Describing Data: Categorical and Quantitative Variables

In order to make sense of data, we need ways to summarize and visualize it.

Summarizing and visualizing variables and relationships between two variables is often known as exploratory data analysis (also known as descriptive statistics).

The type of summary statistics and visualization methods to use depends on the type of variables being analyzed (i.e., categorical or quantitative).

One Categorical Variable

“What is your race/ethnicity?”

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>111</td>
</tr>
<tr>
<td>Black</td>
<td>29</td>
</tr>
<tr>
<td>Hispanic</td>
<td>29</td>
</tr>
<tr>
<td>Asian</td>
<td>2</td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
</tr>
</tbody>
</table>

Display the number or proportion of cases that fall into each category.
The sample proportion ($\hat{p}$) of directors in each category is

$$\hat{p} = \frac{\text{number of cases in category}}{\text{total number of cases}}$$

The sample proportion of directors who are white is:

$$\hat{p} = \frac{111}{175} = .63 \ (63\%)$$

Proportion and percent can be used interchangeably.

<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Asian</th>
<th>Other</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>111</td>
<td>29</td>
<td>29</td>
<td>2</td>
<td>4</td>
<td>175</td>
</tr>
</tbody>
</table>

A relative frequency table shows the proportion of cases that fall in each category.

All the numbers in a relative frequency table sum to 1.

In a bar chart, the height of the bar corresponds to the number of cases that fall into each category.

In a pie chart, the relative area of each slice of the pie corresponds to the proportion/percentage in each category.

Look at the relationship between two categorical variables:

1. Race/Ethnicity
2. Gender
It doesn’t matter which variable is displayed in the rows and which in the columns.

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>52</td>
<td>59</td>
<td>111</td>
</tr>
<tr>
<td>Black</td>
<td>13</td>
<td>16</td>
<td>29</td>
</tr>
<tr>
<td>Hispanic</td>
<td>12</td>
<td>17</td>
<td>29</td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>81</td>
<td>94</td>
<td>175</td>
</tr>
</tbody>
</table>

What proportion of female directors are Hispanic?

A. 12/29
B. 12/175
C. 12/81
D. 81/175
E. 29/175

What proportion of Hispanic directors are female?

A. 12/29
B. 12/175
C. 12/81
D. 81/175
E. 29/175

What proportion of directors are female and Hispanic?

A. 12/29
B. 12/175
C. 12/81
D. 81/175
E. 29/175

In a side-by-side bar chart, the height of each bar corresponds to the number of cases that fall into each category of the table.
In a side-by-side bar chart, the height of each bar corresponds to the number of cases that fall into each category of the table.

A segmented bar chart is like a side-by-side bar chart, but the bars are stacked instead of side-by-side.

Difference in Proportions

A difference in proportions is... the difference in proportions for one categorical variable (e.g., the proportion who are Hispanic) calculated for the different levels of another categorical variable (e.g., gender).

What is the difference in proportion of male directors who are Hispanic and female directors who are Hispanic?

\[
\hat{p}_{MH} - \hat{p}_{FH} = \text{sample proportion of male directors who are Hispanic} - \text{sample proportion of female directors who are Hispanic}
\]

Two-Way Table

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>52</td>
<td>59</td>
<td>111</td>
</tr>
<tr>
<td>Black</td>
<td>13</td>
<td>16</td>
<td>29</td>
</tr>
<tr>
<td>Hispanic</td>
<td>12</td>
<td>17</td>
<td>29</td>
</tr>
<tr>
<td>Other</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>81</td>
<td>94</td>
<td>175</td>
</tr>
</tbody>
</table>

What is the difference in gender proportions among Hispanic directors? \( \hat{p}_{MH} - \hat{p}_{FH} \)

The proportion of male directors who are Hispanic = 0.1794

The proportion of female directors who are Hispanic = 0.1281

One Quantitative Variable

When describing quantitative variables we are interested in the distribution of the values – it’s shape, center, and spread.

Shape: Form of the distribution of values
Center: Main peak
Spread: Relative deviation of the values

To understand these concepts we’ll look at quantitative variables from the student survey.
Dotplot

In a dotplot, each case is represented by a dot and dots are stacked.

Histogram

Create “bins” (i.e., value intervals) and place each case in the appropriate bin based on its value for the variable of interest. The height of the each bar corresponds to the number of cases that have values falling within that particular interval.

Bar Charts vs. Histograms

Although they look similar, a histogram is not the same as a bar chart.

A bar chart is for categorical data, and the x-axis has no numeric scale.

A histogram is for quantitative data, and the x-axis is numeric.

For a categorical variable, the number of bars equals the number of categories, and the number in each category is fixed.

For a quantitative variable, the number of bars (or bins) in a histogram is up to you, and the appearance can differ with different number of bars.

Measures of Center

The sample size, the number of cases in the sample, is denoted by \( n \).

A variable is often denoted by \( x \), and \( x_1, x_2, \ldots, x_n \) represent the \( n \) values of the variable \( x \).

Example: \( x = \) The number of body piercings

\[
\begin{array}{cccc}
\text{gender} & \text{bodypart} & \text{nipple} & \text{face} \\
\text{male} & \text{ear} & 0 & 0 \\
\text{female} & \text{nose} & 1 & 1 \\
\text{male} & \text{nose} & 1 & 1 \\
\text{female} & \text{nose} & 0 & 0 \\
\text{male} & \text{eyebrow} & 1 & 1 \\
\text{female} & \text{ear} & 2 & 2 \\
\text{female} & \text{nose} & 1 & 1 \\
\text{male} & \text{nose} & 0 & 0 \\
\text{female} & \text{ear} & 1 & 1 \\
\text{female} & \text{nose} & 0 & 0 \\
\end{array}
\]

\( x_1 = 6 \)
\( x_2 = 0 \)
\( x_3 = 1 \)
\( x_4 = 5 \)
\( \ldots \)
**Measures of Center**

**Mean**

The sample mean ($\bar{x}$) is the average, and is computed by adding up all the numbers and dividing by the number of cases.

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

**Median**

The sample median ($m$) is the middle value when the data is ordered.

If there are an even number of values, the median is the average of the two middle values.

**Outliers**

An outlier is a value that is notably different from the other values (e.g., much larger or smaller than the other values).

**Resistance**

A statistic is resistant if it is not heavily affected by outliers.

The median is resistant, the mean is not resistant.

<table>
<thead>
<tr>
<th>Number of text messages sent per day:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outlier Included</td>
</tr>
<tr>
<td>Outlier Removed</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

**Outliers**

When calculating statistics that are not resistant to outliers, look for outliers and decide whether the outlier is a mistake.

If not, you have to decide whether the outlier is part of your population of interest or not.

Usually, for outliers that are not a mistake, it’s best to run the analysis twice, once with the outlier(s) and once without, to see how much the outlier(s) affect the results.

**Salary Distribution of NHL Player Salaries**

- Mean ($\bar{x}$) = $2,210,000$
- Median ($m$) = $1,250,000$

Mean is “pulled” in the direction of skewness.
Assignment

Part I
Graded Problems
2.18 and 2.60

Additional Practice Problems (not to be turned in):
2.11 and 2.57

Part II
Goto http://sda.berkeley.edu/cgi-bin/hsda?harcsda=gss10
Find 3 categorical variables and provide the proportion for each category for each variable.
Find 3 quantitative variables and provide the mean & median and whether the distribution of values are symmetric, right skewed, or left skewed.

Getting Variable Statistics from the GSS

Enter the variable name here.*
Make sure these boxes are checked.
If applicable select the appropriate type of chart.
Click on this button and the variable statistics will open up in a new window.

Summary:
One Categorical Variable
Summary Statistics
Proportion
Frequency table
Relative frequency table
Visualizations
Bar chart
Pie chart

Summary:
Two Categorical Variables
Summary Statistics
Two-way table
Difference in proportions
Visualizations
Side-by-side bar chart
Segmented bar chart