Chapter 1: Introduction to Statistics and Data Analysis

Department of Engineering Sciences
Izmir Katip Celebi University

Week 2
2014-2015 Spring

Measures of Location

Measures of location are designed to provide the analyst with some quantitative values of where the center, or some other location, of data is located.

Sample Mean

One obvious and very useful measure is the **sample mean**. The sample mean is simply a numerical average.

Definition

Suppose that the observations in a sample are $x_1, x_2, \ldots, x_n$. The sample mean, denoted by $\bar{x}$, is

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

(Sample) Median

One important measure is the **(sample) median**. The purpose of the (sample) median is to reflect the central tendency of the sample in such a way that it is uninfluenced by extreme values or outliers.

Definition

Given that the observations in a sample are $x_1, x_2, \ldots, x_n$, **arranged in increasing order of magnitude**, the (sample) median, denoted by $\tilde{x}$, is

- the middle observation if n is an odd number,
- the average of the two middle observations if n is an even number.

Alternatively,

$$\tilde{x} = \text{ value at } 0.50(n+1)\text{th } \textbf{ordered} \text{ position}$$

Mode

Another important measure is the **mode**.

Definition

The mode (if one exists) is the most frequent occurring value.

Example. Find the mode(s) if there is (are) any:

a) $3, 2, 1, 2, 1$

b) $1, 3, 5, 4, 6$

c) $1, 3, 7, 3, 2$

d) $3, 7, 5, 5, 3, 7$

Quartiles

Quartiles are descriptive measures that separate large data sets into four quarters:

First quartile, Q_1, separates approximately the smallest 25% of the data from the rest of the data.
Second quartile, Q_2, separates approximately the smallest 50% of the data from the rest of the data and is, actually, the median.
Third quartile, Q_3, separates approximately the smallest 75% of the data from the rest of the data.

$$Q_1 = \text{value at } 0.25(n+1)\text{th } \textbf{ordered} \text{ position}$$

$$Q_2 = \text{value at } 0.50(n+1)\text{th } \textbf{ordered} \text{ position}$$

$$Q_3 = \text{value at } 0.75(n+1)\text{th } \textbf{ordered} \text{ position}$$

Five-Number Summary

The five-number summary refers to the five descriptive measures:

$$\text{minimum} < Q_1 < Q_2 = \tilde{x} < Q_3 < \text{maximum}$$

Note: Do not forget to sort the data (from the smallest to the largest) while calculating elements of five-number summary!

Example. Given the following data:

$$60, 84, 65, 67, 75, 72, 80, 85, 63, 82, 70, 75$$

a) Find the mean.

b) Find the median.

c) Find the mode(s) if there is (are) any.

d) Construct the five-number summary for the given data.

Note: The mean and the median can be quite different from each other! It may be of interest to the reader with an engineering background that the sample mean is the centroid of the data in a sample.

Clearly, the mean is influenced considerably by the presence of the extreme observations (outliers) whereas the median places emphasis on the true "center" of the data set.

In future chapters, the basis for the computation of $\bar{x}$ is that of an **estimate** of the **population mean**. As we indicated earlier, the purpose of statistical inference is to draw conclusions about population characteristics or **parameters** and **estimation** is a very important feature of statistical inference.

Other Measures of Locations

There are several other methods of quantifying the center of location of the data in the sample. We will not deal with them at this point. It is instructive to discuss one class of estimators, namely the class of **trimmed means**.

A trimmed mean is computed by "trimming away" a certain percent of both the largest and the smallest set of values. For example, the 10% trimmed mean is found by eliminating the largest 10% and smallest 10% and computing the average of the remaining values.

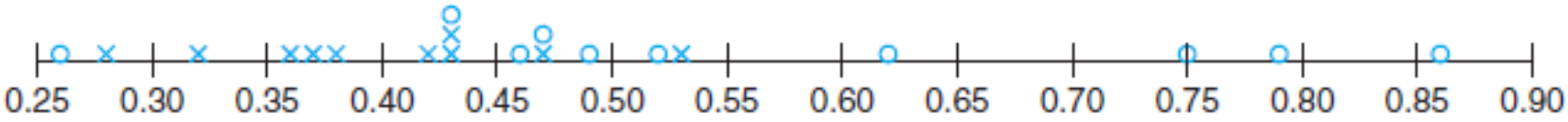
The trimmed mean is, of course, more insensitive to outliers than the sample mean but not as insensitive as the median. On the other hand, the trimmed mean approach makes use of more information than the sample median. Note that the sample median is, indeed, a special case of the trimmed mean in which all of the sample data are eliminated apart from the middle one or two observations.

Example. The following associated with a study conducted at the Virginia Polytechnic Institute and State University on the development of a relationship between the roots of trees and the action of a fungus. Minerals are transferred from the fungus to the trees and sugars from the trees to the fungus. Two samples of 10 northern red oak seedlings were planted in a greenhouse, one containing seedlings treated with nitrogen and the other containing seedlings with no nitrogen. All other environmental conditions were held constant. All seedlings contained the fungus Pisolithus tinctorus. The stem weights in grams were recorded after the end of 140 days. The data are given in the following table:

No Nitrogen	Nitrogen
0.32	0.26
0.53	0.43
0.28	0.47
0.37	0.49
0.47	0.52
0.43	0.75
0.36	0.79
0.42	0.86
0.38	0.62
0.43	0.46

a) Calculate the sample mean and median for the given data.

b) Compute the 10% trimmed mean for both samples.

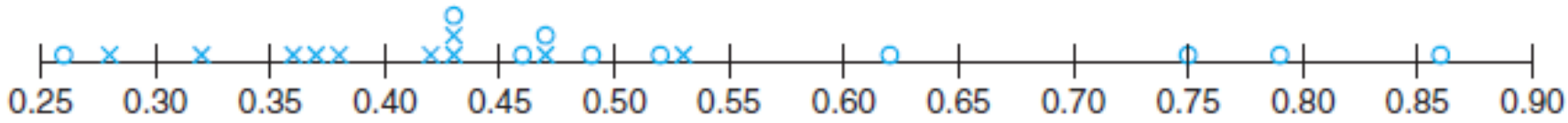


A dot plot of stem weight data

Measures of Variability

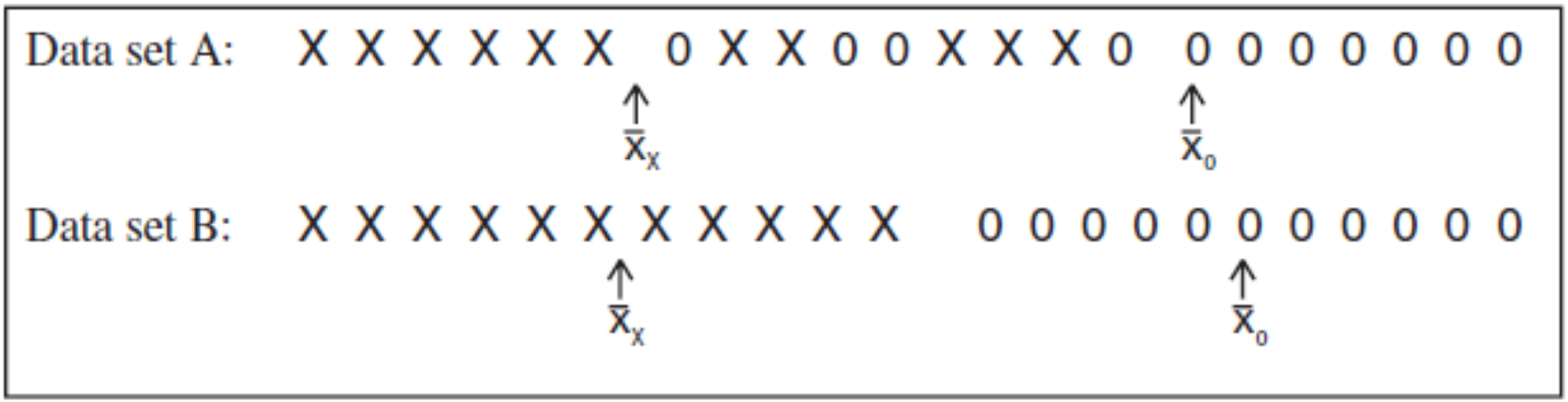
Sample variability plays an important role in data analysis. Process and product variability is a fact of life in engineering and scientific systems: The control or reduction of process variability is often a source of major difficulty. More and more process engineers and managers are learning that product quality and, as a result, profits derived from manufactured products are very much a function of **process variability**.

Even in small data analysis problems, the success of a particular statistical method may depend on the magnitude of the variability among the observations in the sample. Measures of location in a sample do not provide a proper summary of the nature of a data set. For instance, in the previous example, we cannot conclude that the use of nitrogen enhances growth without taking sample variability into account.



It is clear from the figure that variability among the no-nitrogen observations (x) and variability among the nitrogen observations (o) are certainly of some consequence. In fact, it appears that the variability within the nitrogen sample is larger than that of the no-nitrogen sample.

As another example, contrast the following two data sets:



Each contains two samples and the difference in the means is roughly the same for the two samples, but data set *B* seems to provide a much sharper contrast between the two populations from which the samples were taken. If the purpose of such an experiment is to detect differences between the two populations, the task is accomplished in the case of data set *B*. However, in data set *A* the large variability within the two samples creates difficulty. In fact, it is not clear that there is a distinction between the two populations.

Just as there are many measures of central tendency or location, there are many measures of spread or variability.

(Sample) Range and Interquartile Range

The simplest measure of variability is the (sample) range. We again let $x_1, x_2, \ldots, x_n$ denote sample values. Then

$$\text{(Sample) Range} = \max(x_1, x_2, \ldots, x_n) - \min(x_1, x_2, \ldots, x_n).$$

Another measure of variability is the interquartile range (IQR), which measures the spread in the middle 50% of the data, that is,

$$\text{Interquartile Range (IQR)} = Q_3 - Q_1.$$

Sample Variance and Sample Standard Deviation

The sample measures of spread that are used most often are the **sample variance** and the **sample standard deviation**.

Definition

The sample variance, denoted by s^2, is the sum of the squared differences between each observation and the sample mean ($\bar{x}$) divided by the sample size minus 1:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

The quantity $n-1$ is often called the **degrees of freedom associated with the variance** estimate.

Definition

The sample standard deviation, denoted by s, is the positive square root of s^2, that is,

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}.$$

Example. Calculate the variance and standard deviation of the following sample data:

$$3, 0, -2, -1, 5, 10$$

Example. An engineer is interested in testing the "bias" in a pH meter. Data are collected on the meter by measuring the pH of a neutral substance (pH$=$ 7.0). A sample of size 10 is taken, with results given by

7.07 7.00 7.10 6.97 7.00 7.03 7.01 7.01 6.98 7.08

Find the sample standard deviation.

Example. Calculate the sample standard deviations for no-nitrogen and nitrogen groups in the stem weights example and comment on your findings.

Discrete and Continuous Data

Statistical inference through the analysis of observational studies or designed experiments is used in many scientific areas. The data gathered may be **discrete** or **continuous**, depending on the area of application. For example, a chemical engineer may be interested in conducting an experiment that will lead to conditions where amount is maximized. Here, of course, the amount may be in percent or grams/pound, measured on a continuum. On the other hand, a toxicologist conducting a combination drug experiment may encounter data that are binary in nature (i.e., the patient either responds or does not).

Great distinctions are made between discrete and continuous data in the probability theory that allow us to draw statistical inferences. Often applications of statistical inference are found when the data are count data. For example, an engineer may be interested in studying the number of radioactive particles passing through a counter in, say, 1 millisecond. Personnel responsible for the efficiency of a port facility may be interested in the properties of the number of oil tankers arriving each day at a certain port city.

Special attention should be paid to some details associated with binary data. Applications requiring statistical analysis of binary data are voluminous. Often the measure that is used in the analysis is the **sample proportion**. Obviously the binary situation involves two categories. If there are n units involved in the data and x is defined as the number that fall into category 1, then $n - x$ fall into category 2. Thus, $\frac{x}{n}$ is the sample proportion in category 1, and $1 - \frac{x}{n}$ is the sample proportion in category 2.

For example, in a biomedical application, 50 patients may represent the sample units, and if 20 out of 50 experienced an improvement in a stomach ailment (common to all 50) after all were given the drug, then $\frac{20}{50} = 0.4$ is the sample proportion for which the drug was a success and $1 - 0.4 = 0.6$ is the sample proportion for which the drug was not successful.

Actually the basic numerical measurement for binary data is generally denoted by either 0 or 1. For example, in our medical example, a successful result is denoted by a 1 and a nonsuccess a 0. As a result, the sample proportion is actually a sample mean of the ones and zeros. For the successful category,

$$\frac{x_1 + x_2 + \cdots + x_{50}}{50} = \frac{1 + 1 + 0 + \cdots + 0 + 1}{50} = \frac{20}{50} = 0.4.$$

Statistical Modeling and Graphical Diagnostics

Obviously, the user of statistical methods cannot generate sufficient information or experimental data to characterize the population totally. But sets of data are often used to learn about certain properties of the population. Scientists and engineers are accustomed to dealing with data sets. The importance of *characterizing* or *summarizing* the nature of collections of data should be obvious. Often a summary of a collection of data via a graphical display can provide insight regarding the system from which the data were taken.

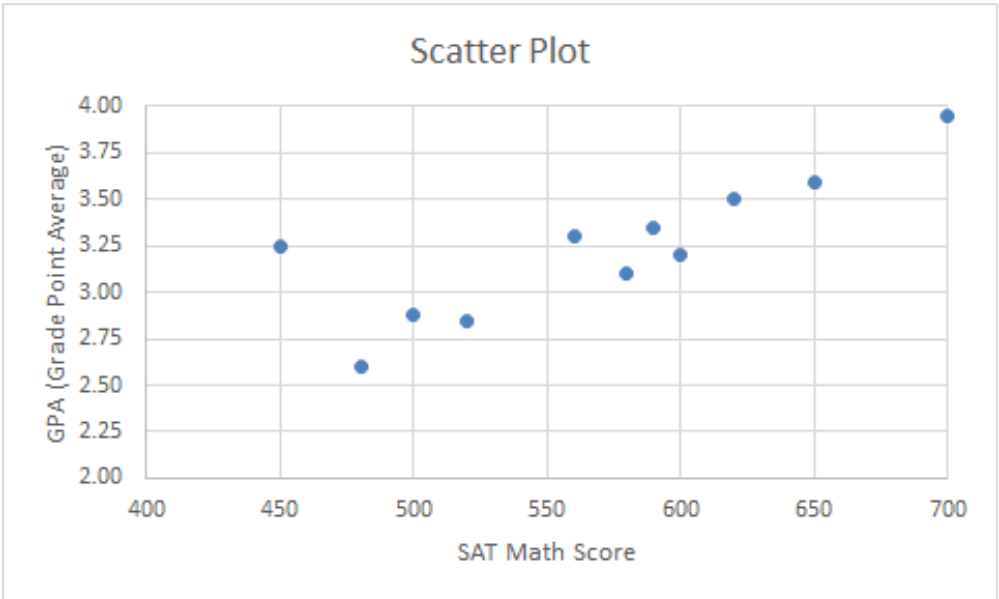
Scatter Plot

A **scatter plot** can be prepared by locating one point for each pair of **two variables** that represent an observation in the data set. The scatter plot provides a picture of the data including:

1. the range of each variable,
2. the pattern of values over the range,
3. a suggestion as to a possible relationship between the two variables, and
4. an indication of outliers (extreme points).

In the following we are given the data and the corresponding scatter plot

SAT Math Score	GPA
450	3.25
480	2.60
500	2.88
520	2.85
560	3.30
580	3.10
590	3.35
600	3.20
620	3.50
650	3.59
700	3.95



Stem-and-Leaf Plot

Statistical data, generated in large masses, can be very useful for studying the behavior of the distribution if presented in a combined tabular and graphic display called a **stem-and-leaf plot** or a **stem-and-leaf-display**.

To illustrate the construction of a stem-and-leaf plot, consider the following data which specifies the "life" of 40 similar car batteries recorded to the nearest tenth of a year:

Car Battery Life

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

The batteries are guaranteed to last 3 years.

We first split each observation into two parts consisting of a stem and a leaf such that the stem represents the digit preceding the decimal and the leaf corresponds to the decimal part of the number. In other words, for the number 3.7, the digit 3 is designated the stem and the digit 7 is the leaf.

We list vertically the four stems 1, 2, 3, and 4 for our data on the left side; the leaves are recorded on the right side opposite the appropriate stem value. The number of leaves recorded opposite each stem is summarized under the frequency column.

Stem-and-Leaf Plot of Battery Life

Stem	Leaf	Frequency
1	6 9	2
2	2 5 6 6 9	5
3	0 0 1 1 1 1 2 2 2 3 3 3 4 4 4 5 5 6 7 7 7 8 8 9 9	25
4	1 1 2 3 4 5 7 7	8
		40

Key: 1|6 stands for 1.6

That stem-and-leaf plot contains only four stems and consequently does not provide an adequate picture of the distribution. To remedy this problem, we need to increase the number of stems in our plot. One simple way to accomplish this is to write each stem value twice as lower (L) and upper (U) and then record the leaves 0, 1, 2, 3, and 4 opposite the appropriate stem value where it appears for the first time, and the leaves 5, 6, 7, 8, and 9 opposite this same stem value where it appears for the second time. This modified double-stem-and-leaf plot is

Double-Stem-and-Leaf Plot of Battery Life

Stem	Leaf	Frequency
1U	6 9	2
2L	2	1
2U	5 6 6 9	4
3L	0 0 1 1 1 1 2 2 2 3 3 3 4 4 4	15
3U	5 5 6 7 7 7 8 8 9 9	10
4L	1 1 2 3 4	5
4U	5 7 7	3
		40

Key: 1U|6 stands for 1.6

In any given problem, we must decide on the appropriate stem values. This decision is made somewhat arbitrarily, although we are guided by the size of our sample. Usually, we choose between 5 and 20 stems. The smaller the number of data available, the smaller is our choice for the number of stems.

For example, if the data consist of numbers from 1 to 21 representing the number of people in a cafeteria line on 40 randomly selected workdays and we choose a double-stem-and-leaf plot, the stems will be 0L, 0U, 1L, 1U, and 2L so that the smallest observation 1 has stem 0L and leaf 1, the number 18 has stem 1U and leaf 8, and the largest observation 21 has stem 2L and leaf 1.

On the other hand, if the data consist of numbers from \$18800 to \$19600 representing the best possible deals on 100 new automobiles from a certain dealership and we choose a single-stem-and-leaf plot, the stems will be $188, 189, 190, \ldots, 196$ and the leaves will now each contain two digits. A car that sold for \$19385 would have a stem value of 193 and the two-digit leaf 85.

Decimal points in the data are generally ignored when all the digits to the right of the decimal represent the leaf such was the case in the example above. However, if the data consist of numbers ranging from 21.8 to 74.9, we might choose the digits 2, 3, 4, 5, 6, and 7 as our stems so that a number such as 48.3 would have a stem value of 4 and a leaf of 8.3.

Note: In order to avoid any confusion, always use a "Key" within your stem-and-leaf display!

Example. Construct a stem-and-leaf display for the nicotine content measured in a random sample of 40 cigarettes. The data are displayed in

Nicotine Data

1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97
0.85	1.24	1.58	2.03	1.70	2.17	2.55	2.11
1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.67	1.37
1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69

Frequency Distribution Table and Histogram

Dividing each class frequency by the total number of observations, we obtain the proportion of the set of observations in each of the classes. A table listing relative frequencies is called a **relative frequency distribution**. The relative frequency distribution for data

Car Battery Life

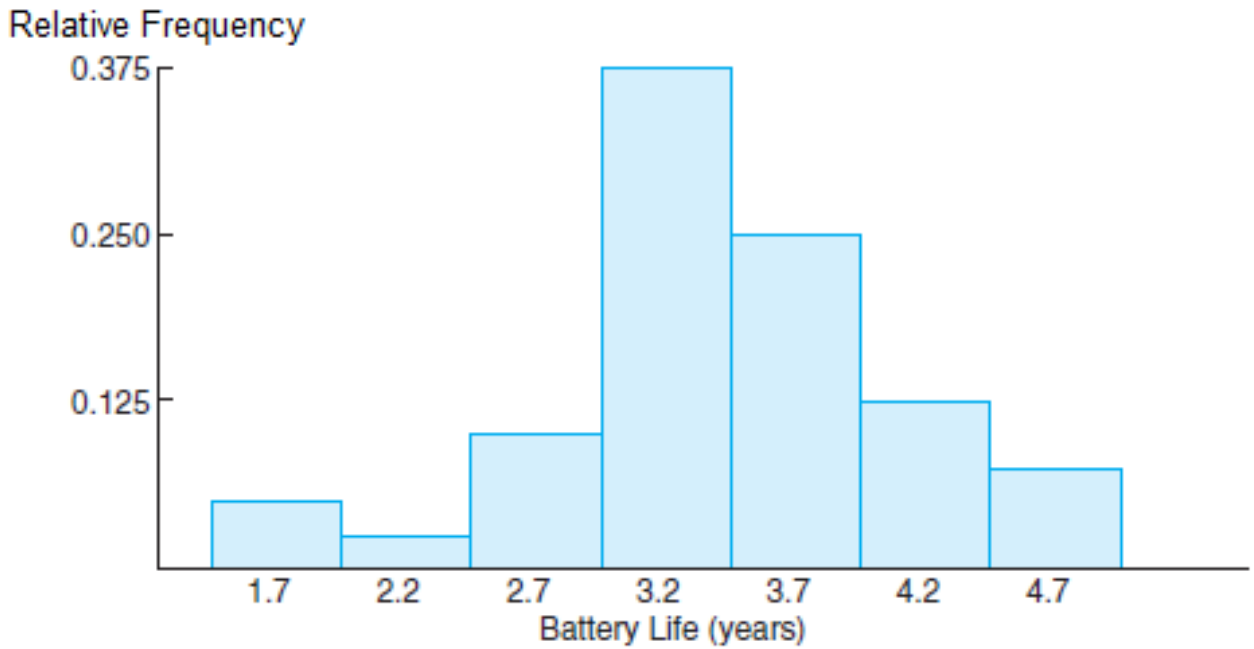
2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6
3.4	1.6	3.1	3.3	3.8	3.1	4.7	3.7
2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1
3.3	3.1	3.7	4.4	3.2	4.1	1.9	3.4
4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5

showing the midpoint of each class interval, is given in

Relative Frequency Distribution of Battery Life

Class Interval	Class Midpoint	Frequency (f)	Relative Frequency
$1.5 - 1.9$	1.7	2	$\frac{2}{40} = 0.050$
$2.0 - 2.4$	2.2	1	0.025
$2.5 - 2.9$	2.7	4	0.100
$3.0 - 3.4$	3.2	15	0.375
$3.5 - 3.9$	3.7	10	0.250
$4.0 - 4.4$	4.2	5	0.125
$4.5 - 4.9$	4.7	3	0.075
		40	1.000

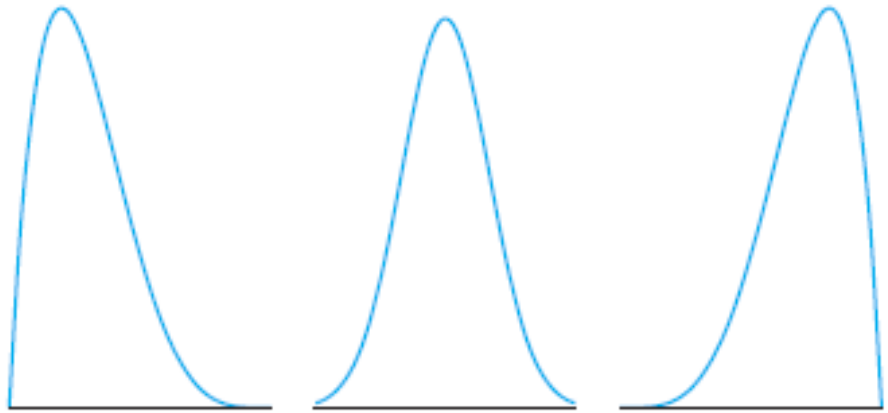
The information provided by a relative frequency distribution in tabular form is easier to grasp if presented **graphically**. Using the midpoint of each interval and the corresponding relative frequency, we construct a **relative frequency histogram**:



Symmetry of Distributions

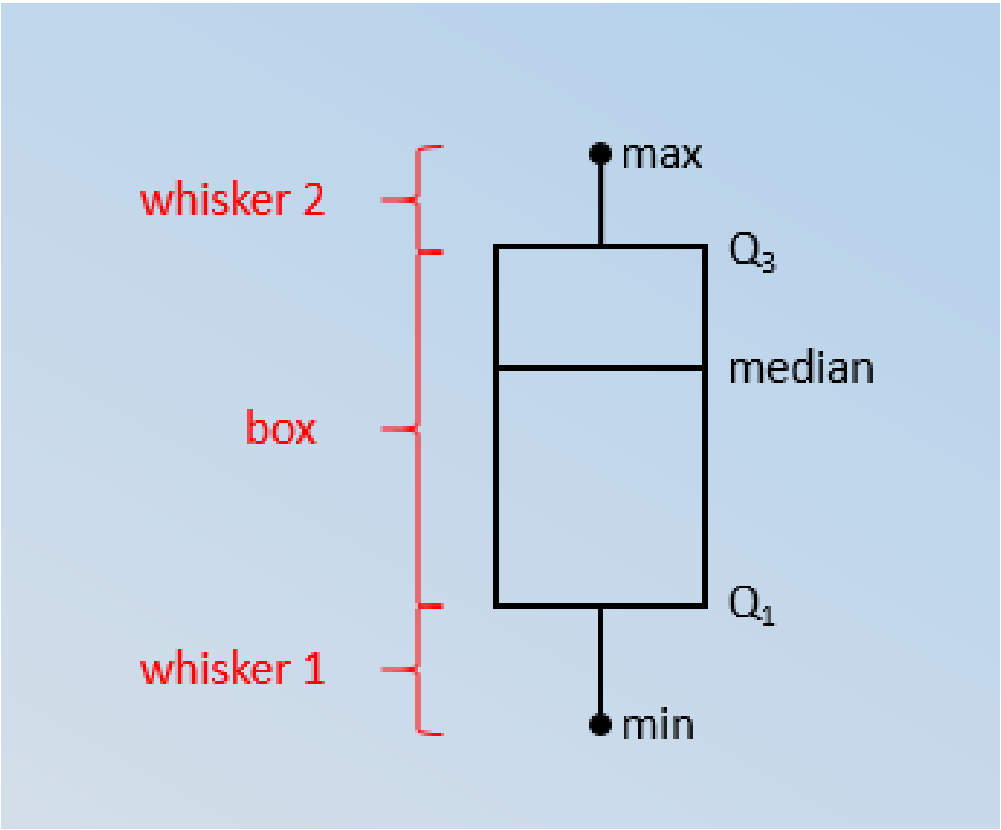
A distribution is said to be **symmetric** if it can be folded along a vertical axis so that the two sides coincide. A distribution that lacks symmetry with respect to a vertical axis is said to be **skewed**.

The distribution in the first illustration is said to be **right-skewed** since it has a long right tail and a much shorter left tail. In the second, we see that the distribution is **symmetric**, while in the third it is **left-skewed**.



Box-and-Whisker Plot or Box Plot

A **box-and-whisker plot** is a graph that describes the shape of a distribution in terms of the **five-number summary**.



Example. The following table gives the weekly sales (in hundreds of dollars) from a random sample of 10 weekdays from two different locations of the same cafeteria.

Location-1:	6	8	10	12	14	9	11	7	13	11
Location-2:	1	19	2	18	11	10	3	17	4	17

Find mean, median, mode, range, IQR for each location and graph the data with a box-and-whisker plot.