CHAPTER 6

DESCRIPTIVE STATISTICS

So far, in this course, we studied probability theory. Now, we will start concentrating on the statistical aspects.

Briefly put, statistics is the science of collecting, organizing, summarizing, analyzing, interpreting, and presenting data

Two important terms in statistics are ***population*** and ***sample***. A ***population*** is a group of observations that includes all of the elements of a set of data. A ***sample*** is any subset of the population. A measurable characteristic of a population is called a ***parameter,*** and a measurable characteristic of a sample is called a ***statistic***. For example, the population average is a parameter whereas a sample average is a statistic.

Usually, the goal of a statistical study is to reach a conclusion about the population. Since typically populations are very large, often, it is not feasible (and even not possible) to compute parameters. Indeed, we may not even have data for every element of the population. For example, if a company's customer service division wanted to learn whether its customers were satisfied, it would not be possible to contact every individual who purchased a product.

Thus, we treat a parameter as a fixed but unknown quantity that has to be ***estimated***. In later chapters, we will learn how to use sample statistics for this purpose. Sample statistics on the other hand (as we shall soon see) can easily be computed. Thus, they are known quantities. However, since they vary from sample to sample (which are chosen at random), they are considered to be random variables.

The branch of statistics that involves collecting, organizing, and summarizing the sample data for subsequent analysis is called ***descriptive statistics***. The set of statistical methods used to make decisions or draw conclusions about populations using the information obtained in samples, is called ***inferential statistics***. We will study descriptive statistics in this chapter, and inferential statistics in the subsequent chapters. As we will see, the passage from descriptive to inferential case needs probability.

In general, descriptive statistics can be classified into three groups:

1) Measures of ***central tendency*** or location

2) Measures of ***variability*** or dispersion

3) Description of the ***shape*** of the distribution.

Which numerical measures will be used to describe the location and the dispersion of the sample data *depends on the shape of the data*. Hence, every descriptive study starts with a **graphical analysis** i.e., with ***plotting the data*** and then goes to **numerical summary.**

Statistical Graphs

The most commonly used statistical graphs for describing the shape of the sample data are the ***histogram*****,** the ***dotplot*****,** and the ***boxplot.***

The Histogram

The first step in getting a histogram is to construct a ***frequency*** or a ***relative frequency distribution*** (***table***). To this end, we divide sample data into smaller subgroups called ***classes***. A class is defined by two numbers: the ***lower class limit*** a and the ***upper class limit*** b. This class contains all data values x such that

$$a \leq x \leq b$$

The difference between two consecutive lower class limits is called the ***class size*** (***width***).

Classes must satisfy the following conditions:

1. Classes must encompass all data values. Thus, the lower limit of the first class must be $\leq$ the smallest data value, and the upper limit of the last class must be $\geq$ the largest data value.
2. There can be no overlaps – classes must be disjoint.
3. There can be no gaps – all values in the data set must belong to a class.
4. Classes should have the same size.
5. We should have a "reasonable" number of classes. The usual interpretation of the term "reasonable" is that the number of classes should be between five and twenty.

The ***frequency*** of a class is the number of elements in that class. Clearly, the sum of all frequencies is equal to the number of data.

The ***relative frequency of a class*** is the frequency of the class divided by the total number of data. Clearly, the sum of all relative frequencies is equal to 1.

EXAMPLES

Example 1.

The $NbOCl_3$ concentration (in gm mole/liter per thousand) measurements obtained in a reactor experiment are as follows:

450	450	470	507	457	605	507	1019	975	858	941	896	778
827	689	557	958	1121	1108	603	582	703	821	937	945	1175
645	757	861	457	529	891	927	1094	648	527	832	951	899

Construct a frequency table.

Class	Frequency
400-499	5
500-599	6
600-699	5
700-799	3
800-899	8
900-999	7
1000-1099	2
1100-1199	3
Total	39

Example 2. For the same data set, construct a relative frequency table.

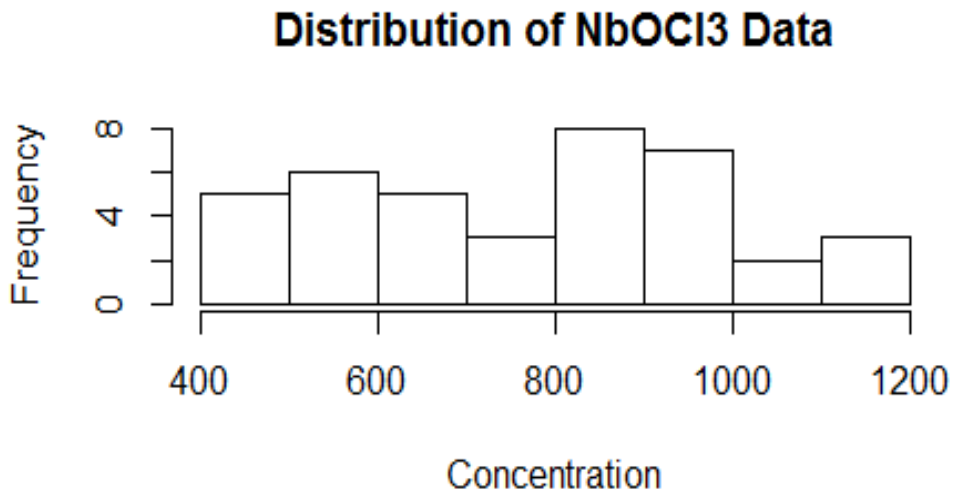
Class	Frequency	Relative Frequency
400-499	5	0.1282
500-599	6	0.1538
600-699	5	0.1282
700-799	3	0.0769
800-899	8	0.2051
900-999	7	0.1795

1000-1099	2	0.0513
1100-1199	3	0.0769
Total	39	1.0000

(Due to round-off the actual sum is 0.9999)

To draw a histogram, we first construct a frequency (relative frequency) distribution of the sample data. We then put the lower class limits on the horizontal axis and the frequencies on the vertical axis, and draw a rectangle over each class with height of rectangle = frequency (relative frequency) of the class. The rectangles must touch each other.

Example. Based on the frequency table given in the previous example, draw a histogram of the $NbOCl_3$ concentration data.



The only problem with a histogram is that we do not see the individual values. For example, looking at the above histogram we know that there are six values between 500 and 600, but we do not know what these values are.

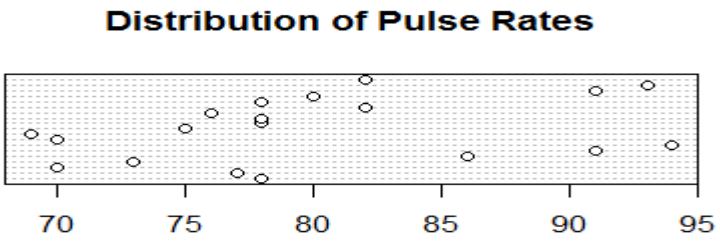
DOT PLOT

A dot plot is a simple histogram-like graph used in statistics for relatively small data sets. Using a number line, we create a scale from the minimum data value to the maximum data value and then place a dot above each observed data point.

Example. Pulse rates of twenty randomly selected adults were as follows:

78	77	70	73	86	91	94	70	69	75
78	78	78	76	82	78	80	91	93	82

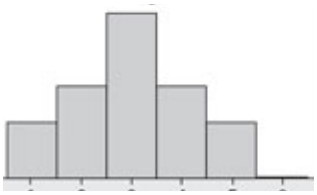
Construct a dot plot.



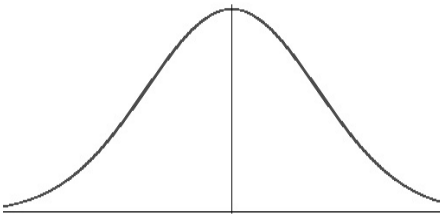
Although a dot plot helps us see individual values, it is limited in the sense that it is only good for data that is not too spread out.

Describing Shape

Now that we can graph the sample data, we can use the graphs to describe the shape of the data. If the histogram looks

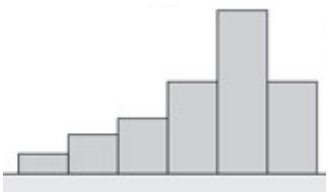


we say the dataset is ***symmetric***. If we approximate the histogram of a symmetric data by a continuous curve, we get

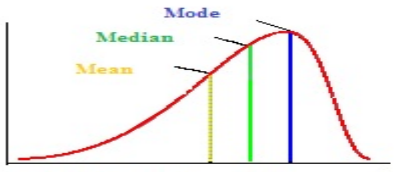


Note that the histogram being symmetric entail the right and the left tails to be identical.

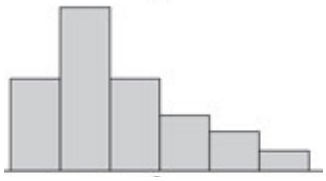
If the left tail is longer than the right tail, we say the dataset is ***left-skewed***.



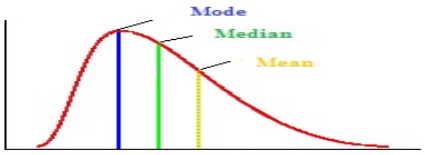
If we approximate the histogram of a left-skewed data by a continuous curve, we get



If the right tail is longer than the left tail, we say the dataset is ***right-skewed***.



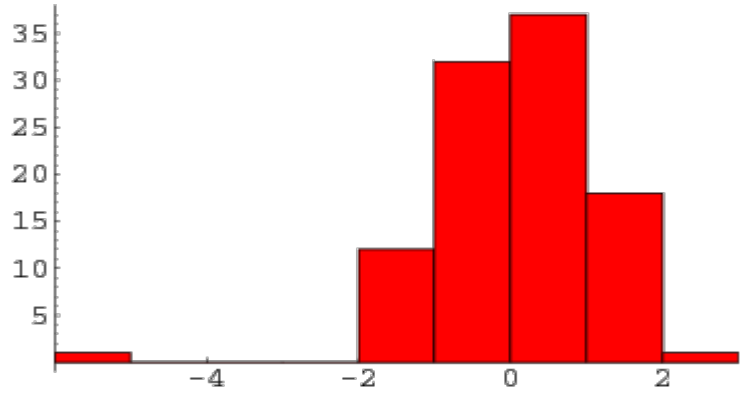
If we approximate the histogram of a right-skewed data by a continuous curve, we get



Outliers

An ***outlier*** is a data point that differs significantly from other observations, in other words, is an observation that lies outside the overall pattern of a distribution. Put differently, outliers are data values that are "too large" or "too small" compared to all the other data values. An outlier may be due to variability in the measurement or it may indicate experimental or a recording error.

Outliers can be spotted in histograms as a rectangle that is far away from all the other rectangles, as the one shown below



We will talk more about outliers later.

Numerical Summaries

Two categories:

Measures of Central Tendency (***Mean***, ***Median***)

Measures of Dispersion (***Variance***, ***Standard Deviation***, ***Range***, ***Interquartile Range***)

Notation:

	Sample	Population
Size	n	N
Mean	$\bar{x}$	μ
Standard Deviation	s	σ
Variance	s^2	σ^2

Let our sample data be $\{x_1, x_2, \ldots, x_n\}$

Sample Mean

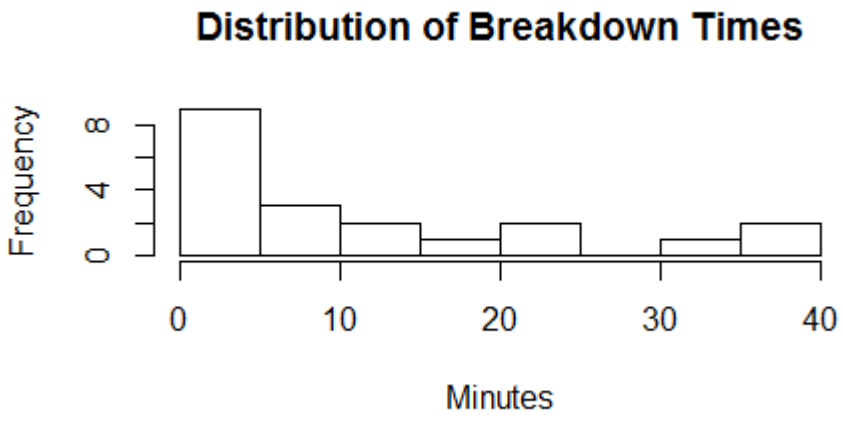
$$\bar{x} = \frac{\sum x_j}{n}$$

As usual, the mean is interpreted as the **c**enter of mass.

Example. The breakdown times of an insulating fluid (in minutes) is as follows:

0.19	0.78	0.96	1.31	2.78	3.16	4.15	4.67	4.85	31.75
12.06	8.27	8.01	7.35	22.52	23.91	17.88	36.71	37.15	14.73

(a) Describe the shape of the data



The data is right-skewed (it has a long right tail).

(b) Find the mean breakdown time.

$$\bar{x} = \frac{\sum x_j}{n} = \frac{0.19 + 0.78 + \cdots + 14.73}{20} = 12.1595$$

Sample Median

Arrange the sample data in ascending order. Median is the value in the middle position. If n is odd, this is the value at the $\left(\frac{n+1}{2}\right)$ position. If n is even, this would be the average of the values in the $\left(\frac{n}{2}\right)$ and $\left(\frac{n}{2}+1\right)$ positions.

The median is interpreted as the middle value: Half of the data values are less than the median and half are greater than the median.

Example. Using the data on the breakdown times of an insulating fluid (in minutes), find the median breakdown time.

We arrange the data in ascending order

0.19 0.78 0.96 1.31 2.78 3.16 4.15 4.67 4.85 7.35

8.01 8.27 12.06 14.73 17.88 22.52 23.91 31.75 36.71 37.15

Taking the average of the values in the tenth and eleventh positions,

$$\text{Median} = \frac{7.35 + 8.01}{2} = 7.68$$

So, 50% of breakdown times are less than 7.68 minutes and 50% are greater than 7.68 minutes.

Comparison of Mean and Median: Mean is sensitive to extreme values whereas median is resistant to extreme values. Mean is always pulled to the side of the longer tail or outlier. So if data are left-skewed, mean < median. If data are symmetric, mean = median. If data are right-skewed, mean >median.

Thus, if data are approximately symmetric with no outliers, we use mean as measure of center. If data are skewed and/or have outliers, we use median as the measure of center.

The Quartiles

The ***first quartile*** q_1 is the median of the data values to the left of the median (median not included) and the ***third quartile*** q_3 is the median of the data values to the right of the median (median not included).

Thus, median divides data into lower 50% and upper 50%, the first quartile into lower 25% and upper 75%, and the third quartile into lower 75% and upper 25%.

Consequently, the median can also be called the second quartile and denoted q_2(and in some books, it is).

Example. Using the data on the breakdown times of an insulating fluid (in minutes), find the median breakdown time.

We arrange the data in ascending order

0.19 0.78 0.96 1.31 2.78 3.16 4.15 4.67 4.85 7.35

8.01 8.27 12.06 14.73 17.88 22.52 23.91 31.75 36.71 37.15

To find q_1, we take the median of

0.19 0.78 0.96 1.31 2.78 3.16 4.15 4.67 4.85 7.35

which is 2.97.

To find q_3, we take the median of

8.01 8.27 12.06 14.73 17.88 22.52 23.91 31.75 36.71 37.15

which is 20.20

Measures of Dispersion

We have four major measures of dispersion: ***range***, ***variance***, ***standard deviation***, and the ***interquartile range***.

The ***range*** of the sample data is simply $\max_j x_j - \min_j x_j$.

It is easy to compute, but it does not give us an idea about dispersion from the central value.

The ***variance***, s^2, remedies this situation:

$$s^2 = \frac{\sum_{j=1}^{n}(\bar{x} - x_j)^2}{n-1}$$

The ***standard deviation*** s is computed as

$$s = \sqrt{s^2}$$

The Interquartile range: IQR

$$IQR = q_3 - q_1$$

Example. The breakdown times of an insulating fluid (in minutes) is as follows:

0.19 0.78 0.96 1.31 2.78 3.16 4.15 4.67 4.85 31.75

12.06 8.27 8.01 7.35 22.52 23.91 17.88 36.71 37.15 14.73

(i) Find the range

Range $= \max_j x_j - \min_j x_j = 37.15 - 0.19 = 36.96$ minutes

(ii) Find the variance

$$s^2 = \frac{\sum_{j=1}^{n}(\bar{x} - x_j)^2}{n-1} = 147.885$$

(i) Find the standard deviation

$$s = \sqrt{s^2} = \sqrt{147.885} = 12.1608$$

(ii) Find the interquartile range

$$IQR = q_3 - q_1 = 20.20 - 2.97 = 17.23$$

Five-Number Summary

A very important numerical summary of the data is called the ***five-number summary***. It gives us

min $\boldsymbol{q_1}$ **Median** $\boldsymbol{q_3}$ **Max**

Example. The breakdown times of an insulating fluid (in minutes) is as follows:

0.19	0.78	0.96	1.31	2.78	3.16	4.15	4.67	4.85	31.75
12.06	8.27	8.01	7.35	22.52	23.91	17.88	36.71	37.15	14.73

Find the five-number summary.

$$\min = 0.19 \;\; q_1 = 2.97 \quad \text{Median} = \; 7.68 \;\; q_3 = \; 20.20 \;\; \max = 37.15$$

Fence Rule for Outliers

There is a simple way of determining whether a data set contains any outliers. To this end, we compute two quantities, the ***lower fence*** and the ***upper fence***.

The ***lower fence*** LF is computed as

$$LF = q_1 - \frac{3}{2}IQR$$

and ***upper fence*** UF is computed as

$$UF = q_3 + \frac{3}{2}IQR$$

Any data value $< LF$ or $> UF$ is an outlier.

Example. For the breakdown times of an insulating fluid (in minutes) data we had

the five-number summary

$$\min = 0.19 \;\; q_1 = 2.97 \quad \text{Median} = \; 7.68 \;\; q_3 = \; 20.20 \;\; \max = 37.15$$

We also know that

$$IQR = q_3 - q_1 = 17.23$$

Thus,

$$LF = q_1 - \frac{3}{2} IQR = -22.85$$

and

$$UF = q_3 + \frac{3}{2} IQR = 46.05$$

Since in this data set there are no values less than the LF or greater than the UF, this data set has no outliers.

Using the Calculator

None of these computations are carried out by hand. We just proceed as follows:

In our calculators we hit the **STAT** key.

We choose the **EDIT** mode (default position).

Then, in column L1 we enter the data values.

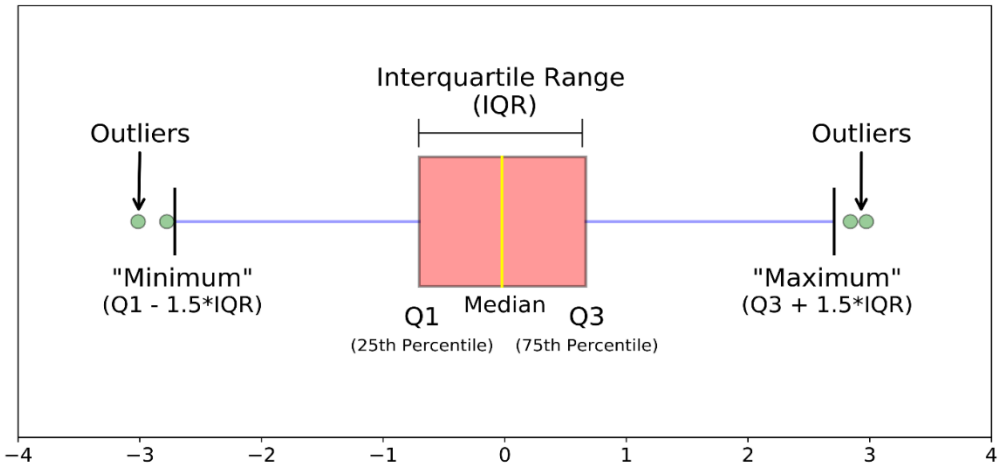
Once we are done, we hit the **STAT** key again.

We now choose **CALC** option and hit **ENTER** twice.

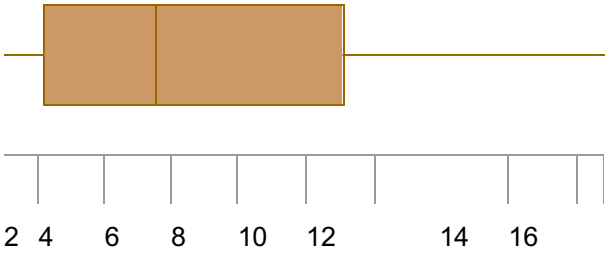
All the required statistical measures will be there.

Boxplot

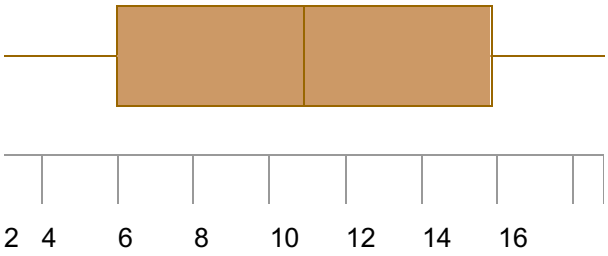
A ***boxplot*** is a graphical manifestation of the five-number summary. We draw a box that extends from q_1 to q_3 (height is immaterial). We mark the position of the median by a vertical line in this box. We then draw a vertical line from the left edge to the minimum value and a vertical line from the right edge to the maximum value. If either the maximum or minimum are outliers, these are then denoted by small circles and the line stops at the value before them.



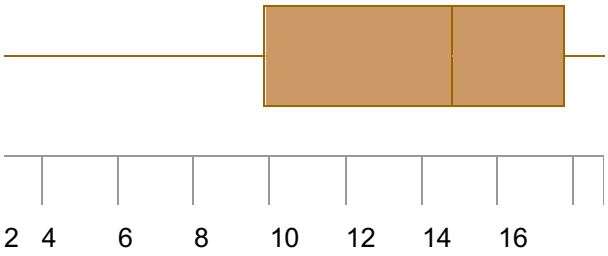
Boxplots are very useful in detecting outliers. However, they also provide information about the shape of a data set. The examples below show some common patterns.



Skewed right



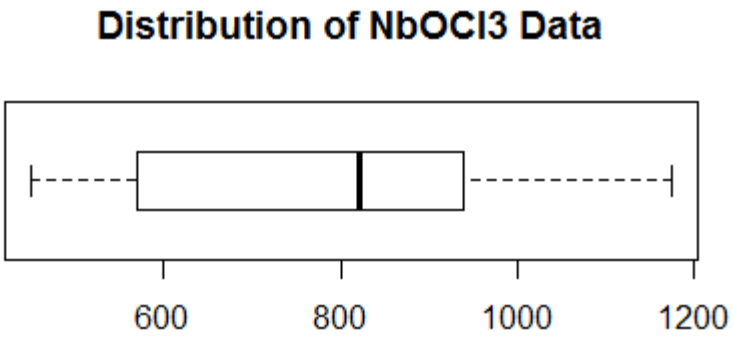
Symmetric

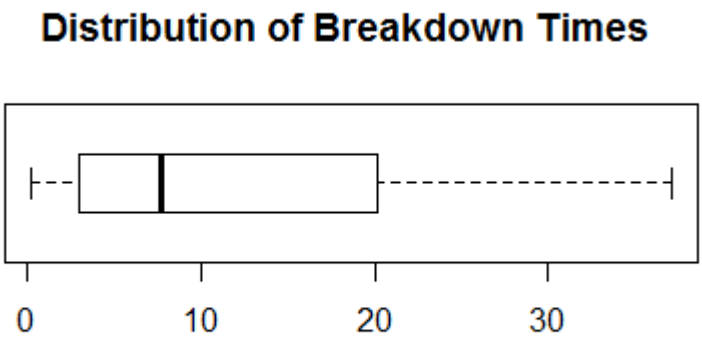


Skewed left

If most of the observations are concentrated on the low end of the scale, the distribution is skewed right; and vice versa. If a distribution is symmetric, the observations will be evenly split at the median.

Now compare them to the boxplots of our examples





How to Use Descriptive Measures

Here is how we incorporate shape in the determination of which measures to use for center and dispersion:

Shape of Sample Data	Measure of Center	Measure of Dispersion
Symmetric with no outliers	Mean	Standard Deviation
Skewed and/or has outliers	Median	IQR

Using R to Obtain Descriptive Graphs and Measures

Selecting Simple Random Samples

Let the number of elements in the population be n, and suppose we want to select a sample of size k with $k \leq n$. We use the ***sample*** function:

>y <- sample ($1{:}\, n$, size = k)

If we want to allow for repetitions, we then write

y <- sample ($1{:}\, n$, size = k, $\text{replace} = \text{TRUE}$)

In either case, when we write

>y

we see the elements of the sample.

Statistical Graphics

Now suppose the data $x_1, \ldots, x_n$ is entered as a vector x:

>x <- $c(x_1, \ldots, x_n)$

To make a frequency table, we write

table (x)

To get a histogram, we write

hist (x).

To give titles, we write **hist** (x, main = "…", xlab = "…"). **Main** gives the title to the entire graph and **xlab** puts a title on the x-axis. We can, if we want fill in the rectangles with any color we want by writing, say,

hist (x, main = "…", xlab = "…", col = "blue")

Here **col** stands for color. We can also make borders a different color:

hist (x, main = "…", xlab = "…", col = "blue", border = "red")

To add the mean of the data set to the histogram we use the ***abline*** function, and write

>**hist** (x, main = "…", xlab = "…", co l= "blue", border = "red")

>**abline** (v=mean(x), col = "…")

To get a dot plot, we write

dotchart (x)

For a boxplot, we write

boxplot (x)

The whiskers extend to $\frac{3}{2}IQR$ in both directions. Points beyond that are plotted as outliers (depicted by a small, empty circle). We write **boxplot** $(x, y, z, ..)$ for side-by-side boxplots.

To have boxplots drawn horizontally, we write

>boxplot(x, horizontal = "TRUE")

Some Functions of Descriptive Statistics

To compute the sum, mean, standard deviation, variance, median, minimum, maximum, and various quantiles we use the following functions:

sum(x)

mean (x)

sd (x)

var (x)

median (x)

min (x)

max (x)

quantile (x, prob = p)

If we want to compute any one of these values for consecutive integers from, say, n to m, we can also write

sum$(n{:}\, m)$

mean $(n{:}\, m)$

sd $(n{:}\, m)$

var $(n{:}\, m)$

median $(n{:}\, m)$

min ($n: m$)

max ($n: m$)

quantile ($n: m$, prob **=** p)

We can also compute several quantiles at one time. For instance, we can write

>quantile (x, probs = $c(p_1, p_2, \dots, p_m)$). So, to compute, the 25th, 50th, and 75th quantiles we write

>quantile (x, probs = c(0.25, 0.50, 0.75))

To compute the trimmed mean, we write

>mean(x, trim = …)

where trim is a number from 0 to 0.5 showing the fraction of observations to be removed from each end.

To compute a weighted average of the numbers $x_1, \dots, x_n$ with weights $w_1, \dots, w_n$, we use the **weighted.mean** function. In this case, we write

> x <- $c(x_1, \dots, x_n)$

>w <- $c(w_1, \dots, w_n)$

>weighted.mean (x, w)

The relation

$$z = (x - \text{mean}(x))/\text{sd}(x)$$

standardizes the entries in x.

The function **summary** (x) gives the five-number summary and the mean. The function ***fivenum***(x) gives just the five-number summary.