Chapter 5
Exploring Data: Distributions

For All Practical Purposes: Effective Teaching

- Students expect that you are knowledgeable of your discipline, but should not expect that you are the all-knowing instructor. If you don't know the answer to a student's question, admit it and let them know you will do your best to find out the answer as soon as possible. Knowing that you are willing to learn should put students at ease.
- Other than knowledge of the topics presented in the course, organization is one of the most crucial elements in your classroom presentations. If your teaching sessions are well organized, students will attend because they know information will be presented in a logical and straightforward manner.

Chapter Briefing

In this chapter, you will be doing *exploratory data analysis*. This combines numerical summaries with graphical display to see patterns in a set of data. One difficulty you may encounter imparting to students is that there are choices that can be made in organizing data. Therefore, there is some subjectivity involved in the organization and interpretation of data. Students should find other topics, such as calculating mean, more straightforward, but tedious. Being well prepared for class discussions with short examples and knowledge of how to organize and interpret data is essential in order to help students focus on the main topics presented in this chapter.

In order to facilitate your preparation, the **Chapter Topics to the Point** has been broken down into the following.

- **Data Sets**
- **Histograms**
- **Stemplots**
- **Mean**
- **Median**
- **Quartiles, Five – Number Summary, and Boxplots**
- **Variance and Standard Deviation**
- **Normal Distributions**
- **The 68 – 95 – 99.7 Rule**

For each of the areas of *displaying data* (histograms and stemplots), *describing center* (mean and median), *describing spread* (quartiles and standard deviation), *quick summary of center and spread* (five – number summary and boxplots), as well as the *normal distribution* and its relation to the *68 – 95 – 99.7 Rule*, examples with solutions that do not appear in the text nor study guide are included in the *Teaching Guide*. You should feel free to use these examples in class, if needed.

Since you may be asked to demonstrate the techniques of this chapter using graphing calculator, the *Teaching Guide* includes the feature **Teaching the Calculator**. It includes brief calculator instructions with screen shots from a TI-83.

The last section of this chapter of *The Teaching Guide for the First-Time Instructor* is **Solutions to Student Study Guide ✎ Questions**. These are the complete solutions to the eight questions included in the *Student Study Guide*. Students only have the answers to these questions, not the solutions.

Chapter Topics to the Point

Data Sets

Throughout the chapter, you will be examining and interpreting **Data**. These numerical facts are essential for making decisions in almost every area of our lives. In a data set there are **individuals**. These individuals may be people, cars, cities, or anything to be examined. The characteristic of an individual is a **variable**. For different individuals, a variable can take on different values.

Teaching Tip

This is most likely not the first time students have been exposed to data sets. At this time you may choose to have students discuss where in their everyday lives data are used. They may respond with information given on the TV news or newspaper. As a society, we are exposed to data everyday. Because many of your students are relatively young, many of them have had the impact of high car insurance rates due to their age. This is an example of how data have been collected and age has been interpreted as a risk factor for causing insurance claims.

Teaching Tip

A good place to start class discussion of data sets is to collect information from the class to be used throughout the discussions of the various topics. You may choose to collect information such as age. In order to make the data set more diverse, you may choose to have students give their age as a one decimal approximation, such as 19.3 yrs. This will allow you the opportunity to discuss rounding as it will be needed in the chapter.

Histograms

The **distribution** of a variable tells us what values the variable takes and how often it takes these values. The most common graph of a distribution with one numerical variable is called a **histogram**.

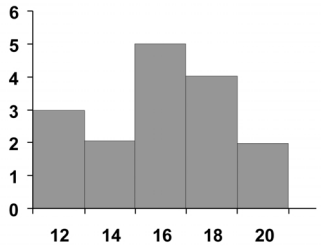
Example

Construct a histogram given the following data.

Value	Count
12	3
14	2
16	5
18	4
20	2

Solution

In this example, the data do not need to be grouped in order to be displayed. Notice that the bars meet halfway between the values of pieces of data on the horizontal axis.



When constructing a histogram, each piece of data must fall into one **class**. Each class must be of equal width. For any given data set, there is more than one way to define the classes. Either you are instructed as to how to define the classes, or you must determine class based on some criteria.

☝Teaching Tip

One difficulty students may encounter in this chapter is defining the classes of equal width and how the intervals affect the actual histogram. You may choose to discuss Example 2 from the text by pointing out how the classes are defined in Step 1 (noting the inequalities) and how they relate to the classes in Step 2. The choices made in Step 1 have the impact on the labeling of the intervals on the actual histogram in Step 3.

Example

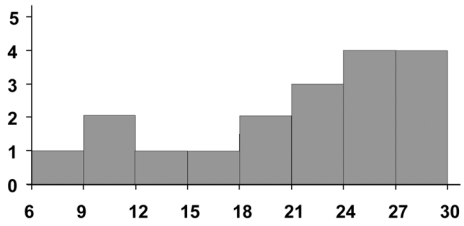
Given the following 18 quiz scores (out of 30 points), construct a histogram.

12	16	13	9	28	10	22	25	29
20	24	27	28	25	24	26	19	30

Solution

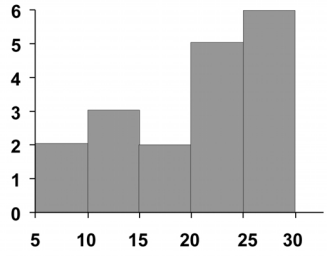
Since there is one student that obtained a perfect score, it makes sense to have the last class end with 30. There are different lengths one could try here for class widths.
One length could be 3 units. Since there are no students that obtained scores in the first two classes, one may opt not to include these classes on the histogram.

Class	Count
1 – 3	0
4 – 6	0
7 – 9	1
10 – 12	2
13 – 15	1
16 – 18	1
19 – 21	2
22 – 24	3
25 – 27	4
28 – 30	4



Another acceptable class width would be 3 units.

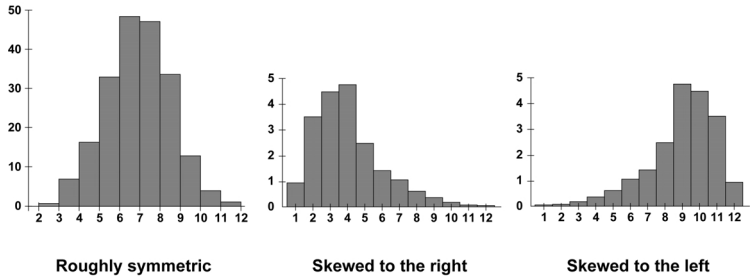
Class	Count
1 – 5	0
6 – 10	2
11 – 15	3
16 – 20	2
21 – 25	5
26 – 30	6



☝Teaching Tip

You may choose to discuss with students as to which histogram they feel is better. Notice in the histogram that the labels are one of endpoints of each of the classes. Students might also ask about where the classes must start. You may tell them that this is generally a matter of examining the data and determining what makes the most sense. The first class does not necessarily have to start at zero, nor does the first bar have to touch the vertical axis.

An important feature of a histogram is its overall **shape**: Although there are many shapes and overall patterns, a distribution may be **symmetric**, **skewed to the right**, or **skewed to the left**.



Teaching Tip

Students often confuse skewed to the right with skewed to the left. If a distribution is skewed to the right, then the larger values extend out much further to the right. If a distribution is skewed to the left then, the smaller values extend out much further to the left. The easiest way to keep the two terms from being confused is to think of the direction of the "tail". If the tail points left, it is skewed to the left. If the tail points right, it is skewed to the right.

Some important features of a distribution are as follows.

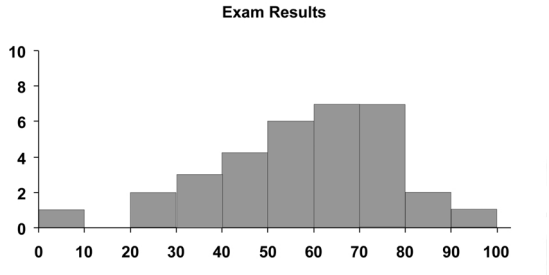
- Another way to describe a distribution is by its **center**. For now, we can think of the center of a distribution as the midpoint.
- Another way to describe a distribution is by its **spread**. The spread of a distribution is stating its smallest and largest values.
- In a distribution, we may also observe **outliers**; that is, a piece or pieces of data that fall outside the overall pattern. Often times determining an outlier is a matter of judgment. There are no hard and fast rules for determining outliers.

Example

Given the following data regarding exam scores, construct a histogram. Describe its overall shape and identify any outliers.

Class	Count	Class	Count
0 – 9	1	50 – 59	6
10 – 19	0	60 – 69	7
20 – 29	2	70 – 79	7
30 – 39	3	80 – 89	2
40 – 49	4	90 – 99	1

Solution



The shape appears to be skewed to the left. The score in the class 0 – 9, inclusive, could be considered an outlier.

Stemplots

A **stemplot** is a good way to represent data for small data sets. Stemplots are quicker to create than histograms and give more detailed information. Each value in the data set is represented as a stem and a leaf. The stem consists of all but the rightmost digit and the leaf is the rightmost digit.
If the data stem (left-hand side) has increasing values in the downward direction, then the plot can be turned $90°$ counter-clockwise in order to resemble a histogram.

Teaching Tip

Students will sometimes need to alter the data (round or even truncate) in order to make their stemplots. You may wish to tell students that they should alter the data in such a way that only one digit becomes the leaf (right-hand side). You may wish to tell students that stemplots can be helpful in organizing larger data sets. This will need to be done later in the chapter.

Example

Recall the 18 quiz scores (out of 30 points) stated earlier. Each score has been converted to a percentage (rounded to the nearest tenth of a percent). Construct a stemplot.

20.0%	53.3%	43.3%	30.0%	93.3%	33.3%	73.3%	83.3%	96.7%
66.7%	80.0%	90.0%	93.3%	83.3%	80.0%	86.7%	63.3%	100.0%

Solution

Given the format of the converted scores, we need to further round to the nearest whole percent. The stemplot would not be meaningful with the tenth of a percent being the leaf.

20%	53%	43%	30%	93%	33%	73%	83%	97%
67%	80%	90%	93%	83%	80%	87%	63%	100%

In the stemplot, the ones digit will be the leaf.

2	0
3	03
4	3
5	3
6	37
7	3
8	0337
9	37
10	0

Mean

A measure of center of data is the **mean**. It is obtained by adding the values of the observations in the data set and dividing by the number of data. The mean is written as $\bar{x}$. The formula for the mean is $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$, where n represents the number of pieces of data.

Example

Calculate the mean of each following data set.

a) 13, 6, 8, 12, 15, 14, 26, 12, 10, 11

b) 20, 61, 3, 2, 4, 5, 10, 7, 2

Solution

a) $\bar{x} = \frac{13+6+8+12+15+14+26+12+10+11}{10} = \frac{127}{10} = 12.7$

b) $\bar{x} = \frac{20+61+3+2+4+5+10+7+2}{9} = \frac{114}{9} \approx 12.7$

Teaching Tip

If you are a teaching assistant, you may take the opportunity to clarify with your faculty member as to their expectations of students using technology in their calculations. Since many calculators have statistical capabilities, you want to make it clear to students as to how much work needs to be shown in homework, quizzes and exams.

Teaching Tip

In the last example, the two data sets yielded approximately the same mean. You may choose to discuss with students the differences between the two sets, noting the outlier in the second data set (part b). You may also choose to discuss the need for rounding. Later in the chapter, students will need to increase accuracy for intermediate calculation.

Median

The **median**, ***M***, of a distribution is a number in the middle of the data, so that half of the data are above the median, and the other half are below it. When determining the median, the data should be placed in order, typically smallest to largest. When there are *n* pieces of data, then the piece of data $\frac{n+1}{2}$ observations up from the bottom of the list is the median. This is fairly straightforward when *n* is odd. When there are *n* pieces of data and *n* is even, then you must find the average (add together and divide by two) of the two center pieces of data. The smaller of these two pieces of data is located $\frac{n}{2}$ observations up from the bottom of the list. The second, larger, of the two pieces of data is the next one in order or, $\frac{n}{2}+1$ observations up from the bottom of the list.

Teaching Tip

Since students will be examining two measures of center in this chapter (mean and median), you may wish to emphasize to students that "median" is a word used on the roadway. The median divides the two sides of the road.

Teaching Tip

Students should get into the habit of organizing the data from smallest to largest. If allowed, students can use certain models of calculators or spreadsheets to readily perform this task. If students are encouraged to use technology, they should be instructed to double-check their data after they have entered it. Checking that the number of pieces of data is correct and scanning the data to make sure it looks correctly entered, prior to organizing it, will save students time in the long run.

Example

Calculate the median of each data set.

a) 13, 6, 8, 12, 15, 14, 26, 12, 10, 11

b) 20, 61, 3, 2, 4, 5, 10, 7, 2

Solution

For each of the data sets, the first step is to place the data in order from smallest to largest.

a) 6, 8, 10, 11, **12**, **12**, 13, 14, 15, 26

Since there are 10 pieces of data, the mean of the $\frac{10}{2} = 5^{th}$ and 6^{th} pieces of data will be the median. Thus, the median is $\frac{12+12}{2} = \frac{24}{2} = 12$. Notice, if you use the general formula $\frac{n+1}{2}$, you would be looking for a value $\frac{10+1}{2} = \frac{11}{2} = 5.5$ "observations" from the bottom. This would imply halfway between the actual 5^{th} observation and the 6^{th} observation. Notice since the 5^{th} observation and the 6^{th} observation were the same, we didn't really need to calculate the median.

b) 2, 2, 3, 4, **5**, 7, 10, 20, 61

Since there are 9 pieces of data, the $\frac{9+1}{2} = \frac{10}{2} = 5^{th}$ piece of data, namely 5, is the median.

Teaching Tip

In determining the median, you may choose to show students to "cover up" then end values and work their way towards the center. In Part a of the last example, we have the following.

6, 8, 10, 11, 12, 12, 13, 14, 15, 26
8, 10, 11, 12, 12, 13, 14, 15
10, 11, 12, 12, 13, 14
11, 12, 12, 13
12, 12
$\frac{12+12}{2} = \frac{24}{2} =$ **12**

In Part b of the last example, we have the following.

2, 2, 3, 4, 5, 7, 10, 20, 61
2, 3, 4, 5, 7, 10, 20
3, 4, 5, 7, 10
4, 5, 7
5

Example

Given the following stemplot, determine the median.

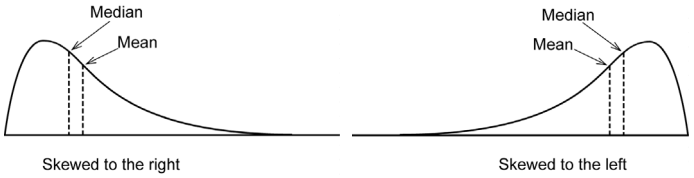
11	029
12	3478
13	034679
14	012359
15	01359
16	09
17	1
18	0

Solution

Since there are 28 pieces of data, the mean of the $\frac{28}{2} = 14^{th}$ and 15^{th} pieces of data will be the median. Thus, the median is $\frac{140+141}{2} = \frac{281}{2} = 140.5$. Notice, if you use the general formula $\frac{n+1}{2}$, you would be looking for the value $\frac{28+1}{2} = \frac{29}{2} = 14.5$ "observations" from the bottom (or top).

Teaching Tip

After mean and median have been discussed, you may wish to revisit skewness of a distribution and how the mean and median are positioned relative to each other. In order to motivate the normal curve, you may also ask students to picture a distribution in which the mean and the median are the same.



Quartiles, Five-Number Summary and Boxplots

One way of describing the spread of data is the **five-number summary**. This summary consists of the median (M), quartiles (Q_1 and Q_3), and extremes (high and low). The **quartiles** Q_1 (the point below which 25% of the observations lie) and Q_3 (the point below which 75% of the observations lie) give a better indication of the true spread of the data. More specifically, Q_1 is the median of the data to the left of M (the median of the data set). Q_3 is the median of the data to the right of M.

A **boxplot** is a graphical (visual) representation of the five-number summary. A central box spans quartiles Q_1 and Q_3. A line in the middle of the central box marks the median, M. Two lines extend from the box to represent the extreme values.

Teaching Tip

In determining the five-number summary, there are four cases to consider in terms of difficulty in locating Q_1, M, and Q_3. Students will have the most difficulty in determining these values when the number of pieces of data is a multiple of four. If you let n be the number of pieces of data, then the order of difficulty would be as follows.

- $3 \equiv n \bmod 4$ Easiest i.e. 7, 11, 15, 19, 23,… pieces of data
- $2 \equiv n \bmod 4$ i.e. 6, 10, 14, 18, 22,… pieces of data
- $1 \equiv n \bmod 4$ i.e. 5, 9, 13, 17, 21,… pieces of data
- $0 \equiv n \bmod 4$ Hardest i.e. 4, 8, 12, 16, 20,… pieces of data

Example

Draw a boxplot for the following data set.

21, 25, 10, 40, 43, 19, 12

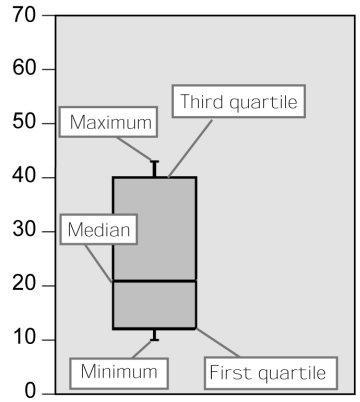
Solution

The first step is to place the data in order from smallest to largest.

10, **12**, 19, **21**, 25, **40**, 43

Since there are 7 pieces of data, the median is the $\frac{7+1}{2} = \frac{8}{2} = 4^{\text{th}}$ piece of data, namely 21. There are 3 pieces of data below the median, M. Thus, the $\frac{3+1}{2} = \frac{4}{2} = 2^{\text{nd}}$ piece of data is the first quartile. Thus, $Q_1 = 12$. Now since there are 3 pieces of data above M, Q_3 will be the 2^{nd} piece of data to the right of M. Thus, $Q_3 = 40$. The smallest piece of data is 10 and the largest is 43. Thus, the five-number summary is 10, 12, 21, 40, 43.

The boxplot is as follows.



Example

Draw a boxplot for the following data set.

31, 16, 11, 18, 10, 9, 12, 15, 15, 17, 20, 25

Solution

The first step is to place the data in order from smallest to largest.

9, 10, 12, 11, 15, 15, 16, 17, 18, 20, 25, 31

Since there are 12 pieces of data, the median is between the 6^{th} and 7^{th} pieces of data.

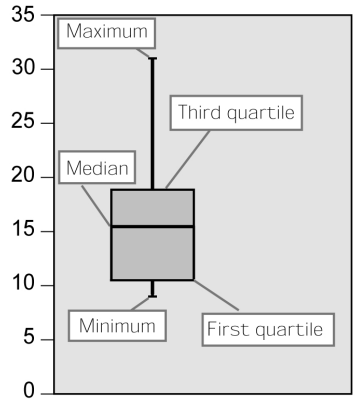
9, 10, 11, 12, 15, **15**, ↓ **16**, 17, 18, 20, 25, 31

Thus the median, M, is $\frac{15+16}{2} = \frac{31}{2} = 15.5$.

There are 6 pieces of data below M. Since $\frac{6+1}{2} = \frac{7}{2} = 3.5$, Q_1 will be the mean of 3^{rd} and 4^{th} pieces of data, namely $\frac{11+12}{2} = \frac{23}{2} = 11.5$. Now since there are 6 pieces of data above M, Q_3 will be the mean of the 3^{rd} and 4^{th} pieces of data to the right of M. Thus, $Q_3 = \frac{18+20}{2} = \frac{38}{2} = 19$.

9, 10, **11**, ↓ **12**, 15, 15, ↓ 16, 17, **18**, ↓ **20**, 25, 31

The smallest piece of data is 9, and the largest is 31. Thus, the five-number summary is 9, 11.5, 15.5, 19, 31. The boxplot is as follows.



Variance and Standard Deviation

Another way of describing the spread of data is **standard deviation**. The standard deviation, s, of a set of observations is the square root of the **variance** and measures the spread of the data around the mean in the same units of measurement as the original data set. The variance, s^2, of a set of observations is an average of the squared differences between the individual observations and their mean value. In symbols, the variance of n observations $(x_1, x_2, \ldots, x_n)$ is

$$s^2 = \frac{\left(x_1 - \bar{x}\right)^2 + \left(x_2 - \bar{x}\right)^2 + \ldots + \left(x_n - \bar{x}\right)^2}{n-1} \text{ or } s^2 = \frac{\sum_{i=1}^{n}\left(x_n - \bar{x}\right)^2}{n-1}.$$

Teaching Tip

If you are a teaching assistant, make sure you are aware of the requirements placed on students regarding how technology should be used in these calculations. If the faculty wishes to integrate forms of technology such as graphing calculators with statistical capabilities or spreadsheets, make sure you can demonstrate their use in the classroom.

☝Teaching Tip

A common student question regards the use of $n-1$ instead of n in the calculation of the sample variance. Its use is based on the fact that we mostly use the sample variance as an estimate of a population variance. Since the population variance is derived from the sample mean and the deviation of each measurement from the sample mean, we could not calculate the population variance if we were missing any one of these measurements (the mean or a single deviation from the sample mean). So, with n pieces of data, only $n-1$ of them vary freely in order for us to calculate the missing piece of data, if we know the mean. $n-1$ is known as the number of **degrees of freedom** of our data set.

☝Teaching Tip

If students are to perform these calculations by hand, you may choose to suggest they put the data in order. With that, they can view the deviations in order. If the sum of the deviations is not zero (or very close due to rounding), the incorrect calculation would be easier to spot.

Example

Given the following data set, find the variance and standard deviation.

6.2, 2.7, 5.4, 8.1, 5.2, 4.9, 7.8

Solution

Placing the data in order (not required, but helpful) we have the following hand calculations.

Observations x_i	Deviations $x_i - \bar{x}$	Squared deviations $\left(x_i - \bar{x}\right)^2$
2.7	$2.7 - 5.757143 = -3.057143$	$\left(-3.057143\right)^2 \approx 9.34612$
4.9	$4.9 - 5.757143 = -0.85714$	$\left(-0.85714\right)^2 \approx 0.73469$
5.2	$5.2 - 5.757143 = -0.55714$	$\left(-0.55714\right)^2 \approx 0.31041$
5.4	$5.4 - 5.757143 = -0.35714$	$\left(-0.35714\right)^2 \approx 0.12755$
6.2	$6.2 - 5.757143 = 0.442857$	$\left(0.442857\right)^2 \approx 0.19612$
7.8	$7.8 - 5.757143 = 2.042857$	$\left(2.042857\right)^2 \approx 4.17326$
8.1	$8.1 - 5.757143 = 2.342857$	$\left(2.342857\right)^2 \approx 5.48898$
sum = 40.3	sum = -0.000001	sum = 20.37714

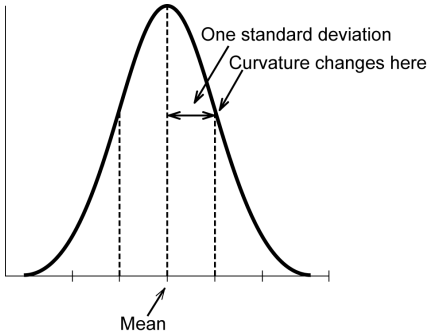
$\bar{x} = \dfrac{40.3}{7} \approx 5.757$ (we used $\bar{x} \approx 5.757143$ in the deviations calculations for better accuracy and rounded to five decimal places in the calculation of squared deviations) and $s^2 \approx \dfrac{20.37714}{7-1} = \dfrac{20.37714}{6} \approx 3.3962$ which implies $s \approx \sqrt{3.3962} \approx 1.843$.

☝Teaching Tip

Although performing the calculations to determine the variance and in turn the standard deviation can be tedious, it is an opportunity to discuss accumulated error caused by rounding at each step.

Normal Distributions

Sampling distributions, and many other types of probability distributions, approximate a bell curve in shape and symmetry. This kind of shape is called a normal curve, and can represent a **normal distribution**, in which the area of a section of the curve over an interval coincides with the proportion of all values in that interval. The area under any normal curve is 1. A normal curve is uniquely determined by is mean and standard deviation. The **mean** of a normal distribution is the center of the curve. The symbol μ will be used for the mean. The **standard deviation** of a normal distribution is the distance from the mean to the point on the curve where the curvature changes. The symbol σ will be use for the standard deviation.



The first quartile is located 0.67 standard deviation below the mean, and the third quartile is located 0.67 standard deviation above the mean. In other words, we have the following formulas.

$$Q_1 = \mu - 0.67\sigma \text{ and } Q_3 = \mu + 0.67\sigma$$

Teaching Tip

Students may get confused as to the use of $\bar{x}$ versus μ for mean and *s* versus σ for standard deviation. The difference is that $\bar{x}$ and *s* are used for **sample** mean and standard deviation, respectively; whereas, μ and σ are used for **population** mean and standard deviation.

Example

The scores on a marketing exam were normally distributed with a mean of 68 and a standard deviation of 4.5.

a) Find the first and third quartile for the exam scores.

b) Find a range containing exactly 50% of the students' scores.

Solution

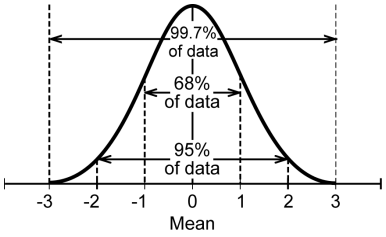
a) The quartiles are $\mu \pm 0.67\sigma = 68 \pm 0.67(4.5) = 68 \pm 3$, or $Q_1 = 65$ and $Q_3 = 71$.

b) Since 25% of the data lie below the first quartile and 25% of the data fall above the third quartile, 50% of the data would fall between the first and third quartiles. We would say an interval would be [65, 71].

The 68 – 95 – 99.7 Rule

In a normal curve, exactly half of the population falls below the mean and exactly half lie above. The **68–95–99.7 rule** applies to a normal distribution. It is useful in determining the proportion of a population with values falling in certain ranges. For a normal curve, the following rules apply:

- The proportion of the population within one standard deviation of the mean is 68%.
- The proportion of the population within two standard deviations of the mean is 95%.
- The proportion of the population within three standard deviations of the mean is 99.7%.



Teaching Tip

You may choose to instruct students that they should always draw a picture, like the one above when answering questions about the normal distribution. The drawing allows students to quickly envision how to exploit symmetry. Students should label mean and calculate values up to three standard deviations from the mean.

Example

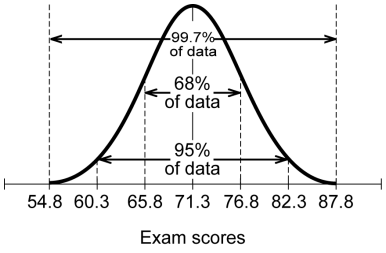
The scores on a marketing exam were normally distributed with a mean of 71.3 and a standard deviation of 5.5.

a) Almost all (99.7%) scores fall within what range?

b) What percent of scores are more than 82?

c) What percent of scores fall in the interval [66, 82]?

Solution



a) Since 99.7% of all scores fall within 3 standard deviations of the mean, we find the following.

$$\mu \pm 3\sigma = 71.3 \pm 3(5.5) = 71.3 \pm 16.5$$

$$71.3 - 16.5 = 54.8 \text{ and } 71.3 + 16.5 = 87.8$$

Thus, the range of scores is 54.8 to 87.8. If scores on the exam are understood to be whole numbers, then the range of scores would be the interval [55, 87].

b) Scores above 82, such as 83 or more are two σ above μ; 95% are within 2σ of μ. 5% lie farther than 2σ. Thus, half of these, or 2.5%, lie above 82.

c) 34% (half of 68%) of the scores would be between 66 and the mean. 47.5% (half of 95%) of the scores would be between the mean and 82. Thus, 34% + 47.5% = 81% of scores fall in the interval [66, 82].

☝Teaching Tip

As the chapter comes to a close, remind students of all the resources they have available to them in preparation for an examination. There are Skills Check exercises (with answers) in the text, Practice Quiz (with answers) in the *Student Study Guide*, flashcards of Review Vocabulary in the *Student Study Guide* as well as web versions for students that have Internet access. Students should be comfortable with organizing data into classes as well as displaying data in the forms of histograms and stemplots. They should be able to give some general features of the graph such as symmetry and skewness as well as be able to judge potential outliers of a data set. Also, students should be able to determine the five-number summary and create a boxplot. Students should also be able to calculate the mean, variance, and standard deviation of a data set and know how technology should be used in these calculations. Finally, students should know the features of a normal curve and how the 68–95–99.7 rule applies to a normal distribution as well as determining quartiles. If review sessions or other materials are made available, write this information on the board and refer to it several times before the examination date.

Teaching the Calculator

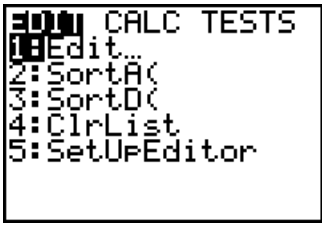
Example 1

Construct a histogram given the following.

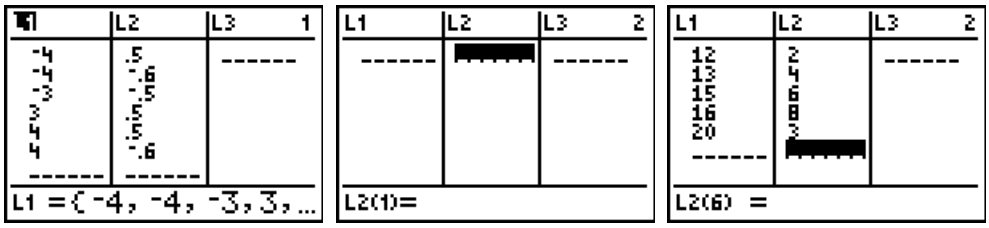
Value	Count
12	2
13	4
15	6
16	8
20	3

Solution

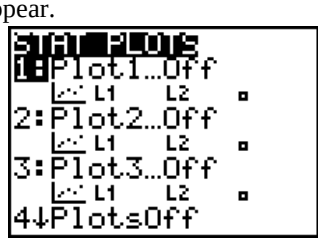
First enter the data by pressing the [STAT] button. The following screen will appear.



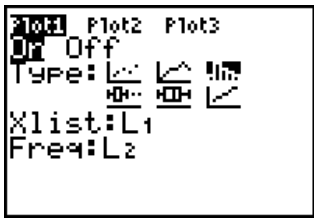
If there is data already stored, you may wish the clear it out. For example, if you wish to remove the data in L1, toggle to the top of the data and press [CLEAR] then [ENTER]. Repeat for any other data sets you wish to clear. Enter the new data being sure to press [ENTER] after each piece of data is displayed.



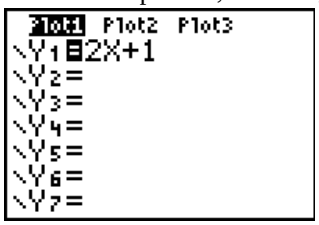
In order to display a histogram, you press [2nd] then [Y=]. This is equivalent to [STAT PLOT]. The following screen (or similar) will appear.

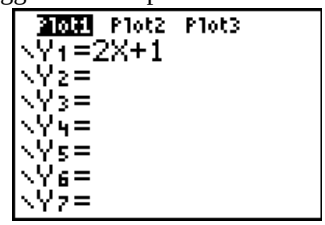


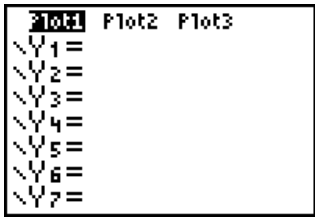
You will need to turn a stat plot On and choose the histogram option ( ). You will also need to make sure Xlist and Freq reference the correct data. In this case L1 and L2, respectively.



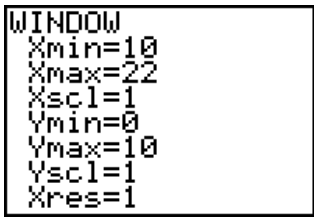
Next, you will need to make sure that no other graphs appear on your histogram. Press [Y=] and if another relation is present, either toggle to = and press enter to deselect or delete the relation.



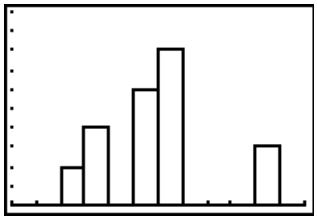




You will next need to choose an appropriate window. By pressing [WINDOW] you need to enter an appropriate window that includes your smallest and largest pieces of data. These values dictate your choices of Xmin and Xmax. Your choice of Xscl is determined by the kind of data you are given. In this case, the appropriate choice is 1. If you are given data such as 10, 12, 14, 16, and values such as 11, 13, and 15 are not considered then the appropriate choice would be 2 in order to make the vertical bars touch. In terms of choices for frequency, Ymin should be set at zero. Ymax should be at least as large as the highest frequency value. Your choice of Yscl is determined by how large the maximum frequency value is from your table.



Next, we display the histogram by pressing the [GRAPH] button.



Notice that the histogram differs slightly from how a hand drawing should be. Ideally, the base of each rectangle should be shifted left by half of a unit.

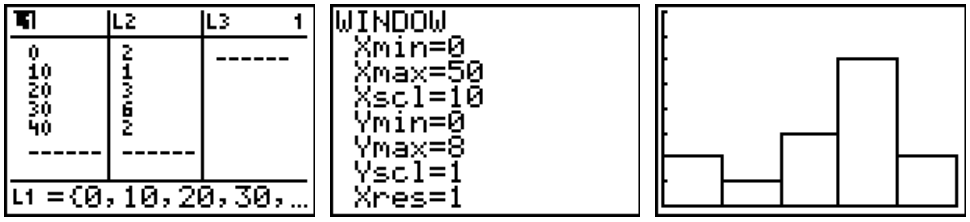
Example 2

Given the following data, construct a histogram.

Class	Count
0 – 9	2
10 – 19	1
20 – 29	3
30 – 39	6
40 – 49	2

Solution

Follow the instructions in Example 1 in order to input data and set up the window in order to display the histogram. The width of the classes should be the Xscl in order to make the vertical bars touch. Also, in a case like this where you are given classes, use the left endpoint of the class as data pieces.



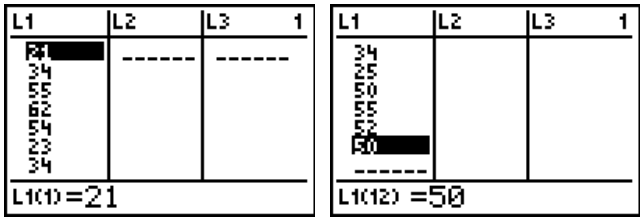
Example 3

Consider the following data.

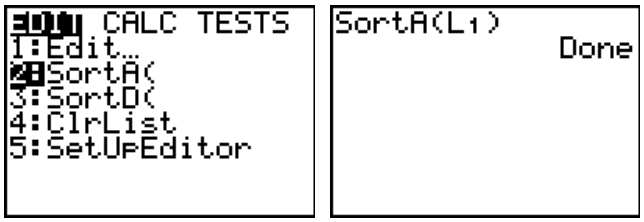
21, 34, 55, 62, 54, 23, 34, 25, 50, 55, 52, 50

- Arrange the data in order from smallest to largest.
- Find the mean.
- Find the standard deviation.
- Find the five – number summary.
- Display the boxplot.

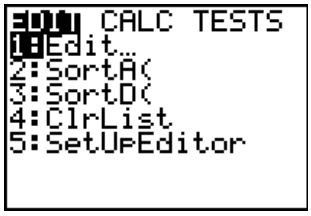
Enter the data, noting that there are 12 pieces of data. Make sure the location of the last entry corresponds to the total number of pieces of data.

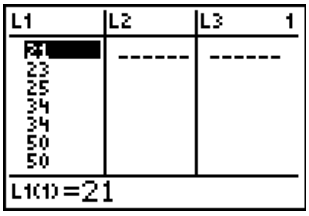


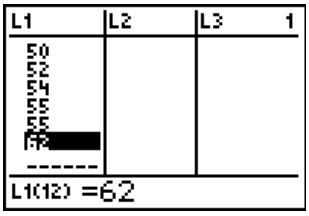
To arrange the data in order from smallest to largest, press the [STAT] button and choose the SortA(option which sorts the data in ascending order. Choose the appropriate data set (in this case L1) and then press [ENTER]. The calculator will display Done indicating the data is sorted.



By pressing the [STAT] button, you can then view the data arranged in order by choosing the Edit option.



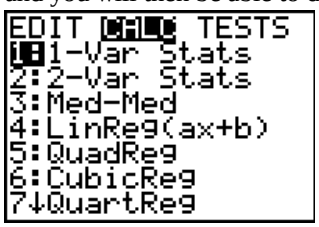


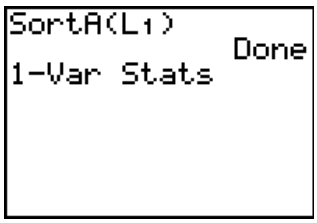


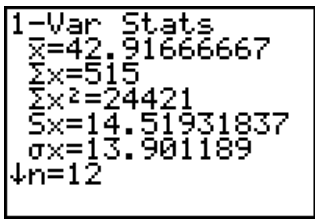
The data arranged from smallest to largest is as follows.

21, 23, 25, 34, 34, 50, 50, 52, 54, 55, 55, 62

To find the mean and standard deviation, press the [STAT] button. Toggle over to CALC and choose the 1-Var Stats option and then press the [ENTER]. You will get your home screen. Press [ENTER] again and you will then be able to determine the mean and standard deviation.

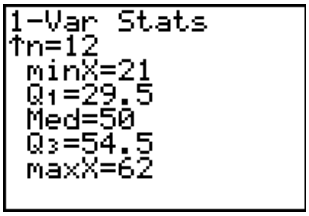






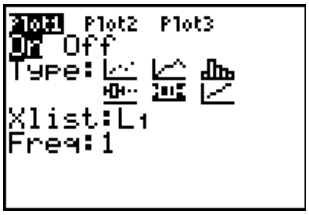
The mean is $\bar{x}$ (approximately 42.917) and the standard deviation is Sx (approximately 14.519).

To determine the five – number summary, from the last screen press the down arrow (▼) five times.

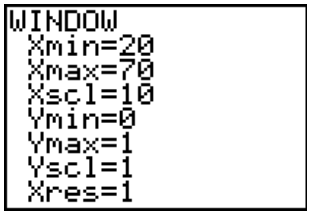


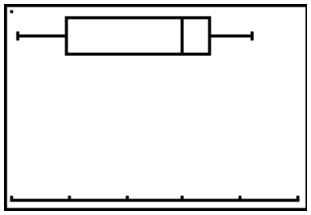
The five – number summary is 21, 29.5, 50, 54.5, 62.

To display the box plot, press [2nd] then [Y=]. This is equivalent to [STAT PLOT]. You will need to choose ⊡ for boxplot. Make sure the proper data are chosen for Xlist and Freq should be set at 1.



Choose an appropriate window for Xmin and Xmax based on the minimum and maximum values. The values you choose for Ymin and Ymax do not have an effect on the boxplot. You may choose values for Xscl and Yscl based on appearance of the axes. Display boxplot by pressing the [GRAPH] button.





Solutions to Student Study Guide Questions

Question 1

Given the following exam scores, describe the overall shape of the distribution and identify any outliers. In your solution, construct a histogram with class length of 5 points.

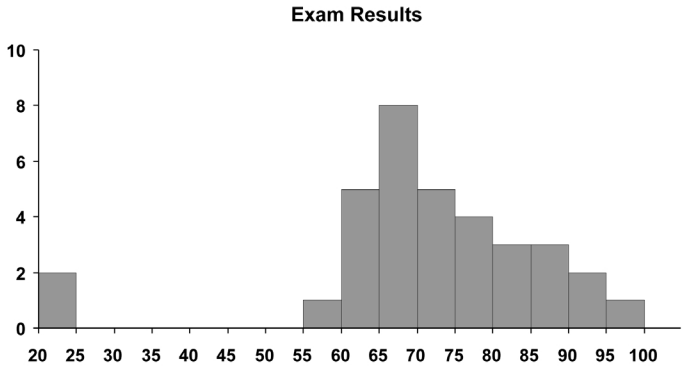
21	22	59	60	61	62	63	64	65
65	66	67	68	68	69	69	70	72
73	74	74	75	76	77	78	80	81
82	85	86	89	91	92	95		

Solution

It is helpful first to put the data into classes and count the individual pieces of data in each class. Since the smallest piece of data is 21, it makes sense to make the first class 20 to 24, inclusive.

Class	Count
20 – 24	2
25 – 29	0
30 – 34	0
35 – 39	0
40 – 44	0
45 – 49	0
50 – 54	0
55 – 59	1
60 – 64	5
65 – 69	8
70 – 74	5
75 – 79	4
80 – 84	3
85 – 89	3
90 – 94	2
95 – 99	1

The distribution appears to be skewed to the right. The scores of 21 and 22 appear to be outliers.



Question 2

The following are the percentages of salt concentrate taken from lab mixture samples. Describe the shape of the distribution and any possible outliers. This should be done by first rounding each piece of data to the nearest percent and then constructing a stemplot.

Sample	1	2	3	4	5	6	7
Percent	39.8	65.7	64.7	20.1	40.8	53.4	70.8
Sample	8	9	10	11	12	13	14
Percent	50.7	68.7	74.3	82.6	58.5	68.0	72.2

Solution

Rounding to full percents, we have the following.

Sample	1	2	3	4	5	6	7
Percent	40	66	65	20	41	53	71
Sample	8	9	10	11	12	13	14
Percent	51	69	74	83	59	68	72

The stemplot is as follows.

2	0
3	
4	01
5	139
6	5689
7	124
8	3

The distribution appears to be roughly symmetric with 20 as a possible outlier.

Question 3

Given the following stemplot, determine the mean. Round to the nearest tenth, if necessary.

1	259
2	3478
3	0334679
4	01259
5	46
6	1
7	3

Solution

$\bar{x} = \frac{12+15+19+23+24+27+28+30+33+33+34+36+37+39+40+41+42+45+49+54+56+61+73}{23} = \frac{851}{23} = 37$

Question 4

Given the following stemplot, determine the median.

1	029
2	3478
3	03345679
4	012359
5	16
6	012

Solution

Since there are 26 pieces of data, the mean of the $\frac{26}{2} = 13^{\text{th}}$ and 14^{th} pieces of data will be the median. This location could also be determined by applying the general formula, $\frac{26+1}{2} = \frac{27}{2} = 13.5$. This implies the median is halfway between the 13^{th} and 14^{th} pieces of data. Thus, $M = \frac{36+37}{2} = \frac{73}{2} = 36.5$.

Question 5

Determine the quartiles Q_1 and Q_3 of each data set.

a) 21, 16, 20, 6, 8, 9, 12, 15, 3, 15, 7, 8, 19

b) 14, 12, 11, 12, 24, 8, 6, 4, 8, 10

Solution

For each of the data sets, the first step is to place the data in order from smallest to largest.

a) 3, 6, **7**, **8**, 8, 9, **12**, 15, 15, **16**, **19**, 20, 21

Since there is 13 pieces of data, M is the $\frac{13+1}{2} = \frac{14}{2} = 7^{\text{th}}$ piece of data, namely 12. Thus, there are 6 pieces of data below M. Since $\frac{6+1}{2} = \frac{7}{2} = 3.5$, Q_1 will be the mean of 3^{rd} and 4^{th} pieces of data, namely $\frac{7+8}{2} = \frac{15}{2} = 7.5$. Now since there are 6 pieces of data above M, Q_3 will be the mean of the 3^{rd} and 4^{th} pieces of data to the right of M. Thus, $Q_3 = \frac{16+19}{2} = \frac{35}{2} = 17.5$.

b) 4, 6, **8**, 8, 10, ↓ 11, 12, **12**, 14, 24

Since there are 10 pieces of data, the mean of the $\frac{10}{2} = 5^{\text{th}}$ and 6^{th} pieces of data will be the median ($M = \frac{10+11}{2} = \frac{21}{2} = 10.5$). Since there are 5 pieces of data below M. We can therefore determine Q_1 to be the $\frac{5+1}{2} = \frac{6}{2} = 3^{\text{rd}}$ piece of data. Thus, $Q_1 = 8$. Now since there are 5 pieces of data above M, Q_3 will be the 3^{rd} piece of data to the right of M. Thus, $Q_3 = 12$.

Question 6

Given the following data, find the five-number summary and draw the boxplot.

12, 11, 52, 12, 15, 21, 17, 35, 16, 12

Solution

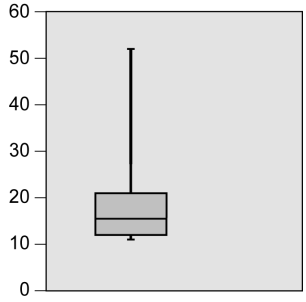
To determine the minimum, maximum, and median, we must put the 10 pieces of data in order from smallest to largest.

11, 12, **12**, 12, 15, ↓ 16, 17, **21**, 35, 52

The minimum is 11, and the maximum is 52. Since there are 10 pieces of data, the mean of the $\frac{10}{2} = 5^{\text{th}}$ and 6^{th} pieces of data will be the median. Thus, $M = \frac{15+16}{2} = \frac{31}{2} = 15.5$. Since there are 5 pieces of data below M, we can therefore determine Q_1 to be the $\frac{5+1}{2} = \frac{6}{2} = 3^{\text{rd}}$ piece of data. Thus, $Q_1 = 12$. Now since there are 5 pieces of data above M, Q_3 will be the 3^{rd} piece of data to the right of M. Thus, $Q_3 = 21$.

Thus, the five-number summary is 11, 12, 15.5, 21, 52.

The boxplot is as follows.



Question 7

Given the following data set, find the variance and standard deviation.

3.41, 2.78, 5.26, 6.49, 7.61, 7.92, 8.21, 5.51

Solution

Observations x_i	Deviations $x_i - \bar{x}$	Squared deviations $\left(x_i - \bar{x}\right)^2$
2.78	$2.78 - 5.89875 = -3.11875$	$(-3.11875)^2 = 9.7266015625$
3.41	$3.41 - 5.89875 = -2.48875$	$(-2.48875)^2 = 6.1938765625$
5.26	$5.26 - 5.89875 = -0.63875$	$(-0.63875)^2 = 0.4080015625$
5.51	$5.51 - 5.89875 = -0.38875$	$(-0.38875)^2 = 0.1511265625$
6.49	$6.49 - 5.89875 = 0.59125$	$(0.59125)^2 = 0.3495765625$
7.61	$7.61 - 5.89875 = 1.71125$	$(1.71125)^2 = 2.9283765625$
7.92	$7.92 - 5.89875 = 2.02125$	$(2.02125)^2 = 4.0854515625$
8.21	$8.21 - 5.89875 = 2.31125$	$(2.31125)^2 = 5.3418765625$
sum = 47.19	**sum =** 0	**sum =** 29.1848875

$\bar{x} = \frac{47.19}{8} = 5.89875$, $s^2 = \frac{29.1848875}{8-1} = \frac{29.1848875}{7} \approx 4.169$ and $s = \sqrt{4.169} \approx 2.04$.

Note: You can maintain less accuracy in the calculations, but since $\bar{x}$ was a terminating decimal, we went ahead and included all values in the table calculations.

Question 8

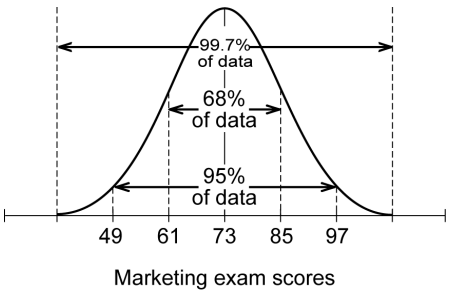
Look again at the marketing exam in which scores were normally distributed with a mean of 73 and a standard deviation of 12.

a) Find a range containing 34% of the students' scores.

b) What percentage of the exam scores were between 61 and 97?

Solution

The following diagram may be helpful in visualizing the intervals.



Marketing exam scores

a) There are different possible answers to this question. Applying only the 68–95–99.7 rule, half of 68%, namely 34%, lies either one standard deviation above the mean or one standard deviation below the mean. Thus, either of the intervals [61, 73] or [73, 85] are valid answers.

b) From part a we know that 34% of the exam scores lie in [61, 73]. In a similar fashion, we can determine that half of 95%, namely 47.5%, of the scores lie in the interval from 73 to 97. Thus, $34\% + 47.5\% = 81.5\%$ of the exam scores were between 61 and 97.