This example suggests the need for what follows in Sections 1.3 and 1.4, namely, descriptive statistics that indicate measures of center of location in a set of data, and those that measure variability.

1.3 Measures of Location: The Sample Mean and Median

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. In Example 1.2, it appears as if the center of the nitrogen sample clearly exceeds that of the no-nitrogen sample. One obvious and very useful measure is the **sample mean**. The mean is simply a numerical average.

Definition 1.1:

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

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

There are other measures of central tendency that are discussed in detail in future chapters. 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 1.2:

Given that the observations in a sample are $x_1, x_2, \ldots, x_n$, arranged in **increasing order** of magnitude, the sample median is

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{if } n \text{ is odd,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}), & \text{if } n \text{ is even.} \end{cases}$$

As an example, suppose the data set is the following: 1.7, 2.2, 3.9, 3.11, and 14.7. The sample mean and median are, respectively,

$$\bar{x} = 5.12, \quad \tilde{x} = 3.9.$$

Clearly, the mean is influenced considerably by the presence of the extreme observation, 14.7, whereas the median places emphasis on the true "center" of the data set. In the case of the two-sample data set of Example 1.2, the two measures of central tendency for the individual samples are

$$\begin{aligned} \bar{x} \text{ (no nitrogen)} &= 0.399 \text{ gram}, \\ \tilde{x} \text{ (no nitrogen)} &= \frac{0.38 + 0.42}{2} = 0.400 \text{ gram}, \\ \bar{x} \text{ (nitrogen)} &= 0.565 \text{ gram}, \\ \tilde{x} \text{ (nitrogen)} &= \frac{0.49 + 0.52}{2} = 0.505 \text{ gram}. \end{aligned}$$

Clearly there is a difference in concept between the mean and median. 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. In a sense, it is the point at which a fulcrum can be placed to balance a system of "weights" which are the locations of the individual data. This is shown in Figure 1.4 with regard to the with-nitrogen sample.

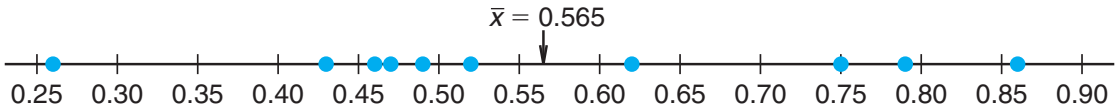


Figure 1.4: Sample mean as a centroid of the with-nitrogen stem weight.

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.

The median and mean can be quite different from each other. Note, however, that in the case of the stem weight data the sample mean value for no-nitrogen is quite similar to the median value.

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. For the most part, alternatives to the sample mean are designed to produce values that represent compromises between the mean and the median. Rarely do we make use of these other measures. However, 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. For example, in the case of the stem weight data, we would eliminate the largest and smallest since the sample size is 10 for each sample. So for the without-nitrogen group the 10% trimmed mean is given by

$$\bar{x}_{\text{tr}(10)} = \frac{0.32 + 0.37 + 0.47 + 0.43 + 0.36 + 0.42 + 0.38 + 0.43}{8} = 0.39750,$$

and for the 10% trimmed mean for the with-nitrogen group we have

$$\bar{x}_{\text{tr}(10)} = \frac{0.43 + 0.47 + 0.49 + 0.52 + 0.75 + 0.79 + 0.62 + 0.46}{8} = 0.56625.$$

Note that in this case, as expected, the trimmed means are close to both the mean and the median for the individual samples. 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.

Exercises

1.1 The following measurements were recorded for the drying time, in hours, of a certain brand of latex paint.

3.4	2.5	4.8	2.9	3.6
2.8	3.3	5.6	3.7	2.8
4.4	4.0	5.2	3.0	4.8

Assume that the measurements are a simple random sample.

(a) What is the sample size for the above sample?

(b) Calculate the sample mean for these data.

(c) Calculate the sample median.

(d) Plot the data by way of a dot plot.

(e) Compute the 20% trimmed mean for the above data set.

(f) Is the sample mean for these data more or less descriptive as a center of location than the trimmed mean?

1.2 According to the journal *Chemical Engineering*, an important property of a fiber is its water absorbency. A random sample of 20 pieces of cotton fiber was taken and the absorbency on each piece was measured. The following are the absorbency values:

18.71	21.41	20.72	21.81	19.29	22.43	20.17
23.71	19.44	20.50	18.92	20.33	23.00	22.85
19.25	21.77	22.11	19.77	18.04	21.12	

(a) Calculate the sample mean and median for the above sample values.

(b) Compute the 10% trimmed mean.

(c) Do a dot plot of the absorbency data.

(d) Using only the values of the mean, median, and trimmed mean, do you have evidence of outliers in the data?

1.3 A certain polymer is used for evacuation systems for aircraft. It is important that the polymer be resistant to the aging process. Twenty specimens of the polymer were used in an experiment. Ten were assigned randomly to be exposed to an accelerated batch aging process that involved exposure to high temperatures for 10 days. Measurements of tensile strength of the specimens were made, and the following data were recorded on tensile strength in psi:

No aging:	227	222	218	217	225
	218	216	229	228	221
Aging:	219	214	215	211	209
	218	203	204	201	205

(a) Do a dot plot of the data.

(b) From your plot, does it appear as if the aging process has had an effect on the tensile strength of this polymer? Explain.

(c) Calculate the sample mean tensile strength of the two samples.

(d) Calculate the median for both. Discuss the similarity or lack of similarity between the mean and median of each group.

1.4 In a study conducted by the Department of Mechanical Engineering at Virginia Tech, the steel rods supplied by two different companies were compared. Ten sample springs were made out of the steel rods supplied by each company, and a measure of flexibility was recorded for each. The data are as follows:

Company A:	9.3	8.8	6.8	8.7	8.5
	6.7	8.0	6.5	9.2	7.0
Company B:	11.0	9.8	9.9	10.2	10.1
	9.7	11.0	11.1	10.2	9.6

(a) Calculate the sample mean and median for the data for the two companies.

(b) Plot the data for the two companies on the same line and give your impression regarding any apparent differences between the two companies.

1.5 Twenty adult males between the ages of 30 and 40 participated in a study to evaluate the effect of a specific health regimen involving diet and exercise on the blood cholesterol. Ten were randomly selected to be a control group, and ten others were assigned to take part in the regimen as the treatment group for a period of 6 months. The following data show the reduction in cholesterol experienced for the time period for the 20 subjects:

Control group:	7	3	−4	14	2
	5	22	−7	9	5
Treatment group:	−6	5	9	4	4
	12	37	5	3	3

(a) Do a dot plot of the data for both groups on the same graph.

(b) Compute the mean, median, and 10% trimmed mean for both groups.

(c) Explain why the difference in means suggests one conclusion about the effect of the regimen, while the difference in medians or trimmed means suggests a different conclusion.

1.6 The tensile strength of silicone rubber is thought to be a function of curing temperature. A study was carried out in which samples of 12 specimens of the rubber were prepared using curing temperatures of 20°C and 45°C. The data below show the tensile strength values in megapascals.

20°C:	2.07	2.14	2.22	2.03	2.21	2.03
	2.05	2.18	2.09	2.14	2.11	2.02
45°C:	2.52	2.15	2.49	2.03	2.37	2.05
	1.99	2.42	2.08	2.42	2.29	2.01

(a) Show a dot plot of the data with both low and high temperature tensile strength values.

(b) Compute sample mean tensile strength for both samples.

(c) Does it appear as if curing temperature has an influence on tensile strength, based on the plot? Comment further.

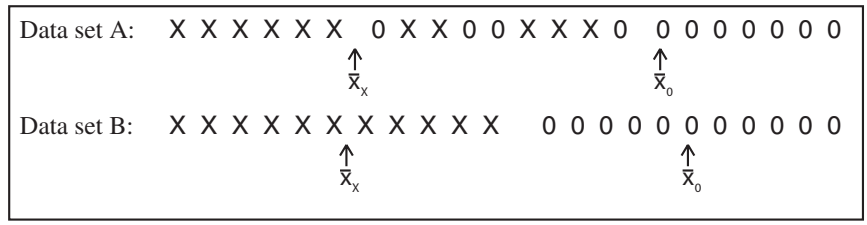
(d) Does anything else appear to be influenced by an increase in curing temperature? Explain.

1.4 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**. As a result, much of Chapters 9 through 15 deals with data analysis and modeling procedures in which sample variability plays a major role. 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 Example 1.2 we cannot conclude that the use of nitrogen enhances growth without taking sample variability into account.

While the details of the analysis of this type of data set are deferred to Chapter 9, it should be clear from Figure 1.1 that variability among the no-nitrogen observations and variability among the nitrogen observations 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. Perhaps there is something about the inclusion of nitrogen that not only increases the stem height ($\bar{x}$ of 0.565 gram compared to an $\bar{x}$ of 0.399 gram for the no-nitrogen sample) but also increases the variability in stem height (i.e., renders the stem height more inconsistent).

As another example, contrast the two data sets below. 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.



Sample Range and Sample Standard Deviation

Just as there are many measures of central tendency or location, there are many measures of spread or variability. Perhaps the simplest one is the **sample range** $X_{max} - X_{min}$. The range can be very useful and is discussed at length in Chapter 17 on *statistical quality control*. The sample measure of spread that is used most often is the **sample standard deviation**. We again let $x_1, x_2, \ldots, x_n$ denote sample values.

Definition 1.3:

The **sample variance**, denoted by s^2, is given by

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

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

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

It should be clear to the reader that the sample standard deviation is, in fact, a measure of variability. Large variability in a data set produces relatively large values of $(x - \bar{x})^2$ and thus a large sample variance. The quantity $n - 1$ is often called the **degrees of freedom associated with the variance** estimate. In this simple example, the degrees of freedom depict the number of independent pieces of information available for computing variability. For example, suppose that we wish to compute the sample variance and standard deviation of the data set (5, 17, 6, 4). The sample average is $\bar{x} = 8$. The computation of the variance involves

$$(5-8)^2 + (17-8)^2 + (6-8)^2 + (4-8)^2 = (-3)^2 + 9^2 + (-2)^2 + (-4)^2.$$

The quantities inside parentheses sum to zero. In general, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ (see Exercise 1.16 on page 31). Then the computation of a sample variance does not involve n **independent squared deviations** from the mean $\bar{x}$. In fact, since the last value of $x - \bar{x}$ is determined by the initial $n - 1$ of them, we say that these are $n - 1$ "pieces of information" that produce s^2. Thus, there are $n - 1$ degrees of freedom rather than n degrees of freedom for computing a sample variance.

Example 1.4: In an example discussed extensively in Chapter 10, 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.

The sample mean $\bar{x}$ is given by

$$\bar{x} = \frac{7.07 + 7.00 + 7.10 + \cdots + 7.08}{10} = 7.0250.$$

The sample variance s^2 is given by

$$s^2 = \frac{1}{9}[(7.07 - 7.025)^2 + (7.00 - 7.025)^2 + (7.10 - 7.025)^2 + \cdots + (7.08 - 7.025)^2] = 0.001939.$$

As a result, the sample standard deviation is given by

$$s = \sqrt{0.001939} = 0.044.$$

So the sample standard deviation is 0.0440 with $n - 1 = 9$ degrees of freedom.

Units for Standard Deviation and Variance

It should be apparent from Definition 1.3 that the variance is a measure of the average squared deviation from the mean $\bar{x}$. We use the term *average squared deviation* even though the definition makes use of a division by degrees of freedom $n - 1$ rather than n. Of course, if n is large, the difference in the denominator is inconsequential. As a result, the sample variance possesses units that are the square of the units in the observed data whereas the sample standard deviation is found in linear units. As an example, consider the data of Example 1.2. The stem weights are measured in grams. As a result, the sample standard deviations are in grams and the variances are measured in grams^2. In fact, the individual standard deviations are 0.0728 gram for the no-nitrogen case and 0.1867 gram for the nitrogen group. Note that the standard deviation does indicate considerably larger variability in the nitrogen sample. This condition was displayed in Figure 1.1.

Which Variability Measure Is More Important?

As we indicated earlier, the sample range has applications in the area of statistical quality control. It may appear to the reader that the use of both the sample variance and the sample standard deviation is redundant. Both measures reflect the same concept in measuring variability, but the sample standard deviation measures variability in linear units whereas the sample variance is measured in squared units. Both play huge roles in the use of statistical methods. Much of what is accomplished in the context of statistical inference involves drawing conclusions about characteristics of populations. Among these characteristics are constants which are called **population parameters**. Two important parameters are the **population mean** and the **population variance**. The sample variance plays an explicit role in the statistical methods used to draw inferences about the population variance. The sample standard deviation has an important role along with the sample mean in inferences that are made about the population mean. In general, the variance is considered more in inferential theory, while the standard deviation is used more in applications.

Exercises

1.7 Consider the drying time data for Exercise 1.1 on page 13. Compute the sample variance and sample standard deviation.

1.8 Compute the sample variance and standard deviation for the water absorbency data of Exercise 1.2 on page 13.

1.9 Exercise 1.3 on page 13 showed tensile strength data for two samples, one in which specimens were exposed to an aging process and one in which there was no aging of the specimens.

(a) Calculate the sample variance as well as standard deviation in tensile strength for both samples.

(b) Does there appear to be any evidence that aging affects the variability in tensile strength? (See also the plot for Exercise 1.3 on page 13.)

1.10 For the data of Exercise 1.4 on page 13, compute both the mean and the variance in "flexibility" for both company A and company B. Does there appear to be a difference in flexibility between company A and company B?

1.11 Consider the data in Exercise 1.5 on page 13. Compute the sample variance and the sample standard deviation for both control and treatment groups.

1.12 For Exercise 1.6 on page 13, compute the sample standard deviation in tensile strength for the samples separately for the two temperatures. Does it appear as if an increase in temperature influences the variability in tensile strength? Explain.

1.5 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 yield is maximized. Here, of course, the yield 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. In Chapter 5, several distinct scenarios, leading to different ways of handling data, are discussed for situations with count data.

Special attention even at this early stage of the textbook 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, x/n is the sample proportion in category 1, and $1 - x/n$ is the sample proportion in category 2. In the 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