Chapter 4 Summarizing Data Graphically

4.2 What are we summarizing?

**Unit:** The item or object we observe. If the object is a person we refer to it as the subject.

**Observation:** information or characteristic recorded for one unit.

**Variable:** A characteristic which can vary from one unit to the next.

**Data Set:** A collection of observations on one or more variables is called a data set.

Types of Variables:

**Qualitative variable:** Variables that classify the units into categories. The categories may or may not have a natural ordering; also called categorical variables.

**Quantitative variables** have numerical values that are measurements or counts. Arithmetic operations on these variables have meaning.

**Quantitative discrete variable** is one for which you can count the number of possible values.

**Continuous variables** can take on any value within an interval.
Type of variables

<table>
<thead>
<tr>
<th>Qualitative</th>
<th>Quantitative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qualitative (type of religion)</td>
<td>Quantitative discrete (number of people)</td>
</tr>
<tr>
<td>Quantitative continuous (length)</td>
<td></td>
</tr>
</tbody>
</table>

**Ex:** The number of incoming people in the bank between 11:00AM and 1:00PM on Friday

**Quantitative discrete**

**Ex:** Average daily low temperature per month, in Chicago

**Quantitative continuous**

**Ex:** You roll two dice and record whether or not the resulting values on the two dice matched.

**Qualitative**
Distribution of a variable provides information on the possible values of the variable and how often each of these values occurs.

(The distribution of a variable can be summarized graphically, numerically and with a model.)

4.3 Displaying Distributions - Qualitative Variables

A pie chart displays the distribution of a qualitative variable by dividing a circle into wedges/slices. The size of the wedge is proportional to the percentage of the data in each category. This is done by having the angle \( = \left( \% / 100 \right) \times 360^\circ \) corresponding to the percentage of items in that category.

A bar graph displays the distribution of a qualitative variable by listing the categories of the variable along one axis and drawing a bar over each category with a height (or length) equal to the percentage of items in that category. The bars should be of equal width.
Exercise: Below we give a few of the 200 observations taken from an observational study of factors related to SAT scores.

We took 4 data values from a data set with 200 subjects.

<table>
<thead>
<tr>
<th>ID</th>
<th>Class Rank</th>
<th>Verbal SAT Score</th>
<th>Verbal SAT &lt;450</th>
<th>Math SAT Score</th>
<th>Grade 8 English</th>
<th>Grade 8 Math</th>
<th>SEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>190</td>
<td>400</td>
<td>Y</td>
<td>410</td>
<td>89</td>
<td>85.4</td>
<td>f</td>
</tr>
<tr>
<td>2</td>
<td>14</td>
<td>660</td>
<td>N</td>
<td>660</td>
<td>97</td>
<td>95.2</td>
<td>f</td>
</tr>
<tr>
<td>3</td>
<td>81</td>
<td>610</td>
<td>N</td>
<td>620</td>
<td>93.2</td>
<td>90.7</td>
<td>m</td>
</tr>
<tr>
<td>4</td>
<td>27</td>
<td>610</td>
<td>N</td>
<td>720</td>
<td>96.2</td>
<td>94</td>
<td>m</td>
</tr>
</tbody>
</table>

(a) indicate whether each variable is quantitative discrete, quantitative continuous, or categorical.

Quantitative discrete: class rank
Verbal SAT score, Math SAT score

Quantitative continuous:
Grade 8 English, Grade 8 math

Qualitative: Verbal SAT <450, sex
(b) In fact 45% of the sample of 200 students had a Verbal SAT less than 450. Make a pie chart and a bar graph showing of this result.

Pie Chart

\[ \frac{45}{100} \times 360^\circ = 162^\circ \]

Bar Graph

Verbal SAT

Percent

0 \quad 0.6 \quad 0.8 \quad 0.8 \quad 0.8

Verbal SAT < 450 \quad \geq 450
4.3.3. Displaying Relationships between Two Categorical Variables:

2 way frequency table: 1st entry in each row gives values of one categorical variable and top entry in each column gives values of the other categorical variable. Entries in each row and column give frequency of combination indicated by row and column values.

Marginal distribution of the row variable: percentage of row variable is found by computing the percentage of each row total based on the grand total (number of observations).

Marginal distribution of the column variable: percentage of column variable is found by computing the percentage of each column total based on the grand total (number of observations).

Conditional distribution of the row variable given the column variable is found by expressing the entries of the original table as percentages of the column total.

Conditional distribution of the column variable given the row variable is found by expressing the entries of the original table as percentages of the row total.
Example: A study was conducted in Juneau, Alaska to assess if gender has an effect on smoking habit. A simple random sample of size 100 was selected from the total adult male population of 2050 and a simple random sample of size 120 was selected from the total adult female population of 1000. The proportion of smokers among 100 males was found to be 0.40, while the total numbers of smokers in the overall sample was found to be 70.

1. Complete the following frequency table for the sampled results.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Smoking Habit</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>Female</td>
<td>30</td>
<td>90</td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>150</td>
</tr>
</tbody>
</table>

2. The marginal distribution of the row variable:

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent</td>
<td>45.45%</td>
<td>54.55%</td>
</tr>
</tbody>
</table>

\[
\frac{100}{220} \times 100\% = 45.45\%
\]

\[
\frac{120}{220} \times 100\% = 54.55\%
\]
3. The marginal distribution of the column variable:

<table>
<thead>
<tr>
<th></th>
<th>Smoker</th>
<th>Non-Smoker</th>
</tr>
</thead>
<tbody>
<tr>
<td>percent</td>
<td>31.82%</td>
<td>68.18%</td>
</tr>
</tbody>
</table>

\[
\frac{70}{220} \times 100\% = 31.82\%
\]
\[
\frac{150}{220} \times 100\% = 68.18\%
\]

4. Use the table below to provide the conditional distribution of smoking habit given gender.

<table>
<thead>
<tr>
<th>Gender</th>
<th>Smoking habit</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Smoker</td>
</tr>
<tr>
<td>Male</td>
<td>40%</td>
</tr>
<tr>
<td>Female</td>
<td>65%</td>
</tr>
</tbody>
</table>

\[
\frac{40}{100} \times 100\% = 40\%
\]
\[
\frac{60}{100} \times 100\% = 60\%
\]
\[
\frac{30}{120} \times 100\% = 25\%
\]