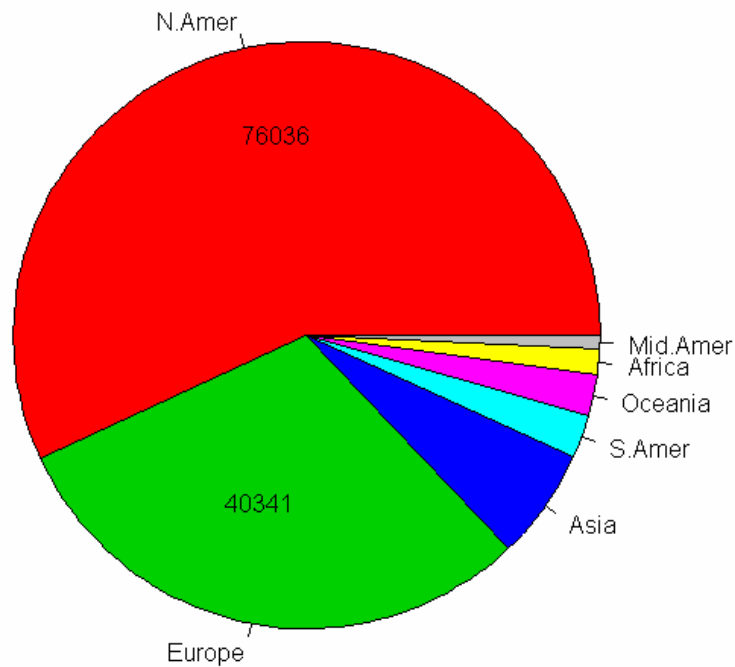


III. GRAPHICAL METHODS

Pie Charts and Bar Charts:

Pie charts and *bar charts* are used for depicting frequencies or relative frequencies. We compare examples of each using the same data.

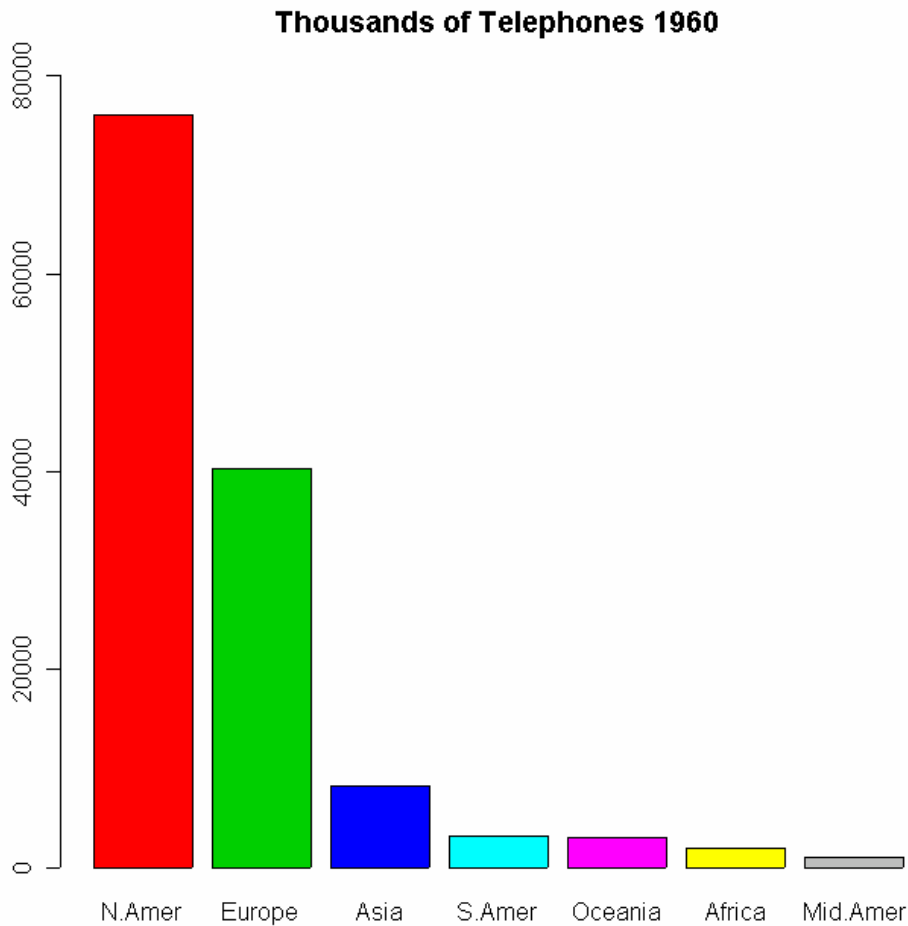
Thousands of Telephones 1960



Sources: AT&T (1961) *The World's Telephones*

R: A language and environment for statistical computing, the R core development team.

This is called a pie chart for obvious reasons. Pie charts are also called *circle graphs*. A bar chart depicting the same data is shown below.



For this example, the bar chart is possibly a better representation of the data for the following reasons:

- In the pie chart, the area corresponding to each region is proportional to the frequency. In the bar chart, the height of the bar is proportional to the frequency. The human eye and brain are better at comparing linear dimensions than areas.
- The bar chart has a vertical axis with a scale of measurement, thereby giving more precise information. In the pie chart, the counts must be placed in or near the areas corresponding to the regions. For some, there is not enough room.

On the other hand, a pie chart may give a more vivid impression of percentages of the total (relative frequencies), although bar charts often have another vertical axis on the right to indicate relative frequencies. From the pie chart above, it is clear that North America had more than half of the world's telephones in 1960. This is not so obvious in the bar chart.

The categories in the bar chart above are arranged from left to right in decreasing order of frequency. Such a bar chart is called a Pareto chart. If the categories are naturally ordered in some other way, their left to right ordering in the bar chart should reflect their natural ordering.

Stemplots (Stem and Leaf Diagrams):

A *stemplot* is a quick way of depicting and tabulating small sets of rounded numerical data. A stemplot can be done by hand. It facilitates calculating the median and other percentiles. The following is a stemplot of the numerical grades of 50 students on a test.

Stem Leaves	Cumulative Frequency
4 7	1
5 448889	7
6 34789	12
7 012234455666888889999	33
8 0022234457799	46
9 0457	50

Each row of a stemplot corresponds to one stem with several leaves attached to it. The stem in this case is the first digit of the grade. The leaf is the second digit. Look at the second row. It tells you that there were six grades in the 50s, namely, 54, 54, 58, 58, 58, and 59. The cumulative frequencies of the stems are listed on the right. They help you find the median. Since there are 50 students, we have to find the 25th and 26th grades from the bottom (which is at the top of the picture). The cumulative frequency for stem 6 is 12 and the cumulative frequency for stem 7 is 33. Therefore, the median is somewhere in the 70s. Counting 13 up from the bottom of the 70s we find that the 25th and 26th grades were both 78. Therefore, the median is 78.

Exercise: The first quartile is 70 and the third quartile is 82.

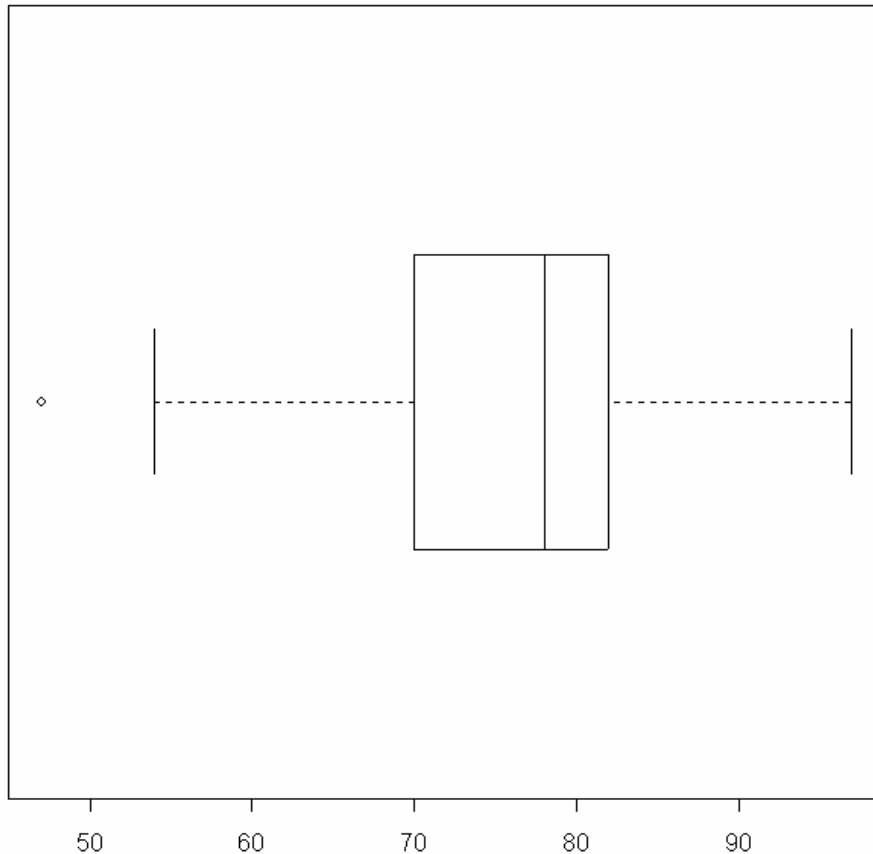
Another feature of a stemplot is that it looks like a horizontal bar chart of the frequencies of the stems. Thus it gives a rough picture of the distribution of values.

Exercise: The variable *Spend* in the teacher pay data file is the state expenditure per pupil on teacher salaries. Construct a stemplot of this data.

Boxplots (Box and Whisker Diagrams):

A *boxplot* is a simple graphical summary of interval data. Below is a computer-generated box plot of the test grades above.

Test Grades



The rectangular middle portion of the diagram is the *box*. The middle line in the box shows the median. The left and right boundaries of the box show the first and third quartiles. The dashed lines extending outward from the box are the *whiskers*. They go out from the quartile to the most extreme data value in that direction which is not more than 1.5 times the *IQR* from the quartile. Any data values that fall outside the ends of the whiskers are simply plotted as points. They are considered outliers. In this case, there is one outlier – the student who made 47 on the test.

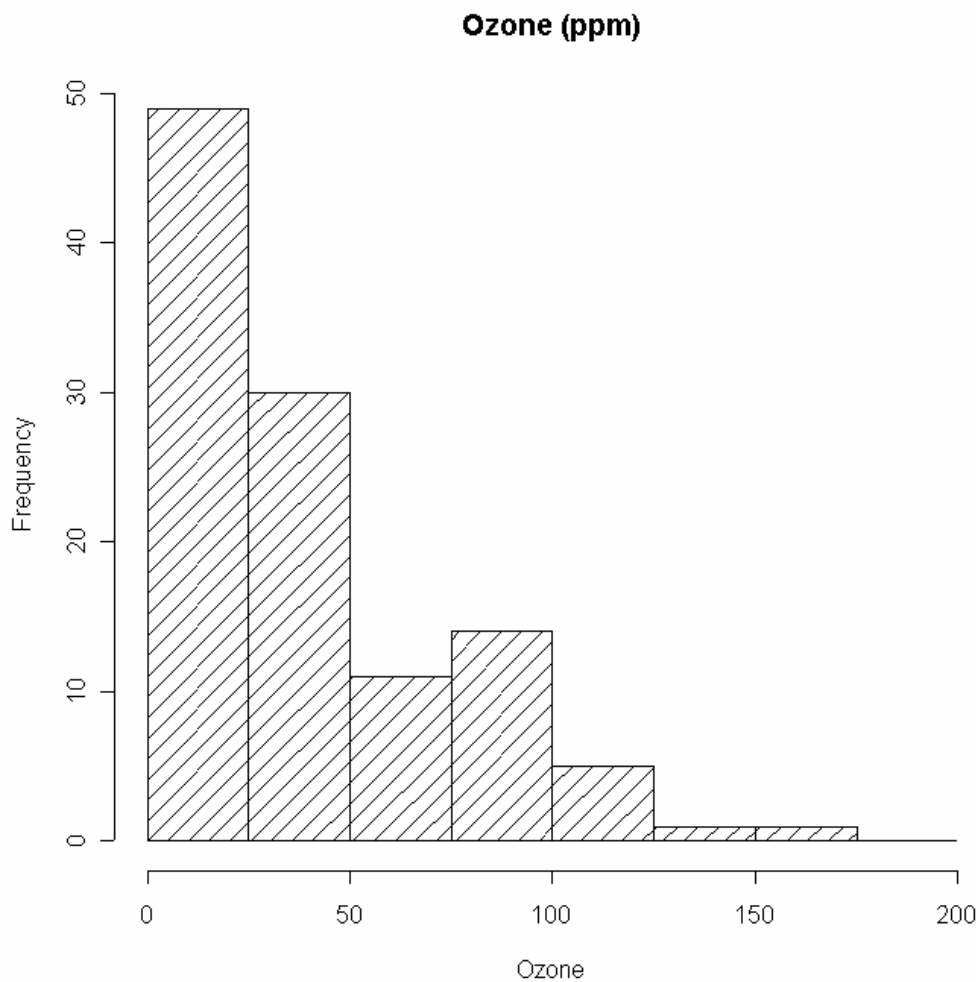
In this box plot, the difference between the third quartile and the median is noticeably smaller than the difference between the median and the first quartile. Also, the left whisker is longer than the right one and there is one extremely small outlier. The distribution of grades appears to be *skewed to the left*. If the boxplot were almost symmetric about its median, we would say that the distribution was not skewed.

Different textbooks and software programs follow different conventions for the length of the whiskers. The one just given is fairly standard.

Exercise: Draw a boxplot of the variable *Spend* in the teacher pay data file.

Histograms:

A histogram looks superficially like a bar chart, but there are important differences. A *histogram* is used for interval or ratio data while a bar chart is used for categorical data. Below is a histogram constructed from 111 daily measurements of ozone concentration in New York City.

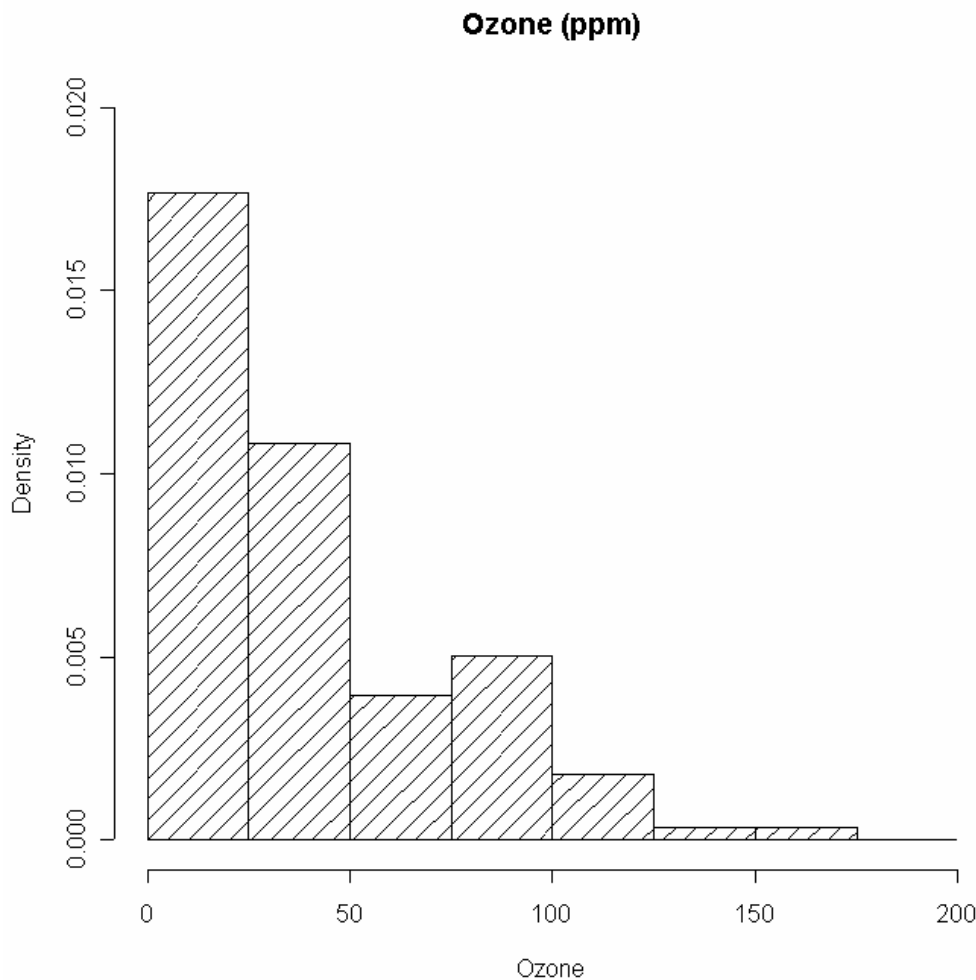


Source: *R: A language and environment for statistical computing*, the R core development team.

These measurements all lie between 0 and 200 parts per million. Notice that the interval between 0 and 200 on the horizontal axis has been subdivided into 8 subintervals of equal length. These are called the *class intervals*, or *bins*. In this case, the class intervals are from 0 to 25, 25 to 50, 50 to 75, etc. The height of the vertical bar above

each class interval is the frequency with which ozone observations occur in that interval. For example, 49 of the measurements are between 0 and 25, 30 of them are between 25 and 50, and so on. A frequency histogram of this type is very similar to a bar chart. As a matter of style, a histogram does not have spaces between the vertical bars, whereas a bar chart for nominal data may have them. (See the example above.)

Another kind of histogram is a *probability histogram*. It subdivides the range of the data on the horizontal axis in the same way. However, the heights of the vertical bars are adjusted so that the *area* of each bar is equal to the relative frequency of its class interval. Below is a probability histogram of the ozone data.



The shape of the figure is the same, but the scale of the vertical axis is different. To find the relative frequency of ozone measurements between 0 and 25 we must find the area of the first vertical bar. The height of the first bar is 0.0177 and therefore its area is

$$0.0177 \times 25 = 0.44.$$

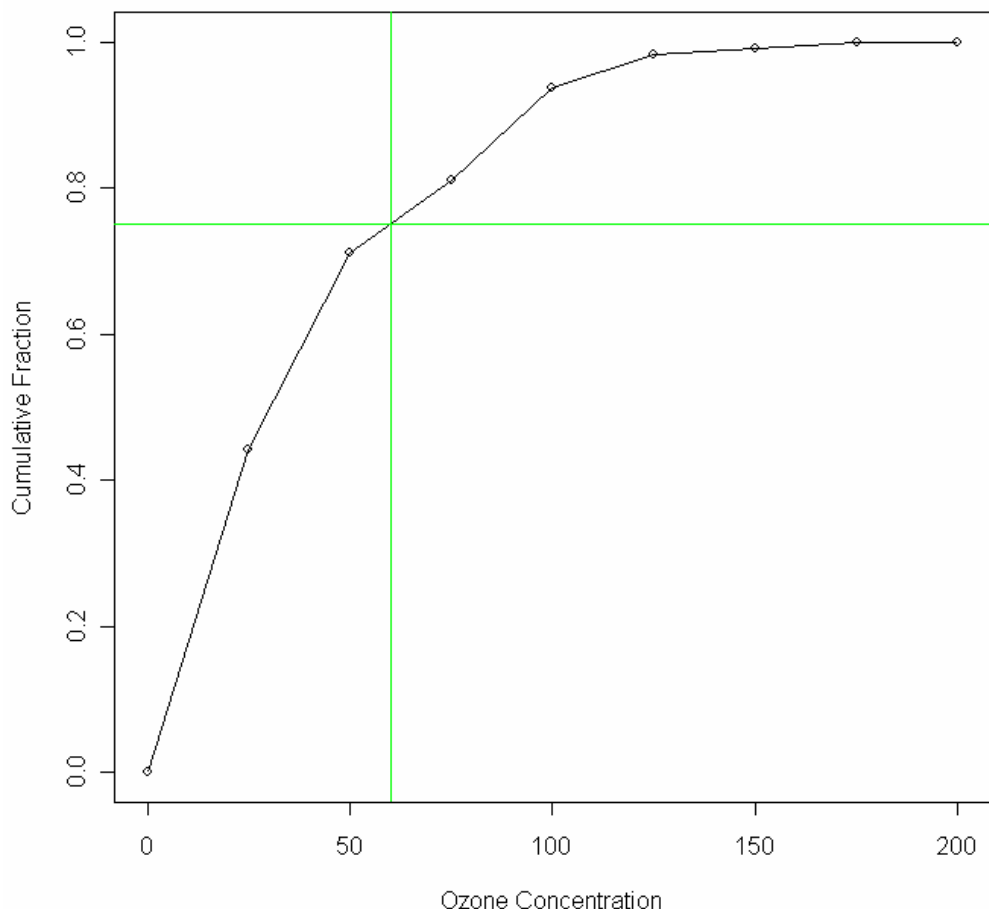
The total area of all the bars in a probability histogram is equal to the sum of the relative frequencies, which is 1.

Probability histograms are not as often used for graphical presentation of data as the ordinary kind because they are harder to understand. However, probability histograms are important because of their relationship to ogives (discussed below) and because they are examples of the essential theoretical concept of a density function.

Exercise: What is the relative frequency of ozone measurements between 75 and 100? What is the approximate relative frequency of ozone measurements between 25 and 75?

Ogives (Cumulative Frequency Polygons):

An *ogive* is a cumulative relative frequency plot for interval or ratio data. It is related to a probability histogram and uses the class intervals from the histogram. To construct an ogive, above the right end-point of each class interval plot the cumulative relative frequency of that interval and all the intervals preceding it. Then join these points by straight line segments. Below is the ogive corresponding to the probability histogram of the ozone data.



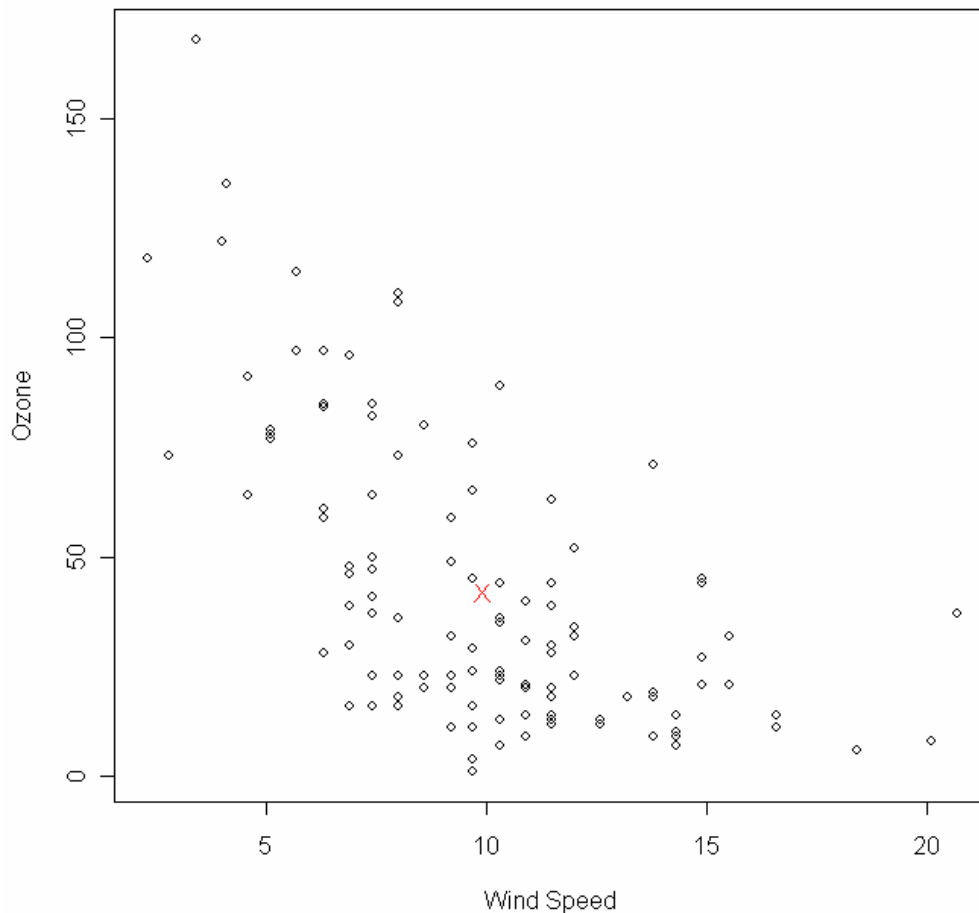
For any given number c on the horizontal scale, the height of the ogive at c is the total area under the probability histogram to the left of c . It is an estimate of the fraction of all the values of the variable that are less than or equal to c . If the data should be lost and the exact fraction could not be calculated, the estimate derived from the ogive could still be used.

In the ogive above, the horizontal line has height 0.75. Therefore, the position where the vertical line meets the horizontal axis is an estimate of the third quartile of the data. The vertical line meets the axis at an ozone concentration of about 60. Therefore, we estimate the third quartile to be 60. In fact, the third quartile of the data is 62, so our estimate is quite accurate.

Exercise: With the ogive plotted above, estimate the median and the 90th percentile of ozone concentrations.

Scatterplots (Scatter Diagrams):

Let x and y denote two variables that are jointly observed. By that we mean that they are defined for the same population. Say that we have observed values of x and y from each of n individuals in the population. Thus we have n pairs of values $(x_1, y_1), \dots, (x_n, y_n)$ for the n individuals. A plot on a rectangular coordinate system of these points is called a *scatterplot* of the data $\{(x_1, y_1), \dots, (x_n, y_n)\}$. Below is a scatterplot of wind speed x and ozone concentration y for 111 days in New York City.

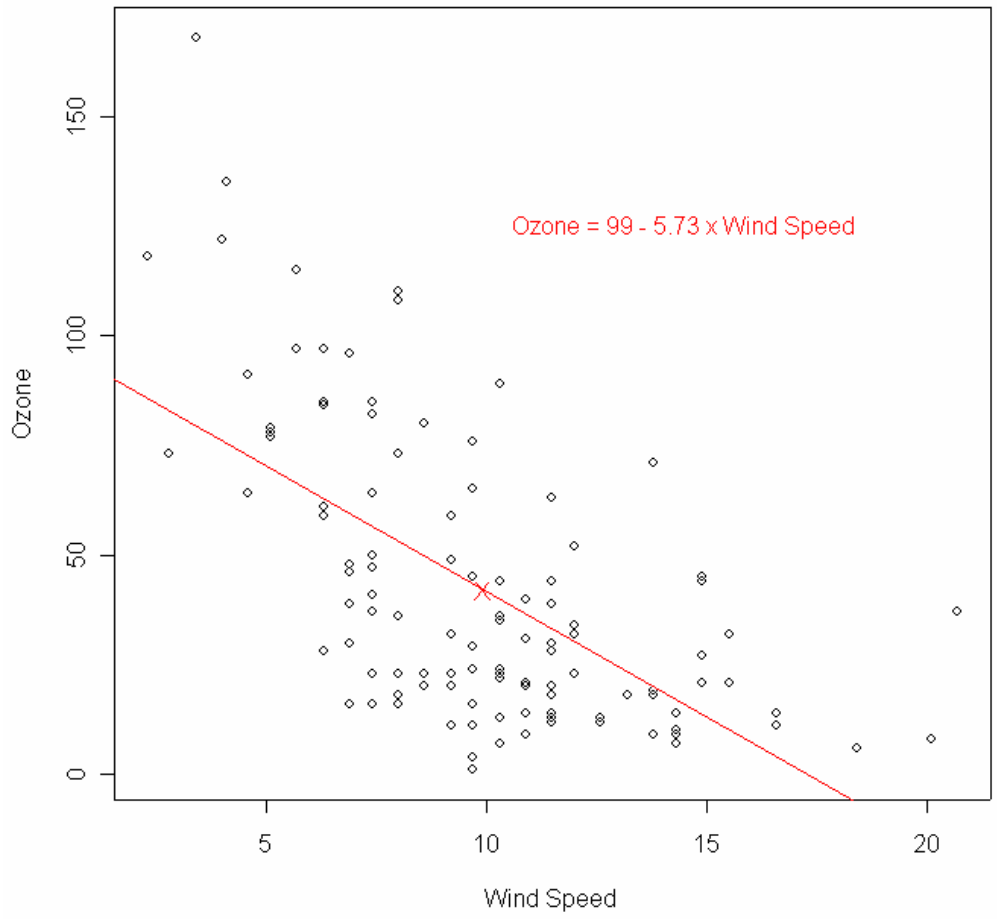


This scatterplot reveals a clear relationship between the two variables. Higher values of wind speed generally go along with lower values of ozone concentration. In this example there is a causal relationship between the two variables. We expect strong winds to disperse concentrations of ozone. In other cases there would be no *a priori* reason to believe in a causal relationship between the variables and we should be very cautious about inferring one.

Sometimes it is imagined that there is an ideal relationship of a simple mathematical form between the variables and because of unexplained random influences the observed values of the variables follow this relationship only approximately. The simplest relationship would be a straight line equation of the form $y = a + bx$, where a and b are constants. Of course, we do not know a and b but we assume that they exist. We could estimate them by drawing a line through the scatterplot that comes as close as possible to fitting the points. The slope of that line would be our estimate of the number b .

Exercise: Draw a line on the scatterplot above that seems to fit the points as nearly as possible and passes through the X near the center of the set of points. Measure the slope of the line. Suppose that tomorrow the wind speed will be 5 mph greater than it is today. By how much would you expect the ozone concentration to increase or decrease?

There is a formal mathematical technique called *least squares* for finding the equation of the line that comes closest to fitting the points in a scatterplot. The details will be discussed later. It can be shown that the line obtained by the method of least squares always passes through the centroid of the data with coordinates (\bar{x}, \bar{y}) . That is the point designated by the X in the scatterplot above. The plot below shows the least squares line and its equation. How does its slope compare to your estimated slope in the exercise above?



Having added the least squares line to the scatterplot, we see that it is implausible that there is a true linear relationship. We see this because most of the points on the extreme left and extreme right are above the line, while most of the points in the middle of the plot are below the line. This indicates that the true relationship, if it exists, is more complicated than the one we have postulated.