Four Step Process

READING:	0.2
EXERCISES:	ALL CHAPTER 0
ASSIGNED:	HW 1
PRODUCER:	DR. MARIO

IMG CREDIT: ALEX RIEGERT-WATERS

Four Steps of Statistical Modeling

1. Choose a form for the model
2. Fit that model to the data
3. Assess how well the model fits the data
    a. Diagnostic Plots
    b. Look for Patterns in the Residuals
    c. Check Assumptions (Randomness, Independence, Normality)
4. Use the model to answer questions

Supplement for Lecture 2

- Download Zip Folder on Course Website for Supplement
- Unzip Folder on Your Computer
- Open the Template.rmd File from the Unzipped Folder
- Install Mosaic Package and Stat2Data Package
- Run First Two Code Chunks and View Dataset

Example: LEGO

Variable of Interest: *Amazon_Price*

Question of Interest:

How well can we predict the price of a LEGO set on Amazon without knowing any other information?

Form of Model

Constant Model:

$$Y = c + \varepsilon$$

The constant ***c*** is called a **parameter**

We use data to replace the unknown ***c*** with a **sample estimate $\hat{c}$**

Fitting the Model to Data

For the constant model, if we want to **estimate** Y, then

$$E[Y] = \hat{y} = \hat{c}$$

The predicted ***y*** is denoted $\hat{\boldsymbol{y}}$.

Good choices for $\hat{c}$

Sample Mean: $\hat{c} = \bar{y} = \frac{\sum y_i}{n}$

Sample Median: $\hat{c} = m_Y$

Assess Fit of Model

Question: Is the model good and which estimator is better?

Calculate residuals for each observation in data

$$residual = \hat{\epsilon} = y - \hat{y}$$

Each observation has a residual so how do we summarize the overall fit?

Assess Fit of Model

Criteria for Assessing Fit (Loss Functions)

- Sum of Errors: $\sum(y - \hat{y})$
- Sum of Squared Errors (SSE): $\sum(y - \hat{y})^2$
- Sum of Absolute Errors (SAE): $\sum|y - \hat{y}|$

Supplement for Lecture 2

- Subset Data to Remove Missing Values
- Estimate Constant Using Sample Mean and Sample Median
- Assess Fit of Both Models Based on Two Different Criteria

Estimator <chr>	Sum_Squared_Errors <dbl>	Sum_Absolute_Errors <dbl>
Mean	3622919	34330.20
Median	3969984	30441.84

Example: LEGO

- Follow-up Question: *Is there a relationship between the theme and the price on Amazon?*
- Alternative: *What effect does the theme of the LEGO set have on the Amazon price?*
- Strategy for Analysis: *Calculate the average price of different themes on Amazon and compare them. Look for statistically significant differences.*

Supplement for Lecture 2

- Mosaic package in R
  - Use of “formulas” in R to express models

$$y \sim x_1 + x_2 + \cdots + x_p$$

  - Use of **data=_____ argument** to specify dataset and eliminate the need to call variables using **data$variable**
  - Has “modified” versions of classic functions that allow us to look at the effect a categorical variable has on a numeric variable

Model with a Binary Predictor

$$Y = f(X) + \epsilon$$

Where

$$f(X) = \begin{cases} \mu_1 & \text{if } X = \text{Friends} \\ \mu_2 & \text{if } X = \text{Marvel} \end{cases}$$

and

$$\mu_1 = Mean\ Price\ for\ Friends$$
$$\mu_2 = Mean\ Price\ for\ Marvel$$

Two-Sample t-Test for Difference in Means

- Hypotheses (Non-directional)

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 \neq \mu_2$$

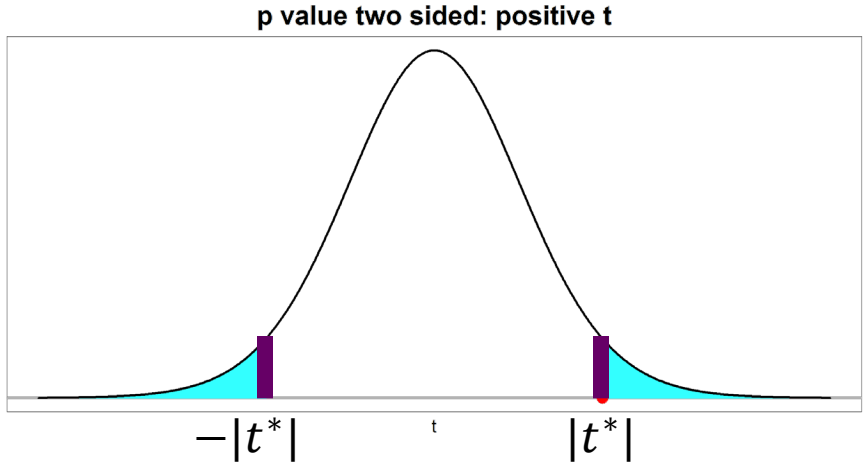
- Test Statistic

$$t^* = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{{s_1^2}/{n_1} + {s_2^2}/{n_2}}}$$

Two-Sample t-Test for Difference in Means

- Calculate P-Value Using t-Distribution

$$t^* \sim t(\Delta df) \qquad \Delta = \frac{\left[s_1^2/n_1 + s_2^2/n_2 \right]^2}{\frac{\left(s_1^2/n_1 \right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2 \right)^2}{n_2 - 1}}$$



- Make Decision
  - P-value < 0.05, then Reject Null and Accept Alternative
  - P-value > 0.05, then Fail to Reject the Null

- Interpret Results in the Context of the Problem/Data

Supplement for Lecture 2

- Welch's Two-Sample t-Test
  - Assume 2 Independent Simple Random Samples from Normal Dist.
  - Don't Assume that Populations Have Equal Variances

- Interpretation of p-value: *Assuming the null hypothesis is true, the* ***p-value*** *measures the percent of all possible test-statistics that are more extreme than the one we observed (Ex: -1.185)*

- *Assess Validity of Assumptions*

Thank You

Make Reasonable Decisions