Objectives

- ★ Calculate and interpret the z-score of a data value as a measure of relative position
- ★ Calculate the five-number summary of a set of data
- ★ Calculate the value of a percentile and the z-score of a data value
- ★ Construct a box plot
- ★ Determine the percentiles and locations of specific data points
- ★ Determine the quartiles of a data set
- ★ Determine the value of a percentile for a given data set
- ★ Find the outliers in a given set of data
- ★ Find the percentile of a particular data value for a given data set
- ★ Find the quartiles of a given data set
- ★ Find the range, mean, median, and mode of a sample of data
- ★ Read and interpret box plots
- ★ Understand the terms: z-score, percentile, and quartile

Oct 2-5:25 PM

Measures of Relative Position

Suppose you want to know where an observation stands in relation to other values in a data set. For example, on many standardized tests such as the SAT, GMAT, and ACT, the test scores themselves are rather meaningless unless they are associated with some measure that tells you how well you did relative to others taking the same test. There are two principal methods of communicating relative position: percentiles and z-scores. Both of these methods are data transformations which change the scale of the data in some way

Oct 2-5:28 PM

Percentiles

The most commonly used measure of relative position is the percentile. In fact, we have already discussed the 50thpercentile; it is the median. For example, in data sets that do not contain significant quantities of identical data, the 30thpercentile is a value such that about 30 percent of the values are below it, and about 70 percent are above it.

Definition

Given a set of data $x_1, x_2, \ldots, x_n$, **the P^{th} percentile is a value, say X, such that approximately P percent** of the data is less than or equal to X and approximately $(100-P)$ percent of the data is greater than or equal to X.

Oct 2-5:29 PM

To determine the P^{th} percentile, perform the following steps.

Procedure: Finding the P^{th} Percentile

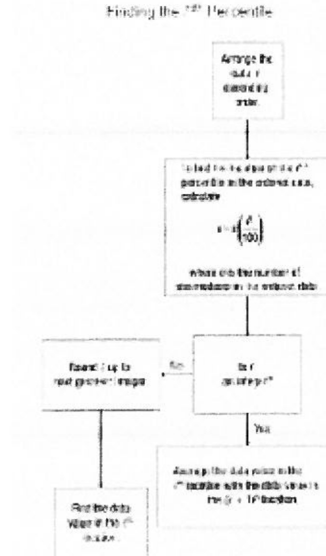
1. Form an ordered array by placing the data in order from smallest to largest.
2. To find the location of the P^{th} percentile in the ordered array, let

$$\ell = n\left(\frac{P}{100}\right)$$

where n is the number of observations in the ordered data.

3. If ℓ is not an integer, then round ℓ up to the next greatest integer. For example, if $\ell = 7.1$, then round ℓ up to 8 and find the data value in the ℓ^{th} location. If ℓ is an integer value, then average the data value in the ℓ^{th} location with the data value in the $(\ell+1)^{th}$ location.

Oct 2-5:31 PM



Oct 2-5:32 PM

It is important to remember that when you find the value of ℓ, this result is not the percentile. It is the *location* of the percentile in the ordered array. Thus, if the result of calculating (and rounding up) ℓ is 15, then the desired percentile would be the fifteenth value in the ordered list.

Interpreting Percentiles

When students take the SAT Reasoning Test, they receive a copy of their scores as well as the percentile they fall into. This percentile can sometimes be confusing. If a student receives a score of 620 on the reading section, they might fall into the 84^{th} percentile. This means **that they received a higher score than 84 percent of the students. The same score on the math section might place the student into the 80^{th}** percentile. Receiving a score of 800 on reading or math will put the student in the 99^{th} percentile. This means that less than 1 percent of the students taking the SAT had the same score.

Oct 2-5:32 PM

Example 4.19

Find the 50^{th} percentile for the following data set.

3, 5, 0, 1, 9, 2, 7

Solution

The number of observations, $n = 7$.

The percentile, $P = 50$.

The location of the percentile, $\ell = 7 \cdot \left(\frac{50}{100}\right) = 3.5$

Since the location of the percentile is not an integer, the value is rounded up to 4.

Thus, the fourth observation in the ordered array is the 50^{th} percentile.

0, 1, 2, [3], 5, 7, 9
↑
fourth observation

Therefore, the median value (which is the 50^{th} percentile) is 3.

Oct 2-5:33 PM

Example 4.20

Suppose the 40 members of your company are given a screening test for a new position. These scores are reported in Table 4.12. To inform potential employees of their screening test performance you may wish to report various percentiles for the test scores. Find the 10^{th} and 85^{th} percentiles for the test.

Table 4.12 – Test Scores

67	43	18	82
45	54	61	55
63	47	21	81
58	46	[illegible]	[illegible]
35	73	69	56
54	80	71	77
27	70	41	29
66	33	44	31
[illegible]	64	52	81
48	55	57	62

Table 4.13 – Ordered Test Scores

18	43	54	66
21	44	55	67
21	45	55	69
27	45	56	70
29	46	57	71
31	47	58	73
32	48	61	77
33	49	62	80
34	52	63	81
41	54	64	82

Oct 2-5:35 PM

Solution

In order to calculate the percentiles, the data must be placed in an ordered array (Table 4.13). To compute the 10^{th} percentile, its position in the ordered array must be determined.

The number of observations, $n=40$.

The percentile, $P=10$.

The location of the percentile, $\ell=40\cdot\left(\frac{10}{100}\right)=4$. Since ℓ is an integer, the 4^{th} and 5^{th} observations in the array must be averaged. Since the fourth data value is 27 and the fifth data value is 29, then the 10^{th} percentile is calculated as follows.

$$10^{th}\text{ percentile}=\frac{27+29}{2}=28$$

To determine the 88^{th} percentile, first calculate its location in the ordered array.

$$\ell=40\cdot\left(\frac{88}{100}\right)=35.2$$

Since the location is not an integer, its value is rounded up to 36. The 36^{th} observation in the ordered array will correspond to the 88^{th} percentile. The 36^{th} value is 73 in Table 4.13, so 73 is the 88^{th} percentile

Oct 2-5:37 PM

Formula: Percentile

The **percentile** of some data value x is given by:

$$\text{percentile of } x=\frac{\text{number of data values less than or equal to } x}{\textbf{total number of data values}}\cdot 100$$

Note that when finding the percentile of a specific value, if there are multiple occurrences of that value in the data, they all need to be counted in the numerator in order to calculate the percentile. To determine the percentile for a score of 56, the number of data values less than or equal to 56 must be counted. Since there are 24 data values less than or equal to 56, the resulting percentile would be

$$\text{percentile of a score of } 56=\frac{24}{40}\cdot 100=60.$$

Oct 2-5:37 PM

Formula: Percentile

The **percentile** of some data value x is given by:

$$\text{percentile of } x = \frac{\textbf{number of data values less than or equal to } x}{\textbf{total number of data values}} \cdot 100$$

Next, compute the percentile for a score of 67:

$$\text{percentile of a score of } 67 = \frac{32}{40} \cdot 100 = 80.$$

Oct 2-5:39 PM

Quartiles

The 25[th], 50[th], and 75[th] percentiles are known as **quartiles** and are denoted as Q_1, Q_2, and Q_3. They serve as markers that divide the data into four equal parts. Q_1 separates the lowest 25 percent, Q_2 represents the median (50[th] percentile), and Q_3 marks the beginning of the top 25 percent of the data.

Formula: Interquartile Range

The **interquartile range** is a measure of dispersion which describes the range of the middle fifty percent of the data.

$$\text{IQR} = Q_3 - Q_1$$

Oct 2-5:40 PM

Box Plots – Graphing with Quartiles

A very important use of quartiles is in the construction of box plots. As the name implies, box plots are graphical summaries of the data which, when constructed, have a box-like shape. They provide an alternative method to the histogram for displaying data. A box plot is a graphical summary of the central tendency, the spread, the skewness, and the potential existence of outliers in the data. Figure 4.11 displays a box plot of the screening test data from Example 4.20.

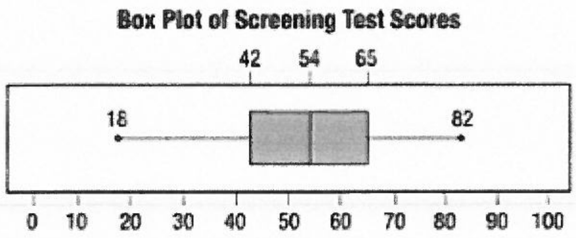


Figure 4.11

Table 4.13 – Ordered Test Scores			
18	43	54	66
21	44	55	67
21	45	55	69
27	45	56	70
29	46	57	71
31	47	58	73
32	48	61	77
33	49	62	80
34	52	63	81
41	54	64	82

Oct 2-5:41 PM

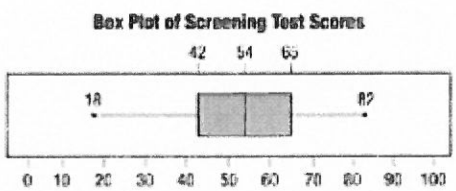


Figure 4.11

The box plot is constructed from five summary measures: the largest data value, the smallest data value, the 25th percentile, the 75th percentile, and the median.

Oct 2-5:44 PM

Although the box plot can be used to display data for a single data set, the histogram is probably more useful for this purpose. The real power of the box plot is the ease with which it allows the comparison of several data sets. Consider the number of wins per season for the New York Yankees, Los Angeles Dodgers, Atlanta Braves, and Chicago Cubs. The four data sets are displayed by box plots in Figure 4.12. It is easy to see from the box plots that the center of the Yankees' number of wins is higher than that for the Dodgers which has a higher center than the center of Braves, and finally the Cubs. Also it appears that the spread of the data, or the variation within the observed values, is not the same for all four data sets. This type of comparison will be used in later chapters to help confirm assumptions which must be made about the data in order to perform statistical inference.

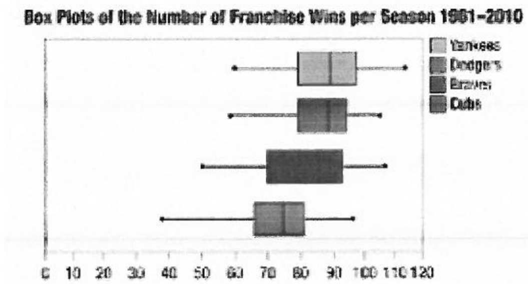


Figure 4.12

Oct 2-5:48 PM

Detecting Outliers

The concept of an outlier is an arbitrary concept. What you consider an outlier and what someone else considers an outlier may not be the same thing. However, one definition of an outlier which has gained some acceptance is developed in the context of a box plot.

Definition

A data point is considered an **outlier** if it is 1.5 times the interquartile range above the 75^{th} percentile or 1.5 times the interquartile range below the 25^{th} percentile.

Oct 2-5:49 PM

For example, suppose test scores of 110 and 2 were added to the screening test data.

Table 4.14 – Ordered New Test Scores

2	43	55	67
18	44	55	69
21	45	56	70
21	45	57	71
27	46	58	73
29	47	61	77
31	48	62	80
32	49	63	81
33	52	64	82
34	54	66	110
41	54		

For the new screening test data, the 25th percentile is 41 and the 75th percentile is 66, meaning that the interquartile range is $66-41=25$. A point is considered an outlier if the data point is:

larger than the 75th percentile +1.5 times the interquartile range.

$$66+1.5\cdot 25=103.5$$

smaller than the 25th percentile −1.5 times the interquartile range.

$$41-1.5\cdot 25=3.5$$

Since 110 is larger than 103.5, it is considered an outlier. Since 2 is smaller than 3.5, it is also considered an outlier. Figure 4.13 shows the box plot of the screening test data with the outliers incorporated. Notice that the whiskers did not change because 82 and 18 are still the largest and smallest observations within 1.5 times the interquartile range from the box.

Oct 2-5:51 PM

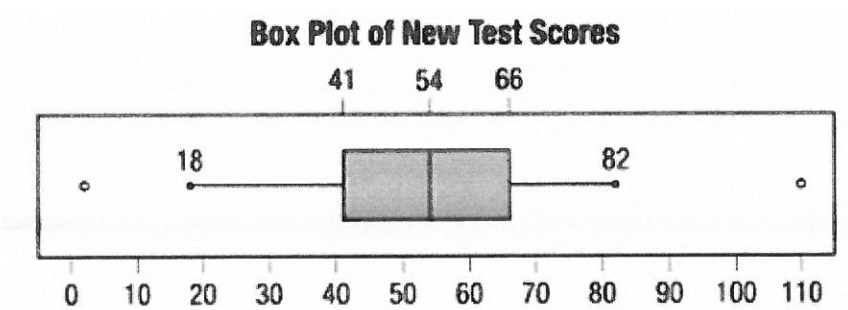


Figure 4.13

Oct 2-5:56 PM

z-Scores

The z-score is a standardized measure of relative position, with respect to the mean and variability (as measured by the standard deviation) of the data set.

Formula: z-Score

The z-score transforms a data value into the number of standard deviations that value is from the mean.

$$z = \frac{x - \mu}{\sigma}$$

Describing a data value by its number of standard deviations from the mean is a fundamental concept in statistics that is found throughout this course. It is used as a standardization technique to describe properties of data sets and to compare the relative values of data from different data sets.

Oct 2-5:57 PM

Example 4.22

Suppose you scored an 86 on your marketing test and a 94 on your management test. The mean and standard deviations of the two tests are given below.

Table 4.15 – Test Scores

Course	Mean	Standard Deviation
Marketing	74	10
Management	82	11

What are the z-scores for your two tests? On which of the tests did you perform relatively better?

Solution

The z-score for the marketing test is $z = \frac{86-74}{10} = 1.20$.

The z-score for the management test is $z = \frac{94-82}{11} \approx 1.09$.

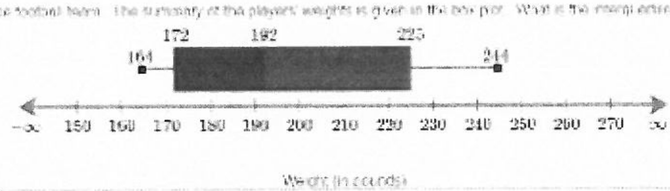
On the marketing test you scored 1.20 standard deviations above the mean, compared to only 1.09 standard deviations above the mean for the management test. Even though the raw score on the management test is larger than the raw score on the marketing test, relative to the means of the data sets, the performance on the marketing test was slightly better. Once again, changing the scale of the data has beneficial effects. It enables the comparison of two measurements that are drawn from different populations.

Oct 2-5:58 PM

If a *z*-score is negative, the data value is less than the mean. Conversely, if the *z*-score is positive, the data value is greater than the mean. The *z*-score is also a unit-free measure. That is, regardless of the original units of measurement (whether the data are measured in centimeters, meters, or kilometers), an observation's *z*-score will be the same.

Oct 2-6:01 PM

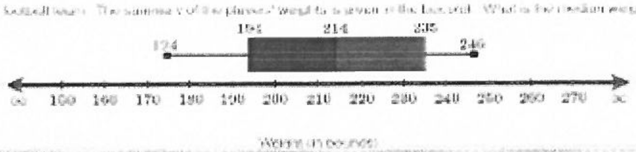
A high school has 30 players on the football team. The summary of the players' weights is given in the box plot. What is the interquartile range of the players' weights?



53 pounds

Oct 2-6:01 PM

A high school has 52 players on the football team. The summary of the players' weights is given in the box plot. What is the median weight of the players?



Answer | Tables | Keypad

214 pounds

Way to go!

Oct 2-6:02 PM

Consider a set of data in which the sample mean is 34.4 and the sample standard deviation is 3.7. Calculate the z-score given that $x = 35.8$. Round your answer to two decimal places.

Answer (How to Enter) | Tables | Keypad | Hide

.38

Good job, KIMBERLY!

Oct 2-6:03 PM

Given the following data, find the diameter that represents the 32nd percentile.

Diameters of Golf Balls

1.56	1.67	1.55	1.32	1.58
1.56	1.31	1.45	1.47	1.54
1.67	1.54	1.37	1.46	1.84

Answer (How to Enter)

1.46

Good job, KIMBERLY!

Oct 2-6:06 PM

The following stem-and-leaf plot represents the test scores for 26 students in a class on their most recent test. Use the data provided to find the quartiles.

Test Scores by Student

Stem	Leaves							
6	0	1	2	2	4	7	8	
7	0	1	2	5	8			
8	3	6	6	7	9	9		
9	0	1	2	2	2	4	5	8

Key: 6 | 0 = 60

Step 1 of 3 : Find the ***second*** quartile.

Answer (How to Enter) Tables

84.5

Good job!

Oct 2-6:09 PM

The following stem and leaf plot represents the test scores for 26 students in a class on their most recent test. Use the data provided to find the quartiles.

Test Scores by Student

Stem	Leaves							
6	0	1	2	2	4	7	8	
7	0	1	2	5	8			
8	3	6	6	7	9	9		
9	0	1	2	2	2	4	5	8

Key: 6 | 0 = 60

Step 2 of 3 : Find the **first** quartile.

Answer (How to Enter) Tables

66

Way to go!

Oct 2-6:14 PM

The following stem-and-leaf plot represents the test scores for 26 students in a class on their most recent test. Use the data provided to find the quartiles.

Test Scores by Student

Stem	Leaves							
6	0	1	2	2	4	7	8	
7	0	1	2	5	8			
8	3	6	6	7	9	9		
9	0	1	2	2	2	4	5	8

Key: 6 | 0 = 60

Step 3 of 3 : Find the third quartile.

Answer (How to Enter) Tables

9[illegible]

Way to go!

Oct 2-6:14 PM

[illegible]

[illegible] got a score of 80.8; this version has a mean of 62.1 and a standard deviation of 11

Carla got a score of 286.4; this version has a mean of 271 and a standard deviation of 22

Vincent got a score of 7.8; this version has a mean of 7.2 and a standard deviation of 0.7

[illegible]

Answer

Tables Keypad

[illegible]

Fantastic, KIMBERLY!

Oct 2-6:16 PM

Key Terms and Ideas

- Frequency Distribution
- Bar Chart
- Relative Frequency
- Cumulative Frequency
- Cumulative Relative Frequency
- Histogram
- Stem and Leaf Display
- Ordered Array
- Dot Plot
- Time Sequence Plot

Oct 2-6:17 PM