Outline

Types of Variables
   Qualitative and Quantitative

Cross-Tabulation

Another Example on the Histogram
   Average and spread in a histogram

Average and Median
Types of Variables

A variable is a characteristic that may change from individual to individual in a population.

For example, in a survey the questions: how old are you? What is your family’s size? What is your gender? What is your marital status? Correspond to the variables: age, size, sex and marital status.

There are different types of variables. Each type is usually handled and analyzed differently.
Qualitative and Quantitative

Variables can be classified as:

- **Quantitative data.** Correspond to observations measured on a numerical scale. This can be:
  
  - **Discrete** when the values can differ by fixed amounts like in size. This is typical of counts.
  
  - **Continuous** differences in values can be arbitrarily small like in age. This is typical of variables that measure physical quantities.
Descriptive Statistics

- **Qualitative data.** Correspond to observations classified in groups or categories like in sex and marital status. Sometimes the groups have some ordering, as in the case of grades. Of particular importance are binary variables that can take only two values.
Classify the following variables:

- Records of whether an electrical device is working or not.
- The depth of the snow pack at a monitoring station in the Sierras.
- The number of female students in AMS 5.
- The final grade of a student in AMS 5.
- The State where a given car is registered.
- The ranking of a school within the State school system.
- The number of calls to 911 in a given month
Cross-Tabulation

In many situations we need to perform an exploratory analysis of data to observe possible associations with a discrete variable. For example, consider measuring the blood pressure of women and divide them in two groups: one taking the contraceptive pill and the other not taking it.

We can produce a table with the distribution of one group in one column and the distribution of the other in another column. This can be used to produce two histograms in order to make a visual comparison of the two groups.
The variable that is used for the cross-tabulation is usually referred to as a **covariable**.

<table>
<thead>
<tr>
<th>blood pressure (mm)</th>
<th>non users %</th>
<th>users %</th>
</tr>
</thead>
<tbody>
<tr>
<td>under 100</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>100–110</td>
<td>20</td>
<td>12</td>
</tr>
<tr>
<td>110–120</td>
<td>31</td>
<td>26</td>
</tr>
<tr>
<td>120–130</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>130–140</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td>140–150</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td>150–160</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>over 160</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>
Descriptive Statistics

![Graph showing blood pressure distribution for women not using the pill]

![Graph showing blood pressure distribution for women using the pill]
We observe that the histogram of pill users is slightly shifted to the right, suggesting an increase in blood pressure among women taking the pill.
Another Example on the Histogram

Data from the 2003 Census produce the following for units (houses and apartments) in the New York City area that are either occupied by the owner or rented out. Total number of units are shown below.

1. The owner-occupied percents add up to 100.2% and the renter-occupied percents add up to 100.0%, why?

2. The percentage of one-room units is much larger for renter-occupied housing. Is that because there is less renter-occupied housing in total?

3. Which are larger on the whole: the owner-occupied units or the renter-occupied units?
## Descriptive Statistics

<table>
<thead>
<tr>
<th>Number of Rooms in unit</th>
<th>Owner Occupied (percent)</th>
<th>Renter Occupied (percent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>0.1</td>
<td>2.8</td>
</tr>
<tr>
<td>3</td>
<td>1.4</td>
<td>22.7</td>
</tr>
<tr>
<td>4</td>
<td>9.7</td>
<td>34.5</td>
</tr>
<tr>
<td>5</td>
<td>23.3</td>
<td>22.6</td>
</tr>
<tr>
<td>6</td>
<td>26.4</td>
<td>10.4</td>
</tr>
<tr>
<td>7</td>
<td>17.5</td>
<td>3.6</td>
</tr>
<tr>
<td>8</td>
<td>10.4</td>
<td>1.2</td>
</tr>
<tr>
<td>9</td>
<td>5.0</td>
<td>0.5</td>
</tr>
<tr>
<td>≥10</td>
<td>6.4</td>
<td>0.7</td>
</tr>
<tr>
<td>Total</td>
<td>100.2</td>
<td>100.0</td>
</tr>
<tr>
<td>Number</td>
<td>72.2 million</td>
<td>33.6 million</td>
</tr>
</tbody>
</table>

April 10th, 2015 | UCSC — AMS5

12/20
The answer to the first question is that there is rounding involved in the calculation of the percentages. As for the second question, the fact that we are taking percentages accounts for the difference in totals, so a smaller total of renter-occupied units does not explain the difference.

What seems to be happening is that units for rent tend to be smaller than units occupied by their owners. This is more clearly seen from the comparison of the two histograms.
Descriptive Statistics

![Histogram for owner-occupied](image)

![Histogram for renter-occupied](image)
Average and spread in a histogram

A histogram provides a graphical description of the distribution of a sample of data. If we want to summarize the properties of such a distribution we can measure the center and the spread of the histogram.

These two histograms correspond to samples with the same center.

The spread of the sample on top is smaller than that of the sample in the bottom.
To obtain an estimate of the center of the distribution we can calculate an average.

The average of a list of numbers equals their sum, divided by how many they are.

Thus, if 18; 18; 21; 20; 19; 20; 20; 20; 19; 20 are the ages of 10 students in this class, the average is given by

\[
\frac{18 + 18 + 21 + 20 + 19 + 20 + 20 + 20 + 19 + 20}{10} = 19.5
\]
In the hospital data that we considered in the previous class the data corresponded to the average length of stay of patients in each hospital in the survey. This means that the length of stay of all patients in a given hospital were added and the sum divided by the number of patients in that hospital.
Average and Median
This histogram corresponds to the rainfall over periods of 10 days in an area of the central plains of Venezuela.

The average or mean rainfall is 37.65 mm. We observe that only about 30% of the observations are above the average.

Notice that this histogram is not symmetric with respect to the average.