

DESCRIPTIVE STATISTICS: MEASURING THE LOCATION OF THE DATA*

Susan Dean
Barbara Illowsky, Ph.D.

This work is produced by OpenStax-CNX and licensed under the
Creative Commons Attribution License 3.0[†]

Abstract

Descriptive Statistics: Measuring the Location of Data explains percentiles and quartiles and is part of the collection coll0555 written by Barbara Illowsky and Susan Dean. Roberta Bloom contributed the section "Interpreting Percentiles, Quartile and the Median."

The common measures of location are **quartiles** and **percentiles** (%iles). Quartiles are special percentiles. The first quartile, Q_1 is the same as the 25th percentile (25th %ile) and the third quartile, Q_3 , is the same as the 75th percentile (75th %ile). The median, M , is called both the second quartile and the 50th percentile (50th %ile).

To calculate quartiles and percentiles, the data must be ordered from smallest to largest. Recall that quartiles divide ordered data into quarters. Percentiles divide ordered data into hundredths. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant.

The **interquartile range** is a number that indicates the spread of the middle half or the middle 50% of the data. It is the difference between the third quartile (Q_3) and the first quartile (Q_1).

$$\text{IQR} = Q_3 - Q_1 \quad (1)$$

The IQR can help to determine potential **outliers**. **A value is suspected to be a potential outlier if it is less than (1.5) (IQR) below the first quartile or more than (1.5) (IQR) above the third quartile.** Potential outliers always need further investigation.

*Version 1.17: May 16, 2012 4:25 am -0500

[†]<http://creativecommons.org/licenses/by/3.0/>

Example 1

For the following 13 real estate prices, calculate the IQR and determine if any prices are outliers. Prices are in dollars. (*Source: San Jose Mercury News*)

389,950; 230,500; 158,000; 479,000; 639,000; 114,950; 5,500,000; 387,000; 659,000; 529,000; 575,000; 488,800; 1,095,000

Solution

Order the data from smallest to largest.

114,950; 158,000; 230,500; 387,000; 389,950; 479,000; 488,800; 529,000; 575,000; 639,000; 659,000; 1,095,000; 5,500,000

$$M = 488,800$$

$$Q_1 = \frac{230500 + 387000}{2} = 308750$$

$$Q_3 = \frac{639000 + 659000}{2} = 649000$$

$$\text{IQR} = 649000 - 308750 = 340250$$

$$(1.5)(\text{IQR}) = (1.5)(340250) = 510375$$

$$Q_1 - (1.5)(\text{IQR}) = 308750 - 510375 = -201625$$

$$Q_3 + (1.5)(\text{IQR}) = 649000 + 510375 = 1159375$$

No house price is less than -201625. However, 5,500,000 is more than 1,159,375. Therefore, 5,500,000 is a potential **outlier**.

Example 2

For the two data sets in the test scores example, find the following:

- The interquartile range. Compare the two interquartile ranges.
- Any outliers in either set.
- The 30th percentile and the 80th percentile for each set. How much data falls below the 30th percentile? Above the 80th percentile?

Example 3: Finding Quartiles and Percentiles Using a Table

Fifty statistics students were asked how much sleep they get per school night (rounded to the nearest hour). The results were (student data):

AMOUNT OF SLEEP PER SCHOOL NIGHT (HOURS)	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE RELATIVE FREQUENCY
4	2	0.04	0.04
5	5	0.10	0.14
6	7	0.14	0.28
<i>continued on next page</i>			

7	12	0.24	0.52
8	14	0.28	0.80
9	7	0.14	0.94
10	3	0.06	1.00

Table 1

Find the 28th percentile: Notice the 0.28 in the "cumulative relative frequency" column. 28% of 50 data values = 14. There are 14 values less than the 28th %ile. They include the two 4s, the five 5s, and the seven 6s. The 28th %ile is between the last 6 and the first 7. **The 28th %ile is 6.5.**

Find the median: Look again at the "cumulative relative frequency" column and find 0.52. The median is the 50th %ile or the second quartile. 50% of 50 = 25. There are 25 values less than the median. They include the two 4s, the five 5s, the seven 6s, and eleven of the 7s. The median or 50th %ile is between the 25th (7) and 26th (7) values. **The median is 7.**

Find the third quartile: The third quartile is the same as the 75th percentile. You can "eyeball" this answer. If you look at the "cumulative relative frequency" column, you find 0.52 and 0.80. When you have all the 4s, 5s, 6s and 7s, you have 52% of the data. When you include all the 8s, you have 80% of the data. **The 75th %ile, then, must be an 8.** Another way to look at the problem is to find 75% of 50 (= 37.5) and round up to 38. The third quartile, Q_3 , is the 38th value which is an 8. You can check this answer by counting the values. (There are 37 values below the third quartile and 12 values above.)

Example 4

Using the table:

1. Find the 80th percentile.
2. Find the 90th percentile.
3. Find the first quartile. What is another name for the first quartile?
4. Construct a box plot of the data.

Collaborative Classroom Exercise: Your instructor or a member of the class will ask everyone in class how many sweaters they own. Answer the following questions.

1. How many students were surveyed?
2. What kind of sampling did you do?
3. Find the mean and standard deviation.
4. Find the mode.
5. Construct 2 different histograms. For each, starting value = _____ ending value = _____.
6. Find the median, first quartile, and third quartile.
7. Construct a box plot.
8. Construct a table of the data to find the following:
 - The 10th percentile
 - The 70th percentile
 - The percent of students who own less than 4 sweaters

Interpreting Percentiles, Quartiles, and Median

A percentile indicates the relative standing of a data value when data are sorted into numerical order, from smallest to largest. $p\%$ of data values are less than or equal to the p th percentile. For example, 15% of data values are less than or equal to the 15th percentile.

- Low percentiles always correspond to lower data values.
- High percentiles always correspond to higher data values.

A percentile may or may not correspond to a value judgment about whether it is "good" or "bad". The interpretation of whether a certain percentile is good or bad depends on the context of the situation to which the data applies. In some situations, a low percentile would be considered "good"; in other contexts a high percentile might be considered "good". In many situations, there is no value judgment that applies.

Understanding how to properly interpret percentiles is important not only when describing data, but is also important in later chapters of this textbook when calculating probabilities.

Guideline:

When writing the interpretation of a percentile in the context of the given data, the sentence should contain the following information:

- information about the context of the situation being considered,
- the data value (value of the variable) that represents the percentile,
- the percent of individuals or items with data values below the percentile.
- Additionally, you may also choose to state the percent of individuals or items with data values above the percentile.

Example 5

On a timed math test, the first quartile for times for finishing the exam was 35 minutes. Interpret the first quartile in the context of this situation.

- 25% of students finished the exam in 35 minutes or less.
- 75% of students finished the exam in 35 minutes or more.
- A low percentile could be considered good, as finishing more quickly on a timed exam is desirable. (If you take too long, you might not be able to finish.)

Example 6

On a 20 question math test, the 70th percentile for number of correct answers was 16. Interpret the 70th percentile in the context of this situation.

- 70% of students answered 16 or fewer questions correctly.
- 30% of students answered 16 or more questions correctly.
- Note: A high percentile could be considered good, as answering more questions correctly is desirable.

Example 7

At a certain community college, it was found that the 30th percentile of credit units that students are enrolled for is 7 units. Interpret the 30th percentile in the context of this situation.

- 30% of students are enrolled in 7 or fewer credit units
- 70% of students are enrolled in 7 or more credit units
- In this example, there is no "good" or "bad" value judgment associated with a higher or lower percentile. Students attend community college for varied reasons and needs, and their course load varies according to their needs.

Do the following Practice Problems for Interpreting Percentiles

Exercise 4

(Solution on p. 6.)

- a. For runners in a race, a low time means a faster run. The winners in a race have the shortest running times. Is it more desirable to have a finish time with a high or a low percentile when running a race?
- b. The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in the context of the situation.
- c. A bicyclist in the 90th percentile of a bicycle race between two towns completed the race in 1 hour and 12 minutes. Is he among the fastest or slowest cyclists in the race? Write a sentence interpreting the 90th percentile in the context of the situation.

Exercise 5*(Solution on p. 6.)*

- a. For runners in a race, a higher speed means a faster run. Is it more desirable to have a speed with a high or a low percentile when running a race?
- b. The 40th percentile of speeds in a particular race is 7.5 miles per hour. Write a sentence interpreting the 40th percentile in the context of the situation.

Exercise 6*(Solution on p. 7.)*

On an exam, would it be more desirable to earn a grade with a high or low percentile? Explain.

Exercise 7*(Solution on p. 7.)*

Mina is waiting in line at the Department of Motor Vehicles (DMV). Her wait time of 32 minutes is the 85th percentile of wait times. Is that good or bad? Write a sentence interpreting the 85th percentile in the context of this situation.

Exercise 8*(Solution on p. 7.)*

In a survey collecting data about the salaries earned by recent college graduates, Li found that her salary was in the 78th percentile. Should Li be pleased or upset by this result? Explain.

Exercise 9*(Solution on p. 7.)*

In a study collecting data about the repair costs of damage to automobiles in a certain type of crash tests, a certain model of car had \$1700 in damage and was in the 90th percentile. Should the manufacturer and/or a consumer be pleased or upset by this result? Explain. Write a sentence that interprets the 90th percentile in the context of this problem.

Exercise 10*(Solution on p. 7.)*

The University of California has two criteria used to set admission standards for freshman to be admitted to a college in the UC system:

- a. Students' GPAs and scores on standardized tests (SATs and ACTs) are entered into a formula that calculates an "admissions index" score. The admissions index score is used to set eligibility standards intended to meet the goal of admitting the top 12% of high school students in the state. In this context, what percentile does the top 12% represent?
- b. Students whose GPAs are at or above the 96th percentile of all students at their high school are eligible (called eligible in the local context), even if they are not in the top 12% of all students in the state. What percent of students from each high school are "eligible in the local context"?

Exercise 11*(Solution on p. 7.)*

Suppose that you are buying a house. You and your realtor have determined that the most expensive house you can afford is the 34th percentile. The 34th percentile of housing prices is \$240,000 in the town you want to move to. In this town, can you afford 34% of the houses or 66% of the houses?

**With contributions from Roberta Bloom

Solutions to Exercises in this Module

Solution to Example 2, Problem (p. 2)

For the IQRs, see the answer to the test scores example. The first data set has the larger IQR, so the scores between Q3 and Q1 (middle 50%) for the first data set are more spread out and not clustered about the median.

First Data Set

- $\left(\frac{3}{2}\right) \cdot (\text{IQR}) = \left(\frac{3}{2}\right) \cdot (26.5) = 39.75$
- $X_{\max} - Q_3 = 99 - 82.5 = 16.5$
- $Q_1 - X_{\min} = 56 - 32 = 24$

$\left(\frac{3}{2}\right) \cdot (\text{IQR}) = 39.75$ is larger than 16.5 and larger than 24, so the first set has no outliers.

Second Data Set

- $\left(\frac{3}{2}\right) \cdot (\text{IQR}) = \left(\frac{3}{2}\right) \cdot (11) = 16.5$
- $X_{\max} - Q_3 = 98 - 89 = 9$
- $Q_1 - X_{\min} = 78 - 25.5 = 52.5$

$\left(\frac{3}{2}\right) \cdot (\text{IQR}) = 16.5$ is larger than 9 but smaller than 52.5, so for the second set 45 and 25.5 are outliers.

To find the percentiles, create a frequency, relative frequency, and cumulative relative frequency chart (see "Frequency" from the Sampling and Data Chapter). Get the percentiles from that chart.

First Data Set

- 30th %ile (between the 6th and 7th values) = $\frac{(56 + 59)}{2} = 57.5$
- 80th %ile (between the 16th and 17th values) = $\frac{(84 + 84.5)}{2} = 84.25$

Second Data Set

- 30th %ile (7th value) = 78
- 80th %ile (18th value) = 90

30% of the data falls below the 30th %ile, and 20% falls above the 80th %ile.

Solution to Example 4, Problem (p. 3)

1. $\frac{(8 + 9)}{2} = 8.5$
2. 9
3. 6
4. First Quartile = 25th %ile

Solution to Exercise (p. 4)

- a. For runners in a race it is more desirable to have a low percentile for finish time. A low percentile means a short time, which is faster.
- b. INTERPRETATION: 20% of runners finished the race in 5.2 minutes or less. 80% of runners finished the race in 5.2 minutes or longer.
- c. He is among the slowest cyclists (90% of cyclists were faster than him.) INTERPRETATION: 90% of cyclists had a finish time of 1 hour, 12 minutes or less. Only 10% of cyclists had a finish time of 1 hour, 12 minutes or longer

Solution to Exercise (p. 5)

- a. For runners in a race it is more desirable to have a high percentile for speed. A high percentile means a higher speed, which is faster.
- b. INTERPRETATION: 40% of runners ran at speeds of 7.5 miles per hour or less (slower). 60% of runners ran at speeds of 7.5 miles per hour or more (faster).

Solution to Exercise (p. 5)

On an exam you would prefer a high percentile; higher percentiles correspond to higher grades on the exam.

Solution to Exercise (p. 5)

When waiting in line at the DMV, the 85th percentile would be a long wait time compared to the other people waiting. 85% of people had shorter wait times than you did. In this context, you would prefer a wait time corresponding to a lower percentile. INTERPRETATION: 85% of people at the DMV waited 32 minutes or less. 15% of people at the DMV waited 32 minutes or longer.

Solution to Exercise (p. 5)

Li should be pleased. Her salary is relatively high compared to other recent college grads. 78% of recent college graduates earn less than Li does. 22% of recent college graduates earn more than Li does.

Solution to Exercise (p. 5)

The manufacturer and the consumer would be upset. This is a large repair cost for the damages, compared to the other cars in the sample. INTERPRETATION: 90% of the crash tested cars had damage repair costs of \$1700 or less; only 10% had damage repair costs of \$1700 or more.

Solution to Exercise (p. 5)

- a. The top 12% of students are those who are at or above the **88th percentile** of admissions index scores.
- b. The **top 4%** of students' GPAs are at or above the 96th percentile, making the top 4% of students "eligible in the local context".

Solution to Exercise (p. 5)

You can afford 34% of houses. 66% of the houses are too expensive for your budget. INTERPRETATION: 34% of houses cost \$240,000 or less. 66% of houses cost \$240,000 or more.

Glossary

Definition 1: Interquartile Range (IRQ)

The distance between the third quartile (Q3) and the first quartile (Q1). $IQR = Q3 - Q1$.

Definition 2: Outlier

An observation that does not fit the rest of the data.

Definition 3: Percentile

A number that divides ordered data into hundredths.

Example

Let a data set contain 200 ordered observations starting with $\{2.3, 2.7, 2.8, 2.9, 2.9, 3.0, \dots\}$. Then the first percentile is $\frac{(2.7+2.8)}{2} = 2.75$, because 1% of the data is to the left of this point on the number line and 99% of the data is on its right. The second percentile is $\frac{(2.9+2.9)}{2} = 2.9$. Percentiles may or may not be part of the data. In this example, the first percentile is not in the data, but the second percentile is. The median of the data is the second quartile and the 50th percentile. The first and third quartiles are the 25th and the 75th percentiles, respectively.

Definition 4: Quartiles

The numbers that separate the data into quarters. Quartiles may or may not be part of the data. The second quartile is the median of the data.