9.1 white text

Probability and Statistics **Unit 7 Guided Notes**

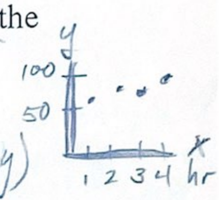
7.1: Correlation

This chapter talks about relationships between variables that are measured on the same individuals.

Terms to know:

1. Response variable: (y) The dependent variable in the study that measures the outcomes.
2. Explanatory variable: (x) The independent variable that influences change in the response variable.

Example 1: For each situation, decide which variable is the explanatory and which is the response.



a) Number of hours studying for an exam and the grade on the exam.

→ explanatory (indep., x) → response (dep., y)

b) The amount of saturated fat in a person's diet and that person's weight.

→ x: expl. (indep.) → y: response (dep)

c) The yield of a crop and the amount of yearly rainfall.

→ response (dep, y) → expl. (indep, x)

When examining the relationship of variables, use the following principles:

- Plot the data (and find numerical summaries)
- Look for overall patterns and deviations from these patterns
- If the pattern is regular, use a mathematical model to describe it.

→ REVIEW term

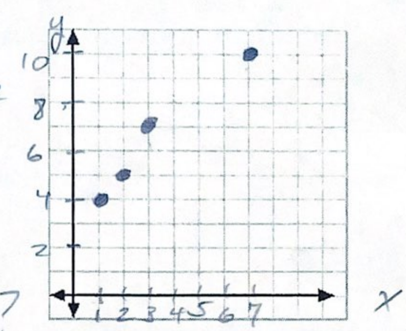
To display the relationship between **two quantitative variables**, use a scatterplot.

- Put the explanatory variable on the x-axis.
- Put the response variable on the y-axis.
- If the relationship cannot be determined, you may choose which variable goes where.

Example 2: Create a scatter plot of the data shown.

make up meaning! if time

x	1	2	3	7
y	4	5	7	10



patterns? could you put a line thru it to hit close to most?

as x incr., y incr. →

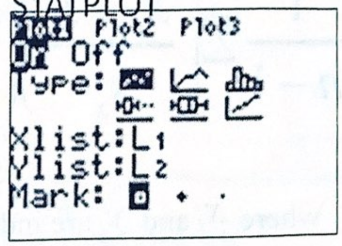
L1	L2
1	4
2	5
3	7
7	10

L1	L2

On the calculator…

Keystroke	Comments
STAT, EDIT	Put in L1 and L2
2^{nd} STATPLOT	Make sure the plot is ON
Choose the scatter plot icon	
ZOOM 9	To fit your data

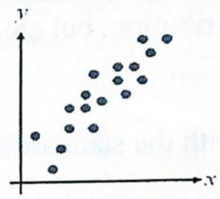
STATPLOT



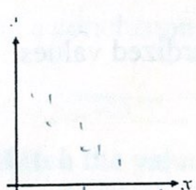
Types of Correlation:

As x incr,

y tends to incr.



Positive Linear Correlation

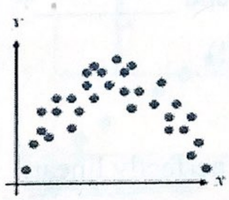


As x incr,

y tends to decr.

Negative Linear Correlation

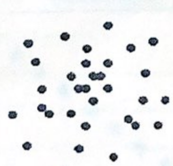
The pattern between x and y does not look like a straight line.



Nonlinear Correlation

quadr?

parabolic?



There is no relationship, no pattern, no trend between the x and y values.

No Correlation

Correlation Coefficient (r): a numerical value used to measure the strength and direction of linear relationships between 2 quantitative variables.

$$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

where x_i and y_i are individual values, n is the # of ~~individuals~~ pairs

(Note: we are using s to describe standard deviation because we are not using a population distribution, but are instead using sample data.)

Do you see the similarity with the standardization process for finding z-scores?

$$z = \frac{x - \bar{x}}{\sigma}$$

r is the AVERAGE of the products of the x and y standardized values.

Important facts about the correlation coefficient (r):

1. r will not change if we switch x and y.
2. Both variables must be quantitative (duh!)
3. $-1 \leq r \leq 1$ (always!)
4. If $r = 1$ or $r = -1$, then the relationship is perfectly linear. pos $r = 1$ neg $r = -1$
5. Value of $r = 0$ means there is no linear relationship (not the best description of form).
6. The correlation coefficient r only measures the strength of a linear relationship.

Example 3: Find the correlation coefficient (r):

x	1	2	3	7
y	4	5	7	10

Use TI-83: **IMPORTANT:** To calculate r, you must first turn on the ***DiagnosticOn*** command found in the Catalog menu.

TI-83/84

Keystroke	Comments
STAT, EDIT	To put in L1 and L2
2nd CATALOG	Scroll down to DiagnosticOn
ENTER, ENTER	Your screen should say "DONE" You only need to do this once (unless your calculator is reset.)
STAT, CALC, 4: LinReg (ax + b)	The value of r is the 4th one down.

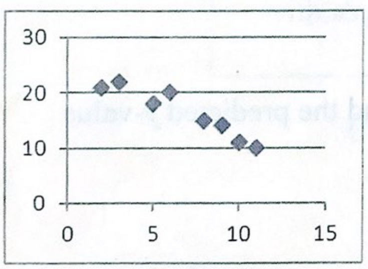
"linear regression line" → next lesson

a) What is the value of r? $r = .982$

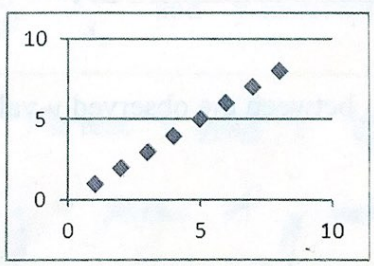
b) Make a conclusion about the type of correlation. str. pos. corr.

close to 1

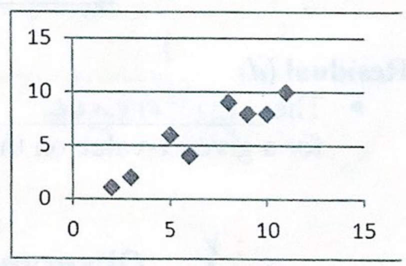
Example 4: **Match** the value of r to each scatter plot. Choices for r: -1, -.8, 0, 0.8, 1



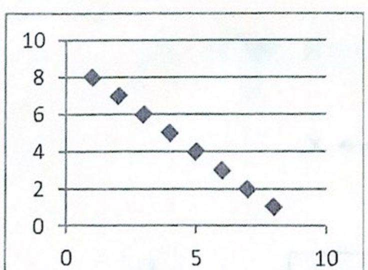
$r = {}^{-}0.8$



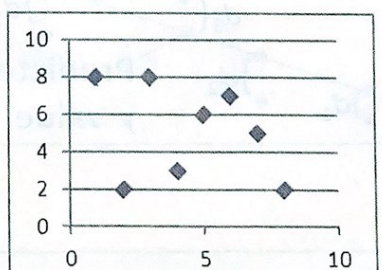
$r = 1$



$r = 0.8$



$r = -1$



$r = 0$

Example 5: a) display the data in a scatter plot on your calc; b) find the value of r;
c) make a conclusion about the type of correlation.

The number of hours 13 students spent studying for a test and their scores on the test.

Hours, x	0	1	2	4	4	5	5	5	6	6	7	7	8
Score, y	40	41	51	48	64	70	73	75	68	93	84	90	95

a)

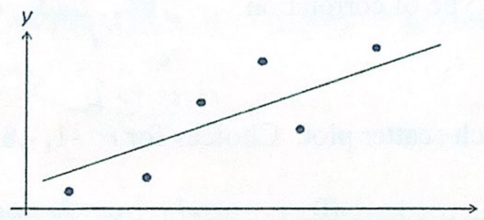
b) $r = 0.923$

c) str. pos. corr.

7.2: Regression Lines

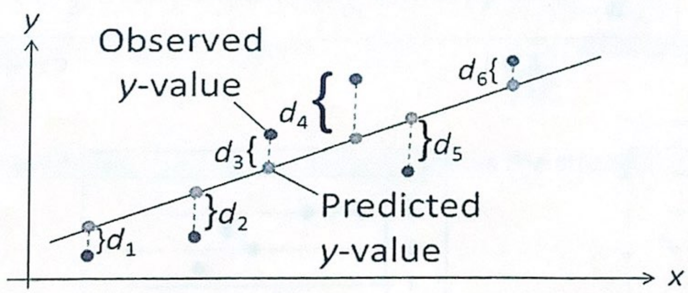
A **regression line** is a straight (not connect-the-points) line that describes how a response variable changes as an explanatory variable changes. We can use a regression line to predict the value of y for a given value of x (or to predict a value of x when given a value for y.)

"line of best fit"



Residual (d)

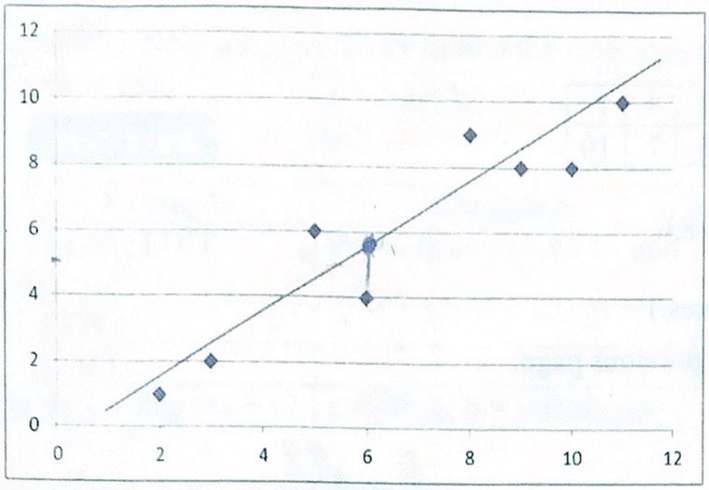
- The difference between the observed y-value and the predicted y-value for a given x-value on the line.



Residual: observed value – predicted value

Sometimes neg.

Example 1: Use the scatter plot and regression line shown to estimate the residual for $x = 6$.



residual: observed − predicted

$\approx 4 - 5.6$

≈ -1.6

Because different people draw different regression lines through data that appears to be linear, we will use a **least squares regression line** to make the sum of the squares of the residuals of the data points as small as possible.

The equation of a least squares regression line will be written as $\hat{y} = ax + b$

with slope $a = r\dfrac{s_y}{s_x}$ and y-intercept $b = \bar{y} - a\bar{x}$

(Note: $\hat{y}$ is read "y hat")

?

$\bar{y}$ = mean of y-values

$\bar{x}$ = mean of x-values

→ regr. line passes thru $(\bar{x}, \bar{y})$

Example 2: We will usually use our TI-83s to calculate $\hat{y}$. This time we will calculate it by hand:

x	1	2	3	7
y	4	5	7	10

Step 1: Make a scatterplot using your TI-83.
Step 2: Verify that the pattern is linear.
Step 3: Find r. (We did this in the 7.1 notes.)
Step 4: Find $\hat{y}$ using the formula on the previous page.

$r = .98$

$\hat{y} = ax + b$

$a = r \cdot \frac{S_y}{S_x}$

$S_x = 2.63$

$S_y = \boxed{}$?

Do Ex. 5

Example 3: Now use your TI-83 to find the least squares regression line. $\hat{y} = ax + b$

x	1	2	3	7
y	4	5	7	10

TI-83/84

Keystroke	Comments
STAT, EDIT	Put in your data for L_1 and L_2
STAT	
CALC	
4: LinReg (ax + b)	a = slope, b = y-intercept

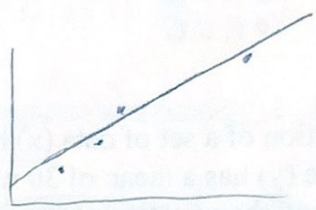
$a = .98795 \approx .99$

$b = 3.28915 \approx 3.29$

Write down the equation here:

$\hat{y} = .99x + 3.29$

Graph this line as y_1 in your TI-83.



Facts about least-squares regression

1. Deciding which variable is independent and which is dependent is essential. You will get a different equation if these are exchanged.
2. There is a close connection between the slope of the regression line and r.
3. The line will always pass through the point $(\bar{x}, \bar{y})$.
4. The square of the correlation, r^2, is the fraction of the variation in the values of y that is explained by the least squared regression line. In other words, r^2 tells us the percent of the data that fits the line well.

What does r^2 tell us in Example 3?

$r^2 = .964429$

So about 96% of the data fits this line well

$a = 50.7287$
$b = 104.0607$

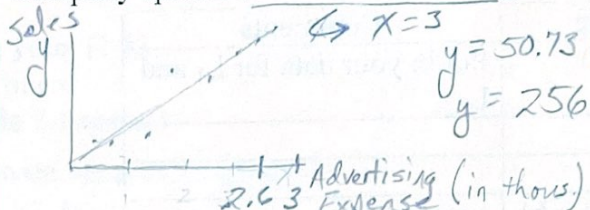
Example 4: Use the data shown.

Advertising expenses, (\$1000), x	Company sales (\$1000), y
2.4	225
1.6	184
2.0	220
2.6	240
1.4	180
1.6	184
2.0	186
2.2	215

a) Find the equation of the regression line.

$\hat{y} = 50.73x + 104.06$

b) Use your regression line to predict the sales when the company spends \$3000 on advertising.



$y = 50.73(3) + 104.06$

$y = 256.25$ thous.

\$256,250

c) Use your regression line to predict the amount of advertising money spent when the sales are equal to \$200,000. → $y = 200$

$\hat{y} = 50.73x + 104.06$

$200 = 50.73x + 104.06$

$-104.06 \qquad -104.06$

$95.4 = 50.73x$

$x = 1.88$ → $\boxed{\$1,880}$

Example 5: A collection of a set of data (x) has a mean 12 with a standard deviation of 1.3. Another variable (y) has a mean of 30 with a standard deviation of 4. The correlation coefficient is 0.91. Find the equation of the linear regression line.

$$\boxed{a = r\frac{s_y}{s_x} \qquad b = \bar{y} - a\bar{x}}$$

$\bar{x} = 12$, $S_x = 1.3$ | $\bar{y} = 30$, $S_y = 4$ | $r = .91$

$a = (.91)\frac{4}{1.3}$ | $b = 30 - 2.8(12)$

$a = 2.8$ | $b = -3.6$

$\hat{y} = ax + b$

$\hat{y} = 2.8x - 3.6$

6.1 White text

7.3 Notes: Confidence Intervals for the Mean (Large Samples)

Objectives:

- **Can you find a point estimate and a margin of error?**
- **Can you construct and interpret confidence intervals for the population mean?**
- **Can you determine the minimum sample size required when estimating μ?**

Essential Idea: Can we estimate the value of the population mean by using sample data?

Point Estimate → but based on sample

- A single value estimate for a population parameter
- Most unbiased point estimate of the population mean μ is the Sample Mean $\bar{x}$.

Example 1: Market researchers use the number of sentences per advertisement as a measure of readability for magazine advertisements. The following represents a random sample of the number of sentences found in 50 advertisements. Find a point estimate of the population mean, μ. *(Source: Journal of Advertising Research)*

9	20	18	16	9	9	11	13	22	16	5	18	6	6	5	12	25
17	23	7	10	9	10	10	5	11	18	18	9	9	17	13	11	7
14	6	11	12	11	6	12	14	11	9	18	12	12	17	11	20	

$\bar{X} = \frac{\Sigma x}{n} = \frac{620}{50} = 12.4$

OR

STAT edit...

The point estimate for the mean length of all magazine advertisements is 12.4 sentences.

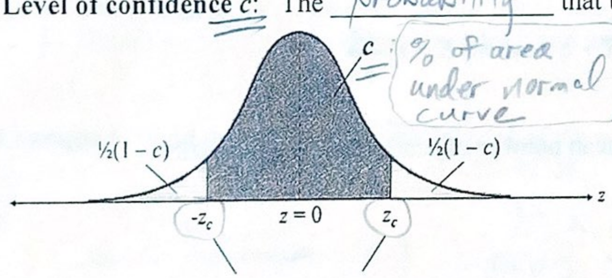
→ not possible (decimal)

What would the probability be that the population mean is **exactly the same** as the point estimate? 0.

10 11 12 13 14 15 → Therefore, we will estimate that μ lies in an interval. something like [12, 13] [11, 14] [10, 15]

How confident do we want to be that the interval estimate contains the population mean μ? 90% 95% 99%

Level of confidence c: The probability that the interval estimate contains the population parameter.



[When $n \geq 30$, sampling distr. of sample means is normal distr.]

Z = -1.64 } 2 dec.

Example 2: Find the critical values z_c for the following confidence levels.

a) $c = 90\%$

c = .90

$\frac{1}{2}(1-.9) =$

.05 .05

2nd VARS: invNorm(.05, 0, 1) area

Call it $\boxed{Z = -1.645}$

$-Z_c = -1.645$; $Z_c = 1.645$

b) $c = 94\%$

.94

.03 .03

invNorm(.03, 0, 1) = -1.881

$-Z_c = -1.881$

$Z_c = 1.881$

c) $c = 97\%$

.97

.015 .015

invNorm(.015, 0, 1) = -2.170

$-Z_c = -2.170$

$Z_c = 2.170$

Sampling error: The difference between the point estimate and the actual population mean.

sample mean $\bar{x}$ $\qquad \mu$

hard to find!
- → $\boxed{\bar{x}-\mu}$
- → (varies from sample to sample)
  each sample has diff. $\bar{x}$
- → μ is usually unknown

can find max. value for error

Margin of error
- The greatest possible distance between the point estimate and the population mean for a given level of confidence, c.
- Denoted by E.

$$E = z_c \sigma_{\bar{x}} = z_c \frac{\sigma}{\sqrt{n}}$$

→ When $n \geq 30$, the sample standard deviation, s, can be used for σ.

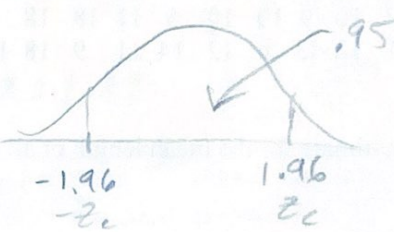
Example 3: Use the magazine advertisement data and a 95% confidence level to find the margin of error for the mean number of sentences in all magazine advertisements. Assume the sample standard deviation is about 5.0. (Recall that $\bar{x} = 12.4$.) → sent.

→ z-score = 1.96 $\quad n = 50 \geq 30$ ✓ normal

$E = z_c \cdot \frac{\sigma}{\sqrt{n}}$ → 5

$E = 1.96 \cdot \frac{5.0}{\sqrt{50}}$

$E \approx 1.4$



You are 95% confident that margin of error is ~1.4 sentences

A c-confidence interval for the population mean μ:

The ______________ that the confidence interval contains μ is c.

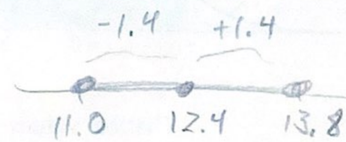
c-confidence interval: $\bar{x} - E < \mu < \bar{x} + E$ where $\boxed{E = z_c \frac{\sigma}{\sqrt{n}}}$

Example 4: Construct a 95% confidence interval for the mean number of sentences in all magazine advertisements. See #3 for E.

$\bar{x} - E < \mu < \bar{x} + E$

$12.4 - 1.4 < \mu < 12.4 + 1.4$

$11.0 < \mu < 13.8$



With 95% confidence, you can say that the population mean μ number of sentences is between 11 and 13.8.

↑ unknown

Example 5: A college admissions director wishes to estimate the mean age of all students currently enrolled. In a random sample of 20 students, the mean age is found to be 22.9 years. From past studies, the standard deviation is known to be 1.5 years, and the population is normally distributed. Construct a 90% confidence interval of the population mean age.

$n = 20$

$\bar{x} = 22.9$

$\sigma = 1.5$

$E = z_c\left(\frac{\sigma}{\sqrt{n}}\right)$

$E = \pm 1.64\left(\frac{1.5}{\sqrt{20}}\right)$

$E = 0.5500 \approx 0.6$

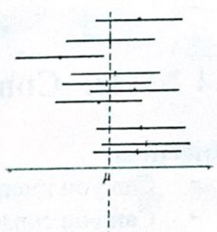
$22.9 - 0.6 < \mu < 22.9 + .6$

$22.3 < \mu < 23.5$

***When rounding, round off to the same number of decimal places given for the sample mean.* **

Interpreting Results for Confidence Intervals:

- μ is a fixed number. It is either in the confidence interval or not.
- Incorrect: "There is a 90% probability that the actual mean is in the interval (22.3, 23.5)."
- Correct: "If a large number of samples is collected and a confidence interval is created for each sample, approximately 90% of these intervals will contain μ.



Using a graphing calculator to find a confidence interval:

1. STAT, EDIT (put in the list if actual data is given)
2. STAT: TESTS
3. 7: Z-Interval
4. Select Data (if you have entered the actual data) OR select Stats if you entered descriptive statistics.
5. Enter the appropriate values (if needed), and select CALCULATE.

Example 6: Find the 95% confidence interval of the population mean from the following sample:

9 20 18 16 9 9 11 13 22 16 5 18 6 6 5

1st (S_x from calc 5.77, $\bar{x} = 12.2$) one-var STAT

(9.3, 15.1) Then

Example 7: A college admissions director wishes to estimate the mean age of all students currently enrolled. In a random sample of 20 students, the mean age is found to be 22.9 years. From past studies, the standard deviation is known to be 1.5 years, and the population is normally distributed. Construct a 90% confidence interval of the population mean age.

(22.3, 23.5)

$n = 20$

$\bar{x} = 22.9$

$\sigma = 1.5$

$c = 0.9$

as conf. c incr. → int. widens → precision decreases
* but can impr. w/ bigger sample

Finding the minimum sample size:

Given a c-confidence level and a margin of error E, the minimum sample size n needed to estimate the population mean μ is

$$n = \left(\frac{z_c \sigma}{E}\right)^2$$

(ALWAYS ROUND up !)

If σ is unknown, you can estimate it using s provided you have a preliminary sample with at least 30 members.

Example 8: You want to estimate the mean number of sentences in a magazine advertisement. How many magazine advertisements must be included in the sample if you want to be 95% confident that the sample mean is within one sentence of the population mean? Assume the sample standard deviation is about 5.0.

$z_c = \pm 1.96$

$E = 1$

$s_x = 5.0$

$$n = \left(\frac{1.96(5.0)}{1}\right)^2 \quad 96.04 \uparrow = 97$$

$\boxed{16.2}$ white text

7.4 Notes: Confidence Intervals of the Mean (Small Samples)

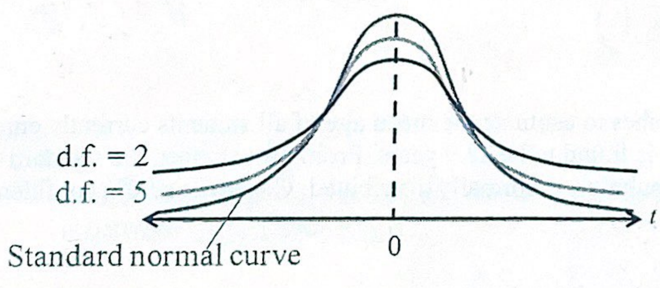
Objectives:

- Can you interpret the t-distribution and use a t-distribution table?
- Can you construct confidence intervals when $n < 30$, the population is normally distributed, and σ is unknown?

The t-distribution: When the population standard deviation is unknown, the sample size is less than 30, and the random variable x is approximately normally distributed, it follows a t-distribution.

Critical values of t are denoted by t_c.

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$



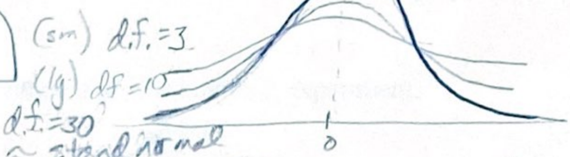
The tails in the t-distribution are "thicker" than those in the standard normal distribution.

p. 325
Wm. Gosset

Properties of t-distributions:

1. The t-distribution is <u>bell-shaped</u> and <u>symmetric</u> about the mean.
2. The t-distribution is a family of curves, each determined by a parameter called the degrees of freedom. The degrees of freedom are the number of <u>free</u> choices left after a sample statistic such as <u>$\bar{x}$</u> is calculated. When you use a t-distribution to estimate a population mean, the degrees of freedom are equal to one less than the sample size.

$$\text{d.f.} = n - 1 \qquad \text{Degrees of freedom}$$



3. The total area under a t-curve is 1 or 100%.
4. The mean, median, and mode of the t-distribution are equal to <u>zero</u>. → true for standard normal curve
5. As the degrees of freedom increase, the t-distribution approaches the normal distribution. After 30 d.f., the t-distribution is very close to the standard normal z-distribution.

Example 1: Find the critical value t_c for a 95% confidence when the sample size is 15. Use Table 5.

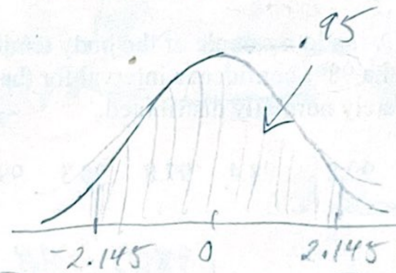
$\rightarrow \boxed{n = 15} \Rightarrow \boxed{d.f. = n - 1 = 14}$

Table 5: t-Distribution

d.f.	Level of confidence, c	0.50	0.80	0.90	0.95	0.98	0.99
	One tail, α	0.25	0.10	0.05	0.025	0.01	0.005
	Two tails, α	0.50	0.20	0.10	0.05	0.02	0.01
1		1.000	3.078	6.314	12.706	31.821	63.657
2		.816	1.886	2.920	4.303	6.965	9.925
3		.765	1.638	2.353	3.182	4.541	5.841
13		.694	1.350	1.771	2.160	2.650	3.012
14		.692	1.345	1.761	2.145	2.624	2.977
15		.691	1.341	1.753	2.131	2.602	2.947
16		.690	1.337	1.746	2.120	2.583	2.921
28		[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
29		.683	1.311	1.699	2.045	2.462	2.756
∞		.674	1.282	1.645	1.960	2.326	2.576

$t_c = 2.145$

Draw a diagram of what this means:



So 95% of area under t-distr. curve for 14 d.f. lies betw. $t = \pm 2.145$

A c-confidence interval for the population mean μ: $\quad \bar{x} - E < \mu < \bar{x} + E \quad$ where $E = t_c \dfrac{s}{\sqrt{n}}$

$E = z_c \dfrac{\sigma}{\sqrt{n}}$ compare

- The ____________________ that the confidence interval contains μ is c.
  (Very similar to constructing a confidence interval using the normal distribution)

$\boxed{n = 16} \Rightarrow \boxed{d.f. = n - 1 = 15}$

Example 2: You randomly select 16 coffee shops and measure the temperature of the coffee sold at each. The sample mean temperature is 162.0°F with a sample standard deviation of 10.0°F. Find the 95% confidence interval for the mean temperature. Assume the temperatures are approximately normally distributed.

$\bar{x} = 162.0 \quad s = 10.0$

✓ Note: Should we use a t-distribution or a normal distribution?

→ $n < 30$, σ is unknown, & approx. normal

Table: $\boxed{t_c = 2.131}$

$E = t_c \dfrac{s}{\sqrt{n}} = 2.131 \cdot \dfrac{10}{\sqrt{16}} \approx 5.3$

$\bar{x} - E < \mu < \bar{x} + E$

$162.0 - 5.3 < \mu < 162.0 + 5.3$

$156.7 < \mu < 167.3$

With 95% confidence, you can say that the mean temperature of coffee sold is between <u>156.7</u> and <u>167.3</u>.

We only have that one TINY table... let's use calc!

Using the graphing calculator to find the confidence interval with a *t*-distribution:

1. STAT, EDIT, enter list (if the actual data is given) ~~→ STAT: calc for x̄, s if needed~~
2. STAT: TESTS
3. 8: TInterval
4. Select Data (if you have entered the original data) OR select Stats if you entered descriptive statistics.
5. Enter the appropriate values

STATS: $n=16$ $\bar{x}=162.0$ $s=10.0$

Example 3: You randomly select 16 coffee shops and measure the temperature of the coffee sold at each. The sample mean temperature is 162.0°F with a sample standard deviation of 10.0°F. Find the 95% confidence interval for the mean temperature. Assume the temperatures are approximately normally distributed.

SKIP ~~1.~~ STAT EDIT

→ 2. STAT → TESTS

3. ↓ 8 TInterval

4. select STATS

5. Enter $\bar{x}$, s, n, $c=.95$ ✓

Interval: (156.67, 167.33) → (156.7, 167.3)

don't need to find d.f. $= n-1$ or t_c :)

DATA:

Example 4: A random sample of the body temperature of 9 adults is taken (in degrees F). The results are below. Find the 98% confidence interval for the population mean body temperature. Assume the temperatures are approximately normally distributed.

→ $n=9$ $\bar{x}=$ $s=$??

99	99.2	98.4	97.8	98.3	99.2	100.1	97.4	98.6

1st ~~STAT → calc~~

calc will find

STAT → CALC to find!

$s = .8$

$\bar{x} = 98.7$

(98.0, 99.3)

Example 5: You randomly select 25 newly constructed houses. The sample mean construction cost is \$181,000 and the population standard deviation is known to be \$28,000. Assuming construction costs are normally distributed, should you use the normal distribution, the *t*-distribution, or neither to construct a 95% confidence interval for the population mean construction cost?

$n=25$

$\bar{x} = 181{,}000$

$\sigma = 28{,}000$

normal distr? t-distr.? Neither?

$n<30$ ✓ σ is unknown ✗ normally distr. ✓

We can use normal distr.