Measures of position: Percentiles and quartiles

Recently, you learned that **measures of position** let you determine the position of a value in relation to other values in a dataset. Along with center and spread, it's helpful to know the relative position of your values. For example, whether one value is higher or lower than another, or whether a value falls in the lower, middle, or upper portion of your dataset.

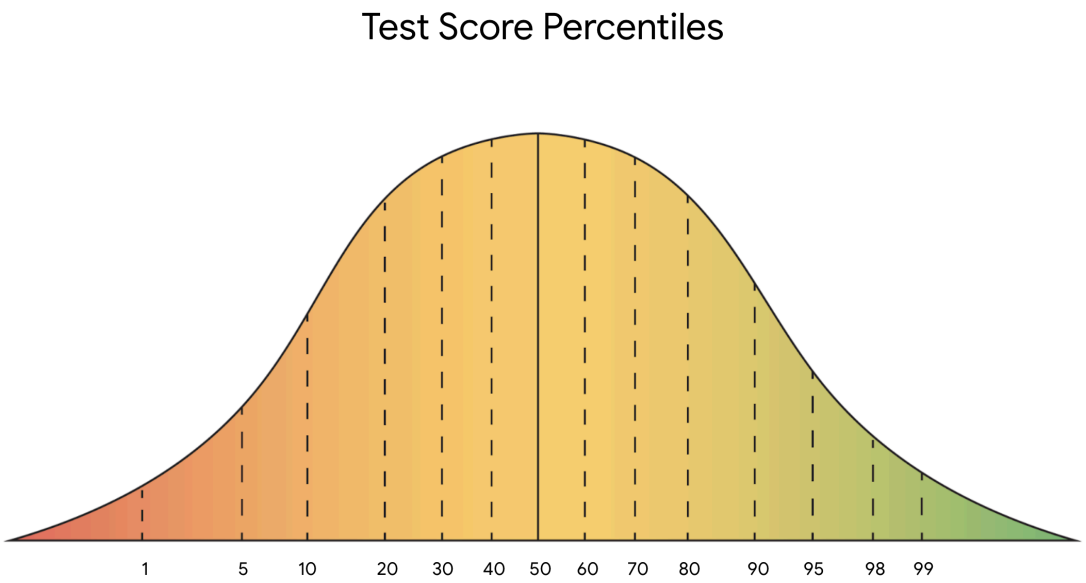
In this reading, you'll learn more about the most common measures of position: percentiles and quartiles. You'll also learn how to calculate the interquartile range, and use the five number summary to summarize your data.

Measures of position

Percentile

A **percentile** is the value below which a percentage of data falls. Percentiles divide your data into 100 equal parts. Percentiles give the relative position or rank of a particular value in a dataset.

For example, percentiles are commonly used to rank test scores on school exams. Let's say a test score falls in the 99th percentile. This means the score is higher than 99% of all test scores. If a score falls in the 75th percentile, the score is higher than 75% of all test scores. If a score falls in the 50th percentile, the score is higher than half, or 50%, of all test scores.



Note: *Percentiles and percentages are distinct concepts. For example, say you score 90/100, or 90%, on a test. This doesn't necessarily mean your score of 90% is in the 90th percentile. Percentile depends on the relative performance of all test takers. If half of all test takers score above 90%, then a score of 90% will be in the 50th percentile.*

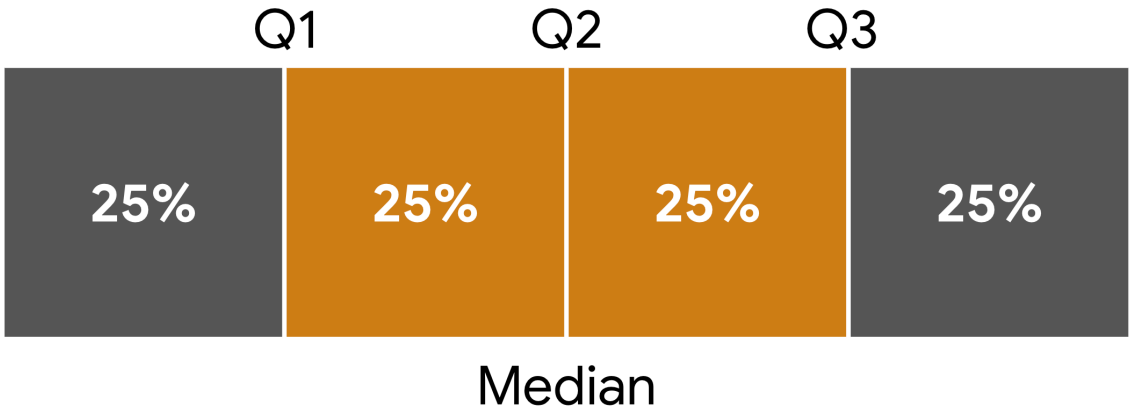
Percentiles are useful for comparing values and putting data in context. For example, imagine you want to buy a new car. You'd like a midsize sedan with great fuel economy. In the United States fuel economy is measured in miles per gallon of fuel, or mpg. The sedan you're considering gets 23 mpg. Is that good or bad? Without a basis for comparison, it's hard to know. However, if you know that 23 mpg is in the 25th percentile of all midsize sedans, you have a much clearer idea of its relative performance. In this case, 75% of all midsize sedans have a higher mpg than the car you're thinking about buying.

Quartile

You can use quartiles to get a general understanding of the relative position of values. A **quartile** divides the values in a dataset into four equal parts.

Three quartiles divide the data into four quarters. Quartiles let you compare values relative to the four quarters of data. Each quarter includes 25% of the values in your dataset.

- The first quartile, Q1, is the middle value in the first half of the dataset. Q1 refers to the 25th percentile. 25% of the values in the entire dataset are below Q1, and 75% are above it.
- The second quartile, Q2, is the median of the dataset. Q2 refers to the 50th percentile. 50% of the values in the entire dataset are below Q2, and 50% are above it.
- The third quartile, Q3, is the middle value in the second half of the dataset. Q3 refers to the 75th percentile. 75% of the values in the entire dataset are below Q3, and 25% are above it.



Example: Car sales

For example, imagine you're a data professional working for an auto dealership. The manager of the sales team wants to compare the performance of each sales representative on the team. The manager asks you to analyze data that provides how many cars each sales representative sold during the past month.

Sales Representative	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**	**#7**	**#8**
Cars Sold	18	13	6	10	15	7	10	9

You can calculate quartiles for your data in four steps:

1. Arrange the values in your dataset from smallest to largest.

[6, 7, 9, 10, 10, 13, 15, 18]

2. Find the median, or middle value, of your entire dataset. This is Q2. There are an even number of values in the dataset, so the median is the average of the two middle values, 10 and 10.

Q2 = (10 + 10) ÷ 2 = 20 ÷ 2 = 10

3. Find the median of the lower half of your dataset [6, 7, 9, 10]. This is Q1. The median is the average of the two middle values, 7 and 9.

Q1 = (7 + 9) ÷ 2 = 16 ÷ 2 = 8

4. Finally, find the median of the upper half of your dataset [10, 13, 15, 18]. This is Q3. The median is the average of the two middle values, 13 and 15.

Q3 = (13 + 15) ÷ 2 = 28 ÷ 2 = 14

Dividing the data into quartiles gives you a clear idea of sales rep performance. You now know that the lower quartile (Q1) of reps sold 8 cars or fewer, and the upper quartile (Q3) sold 14 cars or more. In other words, the lower 25% of reps sold 8 cars or fewer, and the upper 25% sold 14 cars or more. The middle 50% of representatives sold between 8 and 14 cars.

Note: This is not the only way to calculate quartile values. For datasets with many observations, the quartile calculation methodology has a negligible effect on the final calculated values. However, for datasets with few observations, the calculated quartiles may be not insignificant. Numpy's percentile() function, for example, has nine different ways to calculate a given percentile.

Interquartile range (IQR)

The middle 50% of your data is called the **interquartile range,** or **IQR**. The interquartile range is the distance between the first quartile (Q1) and the third quartile (Q3). This is the same as the distance between the 25th and 75th percentiles. IQR is useful for determining the relative position of your data values. For instance, data values outside the interval Q1 - (1.5 * IQR) and Q3 + (1.5 * IQR) are often considered outliers.

Note: *Technically, IQR is a measure of dispersion because it measures the spread of the middle half or middle 50% of your data (between Q1 and Q3). IQR is less sensitive to outliers than the range because it doesn't include the more extreme values in your dataset.*

IQR = Q3 - Q1. In this case, Q3 = 14 and Q1 = 8.

IQR = 14 - 8 = 6

Five number summary

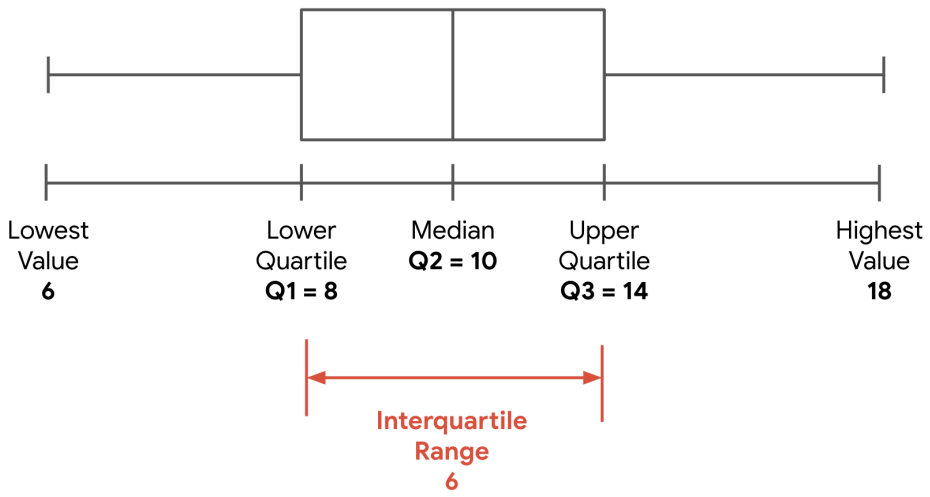
Finally, you can summarize the major divisions in your dataset with the five number summary. The five numbers include:

- The minimum
- The first quartile (Q1)
- The median, or second quartile (Q2)
- The third quartile (Q3)
- The maximum

The five number summary is useful because it gives you an overall idea of the distribution of your data, from the extreme values to the center. You can visualize it with a box plot.

The box part of the box plot goes from Q1 to Q3. The vertical line in the middle of the box is the median (Q2). The horizontal lines on each side of the box, known as whiskers, go from Q1 to the minimum, and from Q3 to the maximum.

The following box plot illustrates the data on car sales. You can find the values on the box plot and determine the interquartile range (IQR). The IQR is the length of the box, or the distance from Q1 to Q3.



Key takeaways

Data professionals use measures of position such as percentiles and quartiles to better understand all types of data, from product sales to household income. Measures of position help you quickly identify the relative location of your data values, and give you a more precise sense of the distribution of your data.