Displaying and Summarizing Categorical Data

Lecture 4  
January 19, 2018

Four Stages of Statistics

- Data Collection ✓
- Displaying and Summarizing Data
  - One Categorical
  - Two Categorical
  - One Quantitative
  - One Categorical and One Quantitative
  - Two Quantitative
- Probability
- Inference

Review: Terminology

- **Population**: any complete collection of people or objects that a statistician is interested in
- **Parameter**: value that describes a characteristic of a population
- **Sample**: set of units selected from a population that a statistician analyzes to better understand the population
- **Statistic**: value calculated from a sample that serves as an estimate of a parameter

Summarizing a Single Categorical Variable

- **Count**: number of observations in a category
- **Proportion**: count in a category divided by total number of observations
- **Percentage**: proportion as decimal times 100%

Note: Most common to work with proportions.

- Two types of proportions:
  - \( p \): Population proportion (parameter)
  - \( \hat{p} \): Sample proportion (statistic) “p-hat”
Summarizing a Single Categorical Variable

- **Frequency Table:** table of counts for each category
  - Counts must sum to total number observations
- **Relative Frequency Table:** table of proportions or percentages for each category
  - Counts must sum to 1 for proportions and 100% for percentages

Raw Data

- **Raw Data:** list of data directly from a source that has not been processed or summarized for use

  - To find counts in Excel:
    - Raw data usually listed in a single column
    - List category names in a separate column
    - Use “COUNTIF” function to find frequency for each category
  
  - To find proportions:
    - Divide each count by total number of observations

Example #1: Frequency Table

- **Scenario:** Random sample of students in business statistics asked what their major is
- **Task:** Generate the frequency table and relative frequency table.

<table>
<thead>
<tr>
<th>Category</th>
<th>Frequency</th>
<th>Proportion</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accounting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Economics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finance</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Marketing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Accounting</td>
<td>Category</td>
<td>Frequency</td>
<td>Formula</td>
<td>Proportion</td>
<td>Formula</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Finance</td>
<td>Accounting</td>
<td>=COUNTIF(A:A,C2)</td>
<td>=G2/SUM(D:D)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Marketing</td>
<td>Finance</td>
<td>=COUNTIF(A:A,C3)</td>
<td>=G3/SUM(D:D)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Economics</td>
<td>Economics</td>
<td>=COUNTIF(A:A,C4)</td>
<td>=G4/SUM(D:D)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Finance</td>
<td>Marketing</td>
<td>=COUNTIF(A:A,C5)</td>
<td>=G5/SUM(D:D)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Excel: Raw Data - Example #1 Frequency Table
Displaying a Single Categorical Variable

- **Pie Chart**: shows proportions of categories of a categorical variable with size of slices corresponding to proportion for each category
  - Should be used when:
    - Responses cannot overlap into more than one category
    - Shows how the whole is divided into slices
- **Bar Graph**: displays counts (or proportions) of a categorical variable where the height of each bar is proportional to the count in the category
  - Can be used any time a pie chart can also be used
  - Should be used when:
    - There is overlap in responses between categories
    - Not all categories are included in the comparison

Example #2: Using Categorical Data

- **Scenario**: Survey of 796 college students found that 288 of them reported binge drinking at some point in the past month.
- **Question**: What type of variable is being studied?
  - **Answer**: ______________________________
- **Question**: How can this data be summarized?
  - **Answer**: ______________________________
  - ____________________ reported binge drinking

Example #3: Using Categorical Data

- **Scenario**: Survey of 796 college students found that 288 of them reported binge drinking at some point in the past month.
- **Question**: How should the proportion .362 be denoted?
  - **Answer**: ______________________________
    - _______ \( \rightarrow \) _______
- **Question**: How should the overall proportion of all college students who binge drink be denoted?
  - **Answer**: ______________________________
    - Actual value _______ \( \rightarrow \) Can’t ____________________

Example #4: Using Categorical Data

- **Scenario**: Survey of 796 college students found that 288 of them reported binge drinking at some point in the past month.
- **Question**: Why is a pie chart appropriate here?
  - **Answer**: ______________________________
- **Question**: Can we conclude that more than one-third of all college students binge drink?
  - **Answer**: _________
    - Good start, but not ____________________
    - Need ____________________
More Than Two Categories

- Most categorical variables we work with have exactly two categories.
- However, there are occasions when we have variables with three or more categories and need to display summaries.

- **Mode**: category with the highest proportion (or count)

Example #5: Mode

- **Scenario**: Survey of adults asks for their favorite color.
- **Question**: What is the mode?
- **Answer**: ____________
  - Category with ____________
  - Has ________________

Example #6: Problems with Pie Charts

- **Scenario**: Survey of adults asks for their favorite color.
- **Question**: What are the problems with using a pie chart?
- **Answer**: Difficult to...
  - ________________
  - ________________
  - ________________

**Takeaway**: Pie charts typically work well for ________________.

Example #7: Using the Pie Chart

- **Scenario**: Survey of adults asks for their favorite color with percentages added to enhance chart.
- **Question**: If 630 people responded that blue was their favorite color, how many chose red?
- **Answer**:
  - Find ________________:
  - Total ________________:

- **Blue** 42%
  - **Red** 10%
  - **Green** 14%
  - **Beige** 3%
  - **Brown** 3%
  - **Purple** 4%
  - **Yellow** 4%
  - **Beige** 2%
  - **Grey** 2%
  - **Other** 12%
  - **White** 2%
Example #8: Reading Bar Graph
- Scenario: Total number of monthly hits on 10 most popular websites that aren’t Google
- Question: What trends exist in the data?
- Answer: ____________________________
  - ____________________________
  - ____________________________
  - ____________________________
  - Steep drop-off after _______ and second large drop after ________

Example #9: Mode and Graph Types
- Scenario: Total number of monthly hits on 10 most popular websites that aren’t Google
- Question: Why would this data not be appropriate to use in a pie chart?
- Answer:
  1. Pie charts show ____________________________
     - This data only looks at ______________________
  2. Pie chart would only show proportion of hits ____________________________
     - Not proportion of ____________________________

Example #10: Determining Graph Type
- Scenario: For the holidays, 54% of shoppers bought a gift card for a store, 36% for a restaurant, and 25% for an online website
- Question: What type of graph would be most appropriate to use to display this data?
- Answer: ____________________________
  - Percentages sum to ____________________________
  - Some people must have bought gift cards ____________

Example #11: Data Collection
- Scenario: Three students perform studies on the proportion of people who wear glasses.
  - A: Observes 21 of 30 people (70%) on the morning bus ride downtown wearing glasses
  - B: Surveys 300 people downtown during the day and finds 200 (66.7%) who respond they wear glasses
  - C: Surveys 30 people downtown during the day and finds 21 (70%) who respond they wear glasses
- Question: What is the problem with collecting this data using an observational study?
- Answer: ____________________________
Example #12: Representative Sample

**Question:** Which sample is probably most representative of the general population?
- A: Observes 21 of 30 people (70%) on the morning bus ride downtown wearing glasses
- B: Surveys 300 people downtown during the day and finds 200 (66.7%) who respond they wear glasses
- C: Surveys 30 people downtown during the day and finds 21 (70%) who respond they wear glasses

**Answer:** 
- A: Likely includes ___________________________; excludes __________________________________________________
- B and C: __________________________________________________
- ___ is better than ___ because the ________________________

Example #13: Role of Sample Size

**Question:** Is one student’s data more convincing that a majority of people wear glasses?
- A: Observes 21 of 30 people (70%) on the morning bus ride downtown wearing glasses
- B: Surveys 300 people downtown during the day and finds 200 (66.7%) who respond they wear glasses
- C: Surveys 30 people downtown during the day and finds 21 (70%) who respond they wear glasses

**Answer:** 
- A: Likely ______________________________________________
- B: ___________ estimate of proportion; smaller __________________________
- C: ___________________________; sample size is ________

Median

**Median:** middle observation in a set of data
- Only defined for ordinal and quantitative data
- Order observations from smallest to largest
  - Odd number → Middle observation
  - Even number → Average of middle two observations

Example #14: Median

**Scenario:** English professor reveals letter grades on most recent paper for all students in the class

A A A B B B B C C C D D F

**Question:** What is the median?

**Answer:** 
- ___ students → _____________ observation
- ___ observations on each side
Example #15: Median

• **Scenario:** Patients asked to rate the amount of pain they are in. Tables below compare patients with sprains with those with broken bones.

<table>
<thead>
<tr>
<th>Broken Bones</th>
<th>Percentage</th>
<th>Sprains</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Little/no pain</td>
<td>10%</td>
<td>Little/no pain</td>
<td>20%</td>
</tr>
<tr>
<td>Moderate pain</td>
<td>45%</td>
<td>Moderate pain</td>
<td>35%</td>
</tr>
<tr>
<td>Severe pain</td>
<td>30%</td>
<td>Severe pain</td>
<td>40%</td>
</tr>
<tr>
<td>Acute pain</td>
<td>15%</td>
<td>Acute pain</td>
<td>5%</td>
</tr>
</tbody>
</table>

• **Question:** What can we say about the medians?
• **Answer:**
  - 
  - 

Summary

• **One Categorical Variable:**
  - **Numerical Summaries:** Count, proportion, percentage
    - Displayed in frequency distributions
  - **Graphical Displays:** Bar graph, pie chart
  - **Mode:** Category with highest proportion
  - **Median:** Middle observation in ordinal and quantitative data