Chapter 2

Key Ideas
Frequency Distribution, Relative Frequency Distribution, Cumulative Frequency Distribution, Histogram, Relative Frequency Histogram, Normal Distribution, Dotplot, Stemplot, Pie Chart, Scatterplot, Time-Series Graph,

Section 2-1: Overview
Once you obtain data from a study, it is often useful to put it into a visual context. This allows people to see what is happening in the dataset instead of seeing a lot of numbers. This chapter deals with a variety of ways to display data to make it easier to understand what the results are saying.

Section 2-2: Frequency Distributions
Frequency Distributions are tables that display data according to frequencies (or counts) of how many data values fall into particular intervals, or categories. They are called distributions because they show the way the observations are distributed among the different groups.

- A basic **frequency distribution** lists the data values (groups) along with their corresponding frequencies.
- A **cumulative frequency distribution** can be useful for ordered data (e.g. data arranged in intervals, measurement data, etc.). Instead of reporting frequencies, the recorded values are the sum of all frequencies for values less than and including the current value.
- A **relative frequency distribution** lists the data values along with the percent of all observations belonging to each group. These relative frequencies are calculated by dividing the frequencies for each group by the total number of observations.

Example: Suppose we take a sample of 200 U.S. households and record the number of people living there. We obtain the following:

Number of People	Frequency
1	10
2	50
3	90
4	40
5	6
6	4

Freq. Distribution

Number of People	Cumulative Frequency
1	10
2	60
3	150
4	190
5	196
6	200

Cumulative Freq. Distribution

Number of People	Relative Frequency
1	5%
2	25%
3	45%
4	20%
5	3%
6	2%

Relative Freq. Distribution

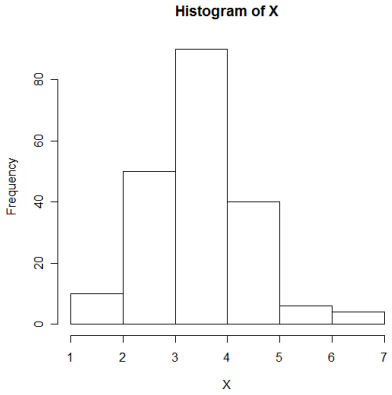
← 10/200

← 50/200

etc.

Section 2-3: Histograms
A **histogram** is a special kind of bar graph that applies to quantitative data (discrete or continuous). The horizontal axis represents the range of data values. The bar height represents the frequency of data values falling within the interval formed by the width of the bar. The bars are also pushed together with no spaces between them.

Example: Number of people in 200 U.S. households (see above).



Note: Here the data values only take on integer values, but we still split the range of values into intervals. In this case, the intervals are [1,2), [2,3), [3,4), etc. Notice that this graph is also close to being bell-shaped. A symmetric, bell-shaped distribution is called a *normal* distribution. These types of distributions will be discussed later.

Also: A **relative frequency histogram** is the same as a regular histogram, except instead of the bar height representing frequency, it now represents the relative frequency (so the y-axis runs from 0 to 1, which is 0% to 100%).

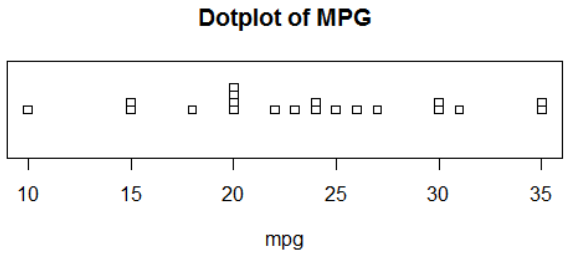
Section 2-4: Statistical Graphs

One drawback to using histograms is that you cannot reconstruct the original data set just by looking at the plot. Here are a few other graphs that allow this to be done.

- **Dotplot**: A dotplot can use either the horizontal or vertical scale to represent the possible data values. Dots are placed above (horizontal) or to the right (vertical) of the line next to the value that observation takes. Dots for repeated data values are stacked on the others.
- **Stemplot (or Stem-and-Leaf Plot)**: Data points are split into a leaf (usually the ones digit) and a stem (the other digits). The plot has a column for stems, and then leaves with a common stem are placed in order to the right of the stem. The result is like a histogram on its side, but the number values of the observations can be read.
- **Scatterplot**: If paired observations (x, y) are taken on each sampled object (e.g. height and weight for each subject), these values can be displayed in a scatterplot. The horizontal axis represents x values, and the vertical axis represents y values. A dot is placed on the graph at the coordinates for each value.
- **Time Series Plot**: A time-series plot is a line graph where the horizontal axis represents time and the vertical axis represents the values of the observations. This is useful (obviously) for data collected over time.

Examples

Dataset #1: Miles per Gallon (MPG) of 20 cars and trucks – 35, 20, 10, 15, 18, 24, 25, 20, 15, 22, 24, 30, 30, 20, 23, 31, 27, 26, 35, 20

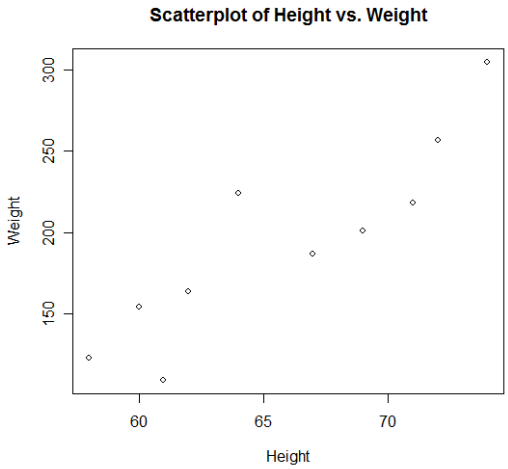


Stemplot of MPG

1 | 0558
2 | 00002344567
3 | 00155

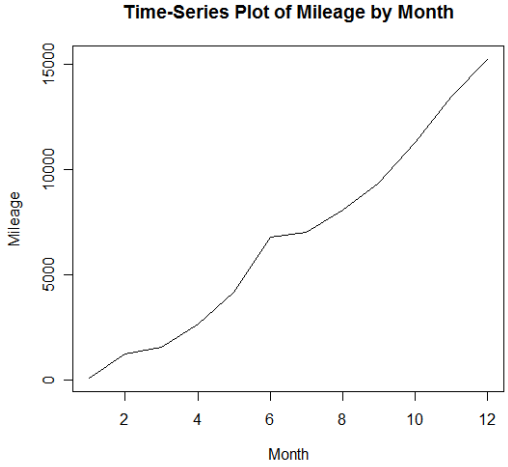
Dataset #2: Height and Weight of 10 People

Height (in.)	Weight (lbs.)
62	164
67	187
74	305
64	224
71	218
69	201
58	123
61	109
60	154
72	257



Dataset #3: My new car's mileage (mi.) over one year

Month	Mileage (mi.)
1	75
2	1236
3	1572
4	2678
5	4203
6	6801
7	7048
8	8103
9	9377
10	11305
11	13501
12	15265



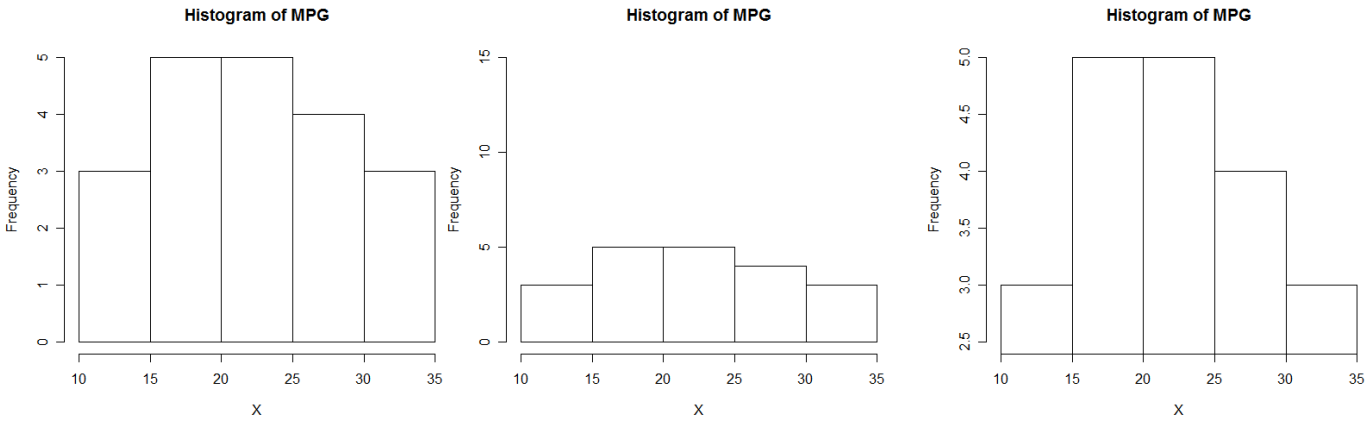
Statistical Graphs for Qualitative Data

There are a couple of graphs that are appropriate for qualitative data that has no natural ordering.

- **Bar Graphs** are like histograms, but the horizontal axis has the name of each category and there are spaces between the bars. Usually, the bars are ordered with the categories in alphabetical order. One variant of a bar graph is called a *Pareto Chart***.** These are bar graphs with the categories ordered by frequency, from largest to smallest.
- **Pie Charts** use a circle with pie wedges drawn that are proportional to the relative frequencies of data values in each of the categories.

The #1 thing to remember when using visual representations of data sets is to represent the data fairly. Scales on the axes should be chosen to fairly display the data without any attempt to create misleading results (see page 62 for further discussion).

Example



Here, we see a fair representation of the data in the 1st histogram. The scale in the 2nd histogram is increased to make differences in the categories look less than they really are. The 3rd histogram has a shortened vertical axis, which makes differences between the categories look much greater than they really are. It is amazing that all 3 plots use the same data set!