DATA ANALYSIS

Week 10: Modeling Relationships

logistics

- now available: formula spreadsheet! all formulas + links in one place
- please review course website for videos + readings before/after class
- Data Around Us! (worth 2.5% of your grade, canvas discussion board + extra credit for Data Detectives)
- Memer of the Semester (worth 1 extra credit point for best memer)

Prep

Before Wednesday

- Review Pearson correlation [10 min] and Linear regression [7.5 min] videos
- Review W7 slides!
- Read Chapter 16 (Section 16.2) from the Gravetter & Wallnau (2017) textbook.
- Read Chapter 15 (Section 15.4) from the Gravetter & Wallnau (2017) textbook.

Before Friday

- Watch Hypothesis Testing (Pearson correlation) [8 min]
- Watch Hypothesis Testing (Linear regression: t-test) [14 min]

After Friday

- Watch Hypothesis Testing (Linear regression: F-test) [13.5 min]
- Read Chapter 10 from the Gravetter & Wallnau (2017) textbook.
- Watch Hypothesis Testing (two-groups: t-test) [11 min]
- Watch Hypothesis Testing (two-groups: F-test) [12 min]
- Work on Problem Set 5!

Formula Spreadsheet

File Edit View Insert Format Data Tools Extensions Help

	A	B	C	D	E	F
1	What	Population or Sample	Math Notation/Formula	Sheets formula	Online Calculator	Notes
2	mean	Population	$\mu = \frac{\sum X}{N}$	=AVERAGE(data_range)		
3	mean	Sample	$\bar{X} = M = \frac{\sum X}{n}$	=AVERAGE(data_range)		
4	median	Both		=MEDIAN(data_range)		
5	mode	Both		=MODE(data_range)		be careful about multiple modes, Sheets will often just return the one
6	variance	Population	$\sigma^2 = \frac{SS}{N} = \frac{\sum(X-\mu)^2}{N}$	=VAR.P(data_range)		

Component	Points
In-class participation and/or attending office hours	2.5
Discussion board participation	2.5
Total	5

2. Win Data Detective (1 point): At different points in the semester, there will be opportunities to find instances of statistics being used in the world around us, and assessing the quality of these statistics via a Canvas discusson board. The two students who are most active and incisive in this discussion board will receive 1 extra credit point each.
3. Win Memer of the Semester (1 point) : Each week, you will have the opportunity to submit a meme via Canvas, that reflects your experience with the course content of that week. Memes should be *original*, i.e., they should be course-specific and something you have created yourself and not simply found on the internet, although you are allowed to use common images/tropes from popular memes as a starting point. Memes also need to have a specific format, with the title of the learning module at the top of the meme (see Canvas). All memes will be gathered and sent to the class anonymously at the end of the semester for a survey, and the student(s) with the average highest score and the best scoring meme will both receive 1 additional point. **Note**: A student can only receive a maximum of 1 point through this mechanism, even if the same student has the highest average score in the context *and* the best scoring meme.

new office hours

- Prof. Kumar
  - Wednesdays, 2-5 pm (Kanbar 217), with some exceptions (e.g., next week!)
  - Thursdays, 2-4 pm (virtual)
- Yanevith
  - Sundays, 3.30-5 pm (Kanbar 101)
- Whitt
  - Tuesdays, 4.15-5.45 pm (Kanbar 101)

review: null hypothesis significance testing

- step 1: state the hypotheses
- step 2: set criteria for decision
- step 3: collect data
- step 4: make a decision!

review: NHST for z-test

step 1: state the hypotheses

step 2: set criteria for decision

step 3: collect data

step 4: make a decision!

$H_0: \mu = 80$
$H_1: \mu \neq 80$
compute μ and $\sigma_M = \frac{\sigma}{\sqrt{n}}$
for sampling distribution under H_0

$\alpha = .05$
find $z_{critical}$ based on one vs. two tailed test

(1) compute $z_{observed} = \frac{M - \mu}{\sigma_M}$

(2) find p-value for z-score

check whether $z_{observed}$ is beyond $z_{critical}$ and p-value < .05. if so, reject null hypothesis!

review: z-test vs. one-sample t-test

z-tests

- **when**: population mean and standard deviation are known
- **want to compare**: sample mean to population mean

one sample t-test

- **when**: population standard deviation is unknown
- **want to compare**: sample mean to population mean

review: NHST for z-test

step 1: state the hypotheses

$H_0: \mu = 80$
$H_1: \mu \neq 80$
compute μ and $\sigma_M = \frac{\sigma}{\sqrt{n}}$ for sampling distribution under H_0

step 2: set criteria for decision

$\alpha = .05$
find $z_{critical}$ based on one vs. two tailed test

step 3: collect data

(1) compute $z_{observed} = \frac{M - \mu}{\sigma_M}$

(2) find p-value for z-score

step 4: make a decision!

check whether $z_{observed}$ is beyond $z_{critical}$ and p-value < .05. if so, reject null hypothesis!

review: NHST for one sample t-test

step 1: state the hypotheses

$H_0: \mu = 80$
$H_1: \mu \neq 80$
compute μ for sampling distribution of means under H_0

step 2: set criteria for decision

$\alpha = .05$
find $t_{critical}$ based on one vs. two tailed test and degrees of freedom = $n - 1$

step 3: collect data

(1) compute and $s_M = \frac{s}{\sqrt{n}}$ for sampling distribution under H_0

(2) compute $t_{observed} = \frac{M - \mu}{s_M}$

(3) find p-value for t-score

step 4: make a decision!

check whether $t_{observed}$ is beyond $t_{critical}$ and p-value < .05. if so, reject null hypothesis!

our models so far...

only one dependent variable (Y), no independent variable: means

- population parameter: μ
- sample statistic: M
- sampling distribution: normal or t
- hypothesis test: z or t

only one dependent variable (Y), one independent variable (X): correlations/slopes

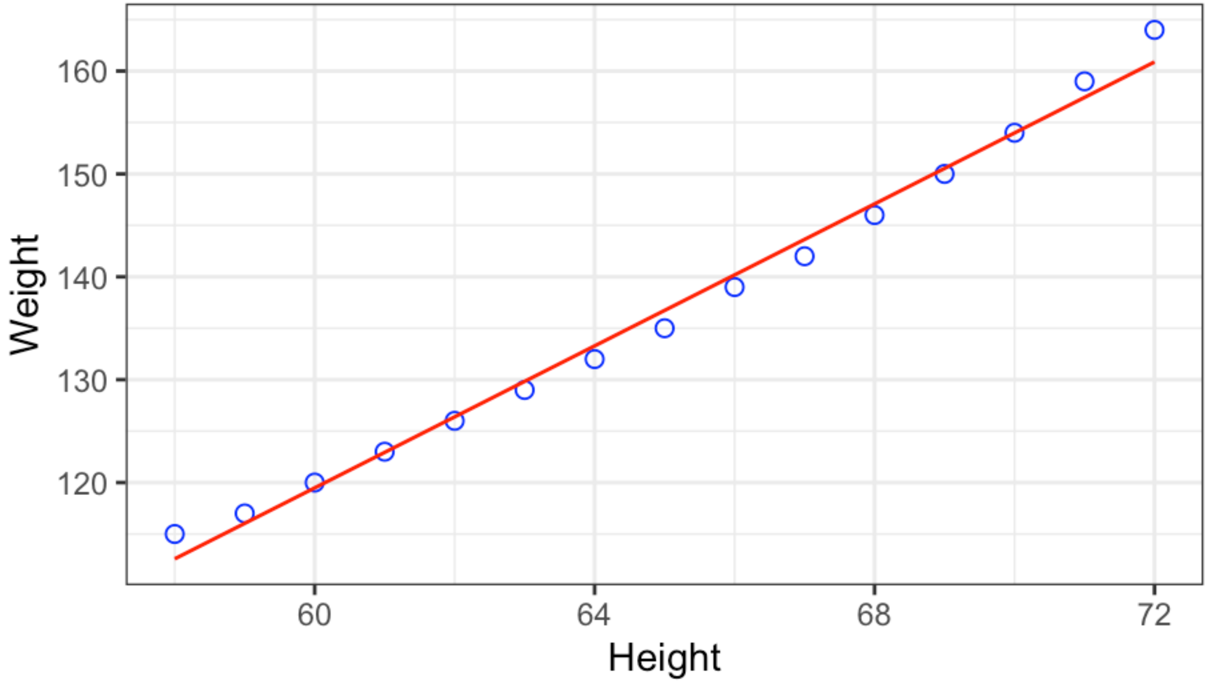
- population parameter: ρ or β
- sample statistic: r or b
- sampling distribution: ???
- hypothesis test: ???

today's agenda

hypothesis testing with one independent variable

review: linear regression

- linear regression attempts to find the equation of a line that best fits the data, i.e., a line that could explain the variation in one variable using the other variable
- Y = bX + a + error
- slope: $b = r \frac{s_y}{s_x}$
- intercept: $a = M_y - bM_x$

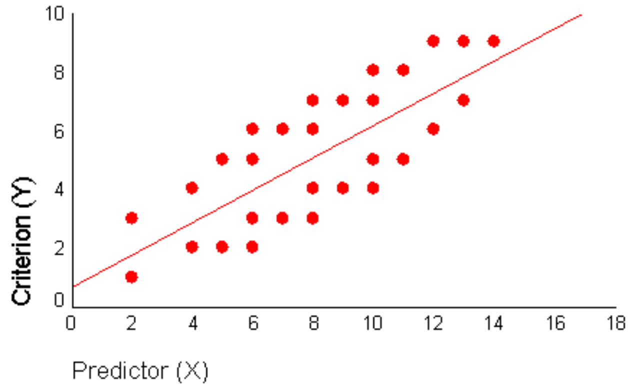


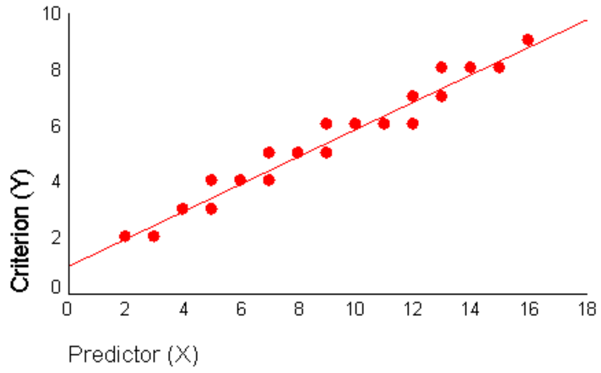
how good is the line of best fit?

- data = model + error
- data = (a + bX) + error
- Y = $\hat{Y}$ + error
- our favorite friend: sum of squared errors (SS)!

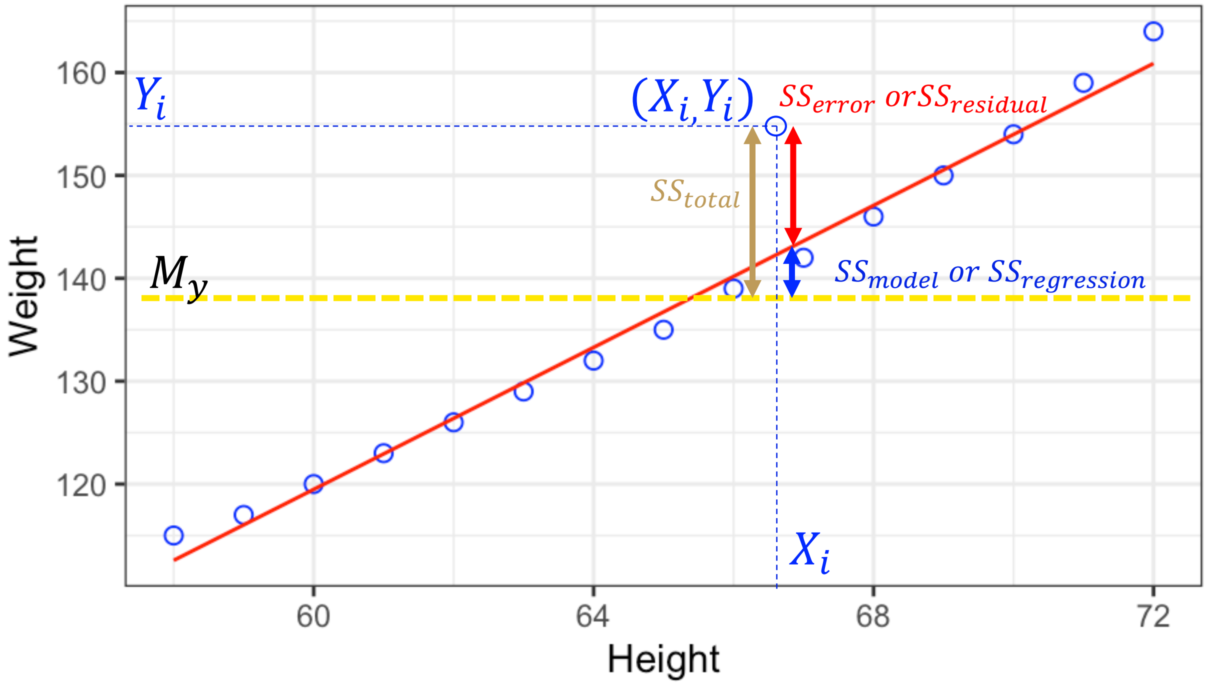
$$\hat{Y} = a + bX = \text{predictions}$$

$$SS_{error} = \sum (Y - \hat{Y})^2$$





understanding goodness/errors



$$SS_{total} = SS_{model} + SS_{error}$$

$$SS_{total} = \sum (Y - M_y)^2$$

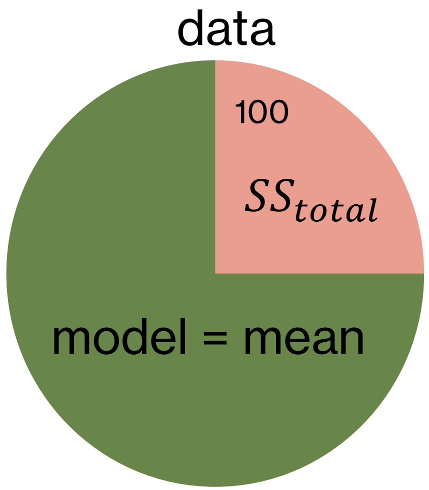
$$SS_{error} = \sum (Y - \hat{Y})^2$$

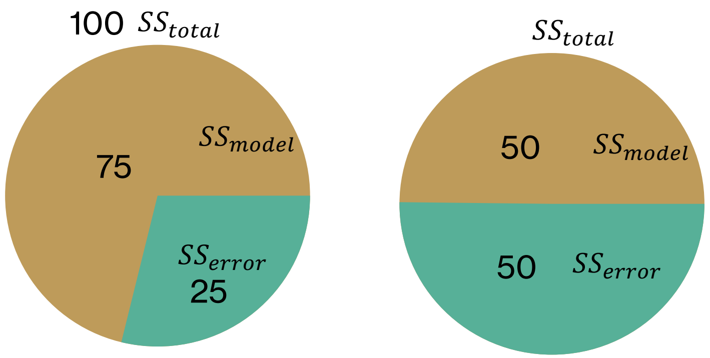
$$SS_{model} = \sum (\hat{Y} - M_y)^2$$

SS_{total} denotes the error left over after the mean has been fit to Y
SS_{error} denotes the error left over after the line Y = a + bX has been fit
SS_{model} denotes the difference, i.e., the error that our line is able to explain vs. what was left over from the mean!

two measures of goodness/errors

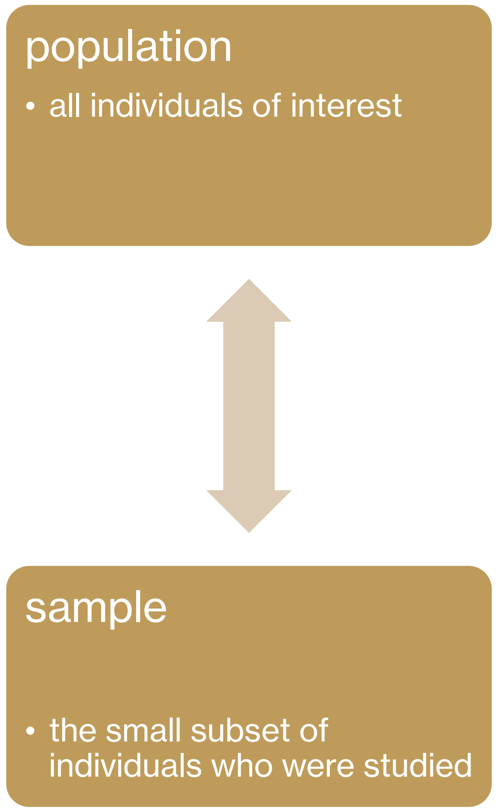
- coefficient of determination (R^2): percentage of variance explained in Y due to X
  - $R^2 = \frac{SS_{model}}{SS_{total}}$
- standard error: "average" error left over in Y
  - standard error of estimate: $SE_{model} = \sqrt{\frac{SS_{error}}{df}} = \sqrt{\frac{SS_{error}}{n-2}}$
  - standard error of correlation: $SE_r = s_r = \sqrt{\frac{1-r^2}{n-2}}$





can we trust our models?

- our goal is to find the best model for our sample of data and generalize to the population
- but how do we know that our sample is representative of the population? how do we know our models are good enough?
- WE ARE HERE!!
- we now have the tools to generalize from samples to populations using NHST!
- we will use SE_{model} and SS_{model} to make inferences



NHST for correlation

step 1:
state the
hypotheses

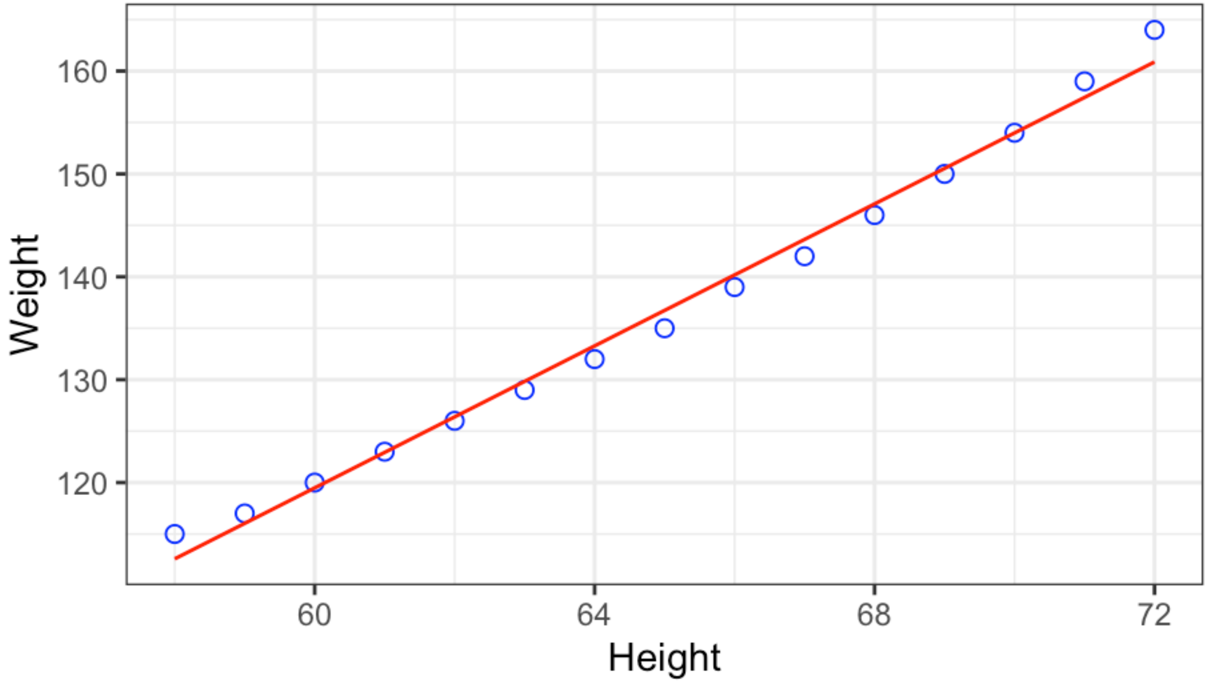
step 2:
set criteria
for decision

step 3:
collect
data

step 4:
make a
decision!

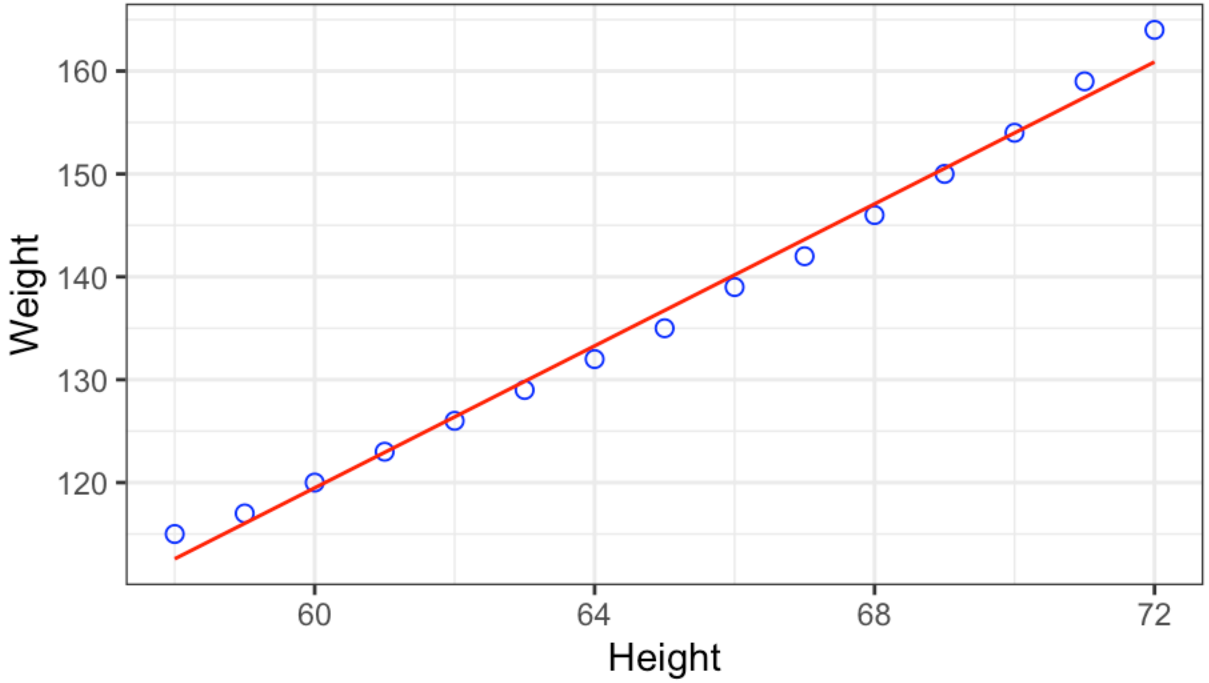
NHST for correlation: step 1a

- stating the hypotheses involves examining the sample statistic is being calculated
  - is it a mean?
  - is it a correlation?
  - is it a slope?
- in this framework, the null hypothesis (H_0) is that the population correlation $\rho = 0$, i.e., there is no relationship between X and Y
  - H_0: $\rho = 0$
  - H_1: $\rho \neq 0$



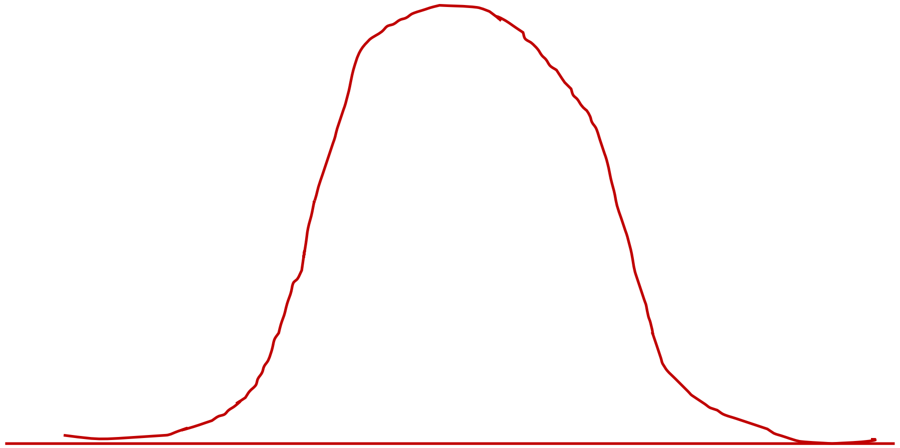
NHST for correlation: step 1b

- if our hypotheses are about correlations, then our sampling distribution should also be for correlations, NOT means
- what is the form of the sampling distribution of slope coefficients?
- the Central Limit Theorem cannot help here (only applies to means!)
- we assume that the sampling distribution of Pearson r follows a t-distribution with n-2 degrees of freedom



NHST for correlation: step 1b

- our sampling distribution of correlations is t-distributed, with n-2 degrees of freedom
- what is the mean of this sampling distribution?
  - $\mu = \rho = 0$ (population correlation)
- what is the standard deviation of this sampling distribution (or standard error)?
  - $SE_r = s_r = \sqrt{\frac{1-r^2}{n-2}}$



NHST for correlation

step 1: state the hypotheses

$H_0: \rho = 0$
$H_1: \rho \neq 0$
compute μ for sampling distribution of correlations under H_0

step 2: set criteria for decision

$\alpha = .05$
find $t_{critical}$ based on one vs. two tailed test and degrees of freedom = $n - 2$

step 3: collect data

(1) compute SE_r for sampling distribution of correlations under H_0

(2) compute $t_{observed} = \frac{r - \rho}{SE_r}$

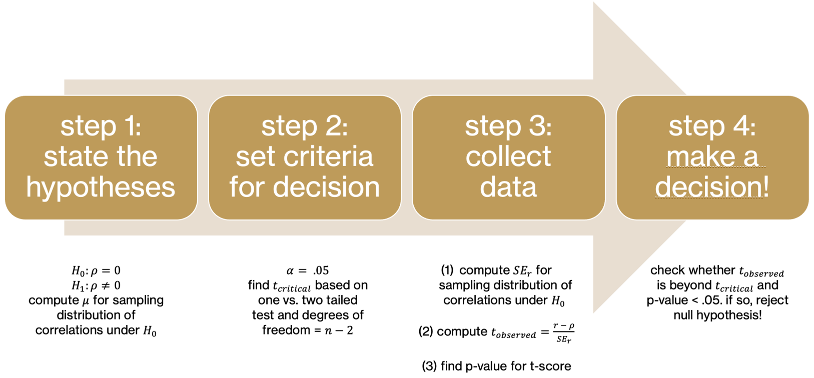
(3) find p-value for t-score

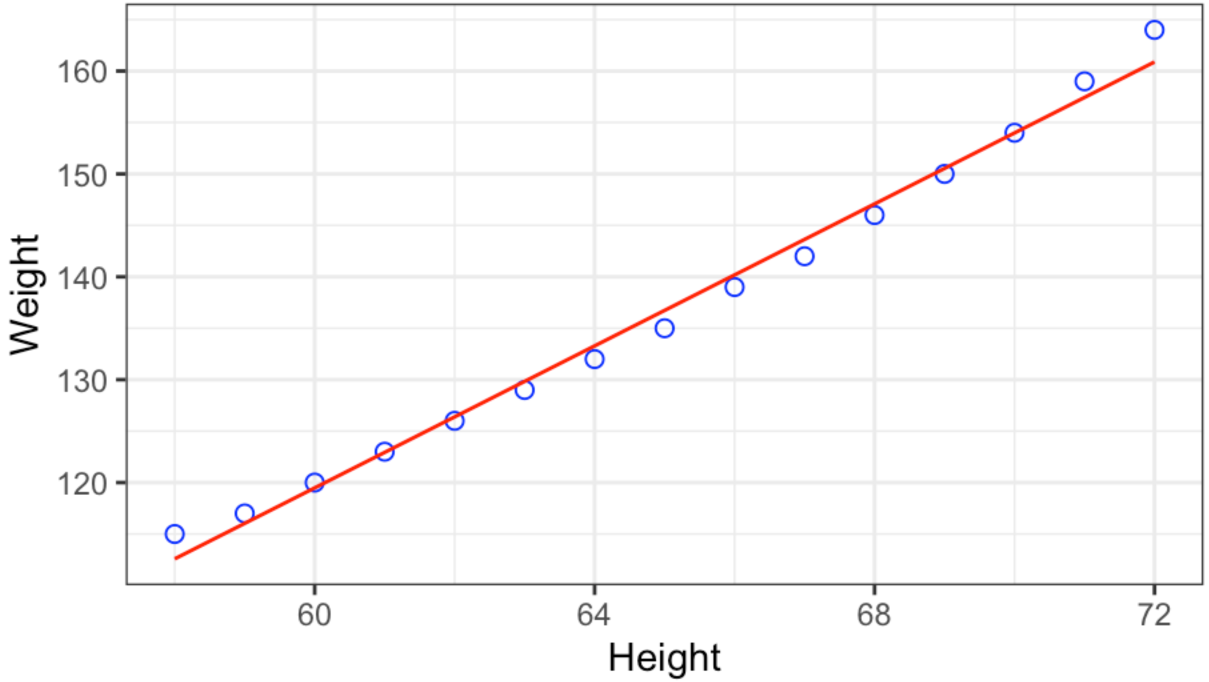
step 4: make a decision!

check whether $t_{observed}$ is beyond $t_{critical}$ and p-value < .05. if so, reject null hypothesis!

activity: NHST for correlation

- [women data]
- compute correlation and perform a hypothesis test!
- use [formula spreadsheet]



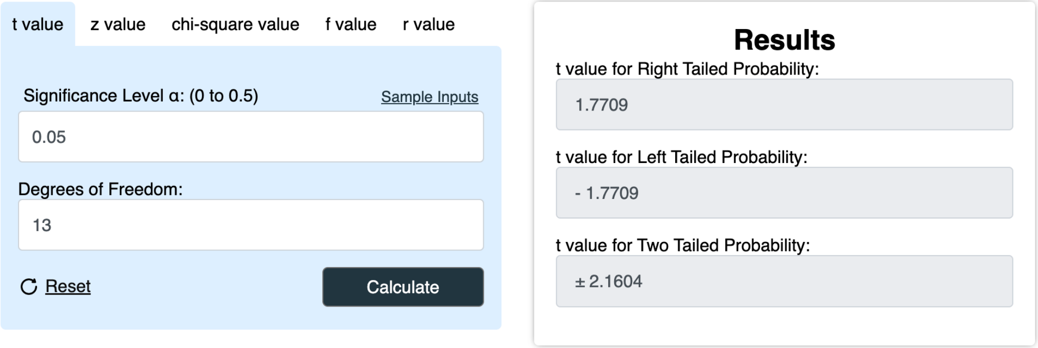


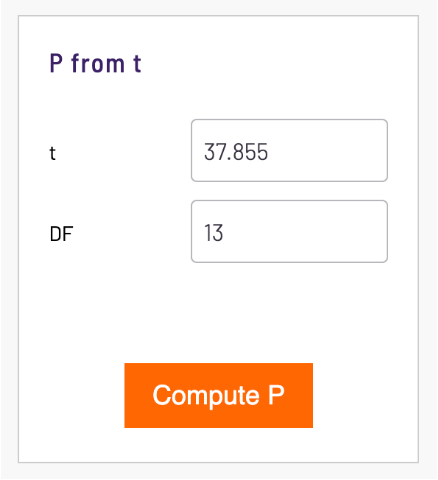
activity: NHST for correlation

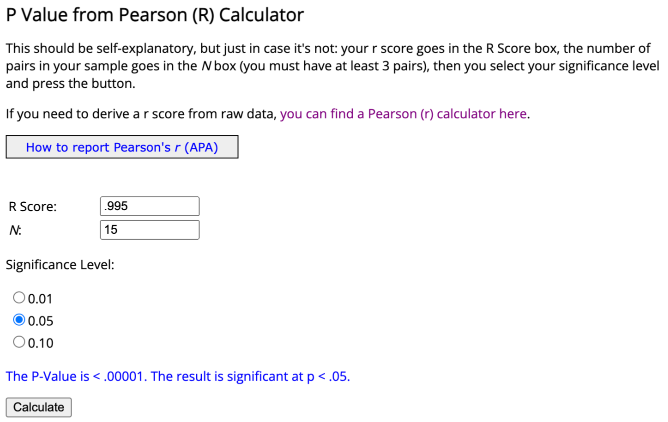
- step 1: state the hypotheses
    - H_0: $\rho = 0$
    - H_1: $\rho \neq 0$
- step 2: set criteria for decision
    - $t_{n-2} = t_{13} = t_{critical} = 2.16 \; at \; \alpha = .05$
- step 3: collect data
    - compute the correlation r = 0.995
    - compute the standard error for correlation

$$SE_r = s_r = \sqrt{\frac{1-r^2}{n-2}} = .026$$

    - compute the t-statistic: $t_{observed} = \frac{r-0}{SE_r} = \frac{.995}{.026} = 37.855$
    - compute p-value: $p_{observed} < .0001$
- step 4: decide!
    - height significantly correlates with weight, r = .995, $t(13)$ = 37.86, p < .001







NHST for linear regression

step 1:
state the
hypotheses

step 2:
set criteria
for decision

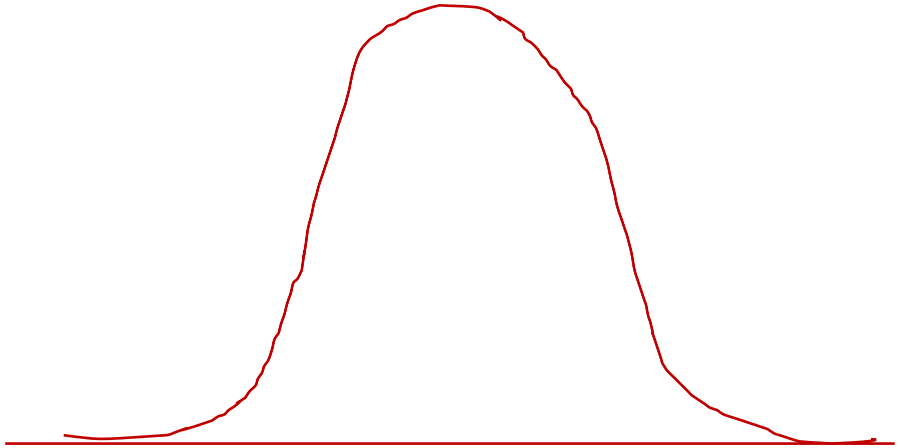
step 3:
collect
data

step 4:
make a
decision!

NHST for linear regression

- step 1: stating the hypothesis
  - H_0: β = 0
  - H_1: β ≠ 0
- assumption: our sampling distribution of slopes is t-distributed, with n-2 degrees of freedom
- what is the mean of this sampling distribution?
- what is the standard deviation of this sampling distribution (or standard error)?
  - $SE_{model} = \sqrt{\frac{SS_{error}}{n-2}}$
  - $SE_b = \frac{SE_{model}}{\sqrt{\sum(X-M_x)^2}}$
  - $SE_a = SE_b \sqrt{\frac{1}{n}\sum X^2}$ (no need to remember/learn, only FYI)

null hypothesis
sampling distribution
of ALL sample slopes



NHST for linear regression (t-test)

step 1: state the hypotheses

$H_0: \beta = 0$
$H_1: \beta \neq 0$
compute μ for sampling distribution of slopes under H_0

step 2: set criteria for decision

$\alpha = .05$
find $t_{critical}$ based on one vs. two tailed test and degrees of freedom = $n - 2$

step 3: collect data

(1) compute SE_b for sampling distribution of slopes under H_0

(2) compute $t_{observed} = \frac{b - \beta}{SE_b}$

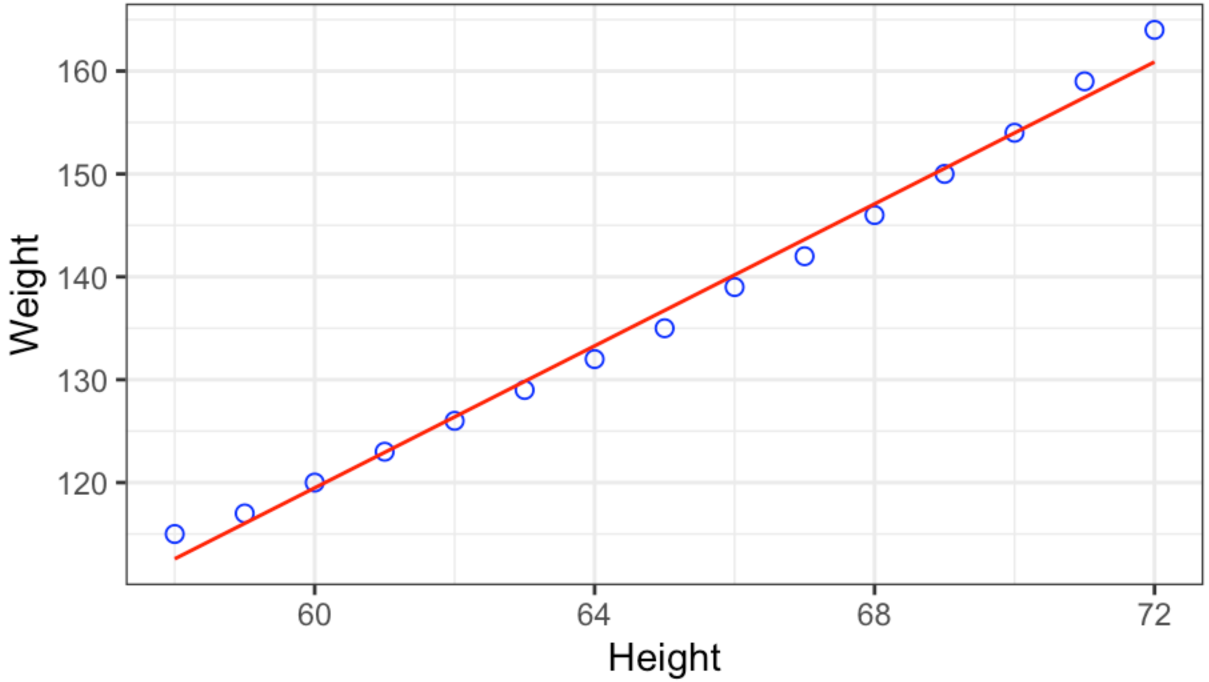
(3) find p-value for t-score

step 4: make a decision!

check whether $t_{observed}$ is beyond $t_{critical}$ and p-value < .05. if so, reject null hypothesis!

activity: women's dataset

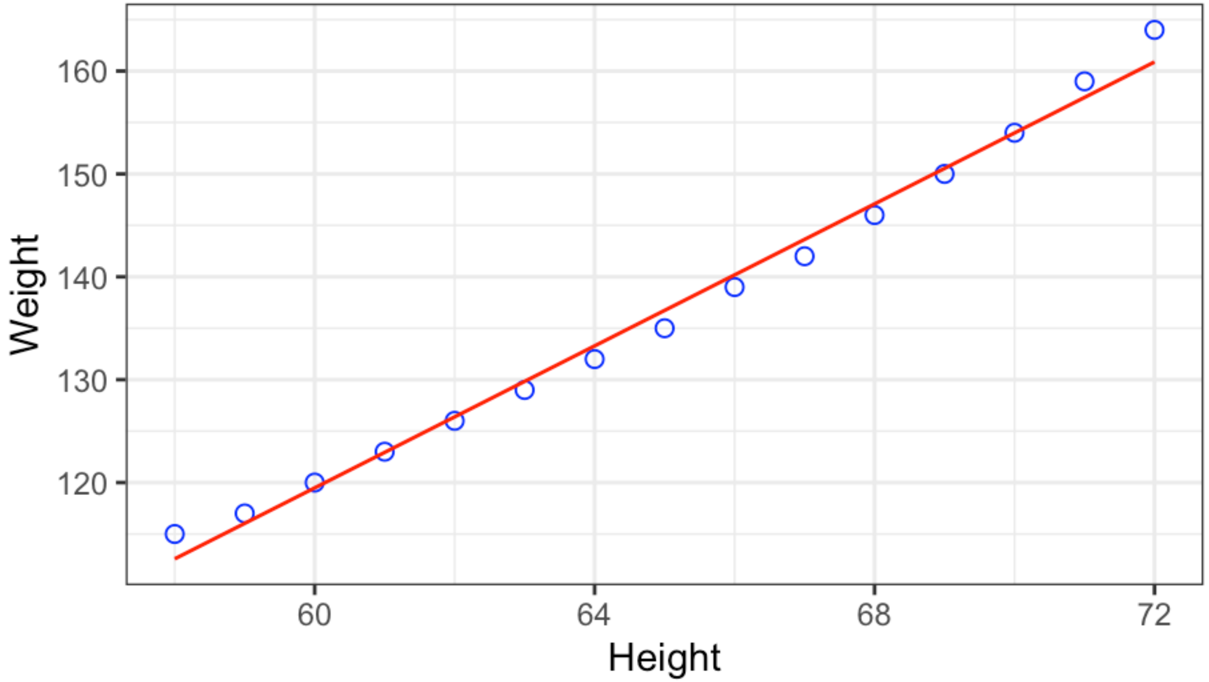
- conduct a two-tailed hypothesis test for the slope from the women's dataset



example: women dataset

- step 1: state the hypotheses
  - H_0: β = 0 (no relationship between height and weight)
  - H_1: β ≠ 0 (non-zero relationship between height and weight)
- step 2: set criteria for decision

$$t_{critical} = t_{n-2} = t_{13} = 2.16 \; at \; \alpha = .05$$

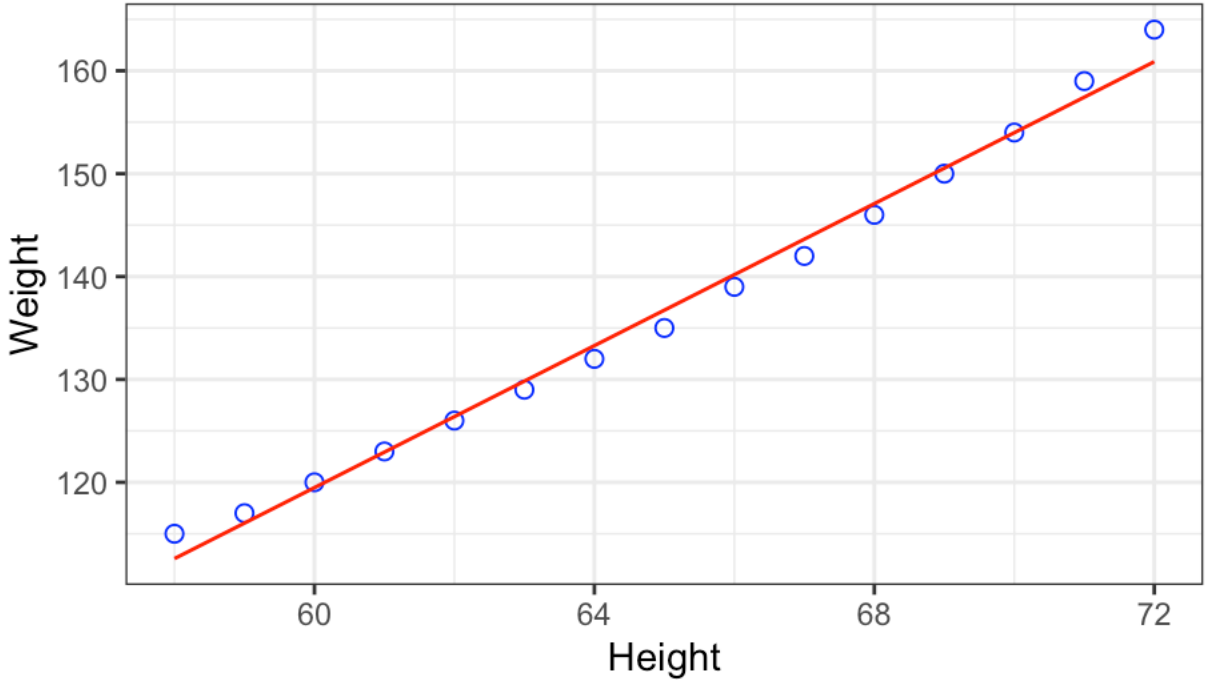


step 3a: women dataset

- collect data
- we compute the slope and intercept
  - $b = r\,\frac{s_y}{s_x}$ = 3.45
  - $a = M_y - bM_x$ = -87.51667
- we compute the standard error for *b*

$SS_{error} = \sum(Y - \hat{Y})^2$ and $SE_{model} = \sqrt{\frac{SS_{error}}{n-2}}$

$SE_b = \frac{SE_{model}}{\sqrt{\sum(X-M_x)^2}}$ = .0911

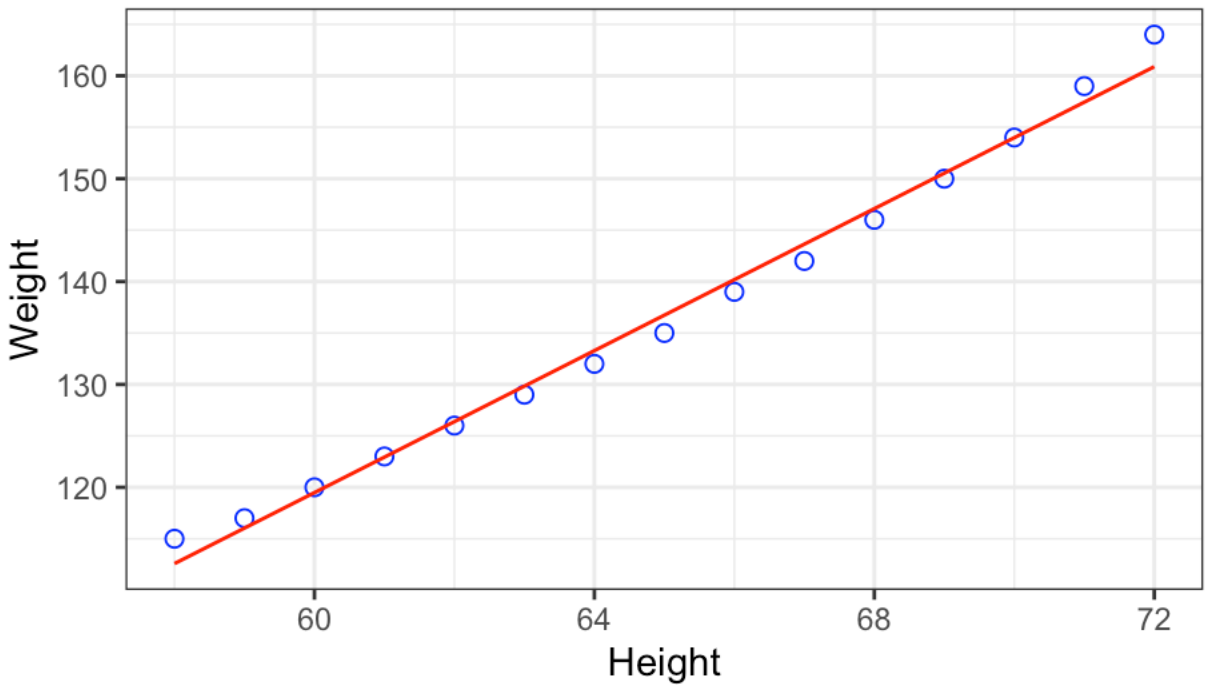


step 3b and 4: women dataset

- step 3: compute $t_{observed}$ and p-value

$$t_{observed} = \frac{b - 0}{SE_b} = \frac{3.45}{.0911} = 37.855$$

  - exactly the <u>same</u> as $t_{observed}$ from correlation because they measure the <u>same</u> thing (a linear relationship)!
  - obtain p-value: $p < .0001$
- step 4: decide!
  - this value is significant, i.e., height significantly predicts weight!



next time

- **before** class
  - *watch*: Hypothesis Testing (Pearson correlation) [8 min]
  - *watch*: Hypothesis Testing (Linear regression: t-test) [14 min]
  - *explore*: PS5 [Chapters 15, 9, and additional problem!]
- **during** class
  - more on regression + two groups

optional content

- any slides after this point are meant for additional exploration
- you will NOT be tested on this content

confidence intervals

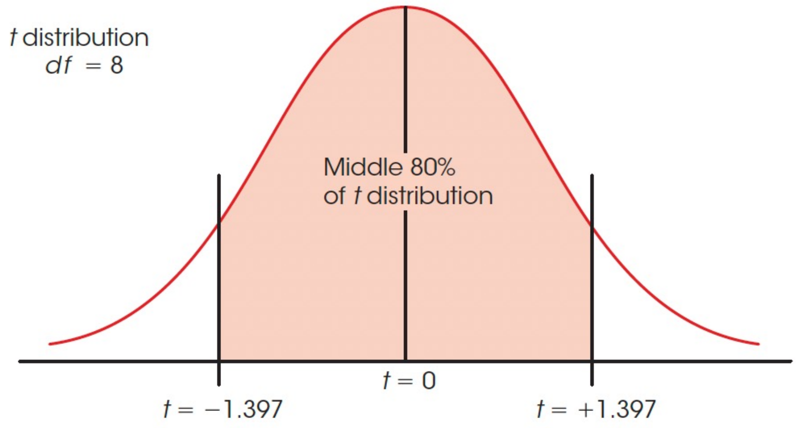
- confidence intervals provide information about the location of the population mean, based on the sample
- given a sample, we can estimate the underlying t-distribution for the sample statistic using the sample size
- we can then estimate what percentage of values will lie within a certain range of t-values
  - for $df = 8, t_{lower} = -1.397 \; and \; t_{upper} = +1.397$
  - 80% confidence interval (80% of the values will lie within this interval)
- confidence intervals specify a range of values within which the population mean will lie, based on the sample mean

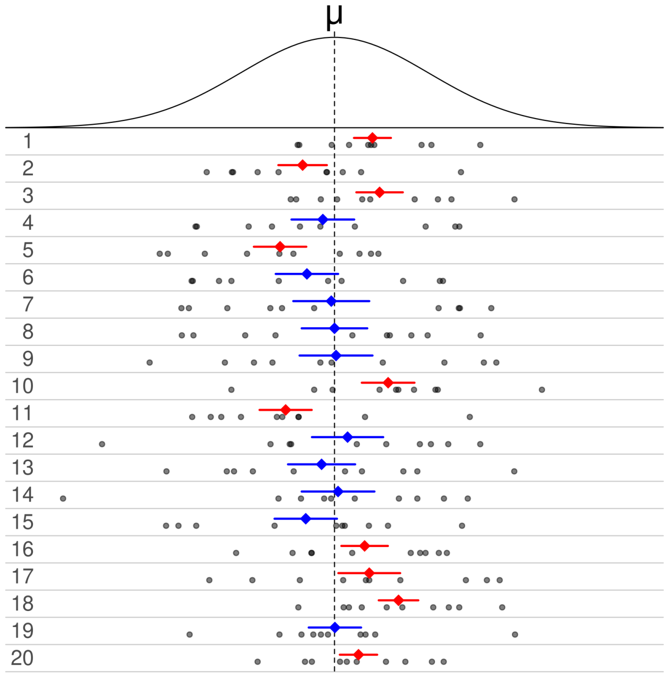
$$t_{lower} = \frac{M - \mu_{lower}}{s_M}, \mu_{lower} = M + t_{lower} s_M$$

$$t_{upper} = \frac{M - \mu_{upper}}{s_M}, \mu_{upper} = M + t_{upper} s_M$$

$$\mu = M \pm t s_M$$

- 80% of the confidence intervals created using samples will contain





example

- research examining the effects of preschool childcare has found that children who spent time in day care, especially high-quality day care, perform better on math and language tests than children who stay home with their mothers (Broberg, Wessels, Lamb, & Hwang, 1997). In a typical study, a researcher obtains a sample of $n = 10$ children who attended day care before starting school. The children are given a standardized math test for which the population mean is $\mu = 50$. The scores for the sample are as follows: 53, 57, 61, 49, 52, 56, 58, 62, 51, 56.
- compute a 95% confidence interval for the population mean for children who attended day care before starting school

frame the problem

- 95% confidence interval means we need the t-value corresponding to the extreme 5%
- $df = n - 1 = 10 - 1 = 9$
- $t_{critical}(9) = \pm 2.2626$
- t-value calculator
- $\mu = M \pm ts_M$
- $\mu = 55.5 \pm 2.26\,(1.34)$
- $CI = [52.46, 58.44]$

CIs: conceptually

- once a confidence interval has been created using a sample, the population mean is either within that interval or not
- the 95% is the long-run probability, i.e., if 100 such confidence intervals were created, 95% would contain the population mean