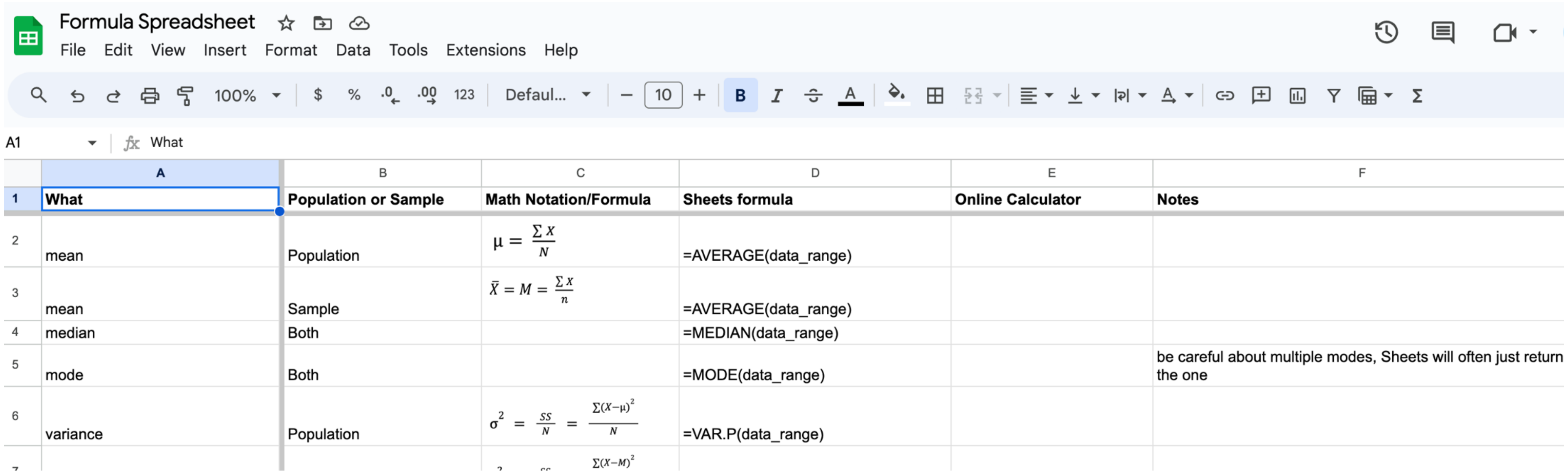
DATA ANALYSIS

Week 10: Modeling Relationships

logistics

- now available: formula spreadsheet! all formulas + links in one place



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

logistics

- midterm 2 is creeping up on us!
- weeks 7-11 content

10	T: March 25, 2025	W10: Modeling Relationships
10	Th: March 27, 2025	W10 continued…
10	Su: March 30, 2025	**Week 10 Quiz due**
11	M: March 31, 2025	**PS4 due / Opt-out Deadline 2**
11	T: April 1, 2025	W11: Special Cases
11	Th: April 3, 2025	W11 continued…
12	M: April 7, 2025	**PS5 + PS4 revision due**
12	T: April 8, 2025	W12: Loose Ends / Exam 2 review
12	Th: April 10, 2025	**Exam (Midterm) 2**

today's agenda

- class survey discussion
- hypothesis testing review
- hypothesis testing for regression

W10 Activity 1: hypothesis testing review

- activity doc
- describe (and/or calculate) the...
    - key variable(s) & research question
    - sample and population
    - sample statistic
    - null & alternative hypothesis
    - sampling distribution
    - critical region, test statistic, p-value
    - statistical significance
    - type I and type II error
    - power
    - effect size

“We surveyed faculty, postdoctoral fellows, and graduate students (N = 1820) from 30 disciplines (12 STEM, 18 SocSci/Hum) (table S1) at geographically diverse high-profile public and private research universities across the United States”

correlation = −0.60 (all fields)
sample size = 30 (all fields)

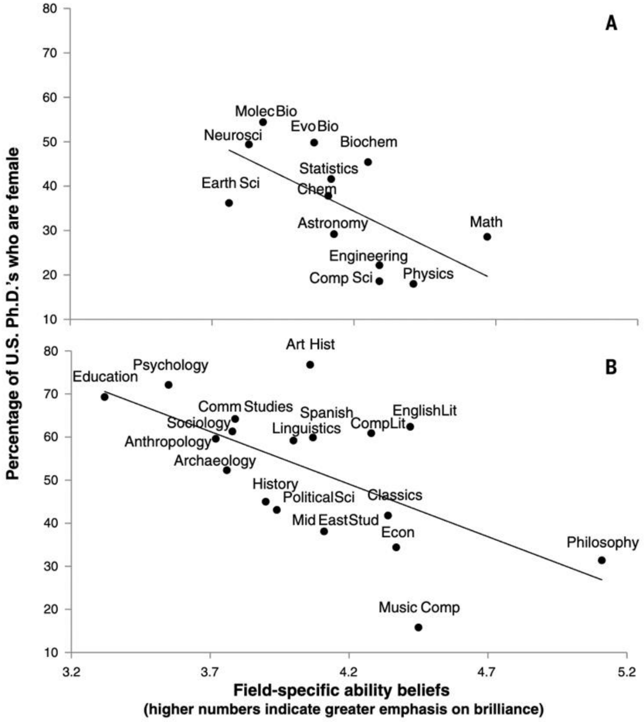


Fig. 1. Field-specific ability beliefs and the percentage of female 2011 U.S. Ph.D.'s in (A) STEM and (B) Social Science and Humanities.

https://doi.org/10.1126/science.1261375

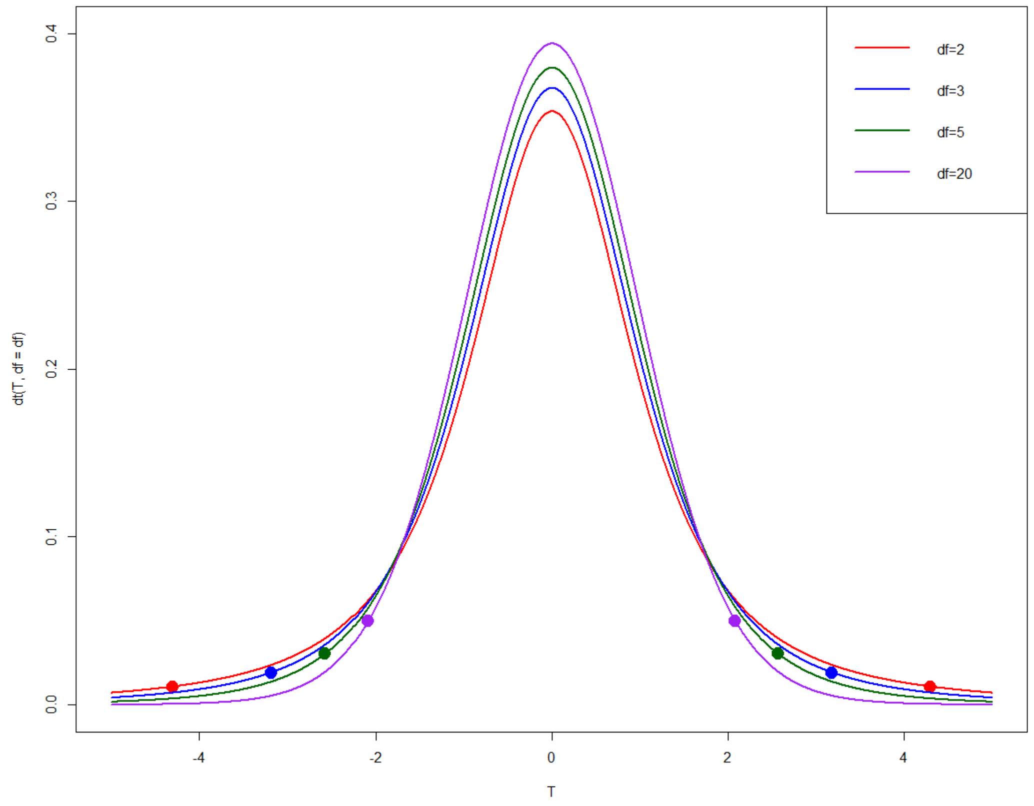
lingering question

As the sample size gets larger, how does the threshold of the correlation value that is **needed** for us to obtain a statistically significant result change?

- There is no consistent relationship between sample size and the critical value for a significant correlation.
- It stays constant
- The correlation threshold gets smaller
- The correlation threshold gets larger

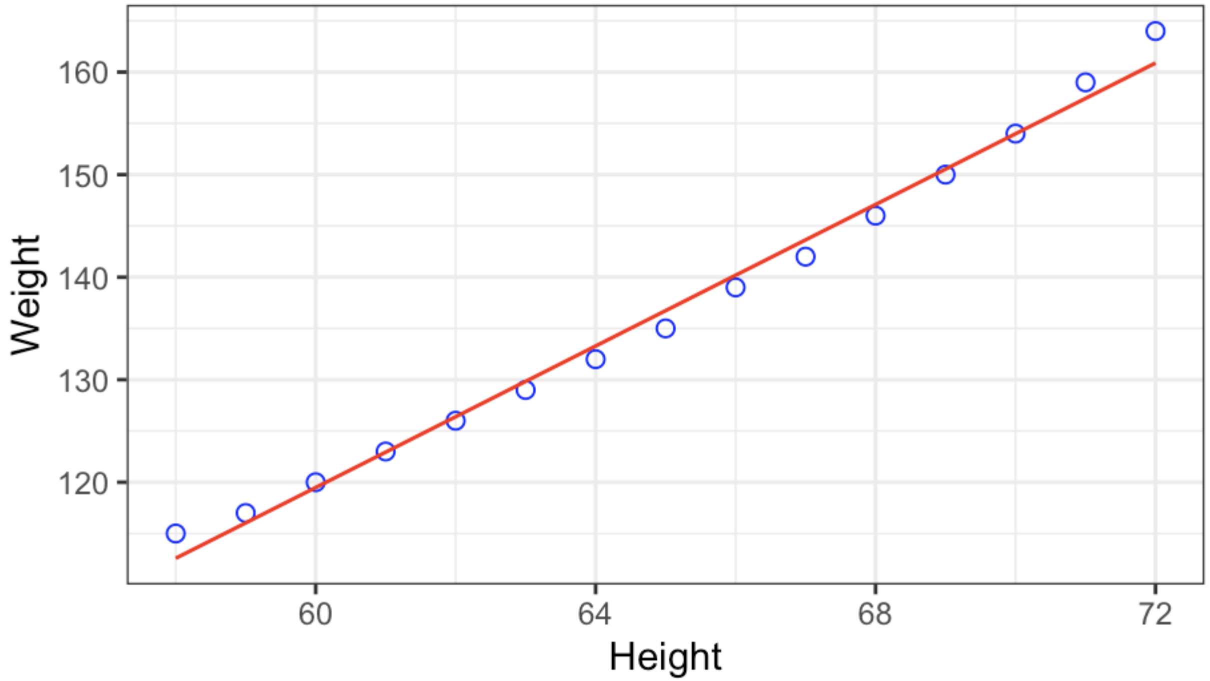
lingering question

- the t-distribution has fatter tails than the normal distribution
- as *n* increases, the t-distribution approaches the normal distribution
- so, with greater sample size, there is less data in the tail
- to get to the critical threshold value, we look at the 5% cut off, but that cutoff has to be moved in as there is less data in the tails!
- i.e., the threshold value <u>decreases</u>



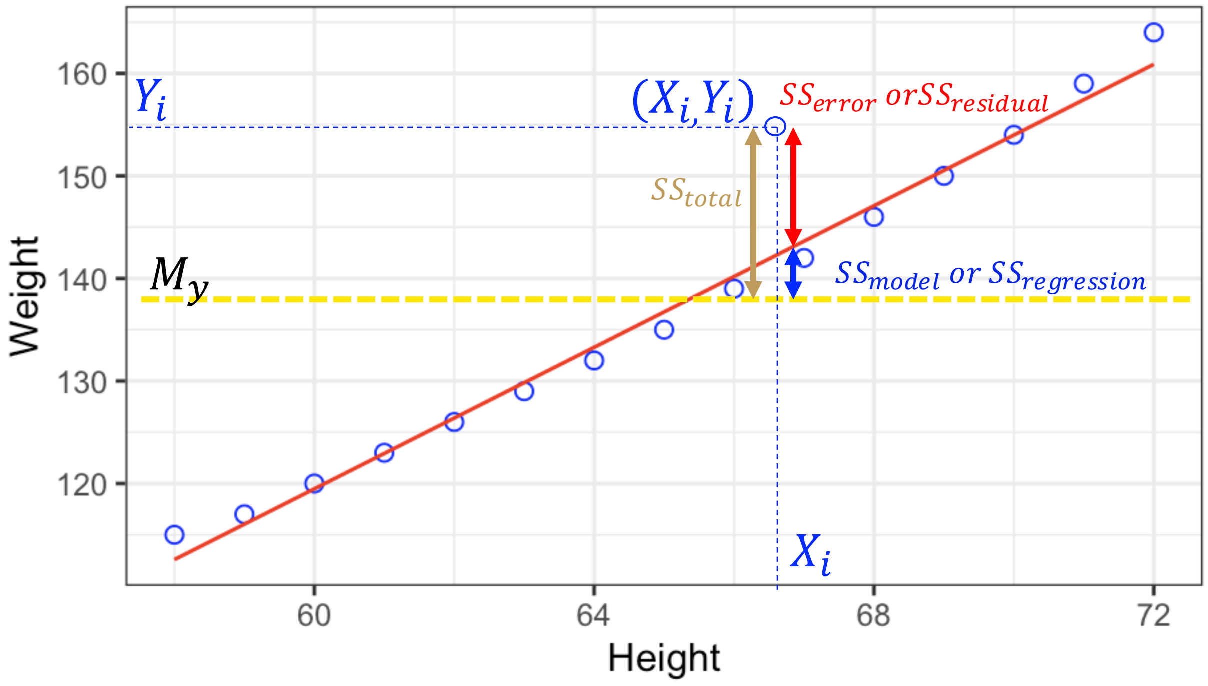
review: linear regression

- 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: $a = M_y - bM_x$



review: model fit

data = model + error
data = (a + bX) + error
Y = $\hat{Y}$ + error



SS_{total} denotes the total error left over after the mean has been fit to Y

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

SS_{error} denotes the error left over after the line $\hat{Y} = a + bX$ has been fit

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

SS_{model} denotes the difference, i.e., the error that our line is able to explain vs. what was left over from the mean!

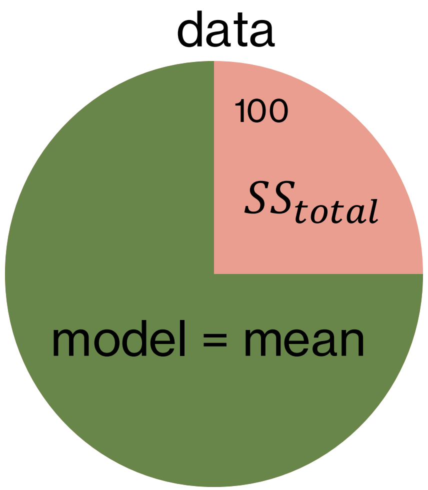
$$SS_{model} = \sum (\hat{Y} - M_y)^2$$

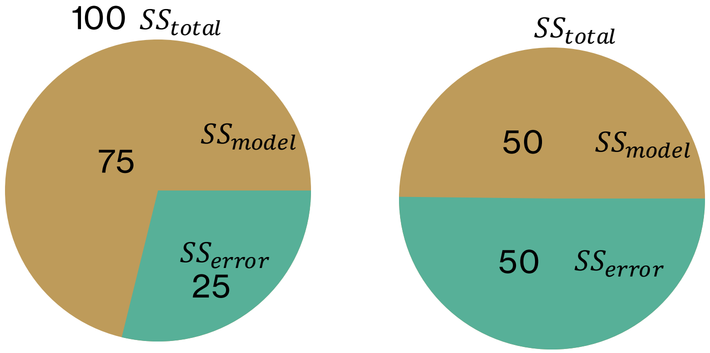
model fit is assessed relative to the mean, i.e., how much better did we do compared to the mean model?

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

two measures of goodness/errors

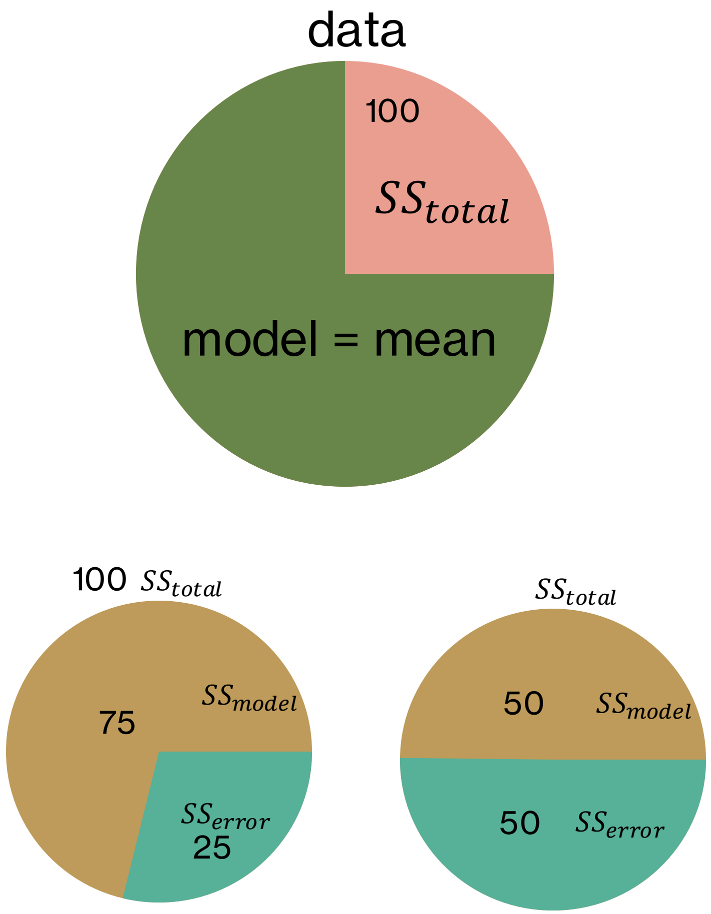
- coefficient of determination (R^2): percentage of variance explained in Y due to X
  - $R^2 = \frac{SS_{model}}{SS_{total}}$
- standard error: "average" error left over in Y
  - standard error of estimate: $SE_{model} = \sqrt{\frac{SS_{error}}{df}} = \sqrt{\frac{SS_{error}}{n-2}}$
  - standard error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



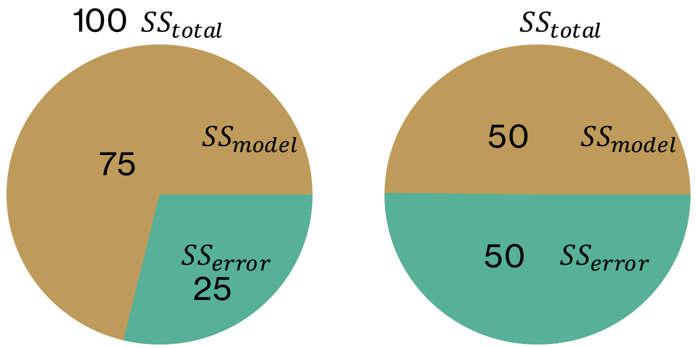


overall test of model: ANOVA

- an **an**alysis **o**f **va**riance (ANOVA) tests whether a variable explains significantly more variance in another variable than chance
    - $SS_{total} = SS_{model} + SS_{error}$
- we can calculate the ratio between the variance explained by the model and the variance expected/left over
    - if $\frac{SS_{model}}{SS_{error}}$ is high, the model explains **more** variance than expected
    - if $\frac{SS_{model}}{SS_{error}}$ is low, the model explains **less** variance than expected
- typically, we want the "average" variance explained, so we also divide this by *degrees of freedom*



F ratio



- The F ratio compares the “average” squared error between model (explained variance) and the natural (unexplained) variance (data = model + error)

$$F = \frac{explained\ variance}{unexplained\ variance} = \frac{MS_{model}}{MS_{error}} = \frac{SS_{model}/df_{model}}{SS_{error}/df_{error}}$$

- obtaining SS_{model} and SS_{error}
  - $SS_{error} = \sum(Y - \hat{Y})^2$ and $SS_{total} = \sum(Y - M_y)^2$
  - $SS_{model} = SS_{total} - SS_{error} = \sum(\hat{Y} - M_y)^2$
- obtaining df_{model} and df_{error}
  - k denotes the number of levels of the independent variable OR number of estimated parameters
  - $df_{model} = k - 1$ (also called df_1 or $df_{numerator}$)
  - $df_{error} = n - k$ (also called df_2 or $df_{denominator}$)

a puzzle

- how many pieces of information do you need to <u>definitely</u> guess the color of the traffic light?
- light is not green
- light is not red
- 2 pieces of information is enough

https://www.youtube.com/watch?v=rATNoxKg1yA

a puzzle

- the mean of quiz scores for 5 students is 9 points.
- what are the scores?
- what if I told you some of the numbers?
- four students’ scores are 8, 10, 8, and 9, what is the score of the fifth student?

degrees of freedom (df)

- main idea: how many pieces of information are needed to obtain a statistic?
- mean = $M = \frac{\sum X}{n}$
  - all values in a dataset are needed
  - why? because changing even a single score would change M
  - $df \ = n$
- standard deviation = $\frac{\sum(X-M)^2}{n-1}$
  - computing M restricts the scores that went into the calculation
  - if M is known, you only need to know n-1 scores to find the last score
  - only $n-1$ scores are free to vary once M is known
  - for SD, effectively only $n-1$ deviations are free to vary
  - $df = n-1$

degrees of freedom (df)

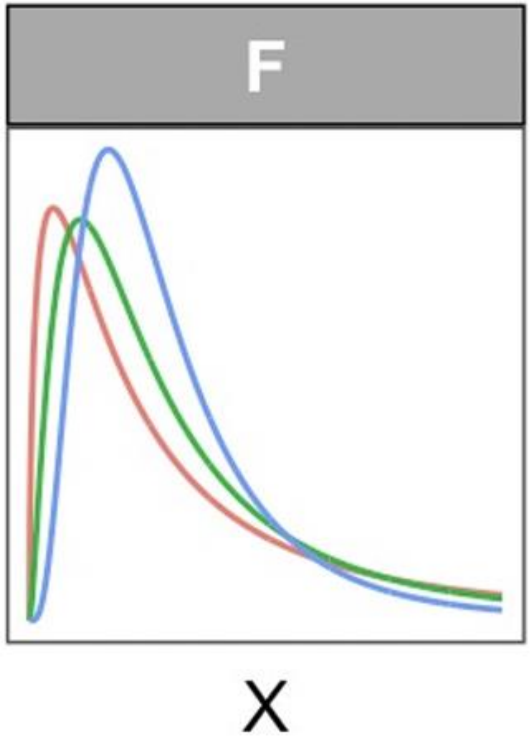
- correlations
  - what is needed to calculate $t_{observed} = \frac{r-\rho}{SE_r}$?
  - *r*, which need two means to be estimated (everything else follows)
  - $df = n - 2$ for t-distribution of correlations
- another way to think about df : number of estimated parameters

degrees of freedom (df)

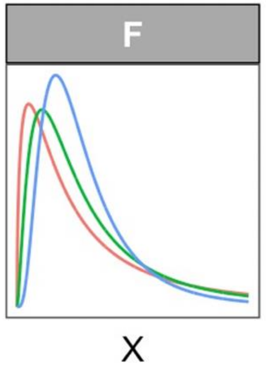
- simple linear regression ($\hat{Y} = a + bX$ where $b = r\,\frac{s_y}{s_x}$ and $a = M_y - bM_x$)
  - $F = \frac{MS_{model}}{MS_{error}} = \frac{SS_{model}/df_{model}}{SS_{error}/df_{error}}$
  - $SS_{model} = \sum(\hat{Y} - M_y)^2$
    - $k = 2$ total estimated parameters (b and a)
    - but knowing b restricts a so we lose one degree of freedom
    - $df_{model} = k - 1$
  - $SS_{error} = \sum(Y - \hat{Y})^2$
    - n observations and 2 total estimated parameters to compute $\hat{Y}$ (b and a)
    - $df_{error} = n - k$

interpreting F values

- The F-distribution is a positively skewed distribution
- defined by two parameters (df_1 and df_2) that determine the exact form/shape
- F-values are typically **<u>non-negative</u>**: why??
  - $F = \frac{MS_{model}}{MS_{error}} = \frac{SS_{model}/df_{model}}{SS_{error}/df_{error}}$
  - $F = 1$: $MS_{model} = MS_{error}$ i.e., the model does not do any better than random chance
  - $F > 1$: more variance explained by model than random chance
- F ratios are enable us to generalize our models to the population (in contrast to R^2 and standard error)



NHST for linear regression (F-test)



step 1: state the hypotheses

$H_0: \beta = 0$
$H_1: \beta \neq 0$

step 2: set criteria for decision

$\alpha = .05$
find $F_{critical}$ based on **right** tailed test and degrees of freedom
$df_1 = k - 1$
$df_2 = n - k$

step 3: collect data

(1) compute SS_{model} and SS_{error}

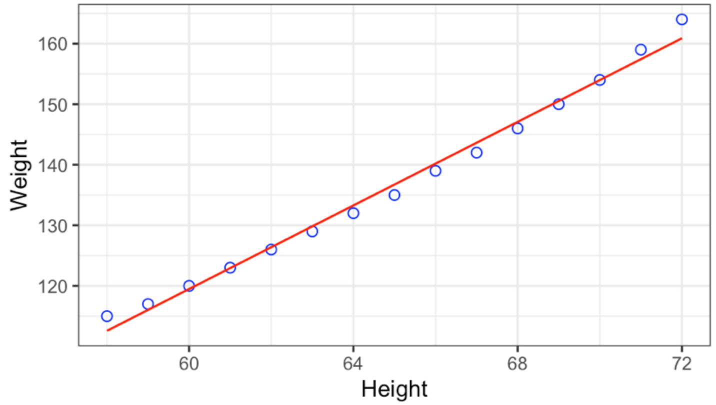
(2) compute $F_{observed} = \frac{MS_{model}}{MS_{error}}$

(3) find p-value for F-score

step 4: make a decision!

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

F-test for women dataset

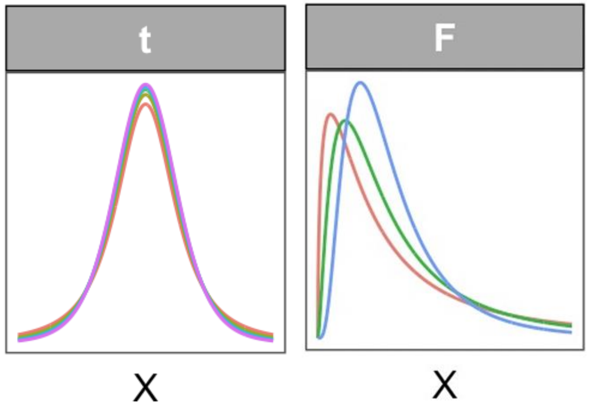


- **step 1: state the hypotheses**
  - H_0: $\beta = 0$, height explains no variance in weights for women
  - H_1: $\beta \neq 0$, height explains some variance in weights for women
- **step 2: set criteria for decision**
  - $\alpha = .05, k = 2, n = 15$
  - $F_{critical}$

  $= F\ (k-1, n-k) = \ F\ (1, 13) = 4.667$
- **step 3: collect data**
  - SS_{error} = 30.23 and SS_{total} = 3362.93
  - thus, SS_{model} = SS_{total} $-$ SS_{error} = 3332.7
  - compute the F-statistic:
  - $F_{observed} = \frac{MS_{model}}{MS_{error}} = \frac{SS_{model}/df_{model}}{SS_{error}/df_{error}} = \frac{3332.7/1}{30.23/13} = 1433$
  - compute p-value: $p_{observed} < \ .0001$
- **step 4: decide!**
  - Height explains significantly more variance in weights than expected by chance, $b = 3.45$, $F\ (1, 13) = 1433, p < \ .0001.$

[Sheets solution](#)
[video tutorial](#)

t and F relationship



- regression test for women dataset
  - $F\,(1, 13) = 1433, p < .0001$
- conduct a correlation test for women dataset ($r = .995$, n = 15)
  - $r = .995, t(13) = 37.86, p < .001$
- what is t^2 ?
- **for the same data, $\boldsymbol{t^2 = F}$!!**
- t tests are in original units of the sample statistic, F tests are in squared error units

F-tables

- F-tests are typically represented in tables

		SS	df	MS	F	p-value
SS_{model}	regression	3332.7	1	3332.7	1433.02	<.0001
SS_{error}	residual	30.23	13	2.33		
SS_{total}	total	3362.93	14			

- knowing parts of the F table are sufficient for completing it!

hypothesis tests in R

```r
data("women")
View(women)

weight_model = lm(data = women, weight ~ height)
summary(weight_model)
car::Anova(weight_model)
```

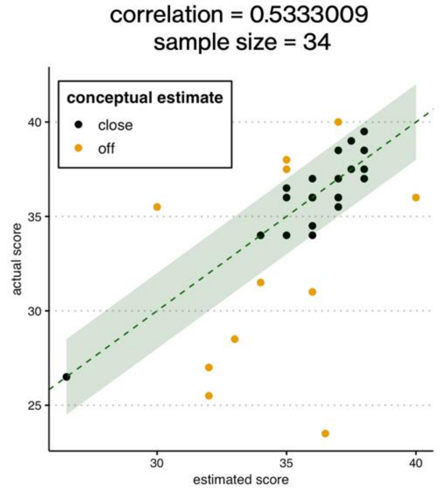
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -87.51667    5.93694  -14.74 1.71e-09 ***
height        3.45000    0.09114   37.85 1.09e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

		SS	df	MS	F	p-value
SS_{model}	IV	3332.7	1	3332.7	1433.02	<.0001
SS_{error}	residual	30.23	13	2.33		

```
Anova Table (Type II tests)

Response: weight
          Sum Sq Df F value    Pr(>F)    
height    3332.7  1    1433 1.091e-14 ***
Residuals   30.2 13                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

review: conceptual exam t-test



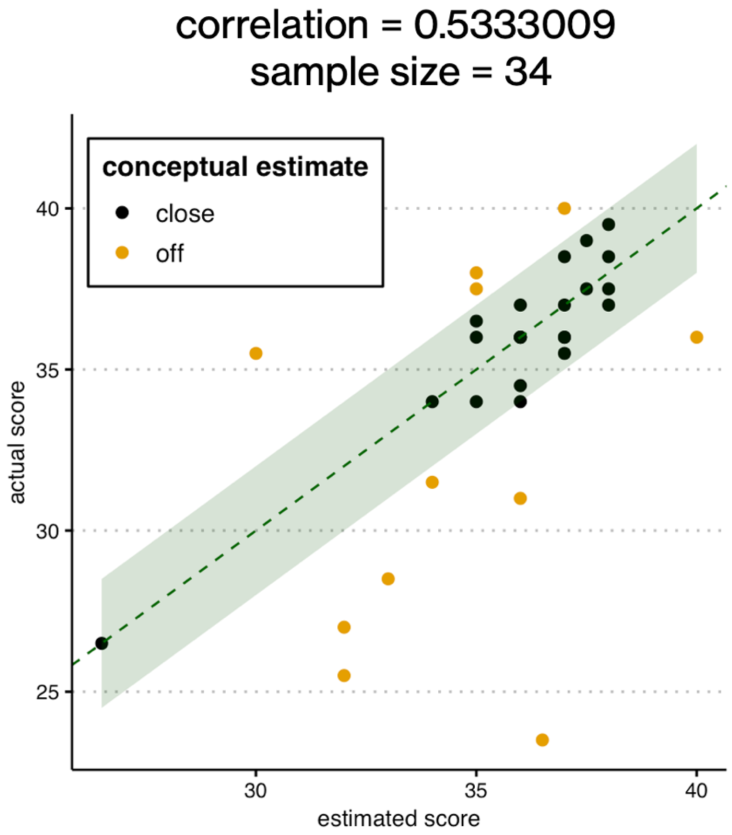
- **step 1: state the hypotheses**
  - H_0: $\rho = 0$, no correlation between estimate and actual score on conceptual exam
  - H_1: $\rho \neq 0$, a correlation between estimate and actual score on conceptual exam
- **step 2: set criteria for decision**
  - $t_{n-2} = t_{32} = t_{critical} = \pm 2.0369 \; at \; \alpha = .05$
- **step 3: collect data**
  - correlation $r = 0.5333009$
  - compute the standard error for correlation

$$SE_r = s_r = \sqrt{\frac{1 - r^2}{n - 2}} = 0.1496$$

  - compute the t-statistic:
  - $t_{observed} = \frac{r-0}{SE_r} = \frac{0.5333009}{0.1496} = 3.57$
  - compute p-value: $p_{observed} = .001$
- **step 4: decide!**
  - estimates significantly correlate with actual scores on the conceptual exam, $r = .53$, $t(32) = 3.57$, $p = .001$

W10 Activity 2: conduct F test

- data



creating F-table

```
> stats_model = lm(data = data_analysis,
+                  m1c_actual ~ m1c_estimate)
> car::Anova(stats_model)
Anova Table (Type II tests)

Response: m1c_actual
             Sum Sq Df F value   Pr(>F)
m1c_estimate 163.78  1  12.718 0.001164 **
Residuals    412.08 32
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- F-tests are typically represented in tables

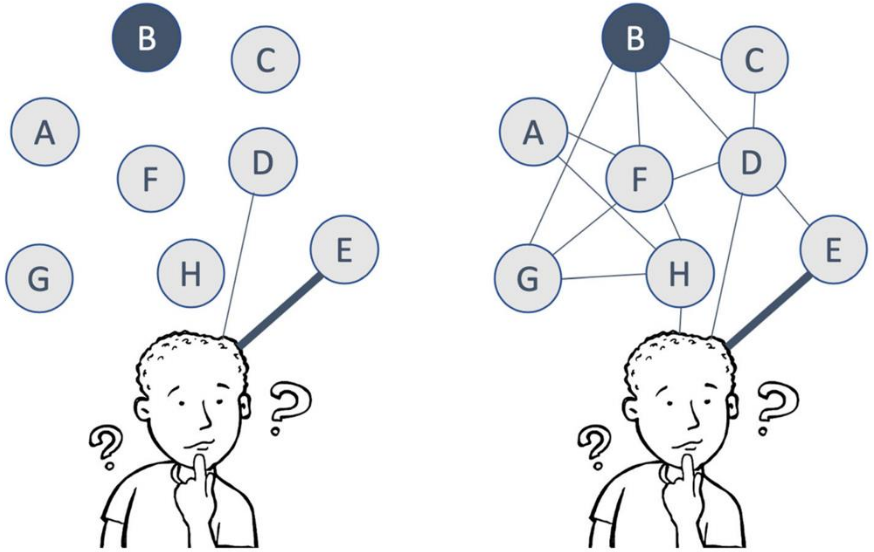
		SS	df	MS	F	p-value
SS_{model}	regression					
SS_{error}	residual					
SS_{total}	total					

practicing connections

- data = model + error
- most sample statistics and statistical tests have the same format
- $\frac{observed}{expected}$
- how are these concepts similar? how are they different?
  - standard deviation
  - z-score
  - t-test
  - F-test

Practicing Connections: A Framework to Guide Instructional Design for Developing Understanding in Complex Domains

Laura Fries[1] • Ji Y. Son[2] • Karen B. Givvin[1] • James W. Stigler[1]



where are we going next?

data = model + error

thus far

- data = mean + error
- data = X (interval/ratio) + error

after break

- data = X (interval/ratio/nominal) + error
- data = X + Y + error
- data (NOIR) = model (NOIR) + error

next time

- new data

Before Tuesday

- Review W7 slides!
- Watch: Hypothesis Testing (Linear Regression). (ok to watch after Tuesday!)

Before Thursday

- Watch: Completing F tables.
- Watch: Hypothesis Testing (Two groups F Test).
- Watch: Hypothesis Testing (One-way ANOVA).

After Thursday

- See Apply section.

Here are the to-do's for this week:

- Submit Week 10 Quiz
- Submit Problem Set 4
- Submit any lingering questions here!
- Extra credit opportunities:
  - Submit Exra Credit Questions
  - Submit Optional Meme Submission