**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 $\alpha$

**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 __________

**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
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 $Y$

  \[
  SSE = \sum_{i=1}^{n} (Y_i - \hat{Y})^2
  \]

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 $\beta_1$ ___________?
    2. What is the __________ of the straight line relationship?
    3. Is the straight line model ________________?
ANOVA Table for Simple Linear Regression

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F-Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>SSR</td>
<td>MSR = \frac{SSR}{1}</td>
<td>( F = \frac{MSR}{MSE} )</td>
</tr>
<tr>
<td>Error</td>
<td>( n - 2 )</td>
<td>SSE</td>
<td>MSE = \frac{SSE}{n - 2}</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>( n - 1 )</td>
<td>SSY</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fundamental Equation of Regression Analysis

\[ SSY = SSR + SSE \]

\[ \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]

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.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Adj SS</th>
<th>Adj MS</th>
<th>F-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>3.459</td>
<td>3.4589</td>
<td>12.50</td>
<td>0.001</td>
</tr>
<tr>
<td>Error</td>
<td>27</td>
<td>7.474</td>
<td>0.2768</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>10.933</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **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.

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Adj SS</th>
<th>Adj MS</th>
<th>F-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>3.459</td>
<td>3.4589</td>
<td>12.50</td>
<td>0.001</td>
</tr>
<tr>
<td>Error</td>
<td>27</td>
<td>7.474</td>
<td>0.2768</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>28</td>
<td>10.933</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Term</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.987</td>
<td>0.570</td>
<td>1.731</td>
<td>0.095</td>
</tr>
<tr>
<td>ACT</td>
<td>0.082</td>
<td>0.0232</td>
<td>3.535</td>
<td>0.001</td>
</tr>
</tbody>
</table>

- **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 $\infty$
    • 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?

- **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 $\beta_1$ 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:

  $$ r = \frac{SS_{XY}}{\sqrt{SS_X \cdot SS_Y}} $$
  $$ r = \frac{S_{XY}}{S_X S_Y} $$
  $$ r = \frac{S_X}{S_Y} \beta_1 $$

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?

<table>
<thead>
<tr>
<th>Speed</th>
<th>Stop. Dist.</th>
<th>$X_i - \bar{X}$</th>
<th>$Y_i - \bar{Y}$</th>
<th>$(X_i - \bar{X})(Y_i - \bar{Y})$</th>
<th>$(X_i - \bar{X})^2$</th>
<th>$(Y_i - \bar{Y})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>118</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>153</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>231</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>319</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\bar{X} = 40$</td>
<td>$\bar{Y} = 177$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Answer:** _______________________________
Example: Correlation Coefficient

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

<table>
<thead>
<tr>
<th>ACT</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>18.436782</td>
<td>1.113598</td>
</tr>
<tr>
<td>0.523824</td>
<td></td>
</tr>
</tbody>
</table>

**Total Variable**

<table>
<thead>
<tr>
<th>Count</th>
<th>Mean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>30</td>
<td>24.333</td>
</tr>
<tr>
<td>GPA</td>
<td>30</td>
<td>2.904</td>
</tr>
</tbody>
</table>

• **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.

<table>
<thead>
<tr>
<th>Term</th>
<th>Coef</th>
<th>SE Coef</th>
<th>T-Value</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.987</td>
<td>0.570</td>
<td>1.73</td>
<td>0.095</td>
</tr>
<tr>
<td>Constant</td>
<td>0.0822</td>
<td>0.0232</td>
<td>3.53</td>
<td>0.001</td>
</tr>
</tbody>
</table>

**Total Variable**

<table>
<thead>
<tr>
<th>Count</th>
<th>Mean</th>
<th>StDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>29</td>
<td>24.172</td>
</tr>
<tr>
<td>GPA</td>
<td>29</td>
<td>2.974</td>
</tr>
</tbody>
</table>

• **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 ________________________________.