# **Measure of location**

When you begin summarizing the data you may want to know the five-number summary.

* The minimum value in the dataset.
* The first quartile, the data where you can say 25% of the data is less than this number (Q1).
* The median, the data where 50% of the data is less than this number(**M or** ).
* The third quartile, the data where you can say 75% of the data is less than this number (Q3).
* The maximum value in the dataset.

To find the above values in a dataset, first sort the data. Then find the median.

# **Median**

To find the median, do the following:

1) Determine the number of values, n, in the dataset.

2) Divide n by 2.

3) If the answer is an integer, i, the median is the average of the ith and (i+1)th value.

4) If the answer is not an integer, then round the value up to the next integer, r. The median is the rth value in the dataset.

Once the Median is determined, you can separate the dataset into two halves.

# **Q1 and Q3**

1) Repeat the process on the lower have of the data set to find the first Quartile, Q1.

2) Repeat the process on the upper half of the data set to find the third Quartile, Q3.

Example 1: Suppose you have a data set of the following values.

2, 5, 6, 7, 1, 9, 11, 5, 6, 8, 2, 1, 6, 6

The sample size is n = 14.

Sort the data

1, 1, 2, 2, 5, 5, 6, 6, 6, 6, 7, 8, 9, 11

The minimum value is 1. The maximum value is 11.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 | 2 | 2 | 5 | 5 | 6 | 6 | 6 | 6 | 7 | 8 | 9 | 11 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 |

To find the median divide 14 by 2 to get 7. The median is the average of the 7th and 8th values.

The data is separated in two halves

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | 1 | 2 | 2 | 5 | 5 | 6 | 6 | 6 | 6 | 7 | 8 | 9 | 11 |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |

To find the first quartile, Q1, take the median of the first seven values.

7/2 = 3.5

round up to 4.

The first quartile is the 4th item, Q1 = 2.

To find the third quartile, Q3, take the median of the second seven values.

7/2 = 3.5

round up to 4.

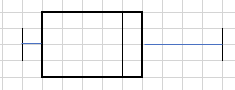
The third quartile is the 4th value of the second set of seven data values, Q3 = 7.

The five number summary is

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 | 2 | 6 | 7 | 11 |

# **Boxplot**

The five-number summary can be displayed in a boxplot



To interpret the chart, we can say that 25% of the data is less than 2. 50% of the data is below 6. 75% of the data is below 7.

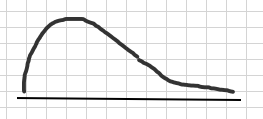
# **Interquartile Range**

To determine the range where 50% of data lies, we calculate the interquartile range. The interquartile range is Q3 – Q1.

IQR = 7 – 2 = 5.

There are three way to discuss the shape of a distribution.

## **Right Skewed**



The box and whisker plot that we created above is an example of a right skewed distribution.

## **Left Skewed**



## **Bell Shaped or Symmetry**



# **Outliers**

When you gather data, your random sample can contain a wide variety of data. For example, if you took a sample of 10 students at a class reunion and asked their average salary. One person in the sample was a successful professional basketball player. His salary was 2,500,000. The other class members had salary between $35,000 thru $100,000. The sample salary values are listed below.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 35,000 | 48,149 | 97,836 | 57,261 | 64,624 | 54,552 | 95,430 | 61,354 | 63,146 | 2,500,000 |

If we take the median of the salaries is the average of the 5th and 6th value.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 35,000 | 48,149 | 54,552 | 57,261 | 61,354 | 63,146 | 64,624 | 95,430 | 97,836 | 2,500,000 |

= 62250

The Q1 is 54,552. The Q3 is 95,430. The interquartile range is Q3 – Q1, 40878.

The 2,500,000 does not affect the median value. When you have values that are too large or too small, we call those values outliners.

Using the five-number summary to determine if a value is an outlier.

If a value is less than Q1 – 1.5(Q3 – Q1), -6765. All values below this value is an outlier.

If a value is less than Q3 + 1.5(Q3 – Q1), 156747. All values above this value is an outlier.

In this data set, 2500000 is an outlier.

# **Percentiles**

You may want to compute the percentile of a data set. For example, the lowest 20% of the data set or the upper 10% of the values.

To compute the percentile, we will use the method outlined below. It is called the n+1 method.

*k = the kth percentile desired.*

*i = index of the position of the data value*

*n = the sample size*

*Step 1: Sort the data from smallest to largest.*

*Step 2: find the index for the kth percent by using this formula .*

*Step 3: if i is an integer, take the ith value in the sort data set for the kth percentile. If i is not an integer, r < i < s where r and s are the nearest integer values. Take the average of the rth and sth values.*

Suppose your data set are the numbers below. Note they have already been sorted and you can read them from left to right.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 30 | 32 | 33 | 34 | 35 | 37 | 39 | 40 | 43 | 49 |
| 52 | 55 | 57 | 57 | 60 | 62 | 63 | 63 | 65 | 66 |
| 69 | 74 | 74 | 78 | 79 | 83 | 84 | 88 | 95 | 97 |

To find the 28th percentile, first note that there are thirty values in the dataset. The index for the 28th percentile is 28\*(31)/100 = 3.858. Since the index is not an integer, we will take the average of the 3 and 4 value.

Twenty-eight percent of data values are below 33.5.

If I wanted to compute the upper 10 percent. I would first take the value and subtract from 100 percent. So, I need to find the 90th percentile. The index is computed as follows.

=27.9

Take the average of the 27th and 28th value.

10% of the data values are above 86, and 90% of the data values are below 86.

# **Finding the Percentile of a data value**

You can look at it another way, suppose you wanted to know what the percentile value 60. First count how many values are below 60, x = 14. Then determine the number of times 60 appears in the dataset, y = 1. Use the following formula to determine the percentile for 60.

(100) = percentile round to the nearest whole number

Round to the nearest integer, therefore, 60 is the 48th percentile.

# **Interpreting Percentiles, Quartiles, and Median.**

I think the authors of Introductory to Statistics give good guideline for interpreting the percentiles. Their interpretation is a follows(Illowsky and Dean, pg. 95).

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
* the percent of individuals or items with data values above the percentile

# References

Illowsky, B., & Dean, S. (2016). *Introductory to Statistics.* Houston: Rice University.