Chapter 1. Picturing Distributions with Graphs

Definition. Individuals are the objects described by a set of data. Individuals may be people, but they may also be animals or things. A **variable** is any characteristic of an individual. A variable can take different values for different individuals.

Definition. A **categorical variable** places an individual into one of several groups or categories. A **quantitative variable** takes numerical values for which arithmetic operations such as adding and averaging make sense. The values of a quantitative variable are usually recorded in a **unit of measurement** such as seconds or kilograms.

Example. Exercise 1.1 page 6.

Categorical Variables: Pie Charts and Bar Graphs

Note. We can better understand data if we have a way to visualize it.

Defnition. The **distribution** of a variable tells us what values it takes and how often it takes these values. The values of a categorical variable are labels for the categories. The **distribution of a categorical variable** lists the categories and gives either the count or the percent of individuals wo fall in each category.

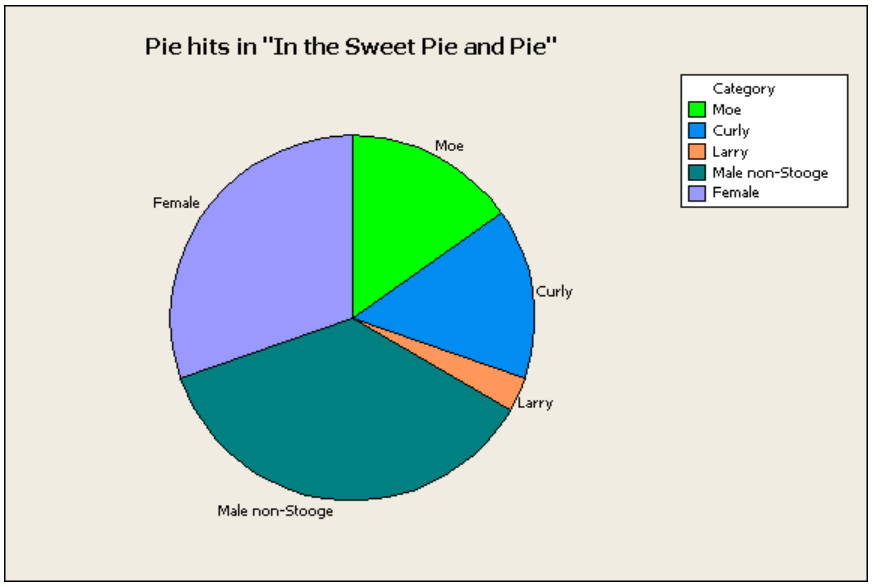
Definition. A *pie chart* reflects the number of individuals falling into different categories by representing the categories as sectors of a circle with the number of individuals in the category reflected by the area of the sector. Pie charts can only be used to illustrate a particular population that has been partitioned into categories. A *bar chart* reflects the number of individuals falling into different categories by plotting the categories along the $x-$axis and the numbers along the $y-$axis. Bar charts can be used to illustrate a partitioned population or data from several different populations (see example 1.3). Pie charts and bar charts are ways of displaying categorical data.

Example S.1.1. A Pie (Chart) in the Face!
You will now collect your first set of data! You will watch a few minutes of the Three Stooges short "In the Sweet Pie and Pie" (short number 58) in which there is a dynamic "pie fight." Watch the film from the time 13:55 (when a waiter enters with a large cake) to the end. Produce categorical data by counting the number of times various people are hit with pies or cakes. Use the categories: Moe, Larry, Curly, male non-Stooge, and female. What are the individuals of the data set? After collecting the data, make a bar chart of the data and use Minitab to make a pie chart of the data. What is the **mode** of the data (i.e., the category containing the most individuals)? What is the average of the distribution?

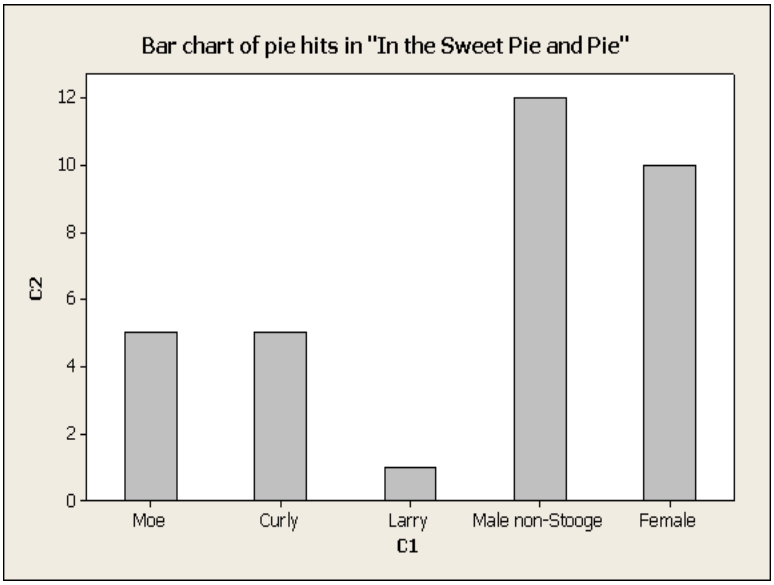
Partial Solution. Start Minitab and input the categories in Column 1 and the corresponding data you have collected in Column 2. The Minitab steps needed to create the pie chart are:

1. Click on the tab `Graph` and select `Pie Chart`.
2. For the `Categorical variable` input `C1` for "Column 1."
3. For the `Summary variables` input `C2` for "Column 2."

You can also click on **Labels** and input a title for the chart if you like. This leads to an output similar to the following:



We can also use Minitab to generate a bar chart:



Quantitative Variables: Histograms

Definition. A *histogram* groups together quantitative variables and reflects the number of individuals in each group along the $y-$axis.

Note. Notice the comment about drawing bar charts versus histograms on page 13: "Draw bar graphs with blank space between the bars to seperate the items being compared. Draw histograms with no space, to indicate that all values of the variables are covered."

Exercise 1.6. Consider the data in Table 1.2 (see page 13). We are told to group the data into classes of width 2 minutes starting at 15 minutes (say (15.0, 17.0], (17.0, 19.0], etc.). We then have the data as:

Class	**Count**
15.0 to 17.0	4
17.1 to 19.0	4
19.1 to 21.0	8
21.1 to 23.0	13
23.1 to 25.0	13
25.1 to 27.0	5
27.1 to 29.0	2
29.1 to 31.0	2

Construct by hand the histogram representing this data.

Example S.1.2. Stooge-o-grams!

Based on data presented in *The Three Stooges—An Illustrated History* by Michael Fleming, if we average the number of slaps per film from the Three Stooges shorts over the 25 years for which they made their short films, we get the following data (rounding to one decimal place):

Year	**Slaps/Film**	**Year**	**Slaps/Film**
1934	33.5	1947	31.9
1935	23.1	1948	9.9
1936	10.5	1949	14.4
1937	6.4	1950	19.6
1938	8.9	1951	21.1
1939	11.6	1952	13.6
1940	14.0	1953	16.0
1941	10.6	1954	8.5
1942	7.6	1955	14.8
1943	12.2	1956	17.8
1944	13.7	1957	6.4
1945	10.2	1958	11.4
1946	15.6		

Make a histogram of this data using classes of width 5 (slaps/film) starting at 5 slaps/film.

Interpreting Histograms

Note. In any graph of data, look for all the **overall pattern** and for striking **deviations** from that pattern. You can describe the overall pattern of a histogram by its **shape**, **center**, and **spread**.

Definition. An important kind of deviation is an **outlier**, an individual value that falls outside the overall pattern.

Definition. A distribution is **symmetric** if the right and left sides of the histogram are approximately mirror images of each other. A distribution is **skewed to the right** if the right side of the historgram (containing the upper half of the observations) extends much farther out than the left side.It is **skewed to the left** if the left side of the histogram extends much farther out than the right side.

Example S.1.3. Stooge-o-gram Interpreted. Describe the histogram of the slaps per film from Example S.1.2.

Solution. The distribution has a single peak in the 10 to 15 slaps/film group. It is skewed to the the right. Most years have less than 15 slaps/film and the histogram extends to the right of its peak farther than it extends to the left. The center of the graph is somewhere between 10 and 15 slaps/film with about 12 of the years to the left and about 12 of the years to the right. Based on the data, one could *guess* that the center is around 12.5. There are 2 observations in the group 30 to 35 slaps/film and these *could* be interpreted as outliers. However, they do fit the general pattern of the distribution in that the distribution trails off on the right side. See Example 1.6 on page 15 for a similar histogram. Figure 1.6 on page 17 gives an example of a symmetric distribution.

Quantitative Variables: Stemplots

Definition. A *stemplot* is a way to represent quantitative data. To make a stemplot:

1. Separate each observaton into a **stem**, consisting of all but the final (rightmost) digit and a **leaf**, the final digit.
2. Write the stems in a vertical column with the smallest at the top, and draw a vertical line at the right of this column.
3. Write each leaf in the row to the right of its stem, in increasing order out from the stem.

Example S.1.4. Stooge Stems.
Consider the slaps per film data from Example S.1.2. Make a stem plot of this data.

Solution. The data is:

Year	Slaps/Film	Year	Slaps/Film
1934	33.5	1947	31.9
1935	23.1	1948	9.9
1936	10.5	1949	14.4
1937	6.4	1950	19.6
1938	8.9	1951	21.1
1939	11.6	1952	13.6
1940	14.0	1953	16.0
1941	10.6	1954	8.5
1942	7.6	1955	14.8
1943	12.2	1956	17.8
1944	13.7	1957	6.4
1945	10.2	1958	11.4
1946	15.6		

The stem is then the integer part of the data and the leaf is the digit in the tenths position.

6	44
7	6
8	59
9	9
10	256
11	46
12	2
13	67
14	148
15	6
16	0
17	8
18	
19	6
20	
21	1
22	
23	1
24	
25	
26	
27	
28	
29	
30	
31	9
32	
33	5

Notice that this would be very tedious (and impractical) if we had a large data set. This stem plot is the same as a histogram

with groups of size 1. Notice that it is difficult to see much of a pattern in the data since each group only contains a few data points. The histogram of Example S.1.2 is preferable in this sense. However, we see from the stem plot that the data points 31.9 and 33.5 do appear to be outliers since they lie rather far from the rest of the data. This was not as clear from the histogram of Example S.1.2.

Note. When making a stemplot, you might desire to *round* data off to the last digit of interest. You might also *split stems* to double the number of stems. We illustrate this with an example.

Example. Exercise 1.36 page 33.

Solution. Rounding to the nearest 100, we get the raw data:

Year	**Recruit.**	**Year**	**Recruit.**	**Year**	**Recruit.**	**Year**	**Recruit.**
1973	200	1980	1400	1987	4700	1994	500
1974	200	1981	1400	1988	1700	1995	300
1975	600	1982	1300	1989	1100	1996	400
1976	300	1983	2200	1990	2400	1997	200
1977	500	1984	1800	1991	1000	1998	400
1978	600	1985	1800	1992	500	1999	400
1979	700	1986	2800	1993	1000	2000	700

As the instructions to the problem say, we "split stems" and group the data into 500 (million) units (fish) wide. We get the stemplot:

0	22233444
0	5556677
1	001344
1	788
2	24
2	8
3	
3	
4	
4	7

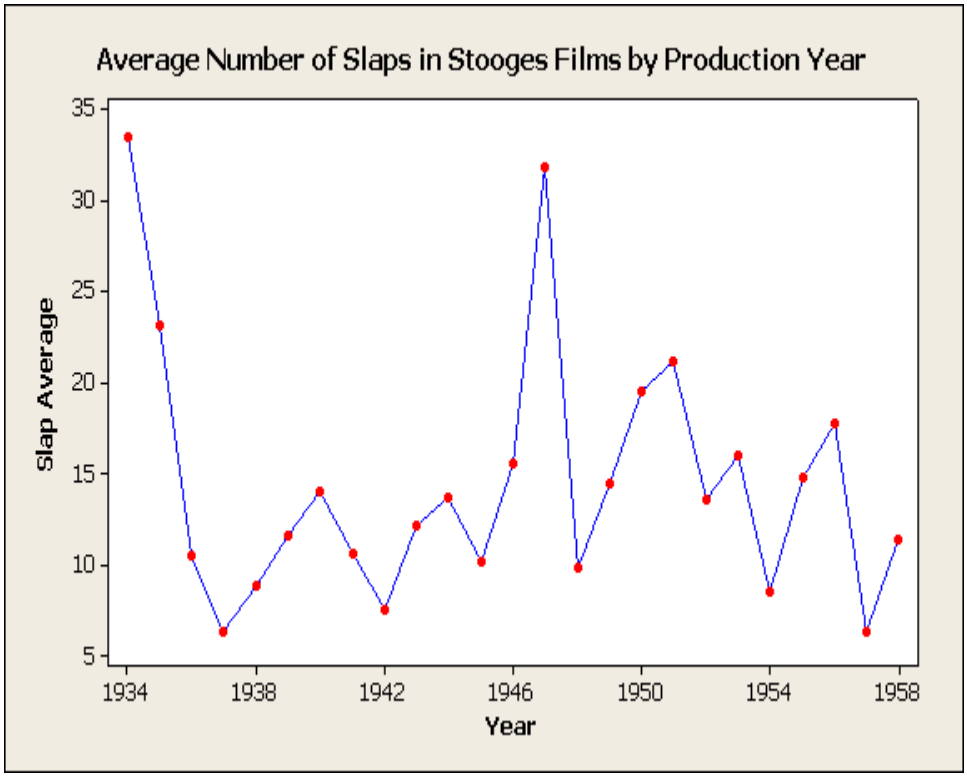
Notice that a stemplot looks like a histogram turned on end.

Time Plots

Definition. A **time plot** of a variable plots each observation against the time at which it was measured. Always put time on the horizontal scale of your plot and the variable you are measuring on the vertical scale. Connecting the data points by lines helps emphasize any change over time.

Example S.1.5. Slapping Time.
Create a time plot of the slaps per film data given in Example S.1.2.

Solution. From Minitab, we get:



Definition. One common overall pattern in a time plot is a **trend**, a long-term upward or downward movement over time.

Example. Exercise 1.43 page 35.

rbg-1-23-2009