BUSN 210 FUN WITH STATS & EXCEL

Chapter 8

construct Confidence Intervals* with Z and t and Determine sample size

*also called Interval Estimation

Chapter 7

In chapter 7 we knew or assumed information about population mean (μ) and population standard deviation (σ) and then we:

made a reasonable statement about $\bar{X}$.

such as:

"we are 95% sure that our $\bar{x}$ (sample mean) will occur between 5954 lbs. and 6047 lbs."

But we had our population parameter, μ, & σ to base it on.

Chapter 8

In chapter 8 we will learn what to do when we don't have information about the population mean (μ) and population standard deviation (σ).

*we may not have information about shape of Distribution, either.

Here we will:

take an $\bar{X}$ (sample mean)

Add some room for error ($z * \frac{\sigma}{\sqrt{n}}$ or $t \frac{s}{\sqrt{n}}$)

and make a statement about μ.

such as:

"we are 95% sure that the population mean (μ) occurs between 2192 and 2626 printed pages." (estimate for printer output).

point Estimate

The statistic, computed from sample information, which is used to estimate the population parameter.

Examples:

$\bar{X}$ is a point estimate of μ (pop. mean)

$\bar{p}$ (sample proportion) is a point estimate of p
(p or π is population proportion)

s is a point estimate of σ (pop. SD)

Point Estimate for μ (pop. mean) is $\bar{X}$ (Sample Mean)

why do we need to use a sample mean, $\bar{X}$, to estimate μ?

Because in many cases we do not have the population mean.

Example:

1. we do not KNOW how many meals married couples eat out each week
   - we can take a sample and our $\bar{x}$ (mean) will be the point estimate for the unknown pop. mean.
2. we do not KNOW what the mean amount of profit is per auto sold in USA.
   - we can take a sample and our $\bar{X}$ will be the point estimate for the unknown pop. mean.

But if we use $\overline{X}$ as a Point estimate for our population Mean, μ, what about sampling error $(\bar{x}-\mu)$? Does $\bar{x}=\mu$ every time? NO!

So we can't just use $\overline{X}$. we will have to add some "error" below and above our $\overline{X}$ to have a range of values that we can use to estimate our μ, population mean.

Confidence Interval (C.I.)

A range of values constructed from sample data so that the population parameter (μ or p) is likely to occur within that range at a specified probability. The specified probability is called the "level of confidence."

also → "confidence level"

also → confidence coefficient

Example:

For building a deck on a house we are 90% sure that the population mean lies between 8 h. 47 mins & 7 h. 13 mins.

The Level of confidence is 90% (There is a 10% risk that the pop. mean is not in our C.I.)

The "confidence Limits" are 8 h. 47 mins and 7 h. 13 mins.

"Confidence Level" =

"confidence level" = "confidence coefficient" =

.90 or 90%

= how likely the population parameter will occur in the range of values 8h. 47 mins to 7h. 13 mins.

"Level of Significance" = α = 1 − "confidence level"
= "Significance Level" =

= 1 − .9 = .10 or 10%

This is the risk that the population parameter is not in our inverval. It is the risk we take of making a mistake with our inferences.

Interval Estimation
also called
Confidence Intervals

② used to estimate the population parameter!!!

① The purpose of interval estimation is to provide information about how close the point estimate is to the value of the population parameter.

Also, said a different way

Pop. mean ⇒ $\bar{x}$ +/- margin of error

Pop. proportion ⇒ $\bar{p}$ +/- margin of error

σ KNOWN or at least reliably estimated

① when there is large amounts of historic data, sigma (σ) may be KNOWN.

② Quality control applications where process is assumed to be operating correctly, it is appropriate to treat sigma as KNOWN

③ Most situations σ is not KNOWN, so we don't use Normal (z) Distribution, we use t-distribution. In this case we will have to use the same sample to estimate μ & σ. We will use X_{bar} & S.

(more later....)

This text book does this: P. 3.75

Interval Estimate of a pop. mean

σ KNOWN or reliably estimated

$$\boxed{\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}$$

construct confidence Limits (Lower & upper End)

Exmple: where confidence level = 95%

$\alpha = 1 - .95 = .05$

$\alpha/2 = .025$

$$\boxed{Z_{\alpha/2} = \text{Upper } Z}$$

$$\bar{X} \pm 1.96 * \frac{\sigma}{\sqrt{N}}$$

* If population distribution is Normal, Formula works correctly
* if Not Normal, it will be approximate, the bigger the n, the more accurate.
* if not, it depends on: 1) pop Distribution & n

Confidence Interval for Pop. mean with σ Known

Formula: $\boxed{\bar{X} \pm Z \underbrace{\frac{\sigma}{\sqrt{n}}}_{\text{Margin of error}}}$ } standard error (standard deviation of sampling distribution of the sample mean)

Something like: within our interval we are 95% sure that the pop. mean, μ, will occur.

Margin of error added to each side of $\bar{X}$ → The error is because we can't just say $\bar{X}$ is μ. We have to say

Z is found by taking Confidence Probability or Percent, dividing by 2 and using the formula:

$= NORMSINV(1-\alpha/2)$

OR

$$= NORMSINV\left(\frac{\text{confidence \%}}{2} + .5\right)$$

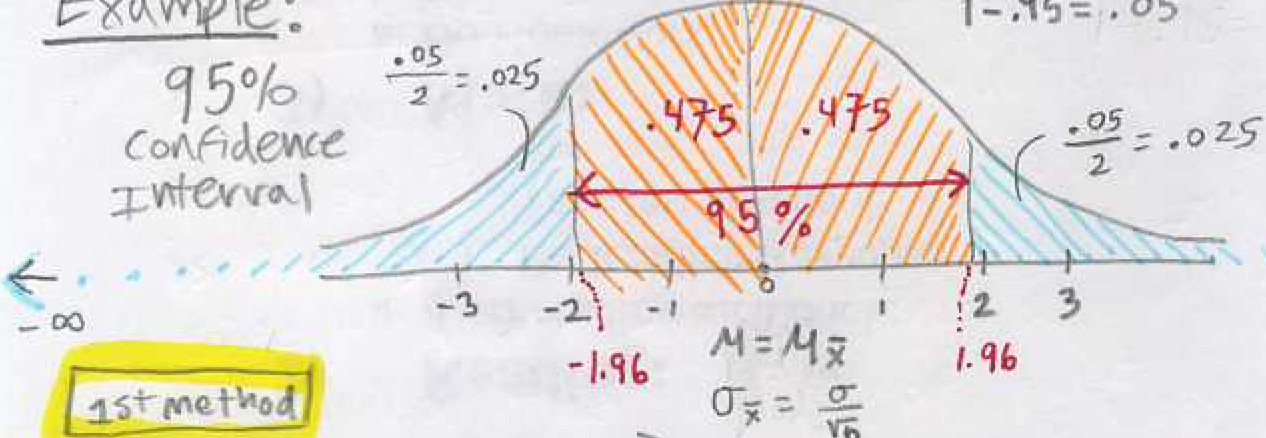
variables:

$\bar{X}$ = sample mean
σ = pop. SD
n = sample size
Z = Z-score or Z-value Standard Normal value

Example:

95% Confidence Interval

$1 - .95 = .05$



The central Limit Theorem says that if our n is large, the Distribution of Sample means will be bell shape

1st method

① $Z = NORMSINV\left(\frac{.95}{2} + .5\right) \approx 1.96$

② then use formula $\bar{X} \pm Z * \frac{\sigma}{\sqrt{n}}$ to find Confidence Limits

2nd method OR find 2 $\bar{X}$ this way:

Lower value:

$= NORMINV\left(\frac{.05}{2}, \bar{X}, \text{standard error}\right)$

upper value:

$= NORMINV\left(1 - \frac{.05}{2}, \bar{X}, \text{standard error}\right)$

3rd method

Excel Function CONFIDENCE

① Margin of Error $=$ CONFIDENCE(alpha, sigma, $\bar{X}$)

② Then use $\bar{X}$ +/- (Margin of Error) to find Confidence Limits.

Example 1 $\boxed{\sigma \text{ KNOWN}}$

Here we are given $\bar{X}$ & expected to estimate a confidence Interval for the population mean, μ.

P. 5

The Solid Construction Company

The Solid Construction Company constructs decks for residential homes. They send out two person teams to build decks. The company conducts a sample of 40 jobs and calculates a mean completion rate of 8 hours to build a typical deck. The standard deviation for the population is known to be equal to 3 hours (from past data).

what is a reasonable range of values for pop. mean

Another way to ask

But what about sampling Error? Does $\bar{X}$ always $= \mu$? we must add some error room" to each side of our $\bar{X}$

1. List variables: $n = 40$
   $\bar{X} = 8$ hours
   $\sigma = 3$ hours
   Standard error $= \frac{3}{\sqrt{40}} = .4743$

2. What is the best estimation for the population mean?
   our best estimate for the population mean is our sample mean, $\bar{X} = 8$ hours. "$\bar{X} = 8$ hours" is a point estimate for our unknown Population mean.

3. Determine a 90% Confidence Interval
   a. State the level of confidence, then divide by two and find probability [redacted] then find z or use NORMSINV function

   confidence Interval = 90%
   $.9/2 = .45$
   α = Risk that μ is not in interval $= .10$

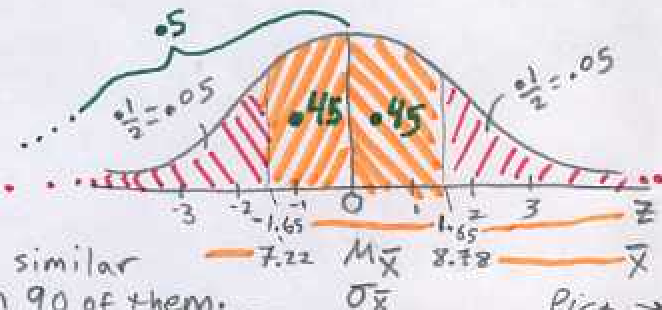
   $\alpha/2 = .10/2 = .05$
   z = NORMSINV$(1 - .05) \approx 1.65$
   or
   z = NORMSINV$(.45 + .5) \approx 1.65$

   b. Using the correct confidence interval formula, calculate the confidence limits

   $\bar{X} \pm z * \frac{\sigma}{\sqrt{n}}$
   $= 8 \pm 1.644853627 * .474341649$
   $= 8 \pm .780222582$

   7.219777418 &
   8.780222582

   minutes
   $.22 * 60 \approx 13$ mins.
   $.78 * 60 \approx 47$ mins.



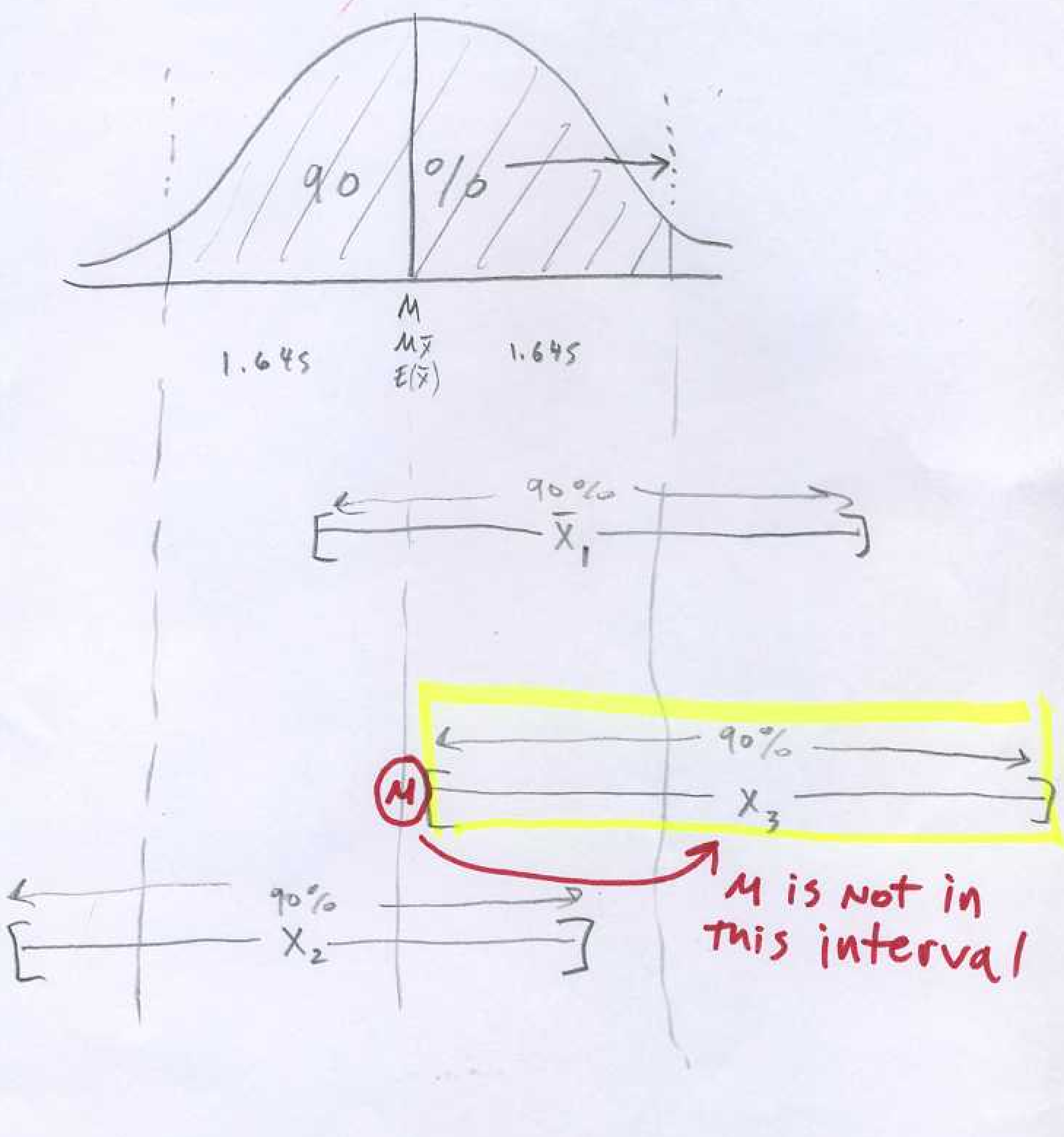
4. Conclusions:
   we are 90% sure that the population mean occurs between 8h. 47 mins & 7h. & 13 mins. If we were to construct 100 similar intervals, we would expect to find pop. mean in 90 of them.

Picture

5. Would it be reasonable to conclude that the population mean is 9 hours? (If SCC claimed 9 hours in an ad, we would treat that as a population mean, mu = 9 hours).
   Because 9 hours is outside our interval, it would not be reasonable to conclude that the pop. mean is 9 hours. SCC's claim would not seem reasonable

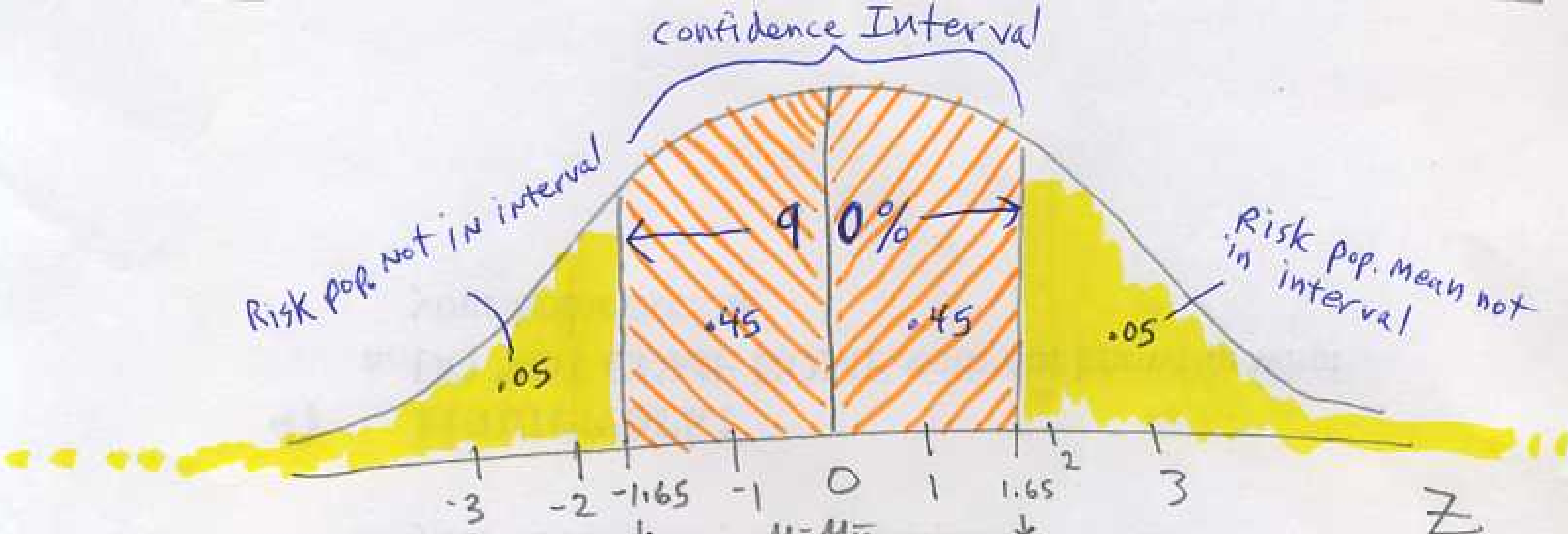
6. What if SCC claimed the population mean is 6 ½ hours? (If SCC claimed 6.5 hours in an ad, we would treat that as a population mean, mu = 6.5 hours).
   Because 6.5 hours is outside our interval, it would not be reasonable to conclude that the population mean is 6.5 hours. SCC's claim would not seem reasonable.

7. What if SCC claimed 7 ½ hours was their average job time? Is this reasonable? (If SCC claimed 7.5 hours in an ad, we would treat that as a population mean, mu = 7.5 hours).
   Because 7.5 hours is inside our interval, it would be reasonable to conclude that the pop. mean is 7.5 hours. SCC's claim seems reasonable.


90%
μ
μ X̄
E(X̄)
1.645
1.645
90%
X̄1
90%
μ
X3
μ is not in this interval
90%
X2

Visual summary for Example 1 σ Known

From $\bar{X}$ 8 hours, we created Confidence Interval so we were 90% sure that population mean was in the Interval 7.22 hours to 8.78 hours.



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

$\sigma_{\bar{x}} = \frac{3}{\sqrt{40}}$

$\bar{X}$ (hours)

{Confidence Limit on Low end} → ? → $8 - 1.65 * \frac{3}{\sqrt{40}}$ } standard error → $8 - .78$ Margin of error → 7.22 hours

{Confidence Limit on upper end} → ? → $8 + 1.65 * \frac{3}{\sqrt{40}}$ } standard error → $8 + .78$ Margin of error → 8.78 hours

Formulas:

$Z = \text{NORMSINV}(1-.05) = 1.65$

$\left\{\text{Confidence Limits}\right\} = \bar{X} \pm Z \frac{\sigma}{\sqrt{n}}$

or

Low value = $\text{NORMINV}(.05, \bar{X}, \frac{\sigma}{\sqrt{n}})$

upper value = $\text{NORMINV}(1-.05, \bar{X}, \frac{\sigma}{\sqrt{n}})$

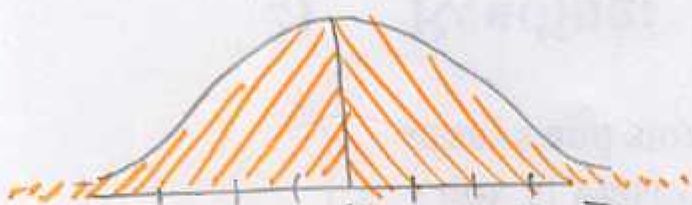
Conclusion: We are 90% sure that the population mean occurs between 7.22 hours and 8.78 hours. A reasonable range of values for the population mean is 7.22 hours to 8.78 hours.

Q: But what if we do not KNOW the population standard deviation, sigma (σ)? Can we still construct confidence Intervals?

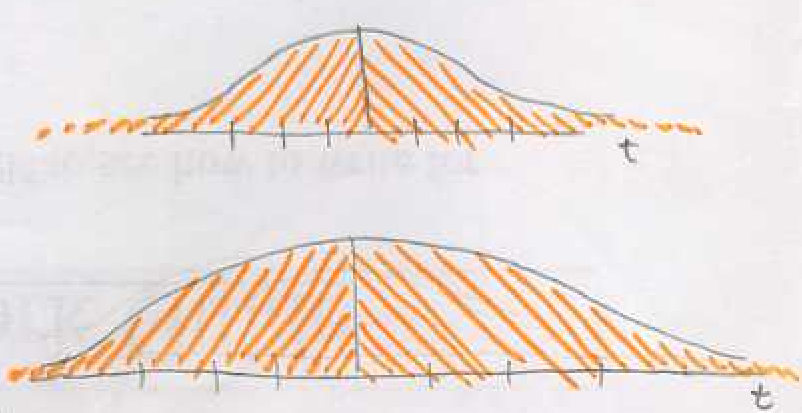
A: Yes we can! But will have to use the point estimate sample standard deviation, S, in place of σ.

When we use the sample standard deviation, S, instead of the population standard deviation, σ, we cannot use the standard Normal curve & the Z-score. When we use the sample standard deviation, s, we must use the $\boxed{\text{t-distribution}}$.

σ KNOWN	σ NOT KNOWN
$Z = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$	$t = \dfrac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$



1 standard Normal curve using Z



many t-distribution curves

t - Distribution (also known as: Student's t Distribution)

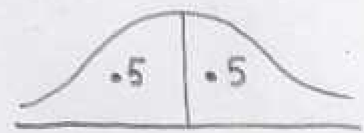
Q: When are we allowed to use the t-Distribution?

A: When the population is Normal (bell) shaped or nearly normal (bell) shaped. (OR n is large)

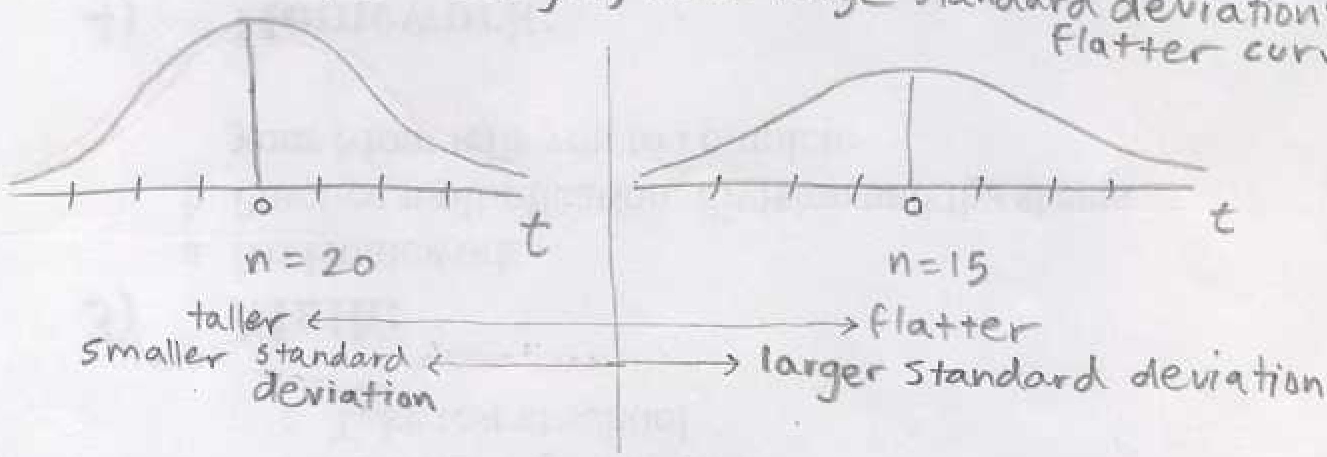
Characteristics of t-Distribution

① Continuous Distribution

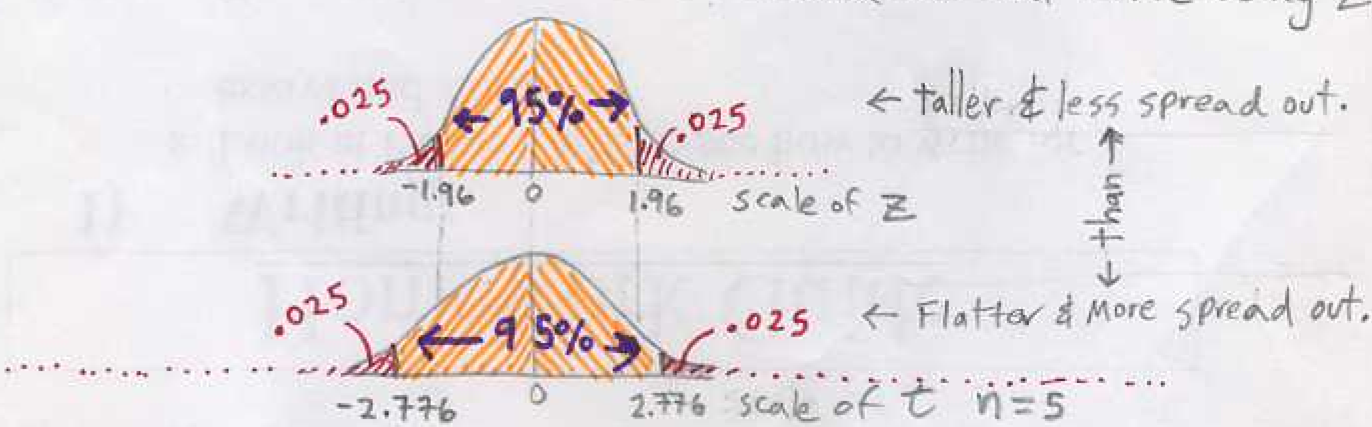
② Bell-shaped and symetrical



③ The are many t-Distributions (one for each n). Smaller sample sizes, n, have large standard deviations & flatter curves.



④ At the center, the t-Distribution is flatter and more spread out than the Standard Normal curve using Z.



⑤ As n increases, t-Distribution approaches Normal standard curve using Z.

confidence Interval for pop. mean with σ Not KNOWN

Formula $\boxed{\bar{X} \pm \underbrace{t \frac{s}{\sqrt{n}}}_{\text{Margin of error}}}$ } standard error or standard deviation of Sampling Distribution of the sample mean

Margin of error ← (we add this to each side of $\bar{X}$)

* Margin of error changes from sample to sample

The error added to each side is used because we can't just say $\bar{X} = \mu$. we have to say something like: "within our interval we are 95% sure that μ (pop. mean) will occur."

① To use the t-Distribution, the population of interest must be approximately bell shaped, or n must be large. (safe $n \geq 50$)

② must have: (plot & look!)

- $\bar{X}$ = sample mean
- s = sample standard deviation
- n = sample size
- df = degree of freedom = n − {number of samples}
- CL = confidence Level = .95
- SL = significance Level = 1 − .95 = .05

③ to find t:

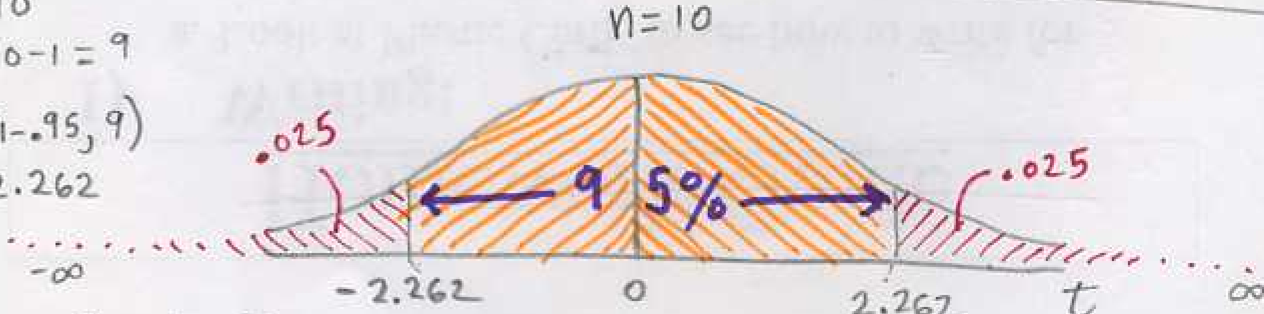
method 1 ① Look up in back of text book with:
① upper t value
② df

OR

method 2 ② =TINV(1 − {confidence Interval Percent}, Degree of Freedom)

Example:
95% confidence Interval
n = 10
df = 10 − 1 = 9

① =TINV(1−.95, 9)
= t = 2.262



② Then use formula $\bar{X} \pm 2.262 * \frac{s}{\sqrt{n}}$

$\boxed{\text{Method 3}}$

use "Data Analysis Add-in" to Generate Descriptive statistics: "confidence Level" = Margin of Error

---

Textbook does this:

$\boxed{\begin{array}{c}\text{Interval Estimate of a pop. mean} \\ \underline{\sigma \text{ Not Known or reliably estimated}}\end{array}}$

$$\left\{\begin{array}{c}\text{Confidence} \\ \text{Interval} \\ \text{for} \\ \text{pop mean} \\ \text{using} \\ \text{t-distribution}\end{array}\right\} \quad \bar{X} \pm t_{\alpha/2} * \frac{s}{\sqrt{n}}$$

$\boxed{t_{\alpha/2} = \text{upper } t \text{ value}}$

$\boxed{\text{why df and not } n\text{?}}$

df = n − # of samples.

1. Because statistics (sample) are being used, you must determire the # of values are free to vary.

Example 7, 4, 1, 8

$$\frac{7+4+1+8}{4} = \frac{20}{4} = 5$$

$$\begin{array}{l} 7-5 = 2 \\ 4-5 = -1 \\ 1-5 = -4 \\ 8-5 = 3 \\ \hline \quad\quad\; 0 \end{array}$$

But if we use one value, say 8, the 3 can't be used. $2-1-4 \neq 0$

2. 1 degree of freedom (loss of one value to vary) is lost in a sampling problem involving "s" of sample because one number, $\overline{X}$, is known (Remember: this is distribution of $\overline{X}$).

Example 2 | σ NOT KNOWN

Printers Manufacturer

Printers Manufacturer is a manufacturer of ink jet printers. They would like to include as part of their advertising the number of pages a user can expect from an ink cartridge. A sample of 10 cartridges was taken with a mean of 2409 pages and a standard deviation of 304 pages.

8. List variables:

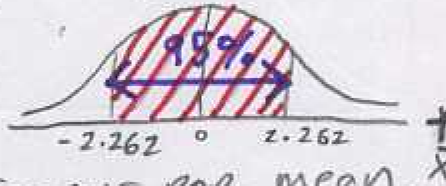
$n = 10$

$df = 10 - 1 = 9$

$\bar{X} = 2409$ pages

$s = 304$ pages

$\frac{s}{\sqrt{n}} = \frac{304}{\sqrt{10}} = 96.13$



9. What is the best estimation for the population mean?

$\bar{X} = 2409$ pages is the best estimate for our pop. mean.
"point estimate"

10. Determine a 95% Confidence Interval

a. State the level of confidence, state the df, then look up t in back of book.

level of confidence = .95 $\quad df = 10 - 1 = 9 \quad \Rightarrow t = 2.262$

Excel = TINV(1-.95, 9)

2409 ± 217.45 Margin of error

b. Using the correct confidence interval formula, calculate the confidence limits

$\bar{x} \pm t \frac{s}{\sqrt{n}}$ $\quad 2409 \pm 2.262 * 96.13$ $\quad$ 2192 & 2626

"The limits for the 95% CI are 2192 pages and 2626 pages.

11. Conclusions:

- The customer can expect an average of 2409 pages/cartridge.

- The typical usage ranges from 2192 pages to 2626 pages (given a 95% CI). The margin of error is .

- The margin of error is 217.45 pages.

- we are 95% sure that the population mean lies between 2192 & 2626 pages.

- If we constructed 100 similar intervals, 95 of them would contain pop. mean.

Example 3 $\boxed{\sigma \text{ Not KNOWN}}$ same example as before, except σ not known.

The Solid Construction Company

The Solid Construction Company constructs decks for residential homes. They send out two person teams to build decks. The company conducts a sample of 40 jobs and calculates a mean completion rate of 8 hours to build a typical deck. The standard deviation in the sample was 3 hours.

1. List variables: $n = 40$
   $\bar{X} = 8$ hours
   $s = 3$ hours

   # of samples = 1
   degree of freedom = $df = 40 - 1 = 39$

   Standard error $= \frac{s}{\sqrt{n}} = \frac{3}{\sqrt{40}} = .4743$

2. What is the best estimation for the population mean?

   our best estimate for the unknown population mean is our sample mean, $\bar{X} = 8$ hours. 8 hours is our point estimate for μ.

3. Determine a 90% Confidence Interval
   a. State the level of confidence [redacted], then find z

   Confidence Interval = 90%

   $= TINV(1 - .90, 39) = 1.684875122 = t$

   b. Using the correct confidence interval formula, calculate the confidence limits

   $\bar{X} \pm t * \frac{s}{\sqrt{n}}$

   $8 \pm 1.684875122 * .474341649$
   $8 \pm .799206444$

   Intervals: 7.20 to 8.80
   or
   7 h. 12 mins. to 8 h. 48 m.

4. Conclusions:

   We are 90% sure that the population mean occurs between 7 hours 20 minutes and 8 hours 48 minutes.

   Notice the interval is wider when we use the t-Distribution as compared to Normal Curve with Z.

5. Would it be reasonable to conclude that the poulatiom mean is 9 hours?

   Because 9 hours is Not inside our interval, it would not be reasonable to conclude that the pop. mean is 9 hours.

6. What if SCC claimed the population mean is 6 ½ hours?

   Because 6 ½ hours is not inside our interval, it would not be reasonable to conclude that the pop. mean is 6 ½ hours.

7. What if SCC claimed 7 ½ hours was their average job time? Is this reasonable?

   It is reasonable to conclude that the pop. mean is 7 ½ hours because it was between our confidence limits of 7.2 hours & 8.8 hours.

Proportions

Proportions: The fraction, ratio, or percent indicating the part of the sample or the population having a particular trait of interest.

Example: A recent survey of Highline students indicated that 98 out of 100 surveyed thought textbooks were too expensive.

---

x = # of successes = particular trait of interest = 98
n = sample size = 100
$\bar{p} = \frac{x}{n}$ = sample proportion

Notes about proportions:

1. The sample proportion is our best estimate of our population proportion, π.
2. Proportions are Nominal Level data.
3. Success or failure sounds Binomial, right? so...

In order to build a Confidence Interval for proportion:

must verify:

1. Are there a fixed # of Trials?
2. Are results independent?
3. Does each trial result in success or failure?
4. π stay same each Trial?
5. $n * \pi > 5$
   $n * (1-\pi) > 5$

construct of confidence Interval for Proportion:

alternative Notation

x = # of successes

n = sample size

$\frac{x}{n} = \bar{p}$ = sample proportion = best estimate for π or P

$$\left\{\begin{matrix}\text{Confidence} \\ \text{Limits}\end{matrix}\right\} = \boxed{\bar{p} \pm Z\sqrt{\frac{\bar{p} * (1-\bar{p})}{n}}}$$

Margin of error

$P = \pi$ = population proportion

sampling error = $\bar{p} - \pi$ or $\bar{p} - p$

$\bar{p}$ = point estimate

---

or books Notations:

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

area on upper side.

---

* margin of error for estimating pop. proportion is almost always less than 0.10. Gallup & Harris commonly use $E = .03$ or $E = .04$

15

Furniture Land South

Furniture Land South surveyed their customers (n = 600) to see if they liked the new line of durable foam furniture decorated in bright colors. 414 said they were excited about the new line. All the binomial tests are met.

$n = 600$
$x = 414$

1 Fixed # trials = yes n = 600
2 independent ✓
3 S/F ✓
4 $\pi = .69$ same each time ✓
5 $\pi * n > 5 \Rightarrow .69 * 600 = 414 > 5$ ✓
6 $n * (1-\pi) > 5 \Rightarrow 600 * .31 = 186$ ✓
we can use Z!

1. List variables:

best estimate for π = $p = \frac{x}{n} = \frac{414}{600} = .69 \rightarrow$ use p for π

success = excited
Failure = not excited

2. What is the best estimation for the population proportion?

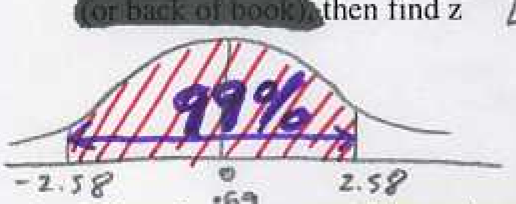
The point estimate p = .69 is the best estimate for our population proportion

3. Determine a 99% Confidence Interval
   a. State the level of confidence, [redacted] then find z

Level of confidence = .99

$.99/2 = .495$

$Z = 2.58$



   b. Using the correct confidence interval formula, calculate the confidence limits

$$p \pm Z\sqrt{\frac{p*(1-p)}{n}} = .69 \pm 2.58\sqrt{\frac{.69*.31}{600}} \Rightarrow .69 \pm .0487$$

.6413 and .7387 are the confidence limits.

4. Conclusions:

- The owner can be 99% sure that the population proportion (% of customers excited about new product) is between .6413 and .7387. The new product will probably be popular.

- .0487 is margin of error

- 99 of 100 similarly constructed intervals would contain pop. proportion.

Confidence Interval:

point estimator +/− Margin of Error

↓

If we know this we can solve for n, sample size.

Point Estimator +/− Error (E)

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

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

sample size for Interval Estimation for pop. mean

$$n = \left(\frac{Z_{\alpha/2} * \sigma}{E}\right)^2$$

sample size for Interval Estimation for pop. proportion

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

These are both estimates Judgment must be used sample size for confidence Interval pop mean $\Big\}$ $N = \left(\frac{Z_{\alpha/2} * \sigma}{E} \right)^2$

Requires estimate: ways to get estimate

1. use estimate for sigma from previous studies (or gov. studies)
2. pilot study before you run experiment
3. estimate large & small values:

$$\frac{\text{large} - \text{small}}{4} = \text{aproximate } \sigma$$

* Always Round up when calculating n!!

Excel: = ROUNDUP (n estimate, 0)

sample size for confidence Interval pop proportion

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

Requires estimate

How to estimate:

1. use estimate from previous studies or avaliable data or gov. data
2. Pilot study before you run experiment
3. Best guess (requires judgment)
4. none of these apply use $p = .50$

* ($.5 * (1-.5) = .25$ and as p increases it will never get bigger than .25). This will always yield largest sample size.