II. Descriptive Statistics
A. Graphs and Tables for Univariate Data

In this section
- Distributions
- Qualitative Data
- Quantitative Data

1. Distributions

We are now to the point in the class where we will begin looking at descriptive statistics. Remember that descriptive statistics are methods of summarizing a set of data. One way to do that is graphically. The main reason statisticians use graphs is to look at the distribution of data.

Distribution – all outcomes for a variable along with how often each outcome occurs

They type of graph used to describe data depends on the number of variables and the type of those variables. In this section we will only look at univariate data which means we will graph one variable.

2. Qualitative Data

The following qualitative data set will be utilized in order to illustrate how some graphs work. The variable of interest in this illustration is type of operation (data is non-numerical and was collected for a one year period from a single hospital). Qualitative data is commonly summarized with either frequencies or percentages which are both identified in the following table.

<table>
<thead>
<tr>
<th>Type of operation</th>
<th>Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abdominal</td>
<td>115</td>
<td>23.1</td>
</tr>
<tr>
<td>Bones and joints</td>
<td>45</td>
<td>9.0</td>
</tr>
<tr>
<td>Eye, ear, nose, and throat</td>
<td>58</td>
<td>11.6</td>
</tr>
<tr>
<td>General</td>
<td>98</td>
<td>19.7</td>
</tr>
<tr>
<td>Neurosurgery</td>
<td>23</td>
<td>4.6</td>
</tr>
<tr>
<td>Other</td>
<td>65</td>
<td>13.1</td>
</tr>
<tr>
<td>Thoracic</td>
<td>20</td>
<td>4.0</td>
</tr>
<tr>
<td>Urologic</td>
<td>74</td>
<td>14.9</td>
</tr>
<tr>
<td>Total</td>
<td>498</td>
<td>100.0</td>
</tr>
</tbody>
</table>
Important characteristics of a pie chart

- Used only for qualitative data
- Only one variable can be included in the graph
- Displays the division of a total quantity
- Users tend to notice largest pieces of the pie
- All categories for the variable must be included in the graph meaning the total percentage must add to be 100%
- Should not include too many categories (You should notice the Other category in the example above. This Other category includes multiple types of surgeries with low frequency. If this was not done, the graph would include many small pieces and become very messy and not very readable.)
- The number of degrees for each wedge should correspond to the percentage (remember 360 degrees in a circle)
Bar Graph

We will use the same data set in order to discuss the properties of a bar graph. The bar graph is the most misused graph, in part because it is very easy to create.

![Bar Chart for Operations](image)

**Important characteristics of a bar graph**
- Used only for qualitative data (This is a common mistake that I see with this graph. These are very easy to make in excel and sometimes they are made with quantitative data. That should never be done because the x-axis is not numerical in a bar graph. It does not matter in what order the bars are displayed. All horizontal dimensions in the graph are based on the idea that there is no numerical meaning to this axis.)
- Can be used for more than one qualitative variable (This is sometimes done with different colors representing different variables)
- Displays frequency or percentage of items in each category
- The length of a bar represents the quantity we wish to compare
- Users tend to notice the tallest bars
- The bars should be of uniform width and uniformly spaced (This goes back to the x-axis being non-numerical. The width of the bars and spacing has no numerical meaning)
- Does not have to include 100% of the data (This is not common, but you can leave data out of a bar graph. If this is done, a note should be included under the graph stating what was left out.)
3. **Quantitative Data**

The goal with these graphs is to look at the distribution for a single quantitative variable. When looking at the graphs presented, it is important to identify if there are extreme values in the data and the shape of the distribution.

**Extreme value** (or **Outlier**) – observations that are separated from the rest of the data set by some margin

**Example**

Identify the extreme value in the following data.

14, 15, 17, 21, 60, 23, 25

**Answer**

The extreme value is 60.

The impact of an extreme value can be great. Suppose we want to estimate the average income for this class. If a billionaire was in the class, then their income would be an extreme value. That would make the average appear very high simply because of one observation.

**Shape** – the pattern displayed when the graph is created (The most common shapes of distributions will be introduced after the graphs have been introduced.)

**Stem-and-Leaf plot**

**Important characteristics of a stem-and-leaf**

- Separates data entries into “leading digits” or “stems” and “trailing digits” or “leaves”
- Organizes and groups data but allows us to recover the original data if desired
- Good for spotting extreme values and identifying shape

**Example**

14 male weights in pounds: 139, 153, 179, 201, 163, 168, 157, 170, 172, 165, 145, 155, 161, 151

In this case the first two digits of each number represent the stems and the last digit represents the leaves. The stem-and-leaf follows.

<table>
<thead>
<tr>
<th></th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>9</td>
<td>5</td>
<td>1 3 5 7</td>
<td>1 3 5 8</td>
<td>0 2 9</td>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Notice in the stem-and-leaf plot it is easy to identify the one extreme value of 201. If a
data point is separated from the rest of the data by at least on empty class (in this case
two) then it is considered an extreme value. The shape is identified by looking at the
pattern created with the numbers. This will be further discussed later in this chapter.

Frequency Distribution

Important characteristics of a frequency distribution

- A summary table in which quantitative data are arranged into conveniently
  established class groupings (the groups must satisfy the following criteria)
  o Should have between 5 and 15 classes
  o Each class grouping should be of equal width
  o Overlapping the classes must be avoided
- Useful when dealing with very large data sets
- Through the grouping process the original data is lost
- **Class midpoint** – the point halfway between the boundaries of each class

In order to illustrate a frequency distribution, we will look at the same data presented
above in the stem-and-leaf (the 14 male weights).

<table>
<thead>
<tr>
<th>Weight</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>130 but less</td>
<td>1</td>
</tr>
<tr>
<td>than 140</td>
<td></td>
</tr>
<tr>
<td>140 but less</td>
<td>1</td>
</tr>
<tr>
<td>than 150</td>
<td></td>
</tr>
<tr>
<td>150 but less</td>
<td>4</td>
</tr>
<tr>
<td>than 160</td>
<td></td>
</tr>
<tr>
<td>160 but less</td>
<td>4</td>
</tr>
<tr>
<td>than 170</td>
<td></td>
</tr>
<tr>
<td>170 but less</td>
<td>3</td>
</tr>
<tr>
<td>than 180</td>
<td></td>
</tr>
<tr>
<td>180 but less</td>
<td>0</td>
</tr>
<tr>
<td>than 190</td>
<td></td>
</tr>
<tr>
<td>190 but less</td>
<td>0</td>
</tr>
<tr>
<td>than 200</td>
<td></td>
</tr>
<tr>
<td>200 but less</td>
<td>1</td>
</tr>
<tr>
<td>than 210</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>14</td>
</tr>
</tbody>
</table>

In the frequency distribution for weight, the same groupings were used as was naturally
created with the stem-and-leaf plot. Notice with a big data set, the stem-and-leaf would
become very messy, but the frequency distribution would simply have bigger
frequencies. The down side is that the actual data values cannot be identified in the
frequency distribution.

Histogram

Important characteristics of a histogram

- The variable (x-axis) in a histogram is quantitative (different than a bar chart). Since
  the variable is quantitative, the width of the bars has numerical meaning and the bars
  always touch. The bars represent the classes identified in the frequency distribution.
- A picture of a frequency distribution
- The actual values identified on the x-axis are the class midpoints
- The y-axis represents the frequency for each class
It is important to understand how the above histogram was created based on the frequency distribution. It is also important to see how it is similar to the stem-and-leaf. Notice the same extreme and the shape is the same if turn the graph as can be seen below.
Shapes of Distributions

**Symmetrical** – both sides are the same when the graph is folded vertically in the center

**Uniform** – every class has equal frequency (bars are the same height)

**Skewed** – one tail is stretched longer than the other (not symmetric). The direction of the skewness is on the side of the longer tail. We start at the highest frequency class (highest bar) and from there the graph tails to the right and to the left.

**Skewed Right** – tail to the right is stretched longer than the tail to the left

**Skewed Left** – tail to the left is stretched longer than the tail to the right
Bimodal – the two classes with largest frequencies are separated by at least one class