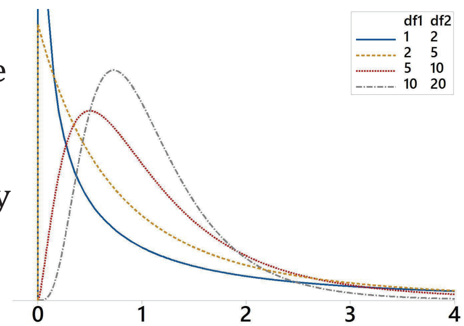
ANOVA Table and Correlation Coefficient

- F-Distribution
- ANOVA Table
- Correlation Coefficient
- Properties of the Correlation Coefficient
- Coefficient of Determination

Lecture 5
Sections 6.1 – 6.5, 7.2

F-Distribution

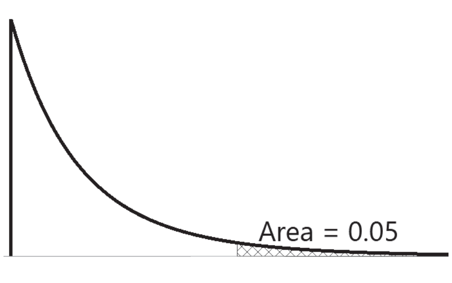
- **F-Distribution:** continuous probability distribution that has the following properties:
  - Unimodal, right-skewed, and non-negative
  - Two parameters for degrees of freedom
    - One for numerator and one for denominator
  - Used to compare two sources of variability
  - To find the critical value, intersect the numerator and denominator degrees of freedom in the F-table (or use Minitab)



- In this course:
  - All tests are upper one-sided
  - Use a 5% level of significance – A different table exists for each α

Example: F-Distribution

- **Question:** What is the critical value for an upper one-sided F-test with 2 and 15 degrees of freedom using $\alpha = .05$?
- **Answer:** ________________
  - Reject H_0 for test statistics ________________________



Denominator Degrees of Freedom	Numerator Degrees of Freedom: 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	161	200	216	225	230	234	237	239	241	242	243	244	245	245	246	246
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74	8.73	8.71	8.70	8.69
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91	5.89	5.87	5.86	5.84
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.70	4.68	4.66	4.64	4.62	4.60
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00	3.98	3.96	3.94	3.92
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57	3.55	3.53	3.51	3.49
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28	3.26	3.24	3.22	3.20
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07	3.05	3.03	3.01	2.99
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91	2.89	2.86	2.85	2.83
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79	2.76	2.74	2.72	2.70
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69	2.66	2.64	2.62	2.60
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60	2.58	2.55	2.53	2.51
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53	2.51	2.48	2.46	2.44
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48	2.45	2.42	2.40	2.38
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42	2.40	2.37	2.35	2.33
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38	2.35	2.33	2.31	2.29
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34	2.31	2.29	2.27	2.25
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31	2.28	2.26	2.23	2.21
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28	2.25	2.22	2.20	2.18
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.20	2.16	2.14	2.11	2.09	2.07
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09	2.06	2.04	2.01	1.99
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	2.07	2.04	2.01	1.99	1.96	1.94
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00	1.97	1.95	1.92	1.90

Types of Variation

- **Explained Variation:** differences in the responses due to the

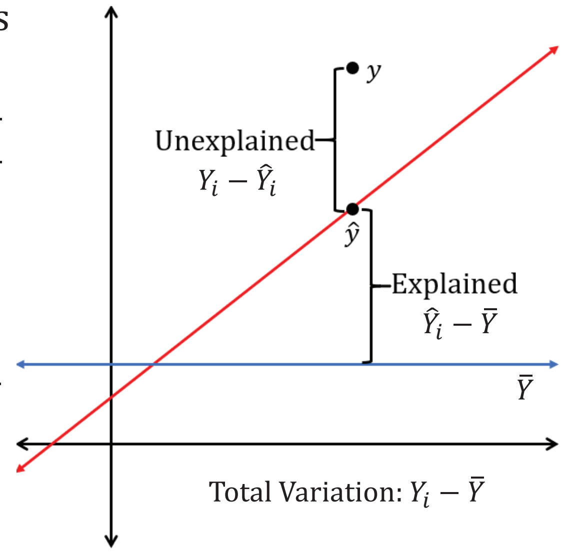
  ____

  ____

  - Sum of squares due to regression (SSR)
- **Unexplained Variation:** differences in the responses due to ______________________________

  ____

  - Sum of squares due to error (SSE)



Sums of Squares

- **Total Sum of Squares:** measures squared distance each response is from the sample mean of the responses
  - Assumes we use $\bar{Y}$ as the naïve prediction for each response instead of considering the relationship Y has with X

$$SSY = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- **Sum of Squares Due to Error:** measures squared distance each response is from its predicted value on the regression line
  - Assumes X is being used to predict Y

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2$$

ANOVA Table for Straight Line Regression

- **Analysis of Variance (ANOVA) Table:** an overall summary of the results of a regression analysis
  - Derived from the fact that the table contains many estimates for sources of variation that can be used to answer three important questions
    1. Is the true slope β_1 ______________?
    2. What is the ___________ of the straight line relationship?
    3. Is the straight line model __________________?

ANOVA Table for Simple Linear Regression

Source	DF	Sum of Squares	Mean Square	F-Statistic
Regression	1	SSR	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Error	$n-2$	SSE	$MSE = \boxed{\frac{SSE}{n-2}}$	
Total	$n-1$	SSY		

$MSE = S^2_{Y|X}$
Square of residual sum of squares

Fundamental Equation of Regression Analysis

$SSY = SSR + SSE$

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$$

Total Unexplained Variation = Regression Variation + Residual Variation

Example: Using the ANOVA Table

- **Scenario:** Use ACT score of 29 college freshmen (without outlier) to describe freshman year GPA.

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	3.459	3.4589	12.50	0.001
Error	27	7.474	0.2768		
Total	28	10.933			

- **Task:** Use the ANOVA table to determine if ACT score is a significant predictor of GPA.
- **Hypotheses:** H_0: __________ vs. H_A: __________
- **Test Statistic:** ____________________
- **Critical Value**: ____________________; **P-Value:** ___________
- **Conclusion:** _______________ and conclude __________

Example: Comparing ANOVA Table and Test for Slope

- **Scenario:** Use ACT score of 29 college freshmen (without outlier) to describe freshman year GPA.

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	3.459	3.4589	12.50	0.001
Error	27	7.474	0.2768		
Total	28	10.933			

Term	Coef	SE Coef	T-Value	P-Value
Constant	0.987	0.570	1.731	0.095
ACT	0.0822	0.0232	3.535	0.001

- **Question:** What is the relationship between the test statistic from the ANOVA table and the test statistic for testing the slope?
- **Answer:** Test statistic from the _____________ is the _______ of the test statistic found from ____ __________________
  - ______________________________

More Sums of Squares

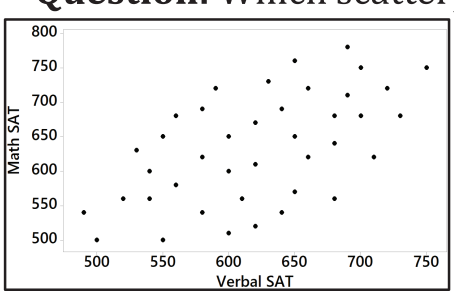
- When studying the relationship between two variables X and Y, there are three necessary sums of squares:
  - $SSX = \sum_{i=1}^{n}(X_i - \bar{X})^2$
    - Sum of squared deviations of predictor values
  - $SSY = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
    - Sum of squared deviations of responses
  - $SSXY = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$
    - Sum of product of joint deviations for each pair of observations

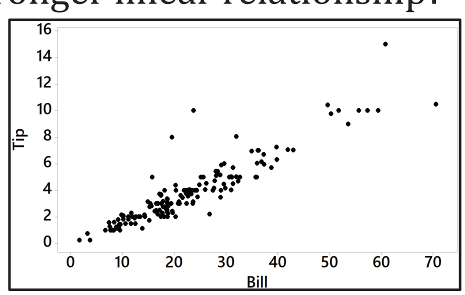
Standard Deviation and Covariance

- **Sample Standard Deviation of Predictor:** $S_X = \sqrt{\frac{1}{n-1}SSX}$
- **Sample Standard Deviation of Response:** $S_Y = \sqrt{\frac{1}{n-1}SSY}$
- **Sample Covariance:** $S_{XY} = \frac{SSXY}{n-1}$
  - Measure of the joint variability between two quantitative variables
  - Sign dictates direction of relationship
  - Unbounded: values range from $-\infty$ to ∞
    - Does not help us interpret strength of relationship

Example: Covariance

- **Scenario:** Verbal SAT score vs. math SAT score on left. Restaurant bill vs. tip on right.
- **Question:** Which scatterplot has the stronger linear relationship?

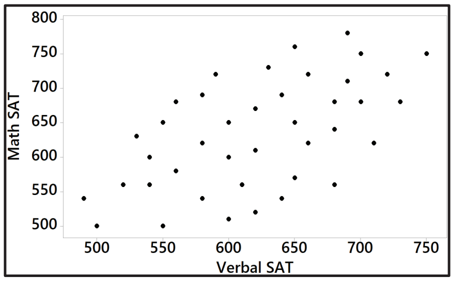




- **Answer:** ____
  - Points are ____

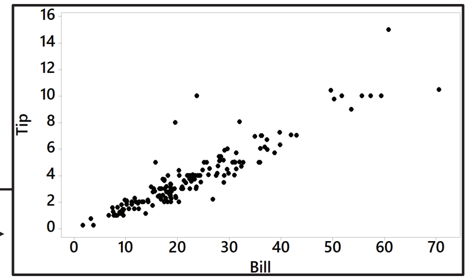
Example: Covariance

- **Scenario:** Verbal SAT score vs. math SAT score on left. Restaurant bill vs. tip on right.
- **Question:** What does the covariance tell us?



$S_{XY} =$ ________

$S_{XY} =$ ________



- **Answer:** ______________________________
  - Covariance will be large if the ______________________________ are large regardless of how __________ the linear relationship is

Correlation Coefficient

- **Correlation Coefficient:** a measure of the strength and direction of the linear relationship between two continuous variables
  1. Ranges from -1 to 1: Larger magnitudes imply stronger relationships
  2. Dimensionless: r is independent of the unit of measurement of X and Y
  3. Follows the same sign as the slope of the regression line: If $\hat{\beta}_1$ is positive, then r is positive, and vice versa

Note: Proofs of properties 1 and 2 require some knowledge of probability theory, covariance, and expectation.

- Can be calculated in three different ways:

$$r = \frac{SSXY}{\sqrt{SSX \cdot SSY}} \qquad r = \frac{S_{XY}}{S_X S_Y} \qquad r = \frac{S_X}{S_Y}\hat{\beta}_1$$

Example: Calculating Correlation Coefficient

- **Scenario:** Record stopping distance for a car at 5 different speeds.
- **Question:** What is the correlation between ACT score and GPA?

Speed	Stop. Dist.	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})(Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$
20	64					
30	118					
40	153					
50	231					
60	319					
$\bar{X} = 40$	$\bar{Y} = 177$					

- **Answer:** ______________________________

Example: Correlation Coefficient

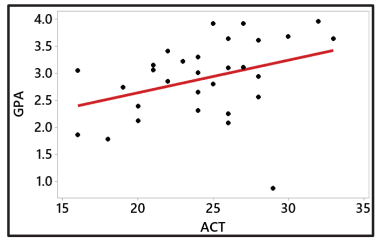
- **Scenario:** Use ACT score of 30 college freshmen to describe their freshman year GPA.

	ACT	GPA
ACT	18.436782	
GPA	1.113598	0.523824

Variable	Total Count	Mean	StDev
ACT	30	24.333	4.294
GPA	30	2.904	0.724

- **Question:** What is the correlation between ACT score and GPA?
- **Answer:**

____



- **Question:** What does the correlation mean?
- **Answer:** ACT score and GPA have a ________ ______________________________________

Example: Correlation Coefficient

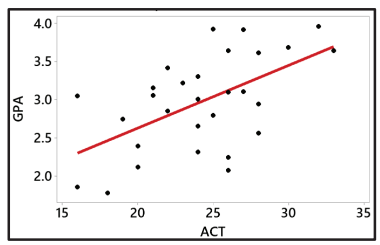
- **Scenario:** Use ACT score of 29 college freshmen (without outlier) to describe freshman year GPA.

Term	Coef	SE Coef	T-Value	P-Value
Constant	0.987	0.570	1.73	0.095
ACT	0.0822	0.0232	3.53	0.001

Variable	Total Count	Mean	StDev
ACT	29	24.172	4.277
GPA	29	2.974	0.625

- **Question:** What is the correlation between ACT score and GPA?
- **Answer:**

____



- **Takeaway:** One outlier can ______________ ________________________ of the correlation.

Proof: Correlation Same Sign as Slope

- **Task:** Prove that the sign of the correlation is always dictated by the sign of the slope.
- **Answer:**
  - Correlation is ____________
  - Standard deviations S_X and S_Y are ____________________ so ________
  - If $\hat{\beta}_1 > 0$, then ____________. Conversely, if $\hat{\beta}_1 < 0$, then ____________.

Example: Perfect Linear Relationship

- **Question:** What happens when there is a perfect linear relationship between X and Y?
- **Answer:**
  - X ________________________ Y every time
  - Every observation lies __________________________________
  - For every point, ____________ so every observation has a residual of ___
  - The sum of squares due to error is $SSE =$ ________________________
  - The coefficient of determination is:

$$r^2 = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

Example: No Linear Relationship

- **Question:** What happens when there is no linear relationship between X and Y?
- **Answer:**
  - No linear relationship means ____
  - The best prediction for every observation is __________________________
  - The total sum of squares is always $SSY =$ ___________________
  - The sum of squares due to error is:

$$SSE = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

  - The coefficient of determination is:

$$r^2 = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

Coefficient of Determination

- **Coefficient of Determination:** the percentage of variability in Y being explained by X

$$r^2 = \frac{SSY - SSE}{SSY}$$

  - The remainder of the variability $1 - r^2$ is due to other factors not being analyzed in the relationship between X and Y

Example: Calculating r^2

- **Scenario:** Use ACT score of 30 college freshmen to describe their freshman year GPA. Given $SSY = 15.191$ and $SSE = 13.240$.
- **Question:** What is the coefficient of determination?
- **Answer:**

  ____

- **Question:** What does the coefficient of determination mean?
- **Answer:** ________________________________ is explained by _____________________.
  - The remaining _________ is due to other factors not being considered in this regression such as ____ ________________________________ etc.

Example: Calculating r^2

- **Scenario:** Use ACT score of 29 college freshmen (without outlier) to describe freshman year GPA.
- **Question:** What is the coefficient of determination?
- **Answer:** ___________________________
- **Takeaway:** By ____________________, the model is able to explain ______________________________
  - It does not have to try to understand why one student's GPA is so ______________________________________.