Chapter 2 - Graphical Summaries of Data

Data recorded in the sequence in which they are collected and before they are processed or ranked are called raw data. Raw data is often difficult to make sense out of (especially if you have very big data sets) so we prefer to organize and summarize our data.

2.1 - Graphical Summaries for Qualitative Data

A frequency distribution is a table that lists all categories and the number of elements that belong to each of the categories.

Reasons for Constructing Frequency Distributions:
1. Large data sets can be summarized.
2. We can gain some insight into the nature of data.
3. We have a basis for constructing important graphs.

The relative frequency of a category is the proportion of items in the category.

Relative frequency of a category = \( \frac{\text{Frequency of that category}}{\text{Sum of all frequencies}} \)

A relative frequency distribution is a table that represents the relative frequency of each category (often times we just add another column with this information in the same table as our frequency distribution).

Practice:
Suppose we ask 30 people if they ever suffer from insomnia. The responses are classified as Never (N), Sometimes (S), Often (O), Always (A), and are recorded below.

<table>
<thead>
<tr>
<th>S</th>
<th>O</th>
<th>S</th>
<th>A</th>
<th>N</th>
<th>N</th>
<th>S</th>
<th>O</th>
<th>A</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>O</td>
<td>O</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>O</td>
<td>N</td>
<td>S</td>
</tr>
<tr>
<td>S</td>
<td>N</td>
<td>A</td>
<td>O</td>
<td>S</td>
<td>S</td>
<td>O</td>
<td>N</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

1. Construct a frequency distribution table for the data above. Compute the relative frequencies and percentages for all categories.

<table>
<thead>
<tr>
<th>Insomnia Occurrence</th>
<th>Tally</th>
<th>Frequency</th>
<th>Relative Frequency</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( \text{Sum} = \) \( \text{Sum} = \) \( \text{Sum} = \)
2. What proportion of the people in this sample suffer from insomnia Often or Always?

A graph made of bars whose heights represent the frequencies or relative frequencies of respective categories is called a **bar graph**. The bars should be of equal width and not touch each other. Make sure to always label both the horizontal and vertical axis.

3. Draw a bar graph for the frequency distribution of the Insomnia Occurrence data.

4. How would a bar graph on the relative frequency distribution of the Insomnia Occurrence data differ from the one you just constructed?

Bar graphs can be constructed vertically (more common) or horizontally. Sometimes we want to compare two or more bar graphs that have the same categories, so we can construct several bar graphs on the same axes, with corresponding bars next to each other. This is called a **side-by-side bar graph**. See example on the right.

A **pie chart** is a circle divided into sectors, each one representing a different category and its size proportional to the proportion size (relative frequency) of that category. Each sector should be labeled with its category and its relative frequency expressed as a percentage.

\[ \text{Angle} = (\text{Relative frequency}) \cdot 360^\circ \]

4. Construct a pie chart for the distribution of the Insomnia Occurrence data.
2.2 - Organizing and Graphing Quantitative Data

Since quantitative data doesn’t have natural categories, we divide the data into **classes**. The classes are intervals of equal width that cover all the values in the data set.

A **Frequency Distribution for Quantitative data** lists all the classes and the number of values that belong to each class. Data presented in the form of a frequency distribution are called **grouped data**.

![Frequency Distribution Table]

Class width = Lower limit of a class – Lower limit of next class

Class midpoint = \( \frac{\text{Lower limit of a class} + \text{Lower limit of next class}}{2} \)

**Practice:**

*Find the class width and class midpoints for each category in the frequency distribution above.*
Constructing a Frequency Distribution:

1. Decide on the number of classes (should be between 5 and 20). The bigger the data set, the more classes you should choose. Rather choose too many classes than too few, but you want to have some large frequencies in some of the classes. (One way to help you choose number of classes is Sturge’s formula: $\# \text{ of classes} = 1 + 3.3 \log n$, where $n$ is the number of observations.)

2. Calculate approximate class width. Round to a convenient number.

   \[
   \text{Class width} = \frac{\text{Largest value} - \text{Smallest value}}{\text{Number of classes that you want}}
   \]

3. Starting point: Begin by choosing a lower limit of the first class, which should be a convenient number less than or equal to your smallest value.

4. Using the lower limit of the first class and the class width, proceed to list all the classes.

   \[
   \text{Lower limit of one class + Class width} = \text{Lower limit for the next class}
   \]

5. Count the number of observations in each class (possibly by tallying) and construct a frequency distribution.

   \[\text{ex.}\]

   How many classes does the frequency distribution have?

   What is the class width?

   Is it clear where a value belongs?

   How big was the sample of pennies?

   Could we have omitted the classes with no frequencies, or do those classes tell us something important?
## Histograms

A **histogram** is a graph in which classes are marked on the horizontal axis and the frequencies, relative frequencies, or percentages are marked on the vertical axis.

The frequencies, relative frequencies, or percentages are represented by the heights of the bars.

In a histogram, the bars are drawn adjacent to each other (that is, there is no space in-between them as in a bar graph).

### Frequency Histograms vs. Relative Frequency Histograms or Percentage Histograms

A **polygon** is a graph formed by joining the midpoints of the tops of successive bars in a histogram with straight lines.

Note that we create one extra class at each end, both of which have zero frequency.
An ogive is a different kind of polygon. A frequency ogive plots cumulative frequencies, and a relative frequency ogive plots cumulative relative frequencies.

The cumulative frequency of a class is the sum of the frequency for that class and all previous classes (with lower limits).

Cumulative relative frequency = \[
\frac{\text{Cumulative frequency}}{\text{Sum of all frequencies}}
\]

Shapes of Histograms

As the number of classes is increased, the polygon eventually becomes a smooth curve.

A high point of a histogram is referred to as a mode. A histogram is unimodal if it has only one mode, and bimodal if it has two distinct modes.


Skewed right. Skewed left.

Uniform or Rectangular.
Practice: The following data give the numbers of computer keyboards assembled at the Twentieth Century Electronics Company for a sample of 25 days.

45  52  48  41  56  46  44  42  48  53  51  53  51
48  46  43  52  50  54  47  44  47  50  49  52

a. Make the frequency distribution table for these data.
b. Calculate the relative frequencies for all classes.
c. Construct a histogram for the relative frequency distribution.
d. Construct a polygon for the relative frequency distribution.
e. Construct a relative frequency ogive
2.3 – More Graphs for Quantitative Data

A **stem-and-leaf plot** allows for a nice overview of quantitative data without losing information on individual observations.

Each value is divided into a stem and a leaf as in the following example, where the rightmost digit is the leaf, and the remaining digits form the stem..

**Data:**

43  58  41  65  49  52  58  60  49

Divide up stem and leaf:

4|3   5|8   4|1   6|5   4|9    5|2   5|8   6|0   4|9

**Stem-and-leaf plot:**

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaves</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>1 3 9 9</td>
</tr>
<tr>
<td>5</td>
<td>2 8 8</td>
</tr>
<tr>
<td>6</td>
<td>0 5</td>
</tr>
</tbody>
</table>

**ex.** List the data represented in the following display.

<table>
<thead>
<tr>
<th>Stem</th>
<th>Leaves</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>5 9</td>
</tr>
<tr>
<td>8</td>
<td>0 2 6 7 7</td>
</tr>
<tr>
<td>9</td>
<td>1 7 8</td>
</tr>
<tr>
<td>10</td>
<td>2 6</td>
</tr>
</tbody>
</table>

**ex.** Prepare a stem-and-leaf plot for the following data. In your final display, you should arrange the leaves for each stem in increasing order.

94   103   79   99   114   89
81   86   81   93   100   96
75   90   88   107  132   95

**ex.** Prepare a stem-and-leaf display for the following data.

1020  1085  1021  1008  1101  974  993  1033  1029
A **dotplot** consists of a graph in which each data value is plotted as a point (or dot) along a scale of values.

Dots that represent equal values are stacked.

Dotplots are helpful in getting a good overview of the data, including finding **clusters** of data, as well as **outliers**.

**Clusters** are areas where the values are more concentrated.

**Outliers** are data values that are extremely large or small compared to the rest of the values in the data set.

Dotplots are also useful for comparing two or more datasets, by creating a dotplot for each dataset, on the same scale, and then place these sets on top of each other. We call this **stacked dotplots**.

The example below is a dotplot display of the ages of actresses at the time they won Academy Award Oscars.

*Ex. Make a dotplot of ages of actors, and make a stacked dotplot to compare the two data sets.*

![Dotplot of Ages of Actresses](image)

<table>
<thead>
<tr>
<th>Table 2-1 Academy Awards: Ages of Best Actresses and Best Actors</th>
</tr>
</thead>
<tbody>
<tr>
<td>The ages (in years) are listed in order, beginning with the first awards ceremony.</td>
</tr>
<tr>
<td><strong>Best Actresses</strong></td>
</tr>
<tr>
<td>22</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>35</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>43</td>
</tr>
<tr>
<td>41</td>
</tr>
<tr>
<td>26</td>
</tr>
<tr>
<td>26</td>
</tr>
<tr>
<td><strong>Best Actors</strong></td>
</tr>
<tr>
<td>44</td>
</tr>
<tr>
<td>38</td>
</tr>
<tr>
<td>41</td>
</tr>
<tr>
<td>49</td>
</tr>
<tr>
<td>44</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>51</td>
</tr>
<tr>
<td>46</td>
</tr>
</tbody>
</table>

*Example borrowed from the Triola Stats book*