

MTH410

Quantitative Business Analysis

Module 2: Descriptive Statistics

Statistical outcomes are only informative if you can communicate them effectively. In Module 2, we will explore graphical methods of data presentation, and we also begin calculating the basic formulas used to provide data for more complex statistical analysis.

Learning Outcomes

1. Construct and interpret visual displays of data.
2. Summarize quantitative data.
3. Evaluate relationships between two data sets.

For Your Success & Readings

For this module, be sure to start the Module 2 Critical Thinking assignment early. You may ask your instructor questions about the Critical Thinking assignment.

The basic calculations in this module provide the foundation for more complex statistical analysis. Therefore, it is important that you understand the rationale behind their calculation and try to remember how to derive the outcomes. You may want to create a separate list of the basic formulas for future reference so that you have it handy as you move through your next modules. The “Using Excel for Descriptive Statistics” document may help.

To navigate through this module successfully, keep the following in mind:

- This week you will complete your first Critical Thinking Assignment. Review the assignment early in the week and contact your instructor if you have any questions or concerns.
- Actively engage in the required discussion question for the week. The weekly discussions are great opportunities to learn from your instructor and fellow course mates. This week’s discussion question asks you to select a type of graph(s) discussed in the book and give two real-world examples of when the graph(s) would be useful to you in a professional application and in your everyday life. Provide graph illustration(s). You should explain only why the type of graph you have selected is the best for your example. Continue with the discussion and share proposed next steps based on your results.

You will be asked to write four managerial reports during the course. Remember to cite all of your references and use APA formatting. An example is provided in the MTH410 Guide to Writing with Statistics. If you have additional questions, consult the **CSU-Global Guide to Writing and APA** (<http://csuglobal.libguides.com/apacitations>).

Required

- Sections 2.1-2.4, 2.6-2.7 in *Introductory Business Statistics*

Recommended

- Katovich, E., & Maia, A. (2018). **The relation between labor productivity and wages in Brazil** (<https://search-proquest-com.csuglobal.idm.oclc.org/docview/2138147722/fulltextPDF/B46BB40ABA38492CPQ/1?accountid=38569>). *Nova Economia*, 28(1), 7–38.
- Laokri, S., Soelaeman, R., & Hotchkiss, D. (2018). **Assessing out-of-pocket expenditures for primary health care: How responsive is the Democratic Republic of Congo health system to providing financial risk protection** (https://search-proquest-com.csuglobal.idm.oclc.org/docview/2056838085?rfr_id=info%3Axri%2Fsid%3Aprimo)? *BMC Health Services Research*, 18(1), 1–19.

1. Tabular and Graphical Presentations

In Module 1, we covered **qualitative (categorical)** and **quantitative** data. The following interactive presents the tabular and graphical methods used in summarizing both qualitative and quantitative data.

Qualitative Data

Tabular Displays Graphical Displays

- Frequency Distribution
- Relative Frequency Distribution
- Percent Frequency Distribution
- Cross tabulation

- Bar Chart
- Pie Chart
- Side-by-Side Bar Chart
- Stacked Bar Chart

Quantitative Data

Tabular Displays Graphical Displays

- Frequency Distribution
- Relative Frequency Distribution
- Percent Frequency Distribution
- Cumulative Frequency Distribution
- Cumulative Relative Distribution
- Cumulative Percent Frequency Distribution
- Cross Tabulation

- Dot Plot
- Histogram
- Stem-and-Leaf Display
- Scatter Diagram

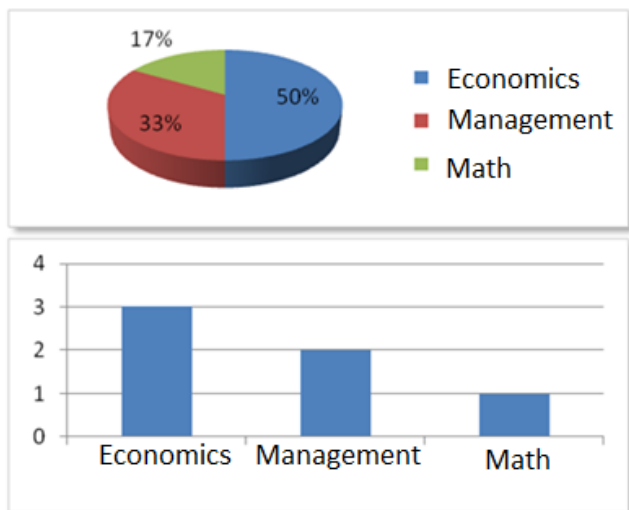
The presentation of statistical data can enhance understanding of the information and is an important component of statistical analysis. For example, qualitative data can be classified or grouped by creating a frequency distribution to reflect the number of times that an item appears in non-overlapping classes. We may also want to understand the relative frequency of an item relative to its class or group, which can be calculated as follows:

Relative frequency of a class = (Frequency of the class) / (Total number of Items)

If you want to show that information as a percentage, you would multiply the relative frequency number by 100.

For example, if you have six textbooks and three are for Economics, one for Math, and two for Management, your relative frequency of textbooks for Economics would be $3/6 = 0.5 = 50\%$.

You could then use a bar graph or pie chart to show the frequency graphically:



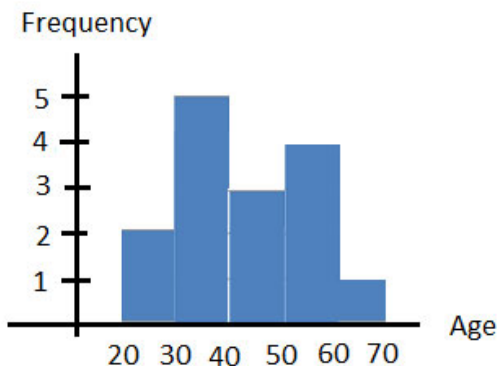
Caution: Three-dimensional displays of data might show misleading results. For example, in a 3-D pie chart, the slices on the far side might appear relatively smaller than those that are closer.

Qualitative data can also be enumerated using a frequency distribution. With qualitative data, however, you need to group the data so that they have an upper and lower bound, and so that the size or width of the groups is the same for each category, and the categories do not overlap. For example, if you were examining the ages of students, you would have the upper bound as the oldest student and the lower bound as the youngest student, and then you might create five groups of ages between these bounds:

20, 26, 30, 37, 37, 38, 39, 42, 43, 48, 51, 52, 52, 55, 69

Class (Age)	Frequency
20 up to 30	2
30 up to 40	5
40 up to 50	3
50 up to 60	4
60 up to 70	1
	Total = 15

The corresponding histogram is shown:



To construct a **relative frequency distribution**, you need to determine the relative frequency of each category according to the following formula:

In the above example, the relative frequency of the 30 to 40 age category is

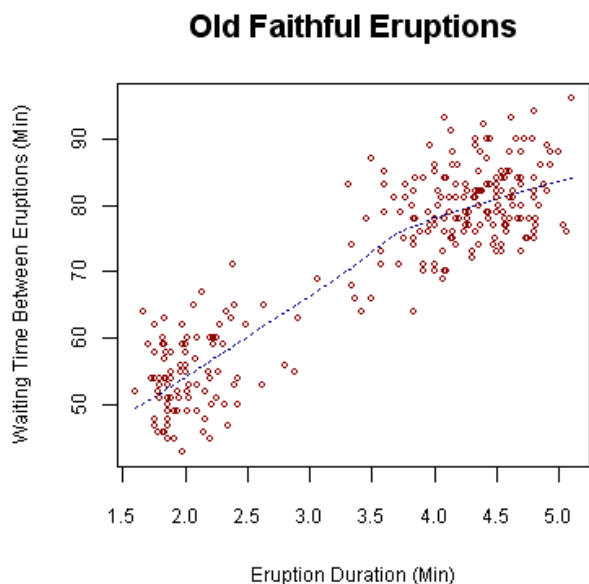
Or, if we wanted to examine the relationship between two variables, we could consider the use of cross tabulations and scatter diagrams. Using our age data, we might want to consider the additional variable of male versus female, or length of time in school. For one variable, we might also want to summarize the data through a stem-and-leaf diagram.

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

Note: with a stem-and-leaf diagram, one can see the raw data. For example, in the diagram above, one can see that there were two people who were 37 years old. One cannot see that information with a frequency histogram.

Relationship between Variables, Correlation Coefficient

We may also want to examine the relationship between two variables. A scatterplot is a graph that can help you visualize a relationship between variables. **This scatterplot** (https://en.wikipedia.org/wiki/Scatter_plot#/media/File:Oldfaithful3.png) describes the relationship between Eruption Duration (min.) versus Waiting Time between Eruptions (min.) of the Old Faithful Geyser.



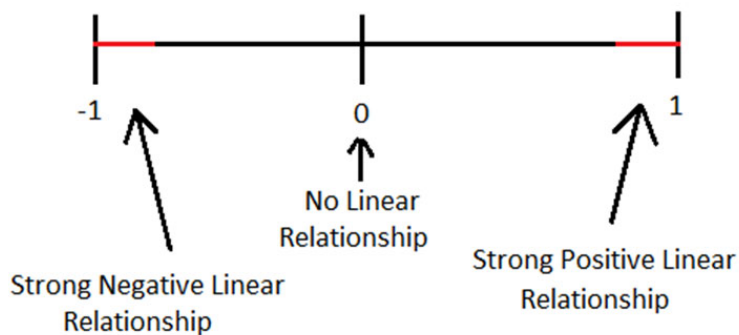
Notice that a longer Waiting Time between Eruptions implies a longer Eruption Duration.

The **correlation coefficient** measures the degree to which two variables vary together or in opposition, with the maximum positive correlation being 1.00. If the two variables covary positively and perfectly, the correlation will equal 1.00. Or, if two variables vary oppositely and perfectly, then the correlation will be equal to -1.00. We therefore have a measure that tells us whether two things covary perfectly, or near perfectly, and whether they vary positively or negatively. If the coefficient is, say, 0.90 or 0.95, then we know that the corresponding variables closely vary together in the same direction; on the other hand, if the coefficient is -0.90 or -0.95, then they vary together in opposite directions. For example, if we know that there is a high correlation between GPA and an age group, we can say that they are positively correlated.

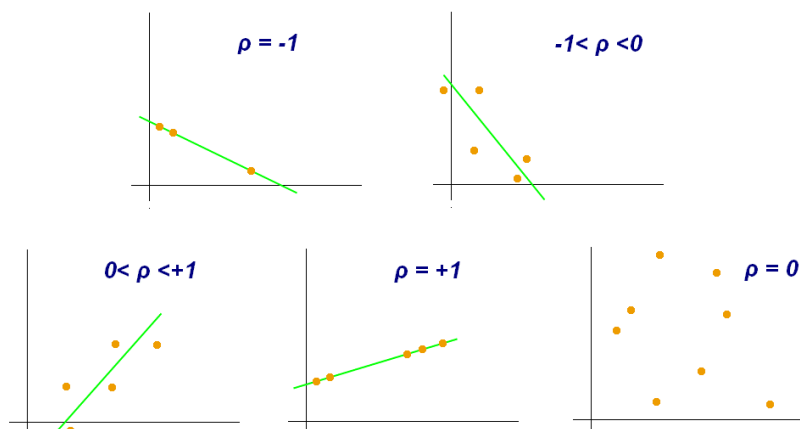
The population correlation coefficient is denoted with the Greek letter, ρ (pronounced “rho”). One obtains such a numerical value if all of the data in the population are used. Otherwise, if not all of the data from the population are obtained, then one uses the sample correlation coefficient, r .

The following number line interval illustrates what the value of the correlation coefficient tends to indicate. A correlation coefficient close to 1 indicates a strong positive linear correlation. This means that the scatter plot tends to follow a linear equation with a positive slope. A correlation coefficient close to -1 indicates a strong negative linear relationship between two variables. A correlation close to 0 indicates a weak linear relationship between two variables.

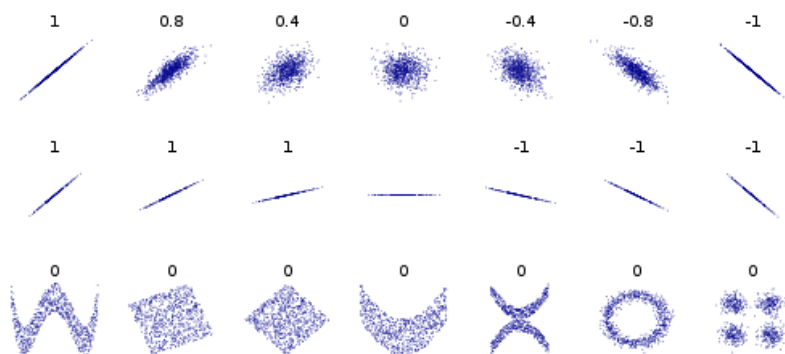
Correlation Coefficient



See the following scatter diagrams that show sample data sets and the corresponding correlation coefficients:



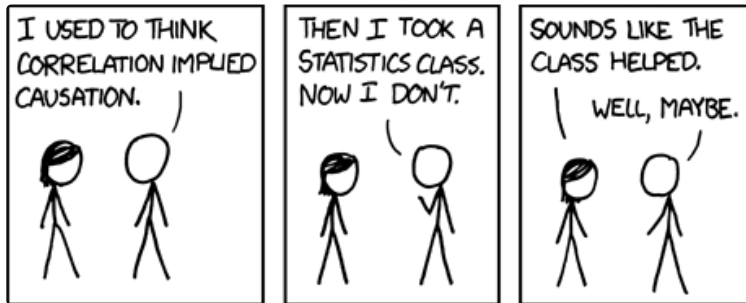
Kiatdd, 2012, **CC BY-SA 3.0** (https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#/media/File:Correlation_coefficient.png)



Notice above that a correlation coefficient of 0 only indicates that there is no linear correlation. There might be another type of correlation. For example, there might be a parabolic relationship between two variables.

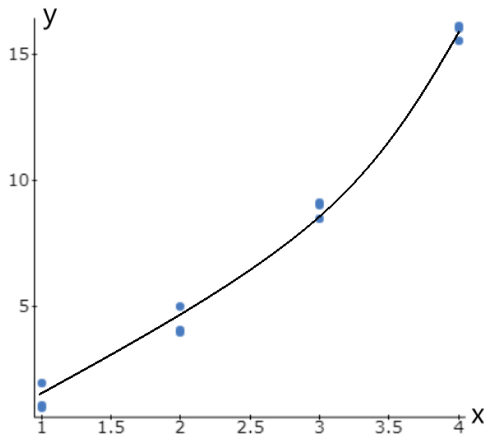
Finding the sample correlation coefficient, r , by using its formula is generally time consuming and prone to error. It is a recommendation to use technology, such as Excel, to compute the correlation coefficient. Select an empty cell in Excel. Then at the top in Excel, click on Formulas, More Functions, Statistical, and then scroll down to CORREL. Then select Array1 for the first row and Array2 for the second row.

In Module 8, we will study linear relationships further. For example, we will discuss a test that can allow us to determine if there is a linear relationship between two variables.



Munroe, n.d. CC BY-NC 2.5 (<http://xkcd.com/552/>)

1.1. Further Explanations on the Correlation Coefficient



Notice that a correlation coefficient of 0 only indicates that there is no linear correlation. There might be another type of correlation that is not linear. The following scatterplot follows a non-linear function:

Let's consider an example: There might be a parabolic relationship between two variables. This means that the scatterplot follows—to some extent—a parabola, or polynomial of degree 2.

A correlation larger than 0.90 is generally considered to be a strong positive linear correlation. A correlation less than -0.90 is generally considered to be a strong negative linear correlation.

An example of a negative correlation can be the cost of certain items and sales. A larger cost tends to decrease sales. A positive correlation can be advertising expenditures and sales. A larger investment in advertising tends to increase sales.

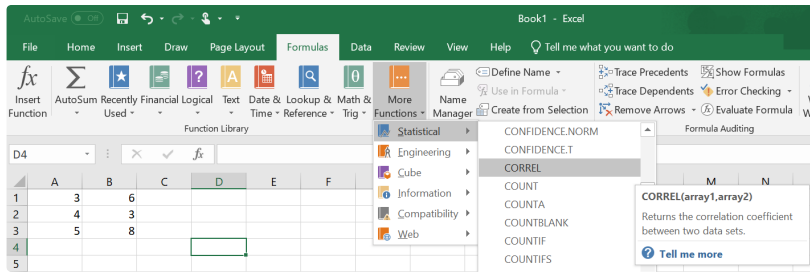
As mentioned previously, it is a recommendation to use technology, such as Excel, to compute the correlation coefficient.

Suppose you want to find the sample correlation coefficient between rows A and B by using Excel.

Click through the tabs below to learn how to do this.

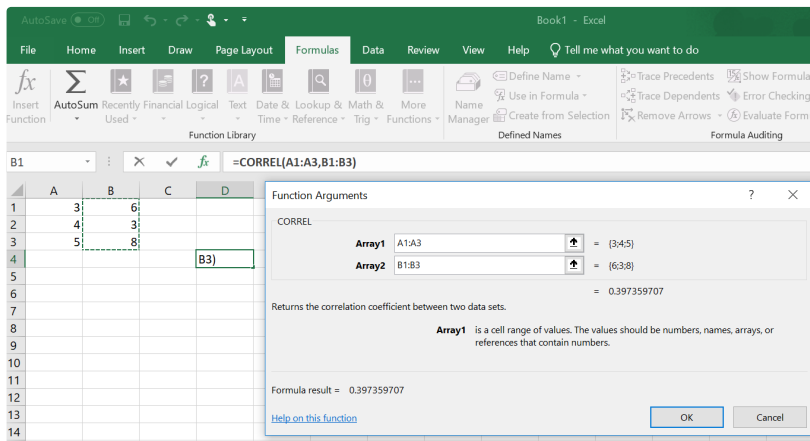
Start with an empty cell

Select an empty cell in Excel. In the example here, cell D4 was selected. Then at the top in Excel, click on Formulas, More Functions, Statistical, and then scroll down to CORREL.








Array 1

Then for Array1, select the numbers in Row A (A1 to A3). For Array2, select the numbers in Row B (B1 to B3) as shown:








Correlation Coefficient

Press “OK” and you will see the correlation coefficient of 0.39736 in cell D4:



AutoSave Off     

File Home Insert Draw Page Layout

fx Σ     

Insert AutoSum Recently Used Financial Logical Text Data

Function Library

D4   *fx* =CORF

	A	B	C	D
1	3	6		
2	4	3		
3	5	8		
4				0.39736

In Module 8, we will further study linear relationships. We will discuss a test that can allow us to determine if there is a linear relationship between two variables.

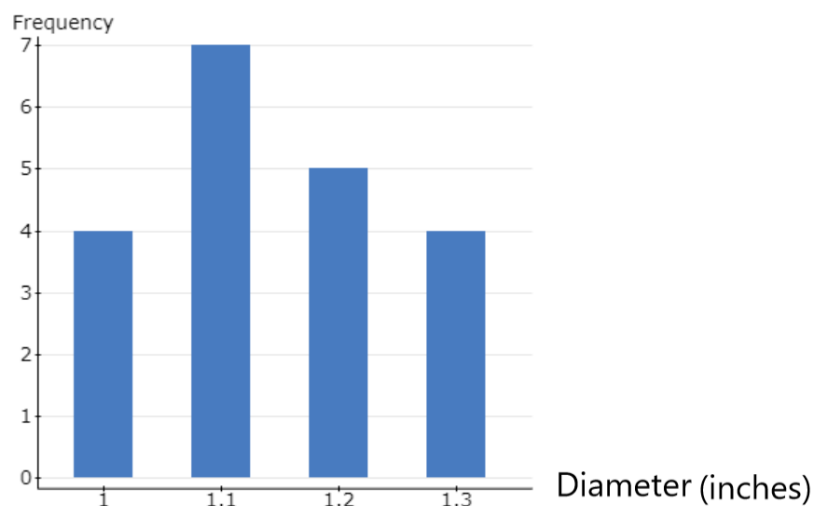
1.2. Relative Frequency Distributions

A relative frequency distribution shows relative frequencies instead of just frequencies. This can give information about each category relative to the other categories.

For example, suppose we are given the diameters (in inches) of 20 tubes made at a factory:

Diameter	Frequency
1	4
1.1	7
1.2	5
1.3	4

The corresponding bar graph is shown:



To construct the relative frequency distribution, use the formula previously mentioned for each category:

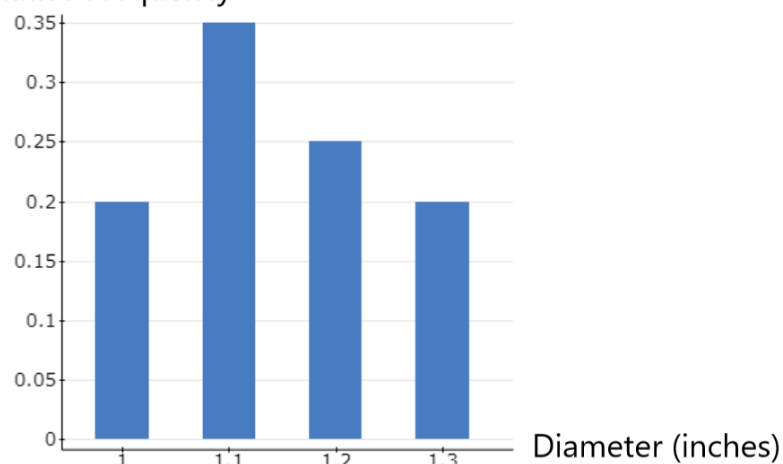
In the above example, the relative frequency of the 1.2 inches category is

The relative frequency distribution is as follows. The Relative Frequency Column is obtained by using the above formula for each diameter category.

Diameter	Frequency	Rel. Freq
1	4	0.2
1.1	7	0.35
1.2	5	0.25
1.3	4	0.2
		Total = 1

The relative frequency bar graph is shown here:

Relative Frequency



Notice that both the relative frequency and frequency bar graphs have the same shape. However, with a relative frequency distribution, or bar graph, we can obtain percentages. For example, 25% of the tubes measured 1.2 inches. This can give valuable information in understanding how categories measure up relative to the others. For example, tubes measuring 1.1 inches (the largest category) in diameter form 35% of the samples. The next highest category, the 1.2 inches, form 25% of the samples. There is a difference of 10% between the 1.1 and 1.3 relative frequencies.

In addition, the sum of all the relative frequencies should be 1, as shown above. In future modules, relative frequency bar graphs can be used to determine probabilities.

Color	Frequency
Red	11
Blue	15
Green	23
Yellow	8

2. Descriptive Statistics

Data can also be summarized through numerical measures that summarize the location, variability, and shape of data distribution. These measures can be taken from samples or from the entire population.

Measures of Location

In statistics, one is often interested in finding a measure of the “center,” “average,” or “typical.” These are often called measures of central tendency. *Review the three measures of central tendency by clicking on the links below.*

MeanMedianMode

The **mean** or average value provides the measure of central location for the data. It is the most commonly used measure of central tendency. The sample mean, \bar{x} , is calculated by summing the values and then dividing them by the sample size. For example, the data set of {0, 1, 3, 5, 7}, would provide a sample mean of $\bar{x} = (0+1+3+5+7)/5 = 3.2$. The formula for the sample mean is

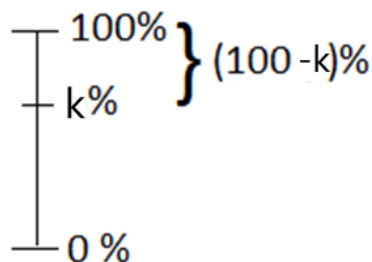
In this case, n is the size of the sample. The Greek letter, Σ , is summation notation. Σ denotes to add all of the data values from the sample. In the above example, $\Sigma x = 0+1+3+5+7 = 16$.

The **population mean** is denoted by the Greek letter μ . The population mean is obtained by sampling the entire population. The formula for the population mean is

N is the size of the entire population.

The **median** is the value of the middle number when the data are arranged from smallest value to largest value. If a data set has an odd number of observations, the median is the middle value. If a data set has an even number of observations, the median is the average of the two middle values. For example, in our data set of {0, 1, 3, 5, 7}, the median would be 3. However, suppose that the set consists of an even number of data points: {0, 1, 3, 5, 7, 10}. There is no “middle” value in the data set. The median is the average of the middle two values:

The **mode** indicates the value or data element that occurs with greatest frequency. If our example data set was modified to contain {0, 1, 1, 3, 5, 7}, the mode would be 1. It is also possible that if there is more than one value with equal frequency such as {0, 1, 1, 3, 3, 5, 7}, the data set would have modes of both 1 and 3 for a bimodal frequency. If there are more than two modes, it could be considered a multimodal frequency. One may also determine the mode for a qualitative data set. For example, the mode of the data set {red, blue, green, red} is red.



A **percentile** divides data into 100 orders. For the k^{th} percentile, approximately $k\%$ of the observations are less than the k^{th} percentile. Also, approximately $(100-k)\%$ of the observations are greater than the k^{th} percentile.

The percentile is a **measure of the location of data**. The percentile determines where the values are relative to the rest of the data.

One often sees percentiles in standardized tests. For example, if a student scored in the 70th percentile, then he/she scored better than approximately 70% of students and worse than approximately 30% of the others.

Suppose data are arranged from smallest to largest. Denote i by the location of the k^{th} percentile. The approximate formula for determining the location, i , of the k^{th} percentile is

Here, n is the number of data points. If i is an integer, then the k^{th} percentile is the i^{th} value in the ordered data set. However, if i is not an integer, then round up i to the next larger integer and round down i to the next smaller integer. The k^{th} percentile will be the average of the two data values in these two positions.

Note: The definition of the percentile is not uniform in all of mathematics. There is no universal definition of what the percentile is. Other textbooks have different ways to compute the percentile. This will produce different answers. The numerical answers in other textbooks will be close to the ones described in our current book.

EXAMPLE

Consider the example of the ages of 15 people:

26, 20, 69, 37, 38, 39, 48, 52, 42, 37, 52, 51, 55, 30, 43

Find the 45th percentile.

Click "Solution" to check your thinking

Solution

First, arrange the data from smallest to largest:

20, 26, 30, 37, 37, 38, 39, 42, 43, 48, 51, 52, 52, 55, 69

Then, apply the above formula to find the approximate location of the 45th percentile:

The value of i is not an integer. Thus, round down and round up 7.2 to produce 7 and 8, respectively. The 7th and 8th values in the ordered data set are 39 and 42, respectively. The average of 39 and 42 is $(39+42)/2 = 40.5$. Thus, the 45th percentile is 40.5.

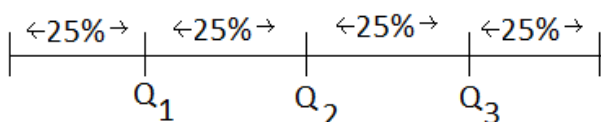
The three **quartiles** are defined in terms of percentiles:

Q_1 = first quartile, or 25th percentile

Q_2 = second quartile, or 50th percentile (or median)

Q_3 = third quartile, or 75th percentile.

Quartiles divide the data into four parts. Each part contains approximately 25% of the data:



EXAMPLE

Find the first quartile, Q_1 , and the third quartile, Q_3 , in the example above.

Click "Solution" to check your thinking

Solution

This is like the previous problem.

Thus, the 25th percentile is the 4th data value. In this case, it is the number 37 ($Q_1=37$).

Thus, the 75th percentile is the 12th data value. In this case, it is the number 52 ($Q_3 = 52$).

Measures of Variability

In addition to measures of location, there are measures of **variability** or **dispersion**. These measures help to identify trends in the data and are a key component of inferential statistics. *Click on the buttons below to learn more about range, interquartile range (IQR), and standard deviation.*

The simplest measure of variability is the **range**, which is calculated by computing the difference between the smallest and largest values.

Example: Recall the example of the ages of 15 people. The arranged data points are as follows:

20, 26, 30, 37, 37, 38, 39, 42, 43, 48, 51, 52, 52, 55, 69

The range is $69-20 = 49$. A disadvantage of the range is that it only considers two data values, the smallest and largest.

Another measure of variability is the **interquartile range (IQR)**. It is defined as the difference between the third and first quartiles:

The IQR is a measure of the range for the middle 50% of data.

Example: Recall the example on the ages of people. The arranged data points are as follows:

20, 26, 30, 37, 37, 38, 39, 42, 43, 48, 51, 52, 52, 55, 69

The first and third quartiles were computed to be $Q_1 = 37$ and $Q_3 = 52$, respectively. Hence, the interquartile range is as follows:

The **five-number summary** of a data set is:

1. the minimum value
2. the first quartile, Q_1
3. the median
4. the third quartile, Q_3
5. The maximum

Example: Recall the example on the ages of people. The arranged data points are as follows:

20, 26, 30, 37, 37, 38, 39, 42, 43, 48, 51, 52, 52, 55, 69

Based on previous work, the five-number summary is

1. The minimum value is 20.
2. the first quartile, Q_1 , is 37.
3. the median is 42.
4. the third quartile, Q_3 , is 52.
5. The maximum is 69.

Standard deviation is often used to show variation. This value is based on the average distance of the data points from the mean. It is calculated as the positive square root of the variance. When the data points are tightly bunched together, and the histogram curve is steep, the standard deviation is small. When the data values are spread apart, and the histogram is relatively flat, that shows you have a relatively large standard deviation. For example, if you have a class of fifteen students with twelve significantly different GPA's, we would see a relatively flat histogram and high standard deviation in relationship to a class of fifteen students with three different GPA's that are relatively close to the mean.

Example: Recall the example of the ages of 15 people: 26, 20, 69, 37, 38, 39, 48, 52, 42, 37, 52, 51, 55, 30, 43

First determine the sample mean:

Then find the sum of the squared deviations from the sample mean:

The **sample variance** is the average of the sum squared deviations from the sample mean:

The terms $(x_i - \bar{x})^2$ are **squared deviations from the mean**. For example, the term $(100 - 100)^2$ is a squared deviation from the mean.

Since the units of the variance are squared, one applies the square root to obtain the standard deviation. The **sample standard deviation** is the square root of the sample variance:

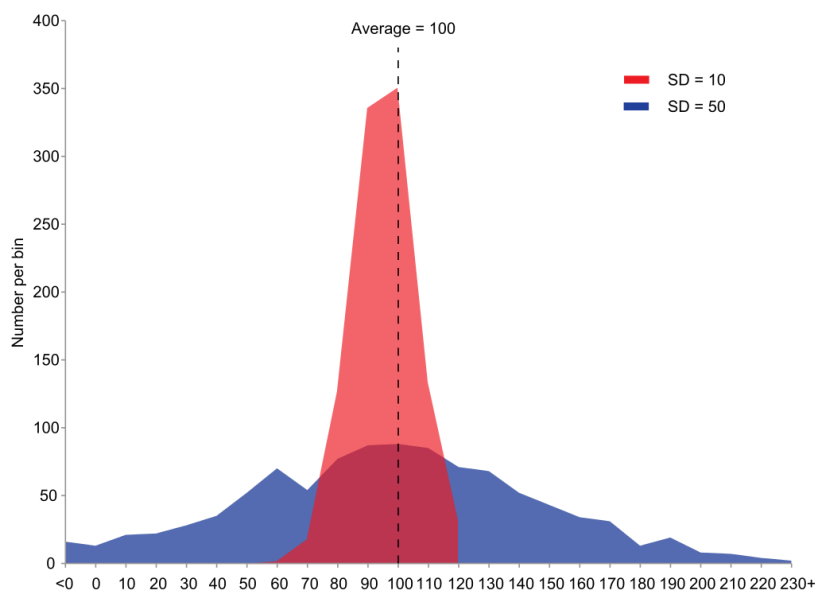
The **population variance** is denoted by using the Greek letter σ^2 .

The population variance is obtained by sampling all of the elements of the population. N is the size of the population and μ is the population mean.

The **population standard deviation** is the positive square root of the population variance:

In practice, one computes the sample standard deviation rather than the population standard deviation. Part of the reason is because one does not have all of the data from the entire population in inferential statistics. Thus, unless specifically instructed, we will compute only the sample standard deviation. Also, note that we divide by $n-1$ and not by n in the formula for the sample standard deviation. There is a mathematical proof stating that s is an “unbiased estimator” of σ . This means that the expected value of s is σ . If the formula for s divided by n , and not by $n-1$, then s would *not* be an “unbiased estimator” of σ .

Consider the data described in the following:



Both sets of data have the same mean. However, the data represented in red have a smaller standard deviation. The standard deviation measures the spread of the data from the mean.

Coefficient of Variation

Another way to measure variation is by using the sample Coefficient of Variation (CV). The formula for the CV is:

$$CV = \frac{s}{\bar{x}}$$
 where s is the sample standard deviation and \bar{x} is the sample mean. The Coefficient of Variation is a measurement of variation relative to the mean.

EXAMPLE

Recall the example on the ages of people. The arranged data points are as follows:

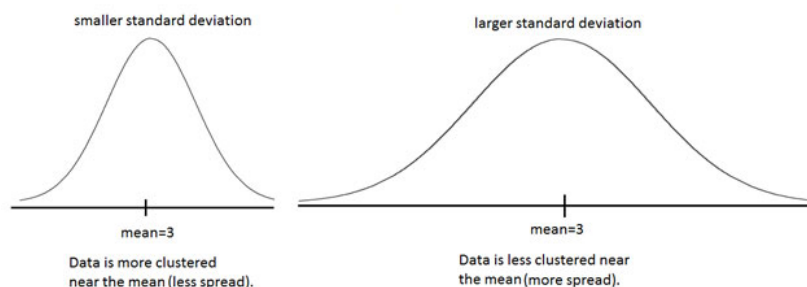
20, 26, 30, 37, 37, 38, 39, 42, 43, 48, 51, 52, 52, 55, 69

Recall that the sample mean and the sample standard deviations were $\bar{x} = 42.6$ and $s = 12.449$, respectively. The Coefficient of Variation is

The Coefficient of Variation is useful in comparing two different, but similar, data sets. For example, suppose we were given another data set of the ages of people and the coefficient of variation is 10.00. Then there is more variation in the first data set, whose coefficient of variation is 29.22.

Computing the value of a standard deviation can be complicated, but its visual representation can help one to understand the basic use of the formula. Suppose the frequency distribution of data is approximately bell-shaped. See the ThoughtCo article “**Introduction to the Bell Curve**” (<https://www.thoughtco.com/introduction-to-the-bell-curve-3126337>) for more information on the bell-shaped curve.

Here is a visual example of the standard deviation for a bell-shaped curve:



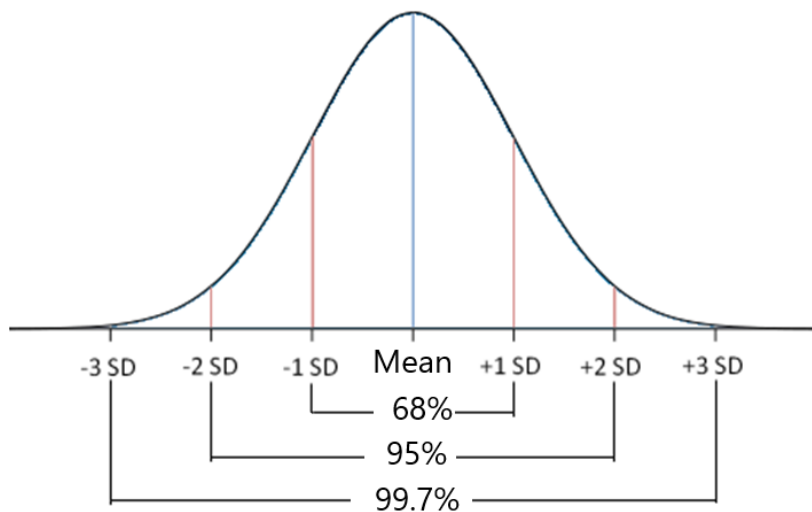
Using z-score to Determine Extreme Values (Outliers)

To determine how far from the mean a standard deviation is, you could use the **z-score**. The z-score of the value x is defined as

$$z = \frac{x - \bar{x}}{s}$$
 where \bar{x} and s are the sample mean and standard deviation, respectively.

For example, suppose $\bar{x} = 20$ and $s = 4$. Then the z-score of $x = 24$ is

A z-score of 1 would mean that the data point 24 is 1 standard deviation away from the sample mean. Consider the bell-shaped curve shown below. Such a curve can be obtained from a frequency distribution. The height of the curve represents the frequency. The **Empirical Rule** states that one standard deviation away from the mean in either direction on the horizontal axis accounts for approximately 68 percent of the data points in this group. Two standard deviations away from the mean account for roughly 95 percent of the data. Three standard deviations account for about 99.7 percent of the data.



Review the bell-shaped curve above. Notice that the set of all z-scores greater than 3 or less than -3 account for approximately 0.03% of the data, since $100\% - 99.7\% = 0.03\%$. Typically, any data value with a z-score less than -3 or greater than 3 is considered an **outlier**.

The following examples show how to use z-score to find an outlier. *Click on each tab to see the example.*

z-score Detection of Outliers—Example 1

Suppose that a set of data has a sample **standard deviation** of $s=3$ with a sample **mean** of $\bar{x}=7$. If $x=10$ is a data value from the set, then its **z-score** is:

Since the **z-score** of 1 is not less than -3 or bigger than 3, then $x=10$ is *not* an outlier.

z-score Detection of Outliers—Example 2

Suppose that a set of data has a sample **standard deviation** of $s=3$ with a sample **mean** of $\bar{x}=7$. If $x=19$ is a data value from the set. Then its **z-score** is:

Since the **z-score** of 4 is bigger than 3, then $x=19$ is an outlier.

EXAMPLE

Recall the example of the ages of people. The arranged data points are as follows:

20, 26, 30, 37, 37, 38, 39, 42, 43, 48, 51, 52, 52, 55, 69

Suppose one wants to determine if there are any outliers in the above observations by using the z-score formula. Recall that the sample mean and the sample standard deviation were $\bar{x}=42.6$ and $s=12.449$, respectively. The z-score of the values 20 and of 69 are

When computing z-scores, round to at least **two decimal places**.

Since neither the smallest and the largest values in the data set are outliers, then the rest of the data will not be outliers either. Hence, there are no outliers based on the z-score.

Using Lower and Upper Limits to Determine Extreme Values (Outliers)

One may also use the Lower Limits and Upper Limits to determine outliers. The Lower and Upper Limits are defined as follows, in terms of the first and third quartiles (Q_1 and Q_3) and the interquartile range (IQR):

An observation is classified as an outlier if its value is less than the Lower Limit or greater than the Upper Limit.

EXAMPLE

Recall the example on the ages of people. The arranged data points are as follows:

20, 26, 30, 37, 37, 38, 39, 42, 43, 48, 51, 52, 52, 55, 69

Suppose one wants to determine if there are any outliers in the above observations by using the Lower and Upper Limits. The first and third quartiles, and interquartile range, were computed to be $Q_1=37$, $Q_3=52$, and $IQR=15$, respectively. Hence, the Lower and Upper Limits are:

Looking at the above data, there are no observations less than the Lower Limit of 14.5. Also, there are no observations greater than the Upper Limit of 74.5. Hence, using the Lower and Upper Limits, there are no outliers.

2.1. More on the Coefficient of Variation

As mentioned previously, an application of the Coefficient of Variation (CV) is to compare variation from different, but similar, data sets.

For example, suppose a manager oversees two stores, Store 1 and Store 2. Store 1 is larger and has more sales than Store 2. Thus, it is expected that Store 1 will have larger mean and standard deviation than Store 2. Hence, it is not informative to compare the standard deviations of both stores. The manager needs a way to compare the variation of both stores in relative terms. The sales for Stores 1 and 2 are shown here, in thousands of dollars, for 7 days:

Store 1		Store 2	
Day	Sales	Day	Sales
1	10	1	7
2	14	2	8
3	15	3	5
4	16	4	6
5	18	5	9
6	20	6	2
7	15	7	6

The mean and standard deviation sales for Store 1 are \$15.43 thousand and \$3.15 thousand, respectively. The mean and standard deviation sales for Store 2 are \$6.14 thousand and \$2.27 thousand, respectively.

Thus, for Store 1, the Coefficient of Variation is

For Store 2, the Coefficient of Variation is

The manager can see that there is more relative variation in Store 2 even though Store 2 has a smaller standard deviation than Store 1.

In general, the CV is recommended when comparing dissimilar data sets. For example, the first set might be restaurant sales and the other can be exam grades.

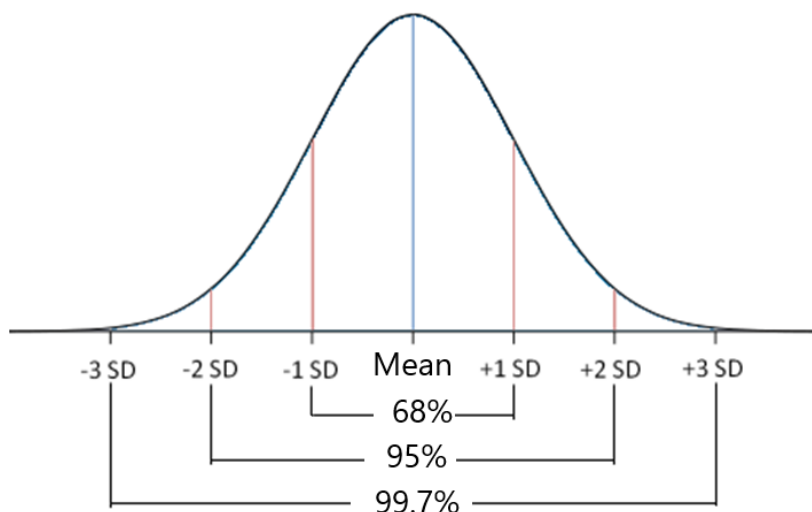
The population Coefficient of Variation is similarly defined:

The site **Statistics How To** (<https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/how-to-find-a-coefficient-of-variation/>) further explains the Coefficient of Variation.

2.2. More on the Empirical Rule

Recall that the **Empirical Rule** states that if the histogram of data is approximately bell-shaped, the following is true:

1. Approximately 68% of the data is within 1 standard deviation from the mean.
2. Approximately 95% of the data is within 2 standard deviations from the mean.
3. Approximately 99.7% of the data is within 3 standard deviations from the mean.



One can use the empirical rule to get a rough idea as to what to expect from bell-shaped data. For example, if the mean of a certain bell-shaped data is $\mu = 100$ with standard deviation $s = 7$, then

Hence, any value less than 79 or greater than 121 should account for approximately 0.03% of the data.

For example, suppose that a histogram is approximately bell-shaped with sample mean $\mu = 15$ and standard deviation $s = 5$. Note that that

Thus, approximately 68% of the data is between 12 and 18. One can even go further and use the symmetry of the bell-shaped curve:

$68/2 = 34\%$
Thus, 34% of data is between 15 and 18. Also,

Thus, approximately 95% of data is between 9 and 21.

Similarly, by symmetry, $95/2 = 47.5\%$. Hence, approximately 47.5% of data is between 15 and 21.

In addition,

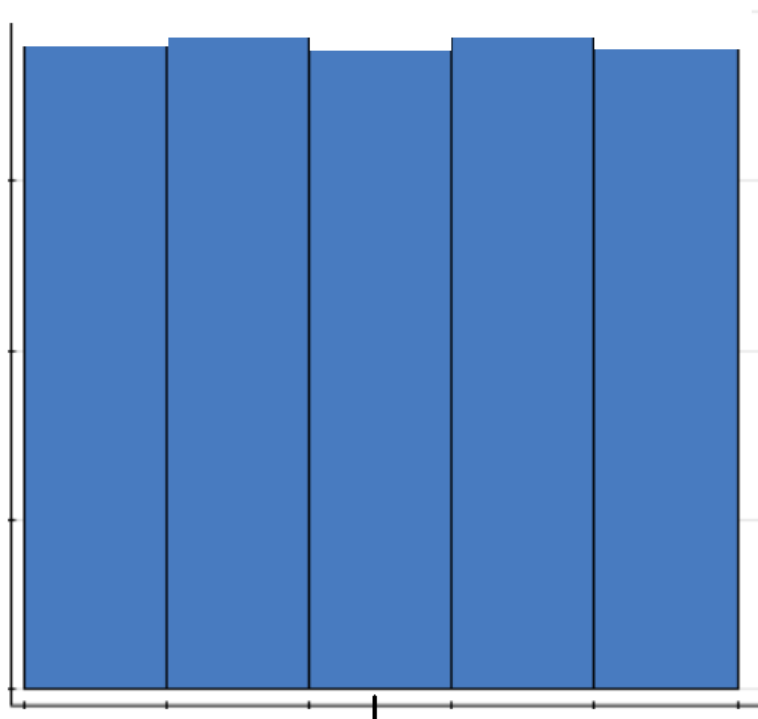
Thus, approximately 99.7% of all data is between 6 and 24.

The **site Statistics How To** (<https://www.statisticshowto.datasciencecentral.com/empirical-rule-2/>) provides a further explanation of the Empirical Rule.

3. Histogram Shapes, Boxplots, Mean from a Histogram

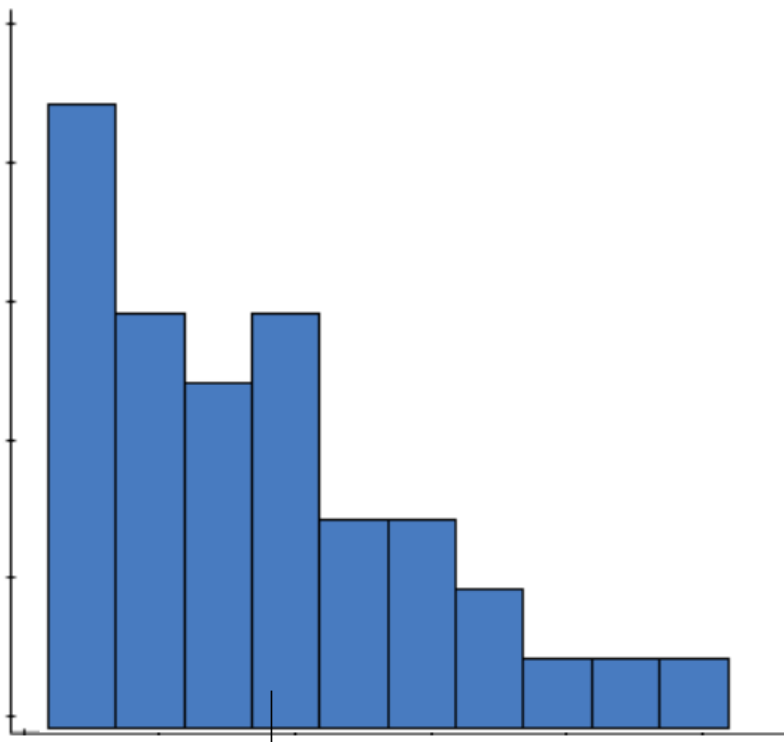
There are 4 basic shapes of a histogram, with labeled on the x -axis:

Approximately Uniform (symmetric):



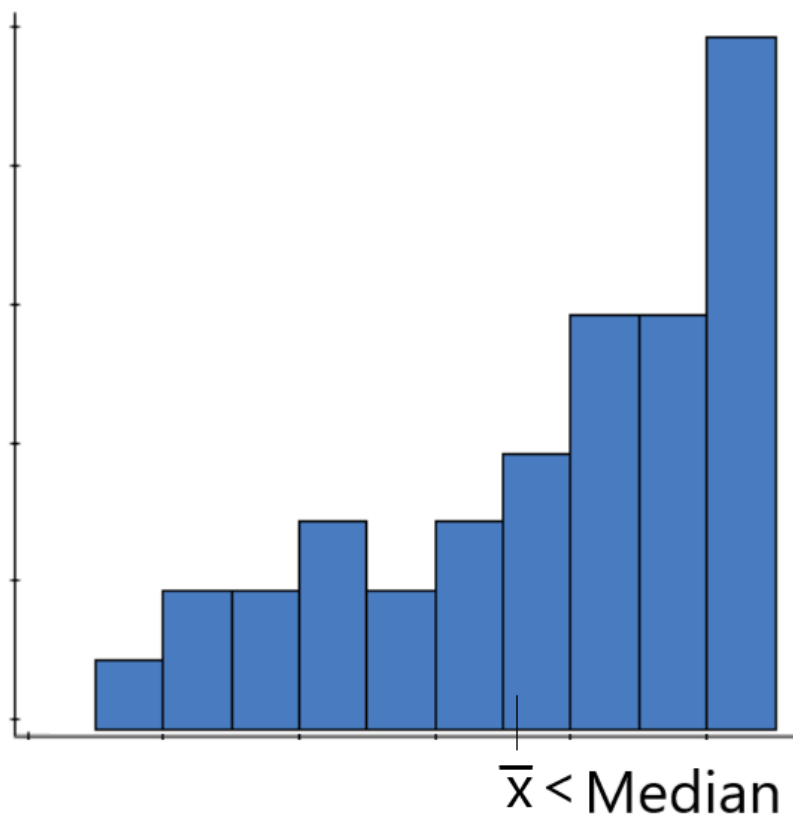
$$\bar{x} \approx \text{Median}$$

Skewed to the right, tail to the right (non-symmetric):

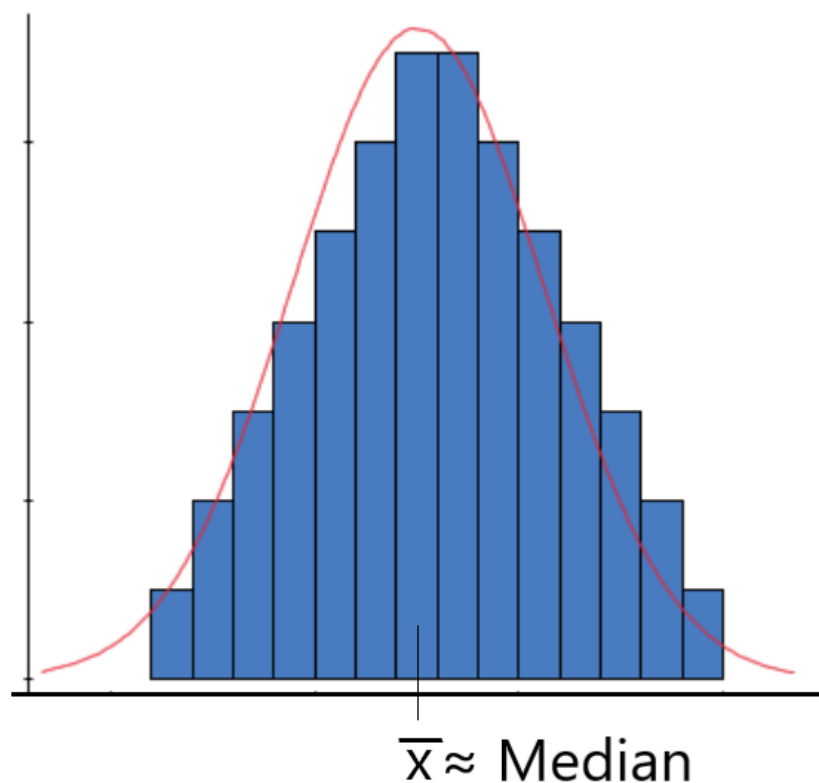


$$\text{Median} < \bar{x}$$

Skewed to the left, tail to the left (non-symmetric):



Approximately bell-shaped (symmetric):



A histogram is **symmetric** if there is a vertical line across it so that the left half is a mirror image of the right half. See the two histograms drawn above (uniform and bell-shaped) for examples of symmetric histograms.

Suppose the histogram is skewed to the left, as above. Since the mean, \bar{x} , is the balance point of the histogram, then a tail to the left causes the value of \bar{x} to be shifted to the left. The median is unaffected by outliers (values to the extreme left or extreme right). That is why \bar{x} is less than the median when a histogram is skewed to the left. Similar reasoning applies to histograms that

are skewed to the right.

See **this site** (<https://www.mathsisfun.com/data/skewness.html>) for an explanation of the skewness of data:

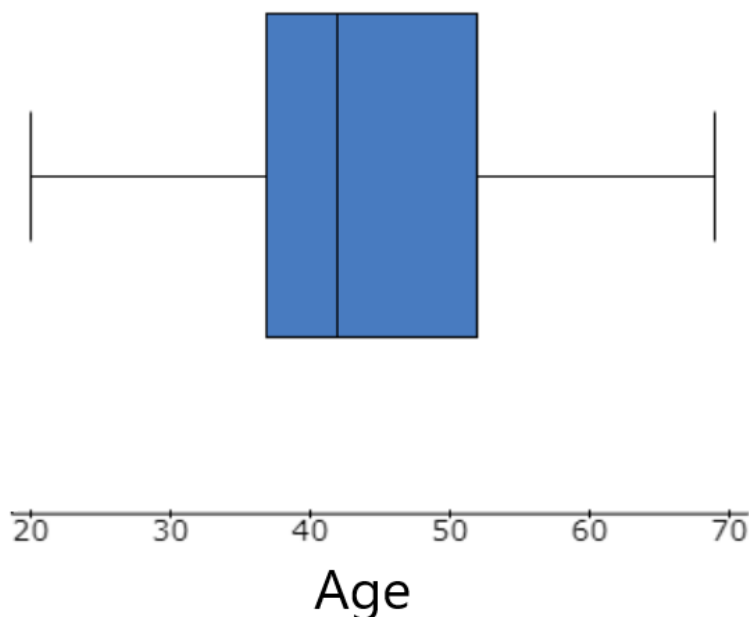
A **boxplot** (or **box-and-whisker plot**) consists of these 5 parts

1. the smallest non-outlier (the left tip of the whisker)
2. the first quartile, Q_1 (the left end of the box)
3. the median (the vertical line inside the box)
4. the third quartile, Q_3 (the right end of the box)
5. the largest non-outlier (the right tip of the whisker)

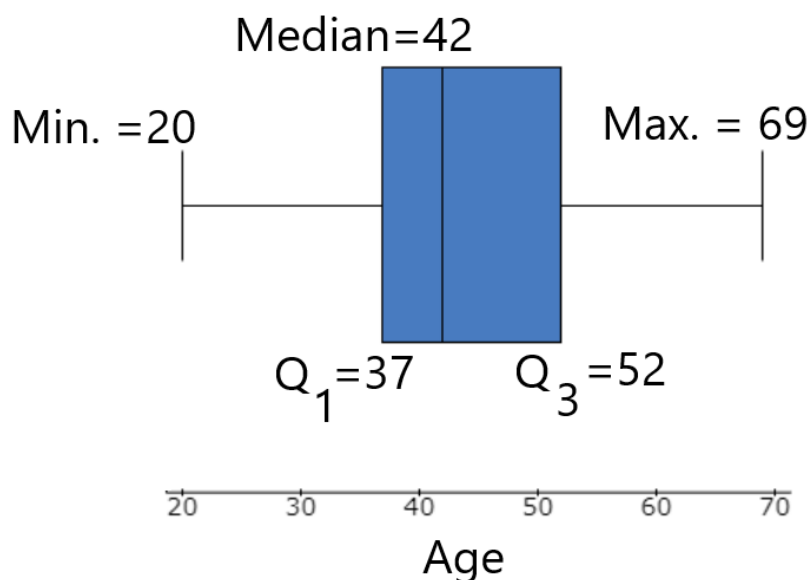
Example: Recall the example on the ages of people. The arranged data points are as follows:

20, 26, 30, 37, 37, 38, 39, 42, 43, 48, 51, 52, 52, 55, 69

The boxplot is as follows:

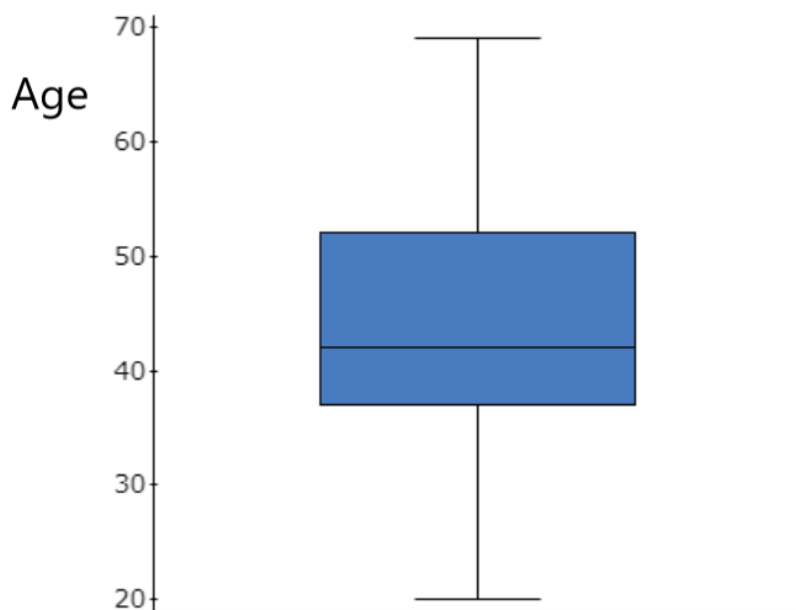


Here is the same boxplot labeled:

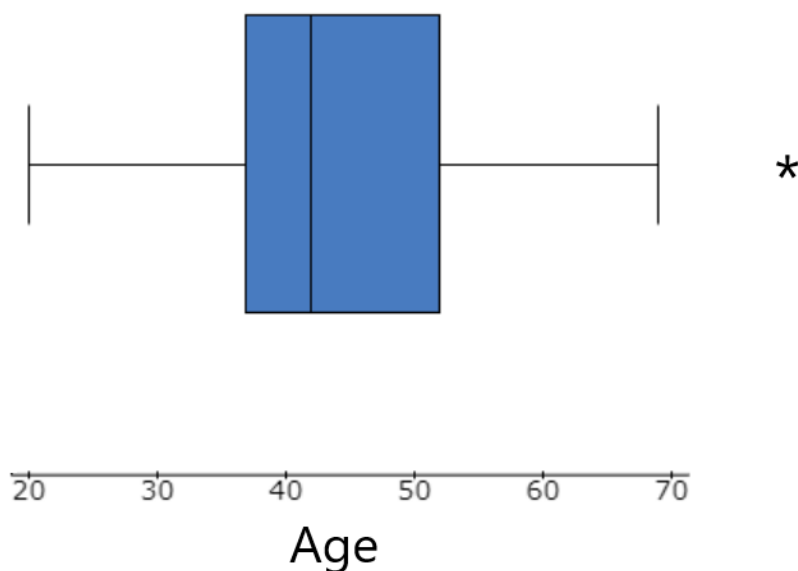


The width of the box is the interquartile range, $IQR = Q_3 - Q_1 = 52 - 37 = 15$. The region inside the box represents approximately 50% of the data. Approximately 50% of the data is between 37 and 52 years of age. The lower and upper whiskers are each approximately 25% of the data.

Boxplots can also be represented vertically, instead of horizontally. For example, the boxplot above can be represented vertically as follows:



Boxplots often omit outliers or place them to the left or right of the whisker. For example, if our data contained a large outlier, it would be denoted with an asterisk (or dot) to the right of the tip of the right whisker:



An advantage that boxplots have over histograms is that they may be used to compare two or more different sets of data side by side.

Refer to **this example linked here from Lumen** (<https://courses.lumenlearning.com/wmopen-concepts-statistics/chapter/interquartile-range-and-boxplots-3-of-3/>). In the first two sets of boxplots, one can compare the ages of actors and actresses who won an Oscar. One can see from the boxplots that the male actors are generally older than the females. In the second example, on the same web page, one can see that the temperature boxplot of Pittsburgh is much wider than that of San Francisco. This means that there is more temperature variation in Pittsburgh than in San Francisco.

Finding the Mean from a Frequency Table

The mean can also be computed from a frequency table. The formula for the mean from a frequency table is as follows:

For example, suppose the weights of certain food cans (in grams) filled at a factory are recorded in the following table:

Weight, x	Number of cans, f
450	3
452	4
453	26
454	58
455	25
456	23
457	11
458	3

In this table, there are no intervals. Each category is a single number. Each term of the form xf is the total weight for the category. For example, 450×3 grams represent the sum of 3 cans weighing 450 grams.

The term $\sum f$ is the sum of all of the frequencies, or all of the cans that were weighed:
 $\sum f = 3 + 4 + 26 + 58 + 25 + 23 + 11 + 3 = 153$

The mean is thus:

EXAMPLE

The following table contains the waiting times, rounded up to the nearest minute, of customers at a call center:

Waiting time	Number of customers, f
1-5.5	10
5.5-10.5	13
10.5-15.5	47
15.5-20.5	13
20.5-25.5	9
25.5-30.5	3

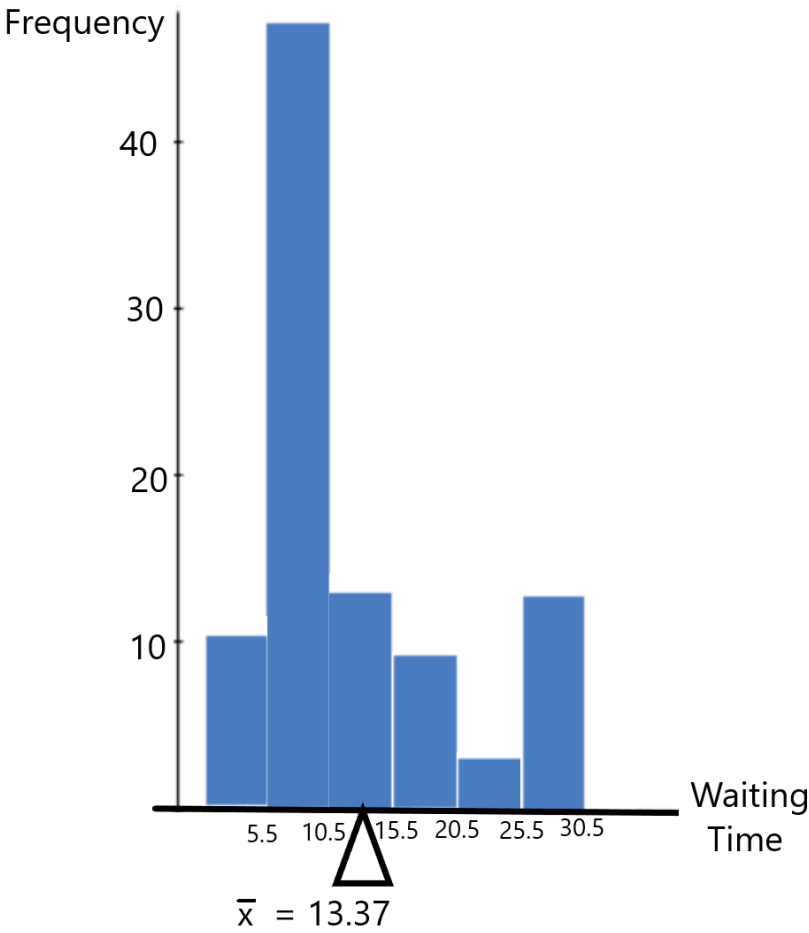
For example, there were 13 customers who waited between 5.5 and 10.5 minutes. Since the waiting times are intervals, one cannot compute the mean. To fix this issue, we use the midpoint, x , as the representative of each interval:

Waiting time	Midpoint, x	Number of customers, f
0.5-5.5	3	10
5.5-10.5	8	13
10.5-15.5	13	47
15.5-20.5	18	13
20.5-25.5	23	9
25.5-30.5	28	3

$\sum f$ is the total number of customers:

The mean is thus:

If you think of the histogram bars as different weights placed side-by-side, the mean is the balance point of the corresponding histogram. This follows from the physics interpretation of as the center of mass. The f and x terms are the weights and corresponding locations on the number line, respectively. The term is the total mass.



Let’s Try an Example:

The following frequency table contains the number of customers served at lunch for 31 days

Number of customers, x	Number of days, f
33	2
35	1
37	4
38	7
39	8
40	6
41	2
42	1

Find the mean number of customers of the above frequency distribution. *Click Answer to see if you are right.*

Answer

- a. 37.75
- b. 37.91
- c. 38.42
- d. 38.36

The answer is 38.42. Is that the answer at which you arrived? Here is how you make this determination:

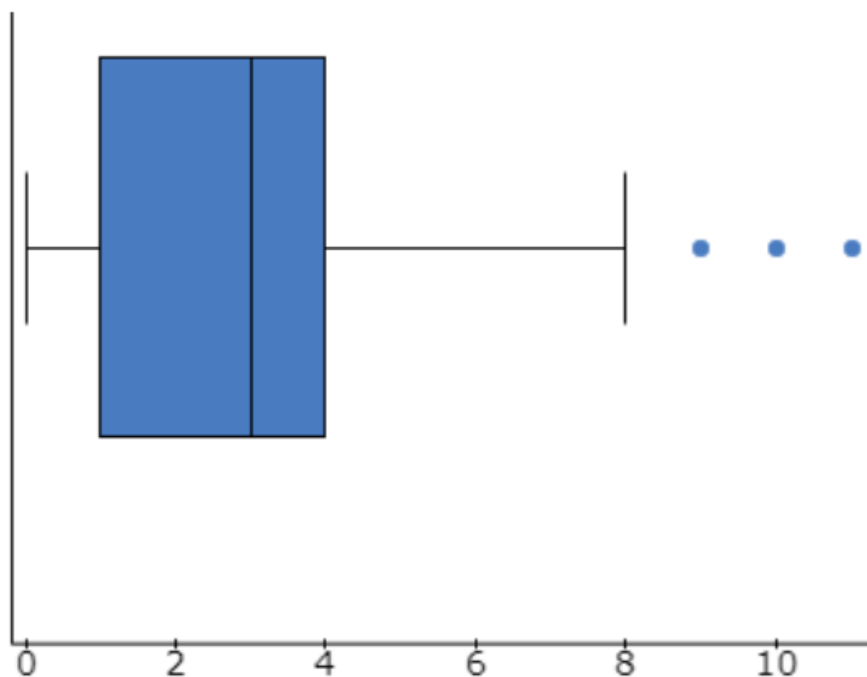
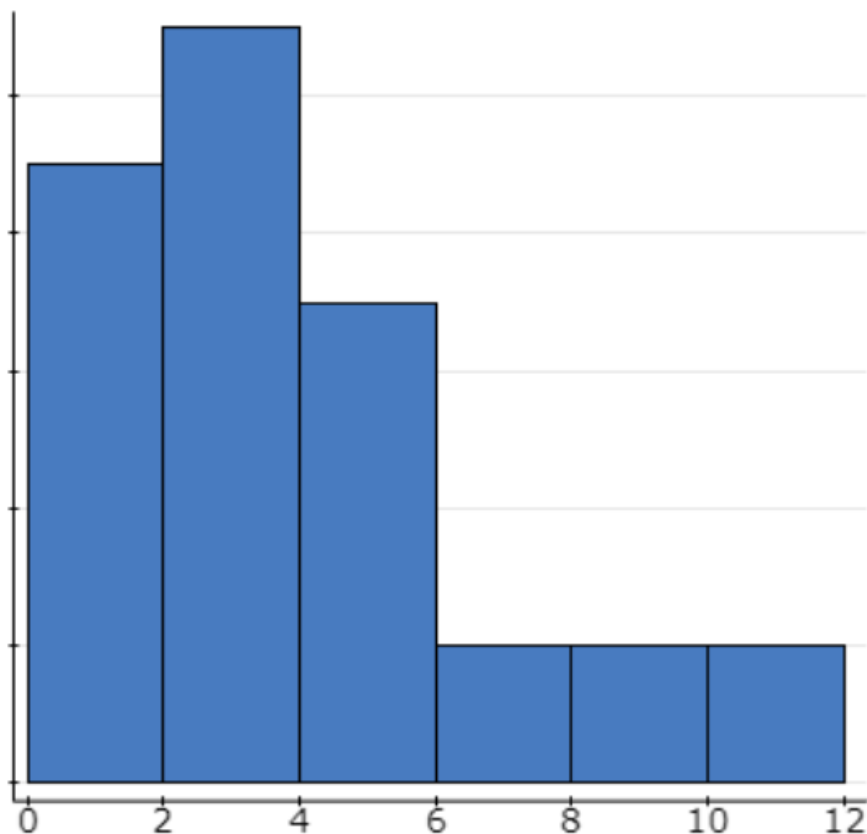
The term $\sum f$ is the sum of all the frequencies, or all the days:

$$\sum f = 2 + 1 + 4 + 7 + 8 + 6 + 2 + 1 = 31$$

The mean is thus:

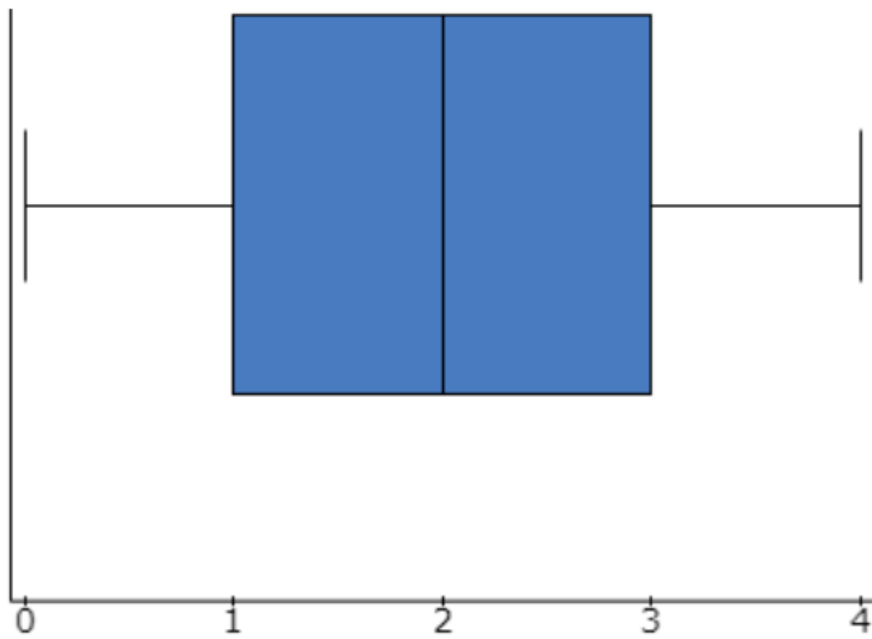
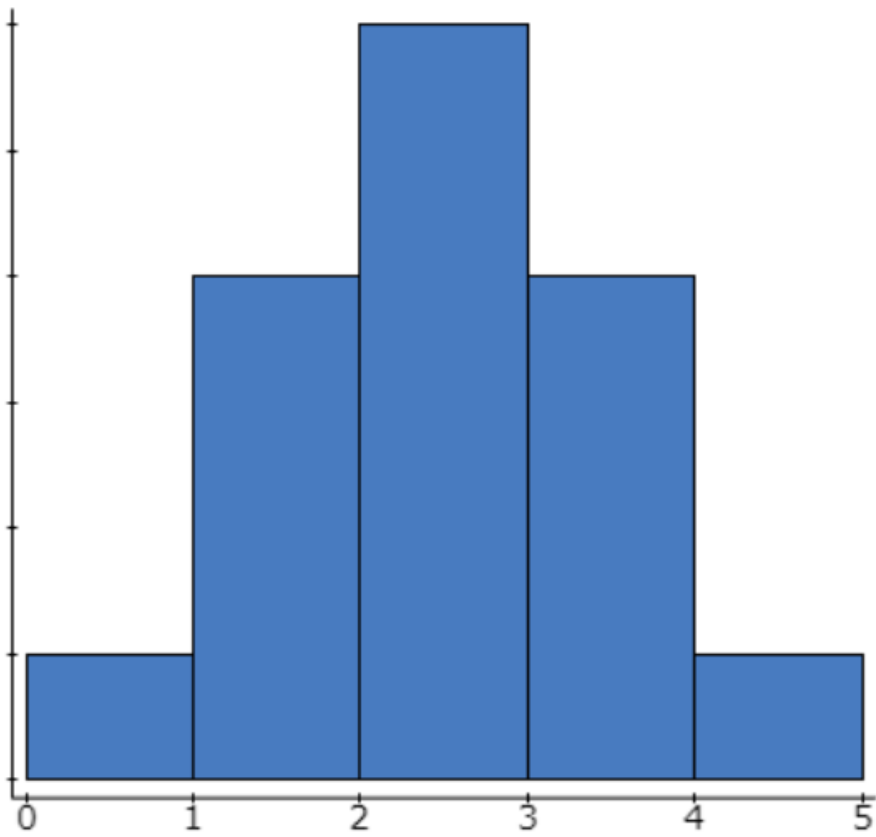
3.1. Further Explanations of Box Plots

The following is a histogram with its corresponding boxplot (box-and-whisker plot). Data that are skewed to the right will have an unusually long right boxplot “whisker.” Notice that the three outliers are denoted with the blue dots.



Similar comments apply to data that are skewed to the left; the left “whisker” will be unusually long.

Approximately symmetric data will have “whiskers” that are approximately the same length, with the median near the center of the box:



As previously mentioned, boxplots make it easy to compare two sets of data in a unique way.

EXAMPLE

Suppose two insurance agencies want to study the ages of customers who made a recent claim.

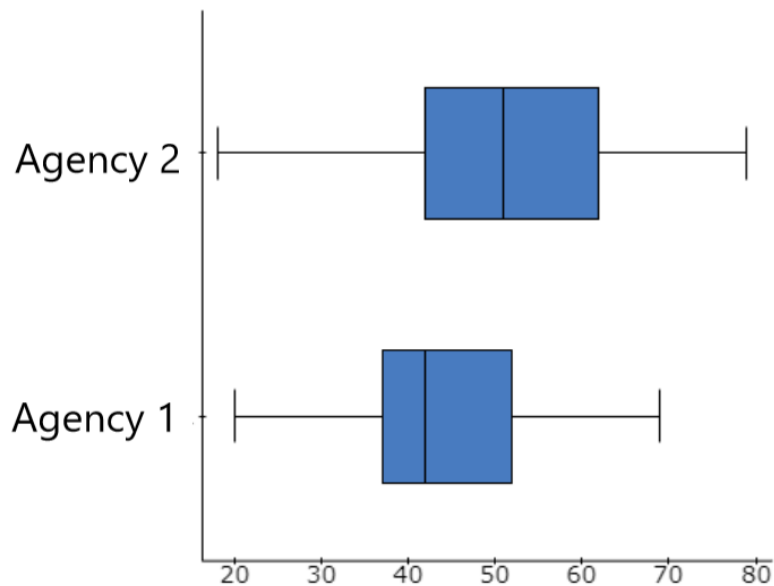
The ages of Agency 1 are as follows:

20, 26, 30, 37, 37, 38, 39, 42, 43, 48, 51, 52, 52, 55, 69

The ages of Agency 2 are:

18, 28, 35, 42, 43, 45, 46, 52, 53, 51, 59, 62, 68, 71, 79

The boxplots are compared here:



We can see that there is a greater variation of ages with Agency 2. The box (middle 50%) of Agency 2 is wider, too. Agency 2 appears to be skewed to the left, because of the longer left whisker. Agency 1 appears to be more symmetrical than Agency 2. Also, the ages of Agency 2 appear to be generally older than those of Agency 1.

3.2. More on Outliers

Detecting outliers is important. One reason is because some measurements, like the mean and standard deviation, are affected by outliers. Another reason is because outliers might indicate something worth studying, instead of discarding.

EXAMPLE

Management at a factory collected data on the number of years its 16 workers have been in Assembly Line A:

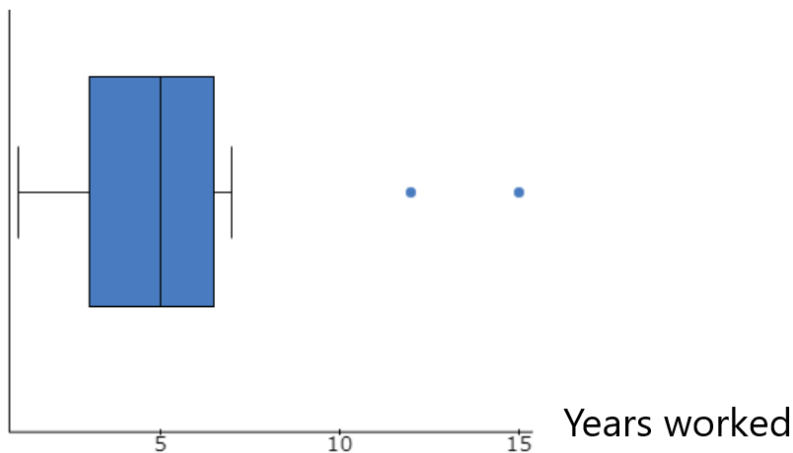
15, 12, 5, 4, 6, 7, 4, 1, 2, 5, 7, 1, 4, 1, 5, 5

The five-number summary is:

The interquartile range is $IQR = Q_3 - Q_1 = 6.5 - 3 = 3.5$

Outliers are determined according to the following formulas:

The boxplot is shown here. The outliers of 12 and 15 are indicated by the two dots next to the boxplot:



Management can see that those who have worked 12 and 15 years are outliers. Rather than ignoring outliers, it might be informative to see why those two workers have worked for many years. For example, management can learn from those two employees what it might take to retain workers.

Note: The population mean and standard deviation of the above data are $\bar{x} = 5.25$ and $s = 3.68$, respectively. The z-score for $x=15$ years and $x=12$ years are:

Using the z-score method, the values of 12 and 15 are not outliers. Thus, the method of outlier determination matters.

4. Summary

We covered various ways to describe data, both graphically and numerically. Being able to use descriptive statistics allows us to gain a better understanding of data. Through this knowledge we can gain additional insight into the usefulness and applicability of the data for decision-making purposes.

Here is the list of the objectives that we have covered and are part of the Mastery Exercises in Knewton Alta:

- Construct and understand frequency tables for a set of business-related data
- Create and interpret histograms
- Create and interpret stem-and-leaf plots
- Create and interpret bar graphs
- Find the mean of a set of data
- Find the median of a set of data
- Find and interpret percentiles and quartiles of a business-related data set
- Find the five-number summary of a business-related data set
- Identify the interquartile range and potential outliers in a set of business-related data
- Construct and understand box-and-whisker plots in business contexts
- Compute z-scores and use them to compare values from different data sets
- Determine if a data set is skewed in business examples
- Compute the sample variance and sample standard deviation in business contexts
- Interpret the standard deviation of a set of business-related data
- Calculate the correlation coefficient using Technology—Excel

Check Your Understanding

Embedded Media Content! Please use a browser to view this content.