 = 0.025$

Degrees of Freedom	Area in Upper Tail					
	.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
.	.	.	.	.	.	.

Small Sample CIs: Apartments

Example: Apartment Rents

95% Confidence Interval

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} \rightarrow 750 \pm 2.131 \frac{55}{\sqrt{16}}$$

$$\rightarrow 750 \pm 29.30$$

$$\rightarrow [\$720.70,\ \$779.30]$$

"We are 95% sure that the true population average rental price of one-bedroom apartments within a half-mile of campus is somewhere between $720.70 and $779.30."

Proportion Intervals

If the goal is to estimate the *population proportion*, ***p***, we can create a $100(1-\alpha)\%$ confident that *p* falls between:

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

for proportion intervals always use z values

where:

$\hat{p}$	is the sample proportion
$(1-\alpha)$	is the confidence coefficient
$z_{\alpha/2}$	is z value with $\alpha/2$ in the upper tail
$\hat{p}(1-\hat{p})$	is an estimate of the std deviation
n	is the sample size

Proportion CI: Clinical Trials

Example: Clinical Trial Results

A new drug called Phe-Myecin is being marketed by a company and to get through the clinical trial requirements, the company must forecast what percentage of the population of drug-takers would experience negative side effects.

A sample of n = 200 patients was taken and 35 people experienced negative side effects. What is a 95% range for what the true population proportion experiencing negative side effects will be?

$\hat{p} = 35/200 = 0.175 \Rightarrow 0.175 \pm (1.96)\sqrt{\frac{0.175 \times 0.825}{200}}$

$\Rightarrow 0.175 \pm 0.053 = [0.122, 0.228]$

Suppose confidence interval of 5%.

$z\frac{\sigma}{\sqrt{n}} < 500 \Rightarrow n = \left(\frac{1.0}{E}\right)^2 \Rightarrow [0.122 \text{ & } 0.228]$

Proportion CI: Clinical Trials

Example: Clinical Trial Results

95% Confidence Interval: $\hat{p} = \frac{35}{200} = 0.175$:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \rightarrow 0.175 \pm 1.96\sqrt{\frac{0.175(1-0.175)}{200}}$$

$$\rightarrow 0.175 \pm 0.053$$

$$\rightarrow [0.122.\ \ 0.228]$$

"We are 95% sure that the true population proportion experiencing negative side effects will be somewhere between 12.2% and 22.8%."

Sample Sizes: Means

To calculate a reasonable sample size to be taken to obtain an estimate of μ, the margin of error needs to be specified.

Margin of Error: $$E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Then the necessary sample size is: $$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$$

Note that a value for σ is necessary here. If σ is unknown, either:

1. Use a reasonable estimate of σ computed in a previous study.
2. Use a pilot study to select a preliminary sample to estimate and use *s* from there.

Sample Sizes: Retail Store

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 to create a 95% confidence interval with a margin of error of $500 or less. How large a sample size is needed for this? (Recall from earlier that s = 4500.)

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

* Always round up when calculating necessary sample sizes.

Sample Sizes: Proportions

To calculate a reasonable sample size to be taken to obtain an estimate of p, the margin of error needs to be specified.

Margin of Error: $$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Then the necessary sample size is: $$n = p(1-p)\left(\frac{z_{\alpha/2}}{E}\right)^2$$

Note that a value for p is necessary here, which isn't known yet. Options are:

1. Use a reasonable estimate of $\hat{p}$ computed in a previous study.
2. Use a pilot study to select a preliminary sample and use that $\hat{p}$.
3. Otherwise, use $p = 0.50$ as a conservative starting point.

Sample Size: Clinical Trials

Example: Clinical Trial Results

Assume the interval result from the clinical trial of Phe-Myecin is too wide to get approval. Now the company needs to report a 95% confidence interval with a margin of error of only ±4% at the most. How large a sample size is needed for this if we do use our prior estimate of $\hat{p}$ = 0.175?

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

$$= 0.175(1-0.175)\left(\frac{1.96}{0.04}\right)^2$$

$$\rightarrow n = 346.64 \uparrow 347^*$$

* Always round up when calculating necessary sample sizes.

Sample Size: Clinical Trials

Example: Clinical Trial Results (limited information)

Now assume the company hasn't done any study yet and is trying to work out a reasonable sample size for the trial. They are told they need to report a 95% confidence interval with a margin of error of only ±4% at the most. How large a sample size is needed for this if we do not use our prior estimate of $\hat{p}$?

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

$$= 0.5(1-0.5)\left(\frac{1.96}{0.04}\right)^2$$

$$\rightarrow n = 600.25 \uparrow 601^*$$

* Always round up when calculating necessary sample sizes.