Notice that the standard deviation is still a form of average distance a value is away from the mean, even though we squared the distances, divided by $n - 1$, and took the square root.

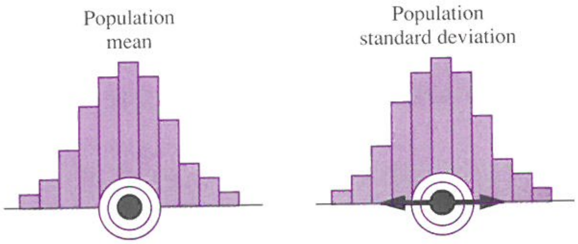
Advantages of Using the Standard Deviation as our Measure of Spread

1. The negative signs are removed without use of the absolute value lines, which greatly reduces the computational difficulties in advanced work.
2. The standard deviation is a more efficient estimator of spread than the average distance from the mean.
3. The standard deviation is a *considerably* more efficient estimator of spread than many other measures considered, such as the median deviation.

   Note: The term *more efficient* means sample values cluster more tightly around the population value—that is, on average, sample values give closer approximations to the population value. For further reading on the standard deviation, refer to section 9.0 under the subheading ''Least-Squares Analysis''; also refer to endnote 16 in chapter 9.

2.4 Estimating Population Characteristics

The purpose of a sample is usually to gather information about a population. Two of the characteristics of a population we most frequently wish to know are the



Unfortunately, small samples (under 30 observations) will sometimes give unreliable approximations of population characteristics, depending on the *shape* of your population histogram. (This will be discussed in greater detail in chapters 7 and 8).

One way to avoid this problem is to keep your sample size at 30 or more observations. If our sample size, n, is kept at 30 or more observations, we do not have to worry about the shape of our population. Results from sample sizes of 30 or more give reliable information about population characteristics for almost any shaped population. So, with this in mind, we can state, *it is preferable to keep your sample size at 30 or more*. More specifically, we can state the following.

> For a valid random sample of 30 or more observations, drawn from almost any population
>
> $\bar{x} \approx \mu$
> The sample average, $\bar{x}$, will be approximately equal to the population average,
> μ (mū).
>
> $s \approx \sigma$
> The sample standard deviation, s, will be approximately equal to the population standard deviation, σ (sigma).

We can also use samples of *smaller than 30* observations, but in that case, we must be assured our population is at least somewhat bell-shaped. In bell-shaped populations, small samples give reliable approximations of population characteristics, or at least reliable enough that after certain adjustments, reasonable conclusions can be drawn. So,

> For a valid random sample of *under 30 observations,* we must be assured our population is at least somewhat bell-shaped.

In the preceding examples, we sometimes used samples as small as five or six observations. If our population was somewhat bell-shaped, this would be perfectly okay. However, if our population was far from bell-shaped (let's say, for instance, extremely skewed), then we can *not* depend on these samples to give reliable estimates of population characteristics.

2.5 Measures of Central Tendency and Dispersion/Spread (Grouped Data)

skip

When we work with large bodies of data, it is sometimes more efficient to group the data into categories. In the following example, we will calculate the mean and standard deviation for such data.

Mean

Example

Say in our nurse-in-training study, the researcher took 36 observations of the nurse. Only now, instead of recording individual times, the researcher chose to record the times as part of a category or group, as follows:

Time Category (in minutes)	**Number of Observations (tally)**
3–5	𝍸 I
6–8	𝍸 𝍸 I
9–11	𝍸 II
12–14	𝍸 IIII
15–17	III

Calculate the mean.

Rationale

Notice when the data is grouped, the identity of the individual values is lost. For instance, in the first tally, we know the nurse was observed on six occasions where it took 3 to 5 minutes to draw the blood, but we do not know the precise time it took on each of these six occasions.

Calculating the mean from this data proceeds as follows. Because it is quite difficult to use a range of time such as 3 to 5 minutes in a calculation, we merely average the two values $(3 + 5)/2 = 4$ and represent the group as 4 minutes. In reality, some measurements in this category are greater than 4 minutes and some less, but with many readings, experience has shown such measurements tend to average out to near 4 minutes.

Notice in the chart below, we used the letter x to represent this category average of 4 minutes and the letter f to represent on how many occasions this occurred, called frequency.

Time Category	x	**Tally**	f **(frequency)**
3–5	4	𝍸 I	6
6–8	7	𝍸 𝍸 I	11
9–11	10	𝍸 II	7
12–14	13	𝍸 IIII	9
15–17	16	III	3
			$\Sigma f = 36$ (or $n = 36$)

Also note that the sum of the frequency column, Σf, is equal to the sample size, n. In other words, we say $\Sigma f = 36$ or $n = 36$. In either case, it tells us that we observed the nurse on a total of 36 occasions. (Note: we also get 36 by adding up the tally slashes.)

To calculate the mean, we simply add up the recorded times which we had observed and divide by 36.

$$\bar{x} = \frac{\Sigma x}{n} = \frac{\overbrace{4 + 4 + 4 + 4 + 4 + 4}^{\text{six readings}} + \overbrace{7 + 7 + 7 + 7 + 7 + 7 + 7 + 7 + 7 + 7 + 7}^{\text{eleven readings}} + \overbrace{10 + 10 + \cdots}^{\text{etc.}}}{36}$$

$$= \frac{336}{36} = 9.333$$

$$= 9.3 \text{ minutes}$$

This is quite tedious however. Another way is to simply multiply 4 × 6, 7 × 11, 10 × 7, 13 × 9, and 16 × 3, then add the results and divide by 36 as follows:

$$\bar{x} = \frac{\Sigma xf}{n} = \frac{4(6) + 7(11) + 10(7) + 13(9) + 16(3)}{36}$$

$$= \frac{24 + 77 + 70 + 117 + 48}{36} = \frac{336}{36} = 9.333$$

$$= 9.3 \text{ minutes}$$

This process is presented more efficiently in chart form in the following solution.

Solution

To calculate the mean, we sum the xf values and divide by 36.

Time Category	x	f	xf
3–5	4	6	24
6–8	7	11	77
9–11	10	7	70
12–14	13	9	117
15–17	16	3	48
			$\Sigma xf = 336$

$$\begin{aligned}\bar{x} &= \frac{\Sigma xf}{n} \\ &= \frac{336}{36} \\ &= 9.333 \\ &= 9.3 \text{ minutes}\end{aligned}$$

Based on these 36 observations, the average time the nurse-in-training took to draw the blood specimens was calculated to be 9.3 minutes. This can be summarized as,

$$\begin{aligned}n &= 36 \text{ observations} \\ \bar{x} &= 9.3 \text{ minutes}\end{aligned}$$

The formula used for grouped data, then, is

Sample average for grouped data

$$\bar{x} = \frac{\Sigma xf}{n}$$

Standard Deviation

To calculate the **standard deviation,** we also use a slightly altered formula as follows:

Sample standard deviation for grouped data

$$s = \sqrt{\frac{n(\Sigma x^2 f) - (\Sigma xf)^2}{n(n-1)}}$$

The formula has been algebraically rearranged to avoid having to calculate each individual distance from the mean. It also avoids having to calculate the mean itself. This greatly accelerates the calculations.

Example

For our 36 observations of the nurse-in-training, calculate the standard deviation.

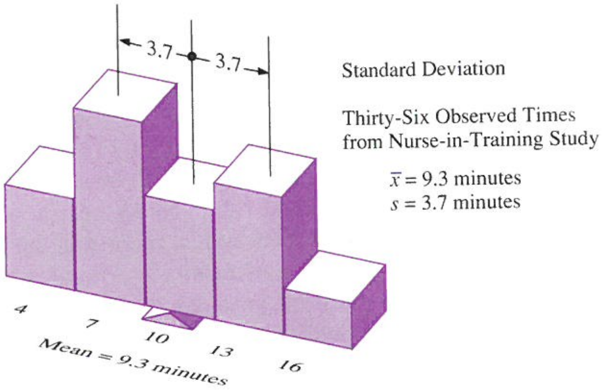
Solution

First, we arrange the data in chart form, just as we did in the preceding example. Only this time, we include an additional x^2f column. The x^2f values are obtained by multiplying the x value by the xf value. For example, in the first row, we obtained an x^2f reading of 96 by multiplying 4 by 24, $4 \times 24 = 96$ (x times xf $= x^2f$).

Time Category	x	f	xf	x^2f
3–5	4	6	24	96
6–8	7	11	77	539
9–11	10	7	70	700
12–14	13	9	117	1521
15–17	16	3	48	768
		36	336	3624
		Σf	Σxf	Σx^2f
		$(n = 36)$		

$$\begin{aligned} s &= \sqrt{\frac{n(\Sigma x^2 f) - (\Sigma xf)^2}{n(n-1)}} \\ &= \sqrt{\frac{36(3624) - (336)^2}{36(36-1)}} \\ &= \sqrt{\frac{130{,}464 - 112{,}896}{36(35)}} \\ &= \sqrt{13.942} = 3.734 \\ &= 3.7 \text{ minutes} \end{aligned}$$

Let's see what the data looks like using a three-dimensional histogram model.



Notice that the sample average and sample standard deviation above ($\bar{x}$ and s) were calculated to be 9.3 and 3.7 minutes, respectively, based on $n = 36$ observations. In a prior section, based on $n = 6$ observations, we had calculated the $\bar{x}$ and s to be 9.0 and 3.9 minutes. Why are the sample results different?

Actually, the two most probable reasons are: first, sample values *approximate* population values—that is, $\bar{x} \approx \mu$ and $s \approx \sigma$, so sample results often differ slightly each time we sample from a population (and the smaller the sample size, the more variation we can expect in sample results); second, the process of grouping itself may have caused us to lose some precision. ■

Advantages of Grouping Data

1. Ease in collection and organization: slash marks on a tally are much easier to record than individual values, and if we rotate the tally $\frac{1}{4}$ turn counterclockwise, we already have, in effect, a histogram.

2. Ease of interpretation: the central tendency and spread of a grouped-data tally can be *estimated* at a glance. This is not true of ungrouped data.

3. Speed: it is usually much faster to calculate the mean and standard deviation of a large body of data when the data is grouped than if we had recorded individual readings. Although with the hand calculator and personal computer this advantage is not as important as it once was.

Disadvantages of Grouping Data

1. Individual values in a sense lose their identity. As a result, we cannot reconstruct the data into different groupings. For instance, what if we wanted to reconstruct the data into the categories, 2–4, 5–7, 8–10 minutes, etc.? Without the individual readings, we lose our ability to do this.

2. Some precision *may* be lost. Remember, each value is recorded as part of a group. If this precision is vital for some subsequent decision-making process, then grouped data should be used with caution. For instance, in our nurse-in-training example, where we calculated the average to be $\bar{x} = 9.3$ minutes, it is theoretically possible the sample average is actually 8.3 or 10.3 minutes, although it is rare that this would happen. Generally, for large groups of data (hundreds of readings or more), we find the results from grouping data to be quite close to the results if we had measured each individual value. Nevertheless, why introduce this uncertainty when it can be avoided by just recording the individual values.* So, in the case where precision of results is critical, as in much of the work in inferential statistics, grouped data for these purposes should be avoided. The exception would be if the data is of a sufficient size (many hundreds or thousands of readings) in which case we would be less likely to find major differences when we calculated the sample average and standard deviation using grouped versus ungrouped techniques.

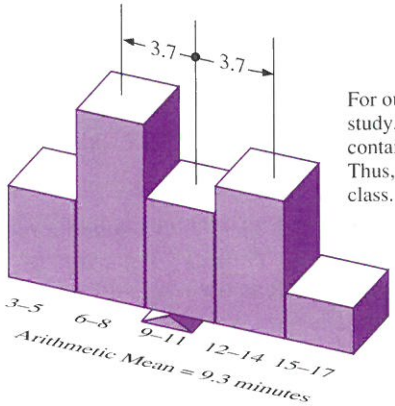
Modal Class

Modal Class

The category that contains the most values.

*Actually, certain techniques are available that allow us to record individual values yet maintain some efficiency of grouping. One such technique is called the stem-and-leaf display, which is presented in section 2.7.

The equivalent of the mode for ungrouped data is the **modal class** for grouped data, which is the category that contains the most values.



For our nurse-in-training study, the category 6–8 minutes contains the most values as shown. Thus, this would be our modal class.

do

2.6 *z* Scores and the Use of the Standard Deviation

In inferential statistics, we often refer to a particular value in terms of its position relative to the mean. In many instances, we use something called a *z* **score.**

***z* score**
The number of standard deviations a value is away from the mean.

Although we will eventually introduce a formula, it is best to first learn *z* scores by estimation.

Example

Suppose we took a sample from a grove of recently planted Indiana poplar trees and measured their heights, with the following results.

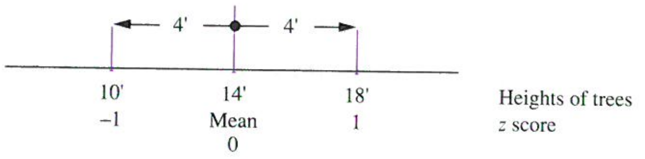
$\bar{x}$ = 14 feet (average height)
s = 4 feet (standard deviation)

Estimate the *z* score for a poplar tree of the following heights.

a. 8 feet
b. 16 feet

Rationale

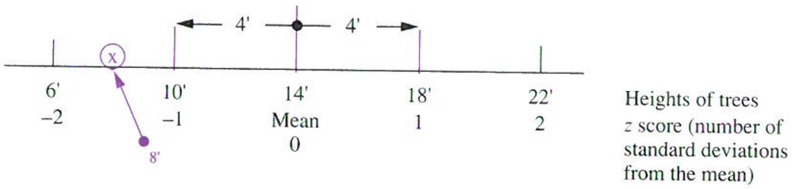
First, we would visually represent the mean and standard deviation of the data on a number line.



Notice that a tree of 18 feet would be four feet or exactly one standard deviation above the mean. Thus, we say a tree of 18 feet has a z score of 1. Remember, a z score is defined as the number of standard deviations a particular value is from the mean. Because a tree of 18 feet is one standard deviation above the mean, it has a z score of 1. Also notice that a tree of 10 feet is 4 feet below the mean. It has a z score of -1.

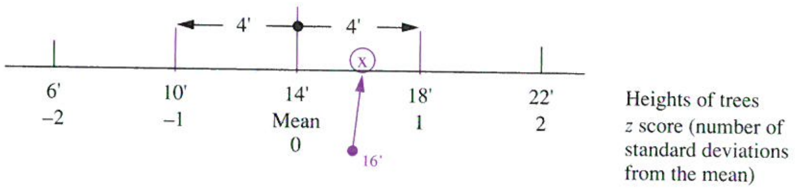
Solution

a. To estimate the z score of a tree with a height of 8 feet, we first must expand our scale to include ±2 standard deviations, then locate the position of 8 feet.



Notice that a tree of 8 feet would be $-1\frac{1}{2}$ standard deviations from the mean, therefore it has a z score of $-1\frac{1}{2}$ (or in decimal form, $z = -1.50$).

b. To estimate the z score of a tree of 16 feet, we locate on the scale where a tree of 16 feet would be.

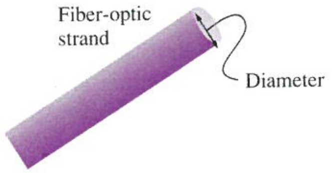


A tree of 16 feet would be $\frac{1}{2}$ of a standard deviation above the mean, therefore it has a z score of $\frac{1}{2}$ (or in decimal form, $z = .50$). ■

Estimating *z* scores is quite important in the study of inferential statistics, so let's look at a second example.

Example

Suppose the diameter of a fiber-optic strand (a glass fiber capable of transmitting hundreds of thousands of times more information than a copper wire) is continuously measured on an electronic assembly line, such that



$\mu = 3.112$ mm* (average diameter)
$\sigma = .008$ mm (standard deviation)

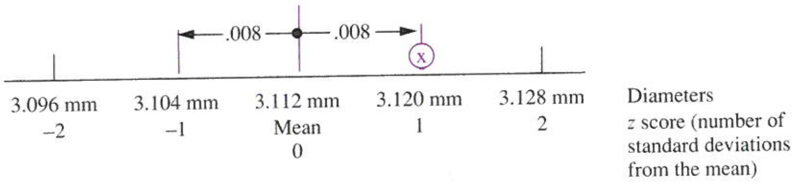
Estimate the *z* score of a fiber-optic strand with the following diameters.

a. 3.120 mm
b. 3.110 mm

Solution

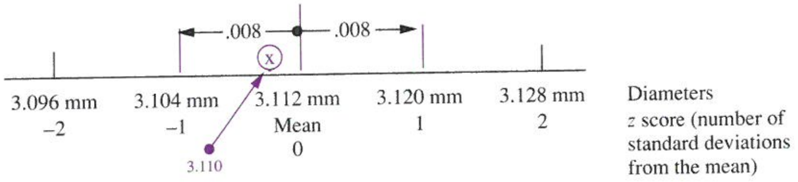
Notice that μ and σ were used to represent the mean and standard deviation. This implies every segment of fiber in the entire population had been measured.

a. To estimate the *z* score of a piece of fiber with a diameter of 3.120 mm, we would visually represent the mean and standard deviation on a number line and locate the position of 3.120 mm.



A fiber with a diameter of 3.120 mm is .008 mm or exactly one standard deviation above the mean, therefore it has a *z* score of 1 (that is, $z = 1.00$).

b. To estimate the *z* score of a piece of fiber with a diameter of 3.110 mm, again we represent the mean and standard deviation on a number line, only this time we locate the position at 3.110 mm.



*Note: 3.112 mm (millimeters) is approximately $\frac{1}{8}$ inch.

Because 3.110 mm is .002 mm below the mean and .002 is exactly $\frac{1}{4}$ of .008, we state that 3.110 mm is $\frac{1}{4}$ of a standard deviation below the mean. Therefore, 3.110 mm has a z score of $-\frac{1}{4}$ (that is, $z = -.25$). ■

In inferential statistics, z score notation is commonly used. From chapter 4 on, we will be using it on a routine basis.

Two Important Findings

Along with the mean, the standard deviation is one of the most important measures we have in inferential statistics and much time has been devoted to its study.

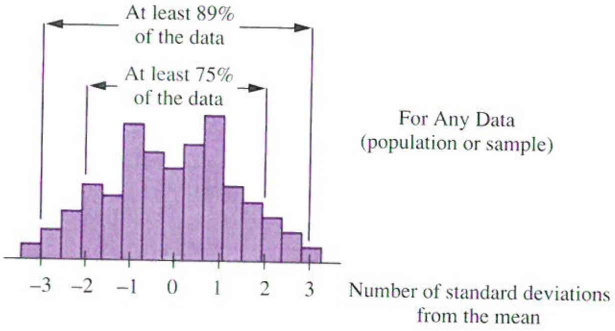
One important finding was made by P. L. Chebyshev (1821–1894).*

For any set of data, whether it be population or sample data, no matter what its shape,

at least 75% of your data will lie within two standard deviations of the mean, and

at least 89% of your data will lie within three standard deviations of the mean.

Pictorially, this might be represented as follows.



A second important finding was discovered seemingly under different circumstances by De Moivre (1733), Laplace (1781), and Gauss (1809), and this is called the normal population distribution. This distribution (a specific form of bell-shaped data) was encountered so frequently in experiments that the name *normal* was adopted sometime in the mid to late 1800s (from unknown origins).

*Chebyshev's Theorem states: at least $1 - \frac{1}{k^2}$ of any collection of data lies within k standard deviations of the mean. For $k = 2$ at least $1 - \frac{1}{2^2}$ of the data lies within 2 standard deviations of the mean. Since $1 - \frac{1}{2^2} = 1 - \frac{1}{4} = \frac{3}{4}$, we can say: at least $\frac{3}{4}$ of the data (or 75%) lies within two standard deviations of the mean. For $k = 3$, at least $1 - \frac{1}{9} = \frac{8}{9}$ of the data (or 89%) lies within three standard deviations of the mean.

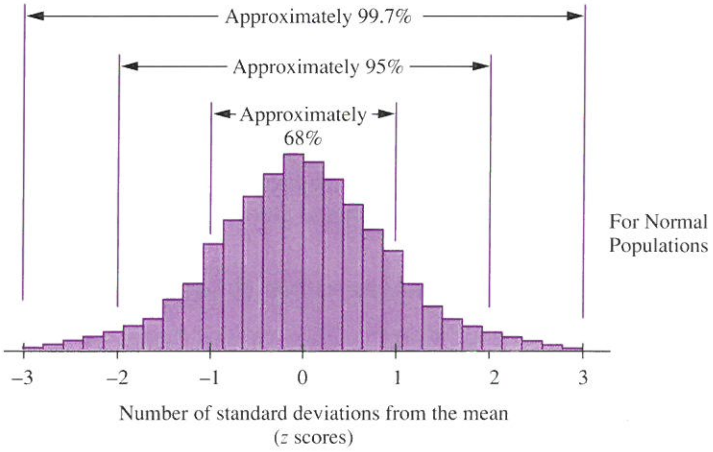
For Normal Population Distributions

Approximately 68% of the data lies within one standard deviation of the mean.

Approximately 95% of the data lies within two standard deviations of the mean.

Approximately 99.7% of the data lies within three standard deviations of the mean.

Pictorially, this might be represented as follows.



The normal distribution is one of the most important distributions we study in inferential statistics and it is wise to take a moment to commit the above approximations to memory.

2.7 Additional Descriptive Topics

Although the material presented thus far provides the main thrust of descriptive techniques needed for future chapters, several variations of these techniques are in popular use and will be presented here. Specifically we shall introduce pictograms, stem-and-leaf displays, and box-and-whisker plots. Quartiles and percentiles are also briefly discussed.

Pictogram

Tallies and histograms are often presented in more dramatic form. One of the most common is a **pictogram** where illustrations or pictures are used to represent data, demonstrated as follows.

Suppose a random sample of small-business owners were asked at what age they first became self-employed. A tally of the results is presented below, left. A pictogram for this data is presented at right.

Ages	Tally	Ages	Pictogram
20–24	||||	20–24	🧍🧍🧍🧍
25–29	𝍸 ||||	25–29	🧍🧍🧍🧍🧍 🧍🧍🧍🧍
30–34	𝍸 ||	30–34	🧍🧍🧍🧍🧍 🧍🧍
35–39	||	35–39	🧍🧍
40+	|	40+	🧍

In essence, pictures are used in lieu of slash marks. In the tally (at left), if each slash mark represents one person in the survey, then in the pictogram, each symbol 🧍 would represent one person in the survey (note, if each slash mark represented 100 people, then each symbol would represent 100 people).

Essentially the pictogram attempts to offer a more visually appealing display of the data and is often used in business charts and in newspaper and magazine articles for dramatic effect.

Stem-and-Leaf Display

As useful as the tally (or pictogram) is in presenting data, often information is lost when we use such processes. For instance, in the above example, the exact ages of the individuals are lost. One slash mark in the category 25–29 could represent a person who is 26 years old or a person 29 years old. We have no way of knowing.

The **stem-and-leaf display** makes an effort to combine the ease and clarity of a tally while maintaining the original information. Suppose the slash marks in the above tally actually represented the following ages.

24 21 23 22 25 29 28 27 27 26 28 25 34 33 30 31 31 32 32 37 35 29 41

A stem-and-leaf display separates each value into a stem (usually the leftmost digit or digits) and a leaf (consisting of a digit or digits to the right of the stem). For instance, the first two ages 24 and 21 could be represented as 2 | 4 1. Each has the same stem, 2 (the leftmost digit) with leaves 4 and 1. Recording all the data in this fashion, we get

Stem	Leaves		
2	4 1 3 2		(last digit 0 to 4)
2	5 9 8 7 7	6 8 5 9	(last digit 5 to 9)
3	4 3 0 1 1	2 2	(last digit 0 to 4)
3	7 5		(last digit 5 to 9)
4	1		

Notice we laid out the stem-and-leaf display in a manner similar to the tally and pictogram, however, now, if needed, the original ages can easily be retrieved, say for instance, if we wish to calculate a precise sample average.

Box-and-Whisker Plot

Another method of presenting data is the **box-and-whisker plot** (sometimes referred to as a **boxplot**). Essentially the technique attempts to divide the data into four equal groupings using the following technique.

1. List the values in increasing order and locate the median (the middle value)

   21 22 23 24 25 25 26 27 27 28 28 29 29 30 31 31 32 32 33 34 35 37 41

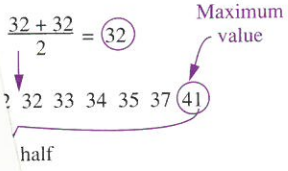
   ↑

   Medi

2. Mark off the lo the median) and locate the
   lef half. Repeat for higher half

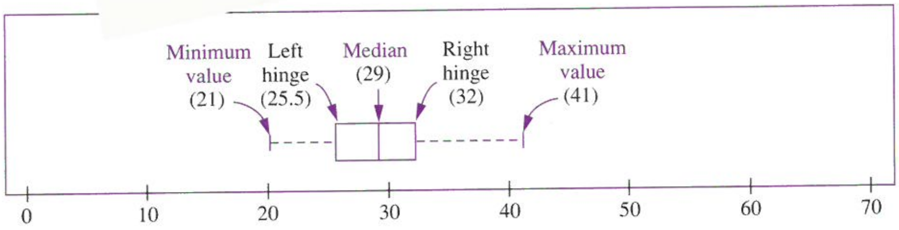
Skip

both



i e, the average of the
t

3. U hinge, and
   ma he box-and-whisker
   plo



The box between the left and right hinges contains the central core of the data (the middle 50% of all the values), with the lower 25% of values represented by the dashed lines to the left of the box and the upper 25% of values represented by the dashed line to the right of the box. The dashed lines are referred to as **whiskers.**

Essentially, the box-and-whisker plot allows us to quickly identify the data's central core (middle 50%) while displaying the span of those values outside the central core.

Quartiles

Like the box-and-whisker plot, quartiles attempt to divide the data into four equal groups (quarters). The procedure is almost identical, only now the left and right hinges (called Q_1 and Q_3 in quartiles) are established as follows: Q_1 is the median of the values to the left of the overall median value of 29 (however, this procedure does *not* include the median value, 29, in the computation). Q_2 is the overall median. Q_3 is the median of the values to the right of the overall median.

$$21\ 22\ 23\ 24\ 25\ \underset{Q_1}{25}\ 26\ 27\ 27\ 28\ 28\ \underset{Q_2}{29}\ 29\ 30\ 31\ 31\ 32\ \underset{Q_3}{32}\ 33\ 34\ 35\ 37\ 41$$

Percentiles

Percentiles is a method used to divide data into 100 approximately equal groupings.

2.8 Writing Research Reports

Research reports are common as a means of formally presenting the results of a research study. Although different formats exist, most include the following information.

Background Statement

Essentially, this is a statement of the problem. Why was the study performed? What questions do you expect to answer? Usually, a brief history of events leading to the commissioning of the study is presented, outlining the need, purpose, and expectations of the study.

For instance, in the Countrygirl Makeup example in section 2.1, it was stated that, ''although successful when introduced in 1980, in recent years, sales have eroded,'' suggesting that a decline in sales has prompted the company to commission the study. Further stated was, ''an independent research firm was commissioned to gather information about the ages of current users,'' implying a need to explore the question of whether the ages of the current target population, young women ages 16 to 24, have shifted.

If prior studies have been performed, the results should be included. If there has been much work done in this area, then these results should be fully presented in a separate section.

Design and Procedures of the Study

Essentially, this section answers the question of *how* the study was conducted. If sampling is to be used, which is often the case, discuss your target population and how you intend to sample from this target population. Also discuss how your sampling techniques provide for (1) internal validity and (2) external validity (a major factor in determining external validity is the component of whether a *random* sample was achieved).

Although questions concerning validity can grow quite complex, you might consider the following:

Internal validity: Were honest, accurate and reliable measurements achieved under the given test conditions?

- Was the measuring scale objective?
- How was the accuracy of the measurements verified?
- Did you lose sample results, say with people refusing to cooperate?
- In more complex experiments, requiring measurements over the passage of time as in educational or medical studies, did outside events affect the measurement? Did the natural maturation of subjects affect the measurement? Did some initial test affect the final test?

External validity: Were you successful in achieving a true *random* sample? Did the test methods influence results?

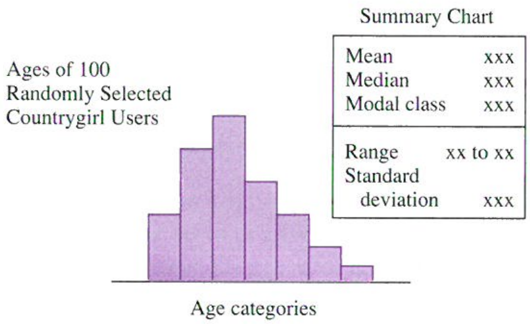
- Did the presence of the experimenter exert an influence?
- Did the test environment create an unnatural setting and thus affect results? In the case where the sampled group is put in the same room and questioned, did one response affect another?

Suggested reading in this area is D. T. Campbell and J. C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Boston: Houghton Mifflin Co., 1963).

Results

Data is summarized visually in the form of histograms, frequency polygons, circle graphs, and charts and tables of vital information, such as means, medians, standard deviations, and so on.

For instance, in the Countrygirl study, a histogram might be presented with a summary chart of vital information as follows:



Summary Chart

Mean	xxx
Median	xxx
Modal class	xxx
Range	xx to xx
Standard deviation	xxx

The goal is to provide a clear, easy-to-read *summary* of results. The reader should be able to look at this summary and within seconds understand the complete test results.

Note: Do *not* include raw data or computer printouts in the body of the report. These should be attached at the end as appendixes.

Analysis and Discussion

The visuals (histograms, etc.) and summary charts and tables are analyzed in regard to the central purpose of the study, which was presented at the beginning of the report under background. Interesting facts and significant findings should be pointed out and discussed.

For instance, in the Countrygirl study, one might give the percentage of users in the sample outside the current age range of 16 to 24 year olds, because a large percentage of users outside our current target age range might have relevance to future advertising and promotion decisions. Compare this to what percentage of the sample was *inside* the target population age range of 16 to 24. Also compare the average and median age from the sample with the average age of the current target population.

In addition, suggest reasons why the data turned out the way it did. Were the results surprising? If so, point out why this might happen.

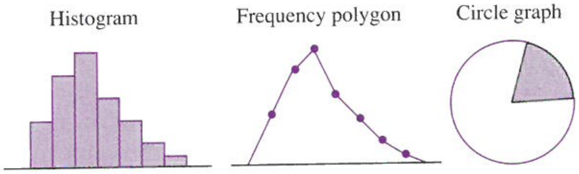
Conclusions and Recommendations

Draw conclusions and suggest future action as if a valid random sample were achieved and the results can be used as representative of the population.

If serious questions exist as to the internal or external validity of the sample, provide a cautionary note and suggest how these problems might be overcome in a future study.

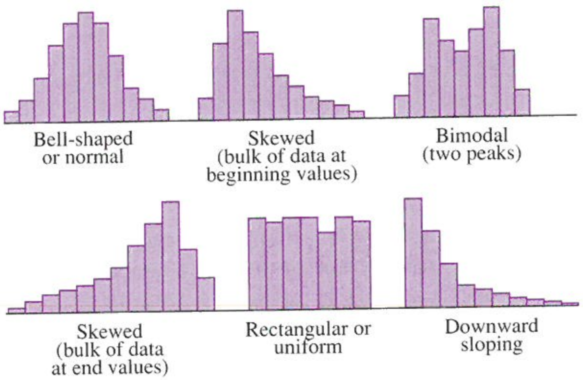
Summary

Graphical representations provide a visual sense of the data, allowing an immediate registration of the data's impact. The following were introduced.



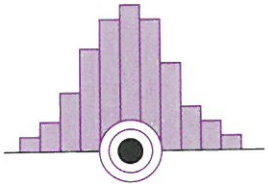
Experience has shown that many *population* histograms take on repeating shapes, that is, there are certain common shapes that seem to reoccur.

Common Shapes of Population Histograms



In addition to graphical representations, we often wish to get some sense of the middle or central value of our data. This is called central tendency and the most frequently used measures of central tendency are given below.

Arithmetic mean: Commonly called the average or mean, this value is calculated by adding up all the values in your data collection and dividing by n, the number of values. To calculate the arithmetic mean of a sample, we use the following formulas.



$$\text{Ungrouped data: } \bar{x} = \frac{\Sigma x \text{ (sum of the values)}}{n \text{ (number of values)}}$$

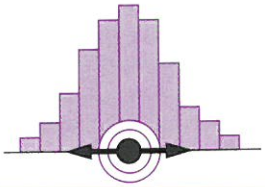
$$\text{Grouped data: } \bar{x} = \frac{\Sigma xf \text{ (sum of the } xf \text{ values)}}{n \text{ (number of values)}}$$

Median: The middle value when data is arranged from lowest to highest value. If there are two middle values, we average the two.

Mode: The most frequently occurring value. Some sets of data may have no mode. If two modes occur, we call this bimodal. If three or more modes occur, we call this multimodal.

Modal class: The equivalent of the mode for grouped data is the modal class, which is the category that contains the most values.

Whereas measures of central tendency attempt to locate the center or middle of the data, measures of dispersion are designed to measure how widely scattered or spread out the data is.



Range: The range is the difference between the high and low value in your data set.

Standard deviation: A form of average distance from the mean.

Along with the mean, the standard deviation is one of the most important measures we have in inferential statistics and much time has been devoted to its study. To calculate the standard deviation of a sample, we use the following formulas.

$$\text{Ungrouped data: } s = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n - 1}}$$

$$\text{Grouped data: } s = \sqrt{\frac{n(\Sigma x^2 f) - (\Sigma xf)^2}{n(n - 1)}}$$

The purpose of a sample is usually to gather information about a population. Two of the characteristics of a population we most frequently wish to know are μ, the population mean, and σ, the population standard deviation. We know from well over a century of practical experience that valid random samples of 30 or more will give reliable information about μ and σ.

More specifically we can state, *for a valid random sample of 30 or more observations, drawn from almost any shaped population.*

$\bar{x} \approx \mu$: the sample average, $\bar{x}$, will be approximately equal to the population average, μ (mū).

$s \approx \sigma$: the sample standard deviation, s, will be approximately equal to the population standard deviation, σ (sigma).

Samples of under 30 observations can also be used, however *for a valid random sample of under 30 observations, we must be further assured our population is at least somewhat bell-shaped.*

In inferential statistics, we often refer to a particular value in terms of its relative position to the mean. In many instances we use something called a z score.

z score: The number of standard deviations a value is away from the mean. For instance, if $\bar{x} = 14$ and $s = 4$, then a value of 18 would be one standard deviation above the mean, expressed as $z = 1.00$.

2,1 - 2,5, 2,7, 2,9, 2,10, 2,17

Exercises

Note that full answers for exercises 1–5 and abbreviated answers for odd-numbered exercises are provided in the Answer Key.

2.1 Countrygirl Makeup is a line of facial products marketed to young women, ages sixteen to twenty-four. Although successful when introduced in 1980, in recent years Countrygirl sales have eroded. An independent research firm was commissioned to gather information about the ages of current users. A nationwide random sample of 50 users yielded the following ages:

19.4	31.3	22.7	27.6	27.9
30.1	23.1	26.4	32.1	22.5
28.2	25.7	33.8	28.9	18.6
26.8	30.5	34.0	21.6	28.2
32.2	27.3	17.5	23.0	32.8
36.0	29.1	42.7	30.5	39.0
26.2	33.2	36.3	22.7	43.1
28.7	26.3	38.6	24.1	21.3
32.1	28.7	25.8	26.0	18.7
18.2	23.9	28.2	20.2	33.1

Age Category	Number of Observations (tally)
15–19	
20–24	
25–29	
30–34	
35–39	
40–44	

a. Tally the data into the age categories given above.
b. Construct separately a histogram and frequency polygon.
c. What is the population?
d. What is the sample? How large is n, the sample size?

2.2 In exercise 2.1, 21 out of 50 possess the attribute of brown hair.

a. What proportion (or fraction) of the sample, p_s, possesses the attribute of brown hair?
b. Convert this fraction to a percentage.
c. Represent this proportion in a circle graph.

2.3 In a medical study, a researcher wished to estimate the average length of time needed for a particular nurse-in-training to draw a series of blood specimens. A sample of the nurse's work over several months yielded the following times: 11, 7, 13, 7, 5, 8, 10, and 15 (in minutes).

Calculate:

a. Mean
b. Median
c. Mode
d. Range
e. Standard deviation

Discuss:

f. What conditions would be necessary for us to use the mean and standard deviation of the sample as representative of the mean and standard deviation of the population?

2.4 Suppose in our nurse-in-training study, the researcher took 49 observations of the nurse. Only now, instead of recording individual times, the researcher chose to record the times as part of a category or group, as follows.

Calculate:

a. Mean
b. Standard deviation
c. Modal class

Construct:

d. Histogram
e. Frequency polygon

Time category (in minutes)	Number of observations (tally)
3–5	𝍸 ||
6–8	𝍸 𝍸 𝍸
9–11	𝍸 𝍸 𝍸 |
12–14	𝍸 |||
15–17	|||

2.5 Suppose we sampled from a grove of recently planted Indiana poplar trees and measured their heights, obtaining the following:

$\bar{x} =$ 14 feet (average height of trees sampled)
$s =$ 4 feet (standard deviation of trees sampled)

Estimate the *z* score for a poplar tree of the following height.

a. 20 feet
b. 10 feet
c. 15 feet

2.6 In a study on shyness at the University of Iowa, a team of psychologists asked 7 participants to rank the anxiety created by being at a party with strangers on a scale from 0 (no anxiety) to 20 (maximum anxiety), yielding the following scores:
17, 19, 16, 14, 19, 15, and 12.

Calculate:

a. Mean
b. Median
c. Mode
d. Range
e. Standard deviation

Discuss:

f. What conditions would be necessary for us to use the mean and standard deviation of the sample as representative of the mean and standard deviation of the population?

2.7 In an educational study of second-graders in Westchester County, N.Y., it took 5 students the following times (in seconds) to put together a simple puzzle: 12, 14, 7, 13, and 9.

Calculate:

a. Mean
b. Median
c. Mode
d. Range
e. Standard deviation

Discuss:

f. What conditions would be necessary for us to use the mean and standard deviation of the sample as representative of the mean and standard deviation of the population?

2.8 A sample of $n = 100$ Countrygirl Makeup users were randomly sampled nationwide, yielding the following ages (taken from the demonstration problem at the beginning of this chapter).

Calculate:

a. Mean
b. Standard deviation
c. Modal class
d. If Countrygirl is currently marketed to 16–24 year olds, and if we managed to achieve a valid random sample above, what effect might these results have on future advertising?

Age category	Number of observations (tally)
15–19	𝍸 𝍸 |
20–24	𝍸 𝍸 𝍸 𝍸 ||||
25–29	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸
30–34	𝍸 𝍸 𝍸 |||
35–39	𝍸 𝍸 |
40–44	𝍸
45–49	|
	Total, 100 users

2.9 At $n = 70$ boutiques randomly selected throughout the New England sales district, the following number of Rolf Laurie designer bed comforters sold last year were

Calculate:

a. Mean
b. Standard deviation
c. Modal class

Construct:

d. Histogram
e. Frequency polygon

Rolf Laurie comforters sold (last year)	Number of New England boutiques (tally)
0–14	𝍸
15–29	𝍸 𝍸 𝍸 ||||
30–44	𝍸 𝍸 𝍸 𝍸 |||
45–59	𝍸 𝍸
60–74	𝍸 |||
75–89	𝍸
	Total, 70 boutiques

2.10 If the average amount wagered per person in state lotteries in a particular year was $\mu = \$100$ with standard deviation $\sigma = \$12$, find the z score for a person who wagered

a. \$130.
b. \$96.

2.11 A vending machine is known to fill cups to an average of $\mu = 7.0$ ounces with standard deviation $\sigma = .4$ ounces. Find the z score for a person that gets a cup with

a. 7.6 ounces.
b. 7.1 ounces.
c. 6.7 ounces.

2.12 Say you took five quizzes in Economics and your average was 76. Four of the five quizzes were graded, 80, 72, 86, and 70; however, one quiz grade was lost. Use the formula for calculating an average to determine the missing grade.

2.13 A young couple, Jason and Rebecca, weighed 170 lbs and 138 lbs, respectively. For their age and body structure, a medical association published the following guidelines:

Male	Female
$\mu = 155$	$\mu = 118$
$\sigma = 12$	$\sigma = 10$

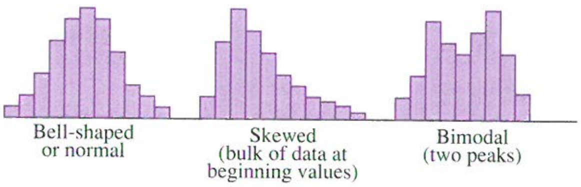
a. Calculate the appropriate z scores for each.
b. According to these guidelines, who would be considered more seriously overweight?

2.14 The Jontnas Company has seven employees with the following annual salaries: \$20,000, \$20,000, \$20,000, \$40,000, \$40,000, \$40,000, and \$240,000.

a. Calculate the mean and median salary.
b. Which do you feel would be a more realistic measure of central tendency, the mean or median salary?

2.15 Experience has shown that many *population* histograms take on a similar appearance. In other words, there are certain common or popular shapes that seem to reoccur.

List population values that might take on each of the following shapes.



2.16 For $n = 200$ randomly selected New England boutiques, 36 carried Rolf Laurie designer blouses.

a. What proportion (or fraction) of the sample, p_s, carries Rolf Laurie designer blouses?
b. Convert this fraction to a percentage.
c. Represent this proportion in a circle graph.

2.17 A West Coast professor's definition of good character includes the following two qualities, ''empathy, meaning regards for the needs, rights, and feelings of others, and self-control, meaning the ability to act with reference to the more distant consequences of current behavior.'' Suppose a test evaluating good character was administered to twenty-one local politicians, resulting in the following scores (10–49 scale):

18 26 23 21 26 29 30 33 32 34
38 38 36 37 41 40 47 41 42 43

a. Construct both a tally and a pictogram (invent your own symbol) using the following categories: 10–19, 20–29, 30–39, and 40–49.
b. Construct a stem-and-leaf display.
c. Construct a box-and-whisker plot.
d. Locate the quartile points, Q_1, Q_2, and Q_3.

(Reference article is *Newsweek*, ''A Sterner Kind of Caring,'' January 13, 1992, p. 68.)

2.18 ''While IQ tests remain excellent predictors of how well one will do in school, they have little or nothing to do with who will earn the most money or prestige, or have the most satisfying social life or relationships,'' according to *New York Times* article, ''New Scales of Intelligence Rank Talent for Living'' (April 5, 1988, p. C1). ''One factor emerging as crucial for life success is what might be called emotional intelligence. How well people manage their emotions determines how effectively they can use their intellectual ability.''

Suppose a test evaluating emotional IQ was administered to nineteen corporate executives, resulting in the following scores (20.0–23.9 scale):

20.7 20.6 21.3 21.8 21.9 21.8 22.4 22.4 22.5 22.4
22.8 22.6 22.9 23.1 23.4 23.5 23.2 23.9 23.7

a. Construct both a tally and a pictogram (invent your own symbol) using the following categories: 20.0–20.9, 21.0–21.9, 22.0–22.9, and 23.0–23.9.

b. Construct a stem-and-leaf display.

c. Construct a box-and-whisker plot.

d. Locate the quartile points, Q_1, Q_2, and Q_3.

Research Projects

The following research projects are offered for class or computer laboratory assignment (note: computer usage is optional). Each requires some preliminary class discussion and the student will issue a two to five page research report for each project.

Class Project A

Use the information and data from the Countrygirl Makeup example in section 2.1.

a. Discuss in class how one might design such a study to provide for (i) internal validity and (ii) external validity.

b. If students have access to a statistical computer package, feed in the raw (ungrouped) data and use the computer to analyze this input—that is, to calculate the mean, median, standard deviation, and other vital information.

c. If students do not have access to a statistical computer package, analyze the data using the grouping techniques as demonstrated in homework exercise 2.8.

d. Issue a research report, as described in section 2.8.

Class Project B

In a medical study, a researcher wished to estimate the average length of time needed for a particular nurse-in-training to draw a series of blood specimens. A sample of 49 observations of the nurse's work over several months yielded the following times (in minutes):

7.3	6.9	15.9	5.4	13.1
9.9	8.4	9.6	4.1	11.1
12.1	14.1	11.3	10.5	4.7
5.7	5.1	10.7	7.3	9.9
12.5	7.3	6.7	7.1	8.1
5.7	16.5	10.8	11.0	8.9
13.7	10.6	7.4	8.5	4.2
15.4	8.1	11.8	9.5	6.3
5.4	3.9	12.0	13.7	10.6
6.6	9.5	8.0	9.4	

a. Discuss in class how one might design such a study to provide for (i) internal validity and (ii) external validity.

b. If students have access to a statistical computer package, feed in the raw (ungrouped) data above and use the computer to analyze this input—that is, to calculate the mean, median, standard deviation, and other vital information.

c. If students do not have access to a statistical computer package, analyze the data using the grouping techniques as demonstrated in homework exercise 2.4.

d. Issue a research report, as described in section 2.8.

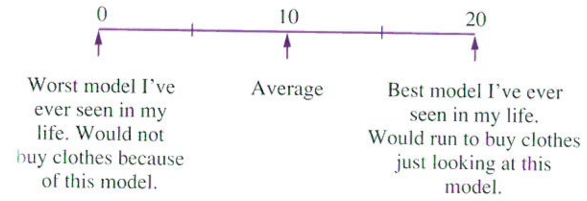
Class Project C

Needed for project: Magazine clippings of two different male models.

Background: A chain of exclusive men's shops in your region, A. L. Lewton, is in need of an on-going model to represent their high-priced line of clothes. Executives of the chain have narrowed the selection to two finalists. Although a professional study was considered, the executives felt the cost was prohibitive and asked your professor to sample the class, feeling that students, young men and women, would be more at the forefront of current taste in clothes. Your professor warned the executives that there might be some questions as to the validity of the results, however the executives insisted.

To gather this data, any plan can be devised. One possible plan might be as follows:

1. Put the following scale on the chalkboard.


0
10
20
Worst model I've ever seen in my life. Would not buy clothes because of this model.
Average
Best model I've ever seen in my life. Would run to buy clothes just looking at this model.

2. Students are informed that magazine clippings of the two male models, labeled Model A and Model B, are shown to them and *each* model is to be rated on the 0 to 20 scale, using the following criteria:

   Suitability for modeling clothes

   - manly structure
   - pleasing looks

   Appropriateness for A. L. Lewton image

   Overall appeal

3. To ensure honest unbiased responses, the vote is to be secret and no one is to discuss the models.
4. One student will be in charge of distributing and collecting the voting slips while the professor walks around the room showing the clippings of the two models, side by side, to various parts of the class.
5. A student then collects and reads the votes aloud as the professor lists the votes on the chalkboard, as follows:

A	**B**
x	x
x	x
x	x
.	.
.	.
.	.

In a computer environment, the students will input the data as the votes are read aloud, first for Model A, then for Model B.

a. For this study, discuss (i) internal validity and (ii) external validity.

b. Feed in raw (ungrouped) data into the computer and calculate the mean, median, standard deviation, and other vital information.

c. If students do not have access to a computer package, analyze the data using the grouping techniques discussed in this chapter.

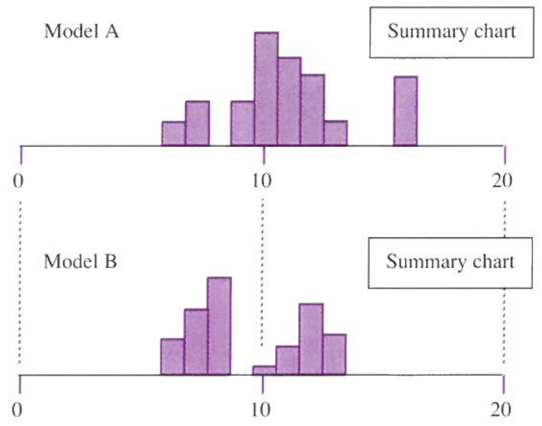
d. Issue a research report, as described in section 2.8.

Hints: Discussion of validity may include:

Was the scale objective? Does a vote from one person of, say, 12 reflect the same meaning as a vote of 12 from another person? Did student comments influence other students? Would different photographs of the models influence the ratings?

Was a random sample from the population achieved? Did the professor influence the vote in any way? Did the environment of the class or laboratory present a proper atmosphere for the voting or affect the vote in any way?

Results might include two histograms with the scales precisely lined up, making it easy for the reader to visually compare the outcome.



Note: these histograms are for demonstration purposes only.