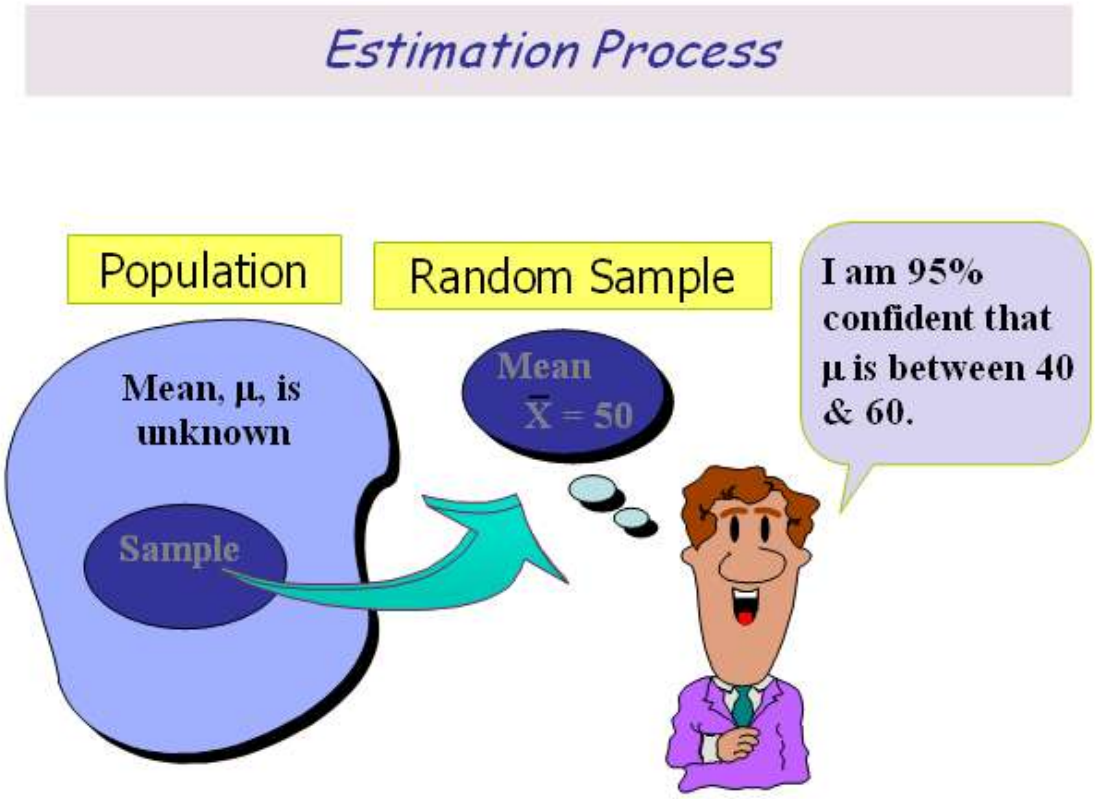


3/28 quiz is moved to April 4th. Will cover chapters 6 and 7.
Last quiz will cover chapters 8 and 9, hypothesis testing & regression.

Confidence Intervals

Confidence intervals are used estimation. Regression model used confidence intervals as well as hypothesis tests. CI and hypothesis tests are very much related. In CI we want to estimate the value of a parameter, μ for instance (μ is the true average bottle contents [population mean]). Example, how much soda put into a bottle. Say we take a sample of 36 bottles and we know the true population standard deviation σ=6 ml (although we would never know this value, addressed below).

From the sample we find that $\bar{X} = 318.7$ml with n=36. $\bar{X}$ is the sample average, it will change from sample to sample. We know that the $\bar{X}$ values are normally distributed because n>30 so Central Limit Theorem applies to the sample means (the point being that we do not know the distribution of the liquid contents of the bottles but we are dealing with the sample means and it's distribution.

This means that $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{36}} = 1$. We know 95% of our sample means will fall within $\pm 2\sigma$, will be within 2σ of $\bar{X}$ in 95% of cases. In this case $\sigma = 1$ ml so a good estimate (95% confidence) for the population mean μ would be 318.7 ± 2 ml or somewhere in the range 316.7 to 320.7 ml. But we are still 5% unsure (in the 95% case). We could go to 99% confidence interval which would be $\pm 3\sigma$ which would make our range 315.7 to 321.7 ml.

So in this way we can make an assertion to within a certain confidence level that the sample mean is within a certain interval. To increase accuracy further we can increase n, the sample size. Increasing n will decrease $\sigma_{\bar{X}}$.

Confidence Interval is always equal to a point estimate $\pm$ a margin of error (MOE).

CONFIDENCE INTERVALS ARE AN ESTIMATION PROCESS

Confidence Intervals give you an Interval of Numbers for a Population Parameter.

We can also increase the accuracy of the confidence interval estimate by increasing the sample size n. This has the effect of reducing the magnitude of the sample variance:

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

Point Estimates

Estimate Population Parameters ...		with Sample Statistics
Mean	μ	$\bar{X}$
Proportion	p	P_S
Variance	σ^2	S^2
Difference	$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$

Trying to find a population parameter and basically we will construct confidence intervals for 2 parameters, μ and p.

When we have numerical data (such as avg height of people, household income, avg weight, avg sqft of house) we will construct a confidence interval for the population mean μ.

When we have categorical data (% people belonging to a certain religion, % people voting republican, % people registered as independents) we will make confidence intervals for the population proportion, p.

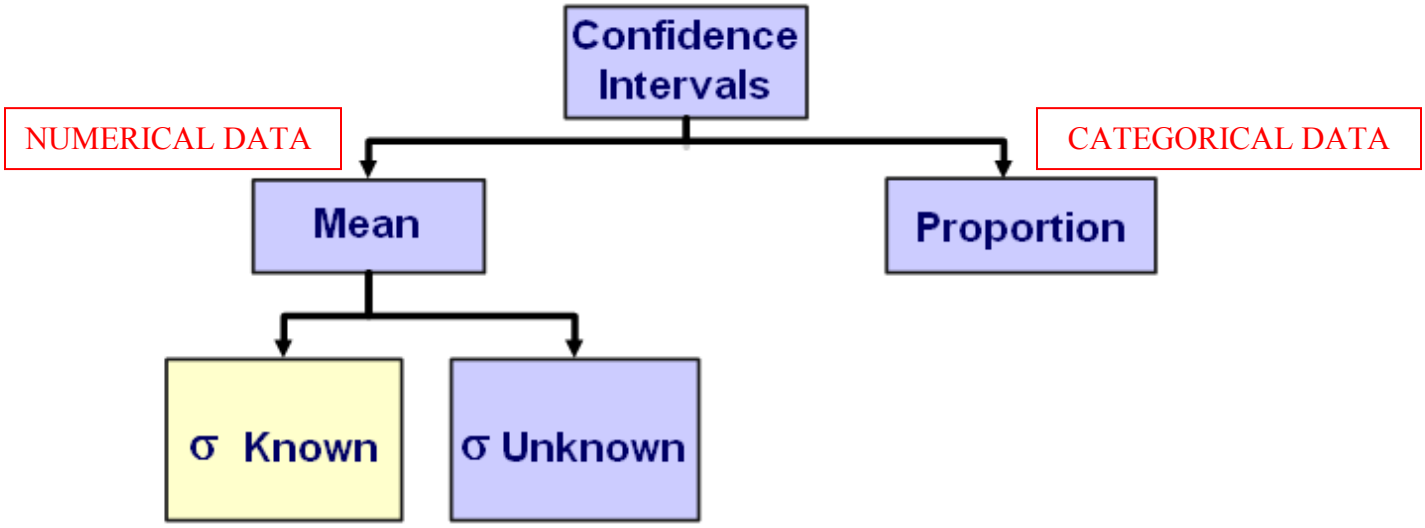
Interval Estimates

- Provides range of values
  - Take into consideration variation in sample statistics from sample to sample
  - Based on observation from 1 sample
  - Give information about closeness to unknown population parameters
  - Stated in terms of level of confidence
    - Never 100% sure

Always the possibility that one particular sample is very unusual.

Takes into account that the $\bar{X}$'s vary. We know the $\bar{X}$'s are normally distributed with $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ and we know 99% of the data is within $\pm 3\sigma$.

Confidence Interval Estimates



The σ known case is not practical. σ is never known, for instructional purposes only.

Confidence Interval for μ (σ Known)

- Assumptions
  - Population standard deviation is known
  - Population is normally distributed
  - If population is not normal, use large sample
- Confidence interval estimate

$$\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

If all assumptions are satisfied we can use this equation.

Confidence Interval is always equal to a point estimate ± a margin of error (MOE).

$$\bar{X} \pm Z\frac{\sigma}{\sqrt{n}}$$

(where Z is the normal distribution critical value for a probability of α/2 in each tail)

EXAM

Important to Know

- When to use Confidence Interval
- Which Confidence Interval to use (σ known, σ unknown, proportions)

Refer to handout **Confidence Intervals. PDF** (end of class 8 notes) and the spreadsheet **CI Formulas.xls.**

CI equation above is based on Z but can be solved for any point.

Elements of Confidence Interval Estimation

- Level of confidence
  - Confidence in which the interval will contain the unknown population parameter
- Precision (range)
  - Closeness to the unknown parameter
- Cost
  - Cost required to obtain a sample of size n

Level of Confidence

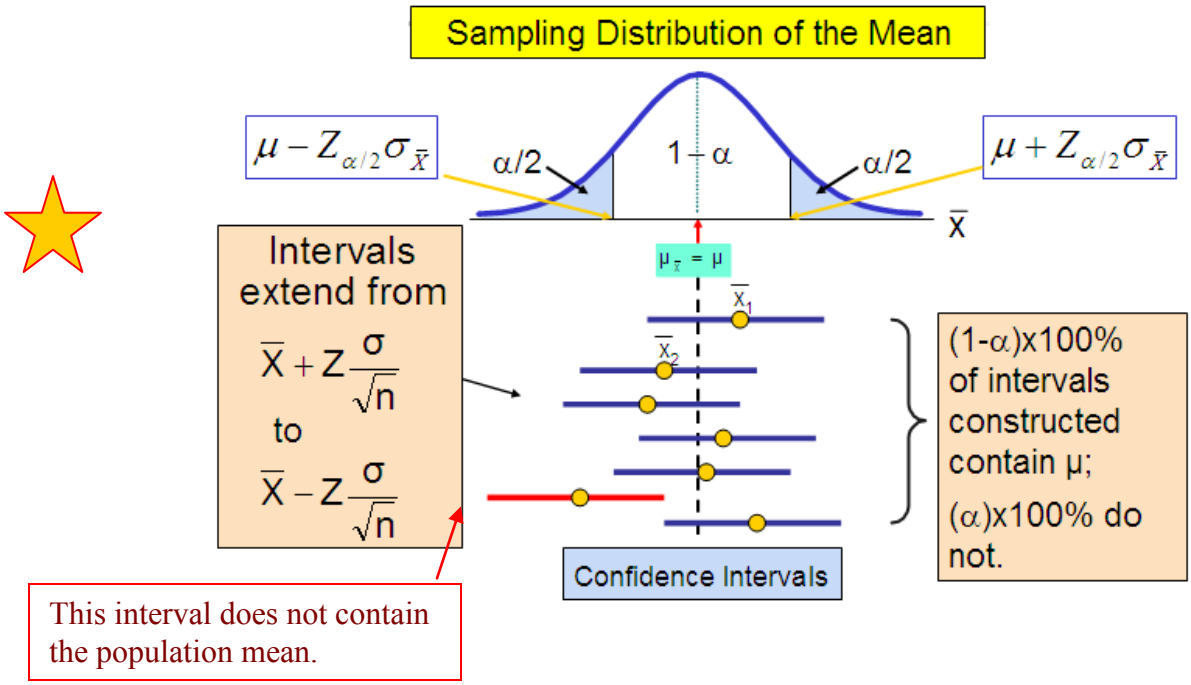
- Denoted by $100(1-\alpha)\%$
- A relative frequency interpretation
  - In the long run, $100(1-\alpha)\%$ of all the confidence intervals that can be constructed will contain the unknown parameter True population parameter such as μ
- A specific interval will either contain or not contain the parameter
  - No probability involved in a specific interval

This is saying that μ is a real parameter, not a random process.

95% CI is most common. α is confidence (even legal). If α=5% then we are using 95% confidence intervals or $\pm 2\sigma$.

(1-α)% is called the Confidence Level.

95% CI → α=1-.95=.05=5% Confidence Level



Dotted line is the true population mean.

Each interval is $\pm Z\frac{\sigma}{\sqrt{n}}$ centered around $\bar{X}$.

- Cc

$\frac{\alpha}{2} = .($

Z units:

X units:

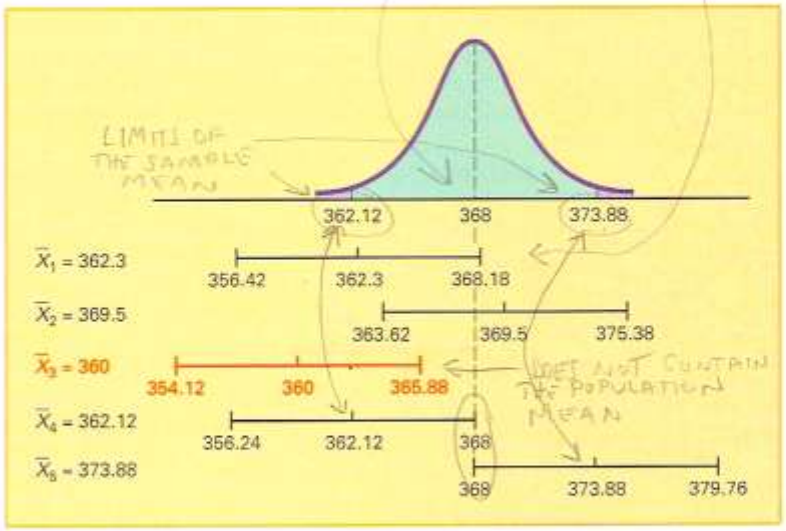
Commonl

Cont L
8
9
9
9
9
9
9

Suppose, for example, that a sample of size $n = 25$ boxes has a mean of 362.3 grams. The interval developed to estimate μ is 362.3 ± (1.96)(15)/($\sqrt{25}$) or 362.3 ± 5.88. The estimate of μ is

$$356.42 \le \mu \le 368.18$$

Because the population mean μ (equal to 368) is included within the interval, this sample has led to a correct statement about μ (see Figure 7.1 below).



To continue this hypothetical example, suppose that for a different sample of $n = 25$ boxes, the mean is 369.5. The interval developed from this sample is 369.5 ± (1.96)(15)/($\sqrt{25}$) or 369.5 ± 5.88. The estimate is

$$363.62 \le \mu \le 375.38$$

Because the true population mean μ (equal to 368) is also included within this interval, this statement about μ is correct.

The value of Z selected for constructing a confidence interval is called the **CRITICAL VALUE**.

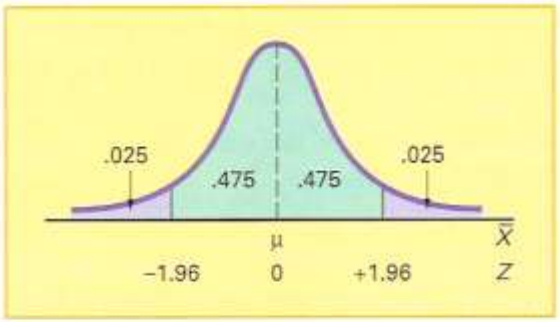


FIGURE 7.2
Normal curve for determining the Z value needed for 95% confidence

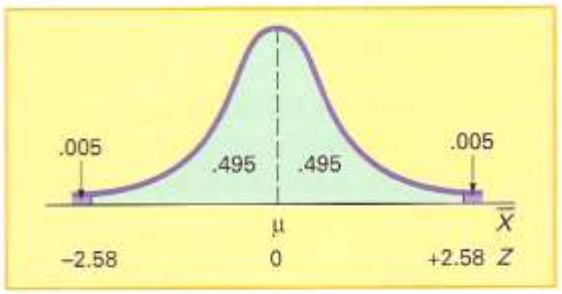
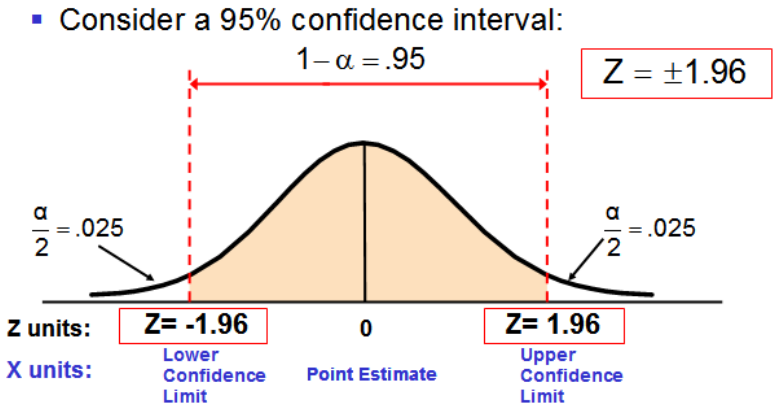


FIGURE 7.3
Normal curve for determining the Z value needed for 99% confidence

Selection of Z Critical Value



In this example we are working to a 95% Confidence interval.

① Solve the Confidence Level, (1 - α), for the confidence α.
② 1 - α = .95 → α = .05
③ Determine the Upper Tail Area: $\frac{\alpha}{2} = \frac{.05}{2} = .025 = 2.5\%$
④ Find the Z value which gives the cumulative probability below $\frac{\alpha}{2}$.

$$1 - \frac{\alpha}{2} = 1 - \frac{.05}{2} = 1 - .025 = .9750 = 97.5\% \quad \Rightarrow \quad Z = 1.96$$

Key Points

- You are finding the point on the Z axis which give the probability up to the $\frac{\alpha}{2}$ point.
- The upper and lower Z's are symmetrical.

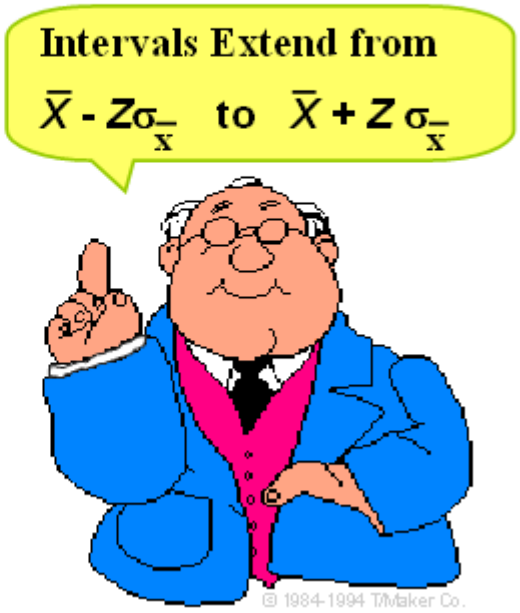
Commonly used confidence levels are 90%, 95%, and 99%

Confidence Level	Confidence Coefficient, $1-\alpha$	Z value
80%	.80	1.28
90%	.90	1.645
95%	.95	1.96
98%	.98	2.33
99%	.99	2.57
99.8%	.998	3.08
99.9%	.999	3.27

When working with this table remember that you are indexing in with the 1- α value, not the α/2 value!

Factors Affecting Interval Width (Precision)

- Data variation
  - Measured by σ
- Sample size
  - $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
- Level of confidence
  - $100(1-\alpha)\%$



$Z_{\frac{\alpha}{2}}$, $\sigma_{\bar{X}}$, n all effect accuracy.

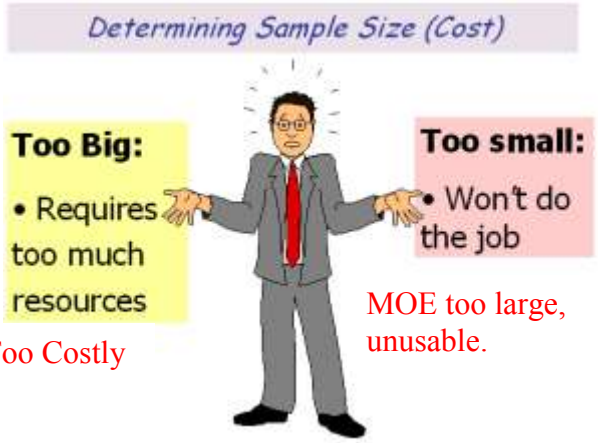
Increasing n will decrease the margin of error. But n is under the radical so in order to

Reduce MOE by $\frac{1}{2}$ we have to take $4n$ samples.

Reduce MOE by $\frac{1}{3}$ we have to take $9n$ samples.

Reduce MOE by $\frac{1}{10}$ we have to take $100n$ samples.

The square root term requires much larger sample sizes in order to reduce error / improve accuracy.



Too Costly

MOE too large, unusable.

Determining Sample Size for Mean

What sample size is needed to be 90% confident of being correct within ± 5? A pilot study suggested that the standard deviation is 45.

Not really correct because the pilot calculation is also a sample.

This is an approximation only.

$$n = \frac{Z^2\sigma^2}{\text{Error}^2} = \frac{1.645^2\left(45^2\right)}{5^2} = 219.2 \cong 220$$

In this example we need σ to calculate sample size n. This is not practical.

Round Up

They are solving $\text{MOE} = \pm Z\frac{\sigma}{\sqrt{n}}$ for n. In this example I think they are using the 95% Z value instead of the 90% value they say they are solving for.

The above all assumed we knew the value of σ, an impractical procedure. Now we move to the case where we do not know σ to begin with. Even if you substitute S for σ this is still only an approximation.

Z Critical Value: $1-\frac{\alpha}{2} = 1-\frac{.10}{2} = 1-.05 = .95 = 95\% \quad \Rightarrow \quad Z = 1.645$

Confidence Interval for μ (σ Unknown)

- Assumptions
  - Population standard deviation is unknown
  - Population is normally distributed
  - If population is not normal, use large sample
- Use Student's *t* Distribution
- Confidence Interval Estimate

n-1

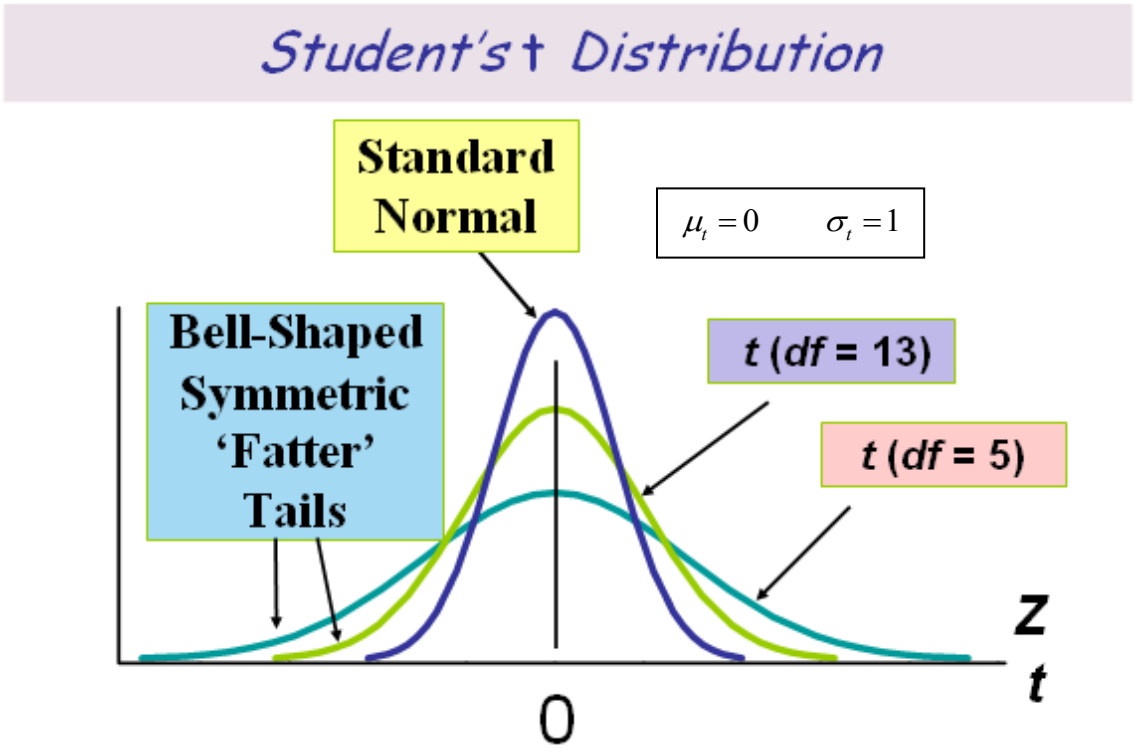
$$\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}$$

Here we have substituted sample S for population sigma and the student-t distribution for Z. The sample S can vary from sample to sample so this introduces extra uncertainty. Note that we use the t distribution with n-1 degrees of freedom, NOT n !!!

- Confidence Interval Estimate:

$$\bar{X} \pm t_{n-1}\frac{S}{\sqrt{n}}$$

Where t is the critical value of the t distribution with n-1 d.f. and an area of α/2 in each tail.



Student-t distribution has a mean of 0 and a standard deviation of 1. It also has a degrees-of-freedom parameter. Low df means flat curve/fat tails, high df means peak. Area under the curve is 1. As the number of degrees of freedom increases, the T Distribution gradually approaches the standard normal distribution. This is because S becomes a better estimate of σ as the sample size gets larger.

WITH A SAMPLE SIZE OF ABOUT **120 OR MORE**, S ESTIMATES σ PERCISELY ENOUGH THAT THERE IS LITTLE DIFFERENCE BETWEEN THE t AND Z DISTRIBUTIONS.

The validity of the confidence interval should be of concern primarily when dealing with a small sample size and a skewed population distribution.

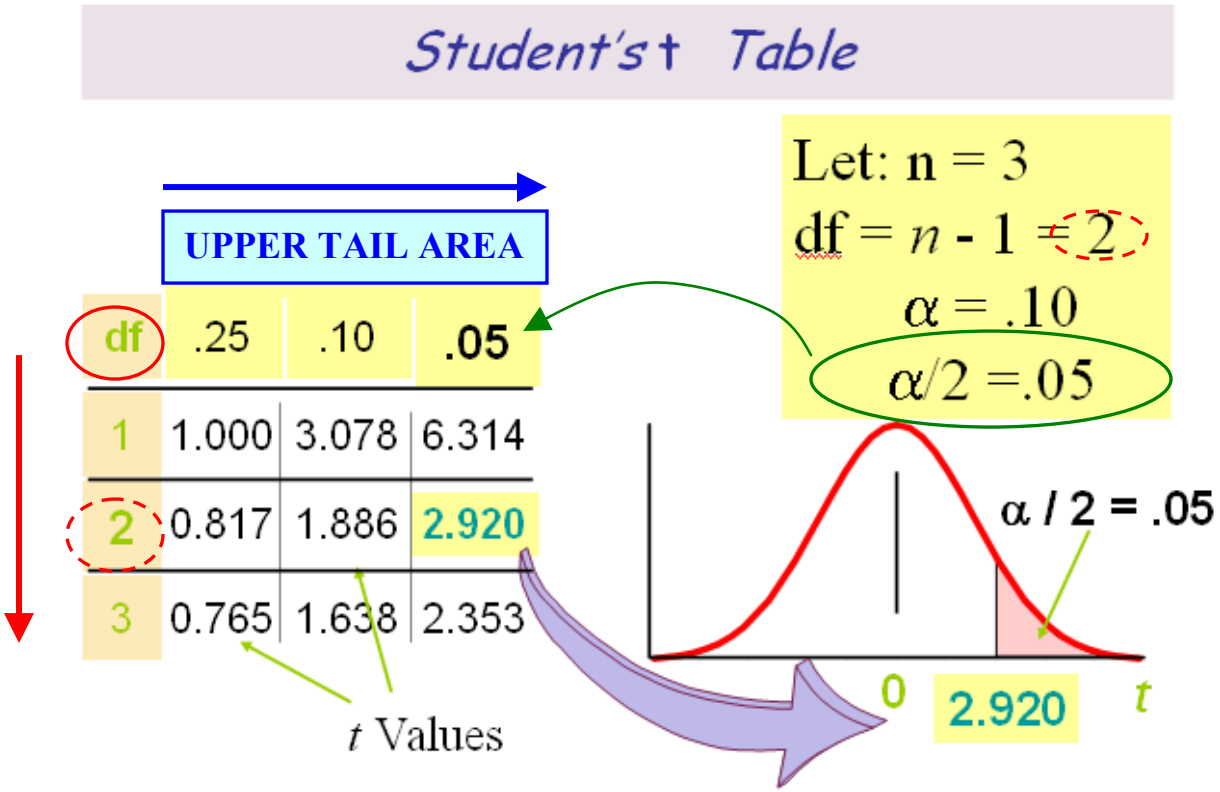
Degrees of Freedom (df)

- Number of observations that are free to vary after sample mean has been calculated
- Example
  - Mean of 3 numbers is 2

$$X_1 = 1 \text{ (or any number)}$$
$$X_2 = 2 \text{ (or any number)}$$
$$X_3 = 3 \text{ (cannot vary)}$$

EXAM

Know how to find the value of t.
Can use tables or spreadsheet.



The body of the table contains t values not probabilities (as in the z case).

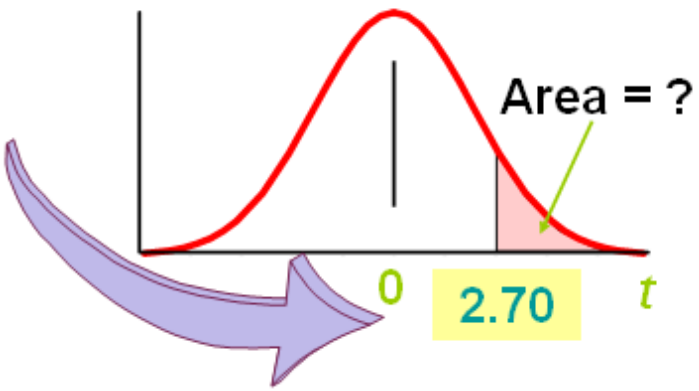
Degrees of Freedom is always n – 1.

EXAM Do not forget to take 1 away from n to find degrees of freedom!

Student's t Table in Excel

Let: n = 3
df = n - 1 = 2

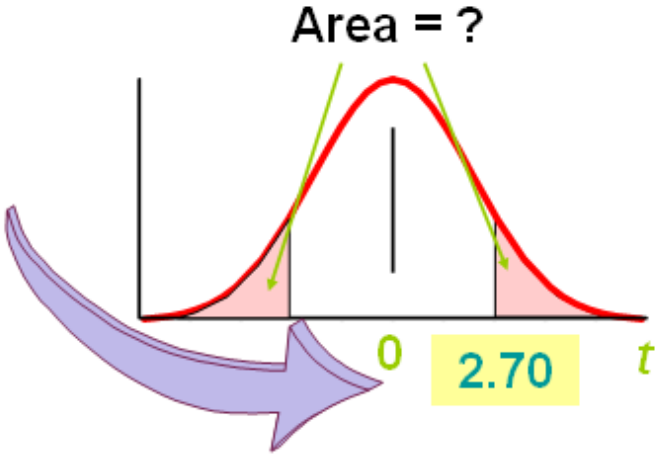
=TDIST(t,df,tails)
=TDIST(2.7,2,1)
Answer: 0.057



Student's t Table in Excel

Let: n = 3
df = n - 1 = 2

=TDIST(t,df,tails)
=TDIST(2.7,2,**2**)
Answer: 0.114



Student's t Table in Excel

$= \alpha$

=TINV(probability,df)

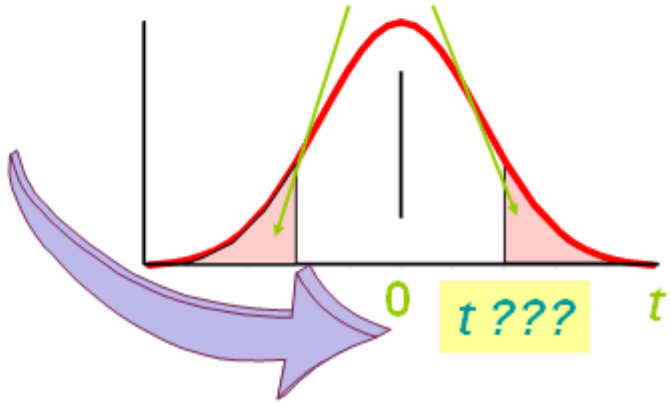
=TINV(0.07,2)

Answer: 3.578

Let: n = 3

df = n - 1 = 2

Area = 0.07



Example

A random sample of $n = 25$ has $\bar{X} = 50$ and $S = 8$.

Set up a 95% confidence interval estimate for μ

$$\bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \le \mu \le \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}$$

$$50 - 2.0639\frac{8}{\sqrt{25}} \le \mu \le 50 + 2.0639\frac{8}{\sqrt{25}}$$

$$46.69 \le \mu \le 53.30$$

Upper tails will each have area α /2 = (1-.95)/2=.025

Confidence Interval Estimate for Proportion

- Assumptions
  - Two categorical outcomes
  - Population follows binomial distribution
  - Normal approximation can be used if $np \geq 5$ and $n(1-p) \geq 5$
  - Confidence interval estimate

$$p_S - Z_{\alpha/2}\sqrt{\frac{p_S(1-p_S)}{n}} \leq p \leq p_S + Z_{\alpha/2}\sqrt{\frac{p_S(1-p_S)}{n}}$$

Collect CATEGORICAL data and calculate a proportion.
Two categorical outcomes which follow a binomial distribution which can be approximated by a normal distribution when np > 5 and n(1-p) > 5. Using these properties we can also make a CONFIDENCE INTERVAL using a Z value.

Example

A random sample of 400 Voters showed 32 preferred Candidate A. Set up a 95% confidence interval estimate for *p*.

$$p_s - Z_{\alpha/2}\sqrt{\frac{p_s(1-p_s)}{n}} \le p \le p_s + Z_{\alpha/2}\sqrt{\frac{p_s(1-p_s)}{n}}$$

$$.08 - 1.96\sqrt{\frac{.08(1-.08)}{400}} \le p \le .08 + 1.96\sqrt{\frac{.08(1-.08)}{400}}$$

$$.053 \le p \le .107$$

Example:

$p_s = \frac{32}{400} = .08$ or 8%. So what is a good estimate of p?

p_s is normally distributed around the true mean p with standard deviation

$\sigma = \sqrt{\frac{p(1-p)}{n}}$. We will approximate p by using the sample proportion so that

$\sigma_{p_s} = \sqrt{\frac{p_s(1-p_s)}{n}}$. Now $MOE = 2 \times \sqrt{\frac{.08(1-.08)}{400}} = .027$.

95% Confidence $\rightarrow \alpha = .05, \rightarrow Z_{\frac{.05}{2}}$ means this is a Z with $\frac{.05}{2} = .025$ area to the right.

In this case we would use the Confidence Interval for Proportions spreadsheet.

Determining Sample Size for Proportion

Out of a population of 1,000, we randomly selected 100 of which 30 were defective. What sample size is needed to be within ± 5% with 90% confidence?

$$n = \frac{Z^2 p(1-p)}{\text{Error}^2} = \frac{1.645^2 (0.3)(0.7)}{0.05^2}$$

$$= 227.3 \cong 228$$

Round Up

Z Critical Value: $1 - \frac{\alpha}{2} = 1 - \frac{.10}{2} = 1 - .05 = .95 = 95\% \quad \Rightarrow \quad Z = 1.645$

In general, it is usually not easy to specify the two factors needed to determine sample size.

For the sampling error you should be thinking not of how much sampling error you would like to have (none) but of how much can be tolerated while still permitting you to draw adequate conclusions from the data.

The population standard deviation, σ, is rarely known. In some instances the population standard deviation can be estimated from past data. Sometimes the best estimate is an educated guess taking into account the range and distribution of the variable. If the data can be assumed to be normal it is known that the range is approximately 6σ, so an estimate of the standard deviation σ can be made as range/6. In other cases a pilot study may be required in order to determine the sample standard deviation.

REFER TO "CONSTRUCTING CONFIDENCE INTERVALS WITH EXCEL" TUTORIAL.

Example

7.55 pg 315 Which spreadsheet is called for? Total population is 1546

$n = 50$ $\bar{X} = \$252.28$ S=$93.67

Setup a 95% confidence interval estimate of the total estimated value.

n>30, central limit theorem applies.

Confidence Interval for Known Std Dev			
Sample size	50		Inputs
Sample Mean	252.28		
Sample Std Dev	93.67		
Confidence level	0.95		
Point Estimate	252.28		Results
Alpha	0.05		
Margin of error	26.62072		
Interval	225.6593	278.9007	

Example

7.49 pg 307

Sample size: 611 # responded: 434

a) Construct 95% CI for the proportion who responded within an hour or two.

Confidence Interval for Proportions			
Sample size	611		Inputs
Sample proportion	0.7103		
Confidence level	0.95		
Point estimate	0.7103		Results
Alpha	0.05		
Margin of error	-0.0360		
Interval	0.6743	0.7463	

EXAM On exam, if given a list of numbers just use Excel to calculate the stdev() and average() [mean] then proceed with the analysis.