

# Lesson 5: Descriptive Statistics

*David G Radcliffe*

The following topics are covered in this lesson:

1. Mean, median, and mode
2. Quartiles and percentiles
3. Variance and standard deviation
4. Frequency tables
5. Stem-leaf diagrams and histograms
6. Populations and samples

## ***Mean, median, and mode***

We are interested in using statistics to summarize data sets. By a data set, we mean any list of numbers, such as the scores on an examination, or the daily high temperatures at a specific location.

The mean, median, and mode are three different ways to describe the central (or most typical) value of a data set.

The **mean** value, or **average**, is found by adding up all of the observations, and dividing by the number of observations. For example, the mean value of {3, 7, 4, 4, 10} is

$$(3 + 7 + 4 + 4 + 10) / 5 = 5.6 .$$

The mean of a data set  $X$  is denoted  $\bar{X}$ . The Excel function AVERAGE() computes the mean of a list of values.

The **median** is the middle value. The median can be found by sorting the values from least to greatest, and selecting the number in the middle of the list. In a sorted list with  $N$  values, the median will occur in position  $(N+1)/2$ . The Excel function MEDIAN() computes the median of a list of values.

**Example:** The median of {3, 5, 9, **10**, 14, 17, 25} is **10**. We have  $N = 7$  and  $(N+1)/2 = 4$ , so the median is the 4<sup>th</sup> number in the sorted list.

**Example:** The median of {16, 24, 28, 32, 37, 80} is **30**. We have  $N = 6$  and  $(N+1)/2 = 3.5$ , so the median is the number in position 3.5. But since 3.5 is not a whole number, we take the average of the 3<sup>rd</sup> and 4<sup>th</sup> numbers in the list:  $(24+28)/2 = 30$ .

**Important:** The data values must be sorted before calculating the median!

The **mode** is the most frequent value of the data set, or the value that occurs the greatest number of times. It is possible for a data set to have more than one mode, but if there are no repeated values then the mode is not defined. The Excel function MODE() computes

the mode of a list of values. If the list has more than one mode, then Excel returns the mode that appears first.

**Example:** The modes of {3, 4, 6, 7, 4, 3, 7, 7, 9, 4, 6} are 4 and 7, because the data set contains three 4's and three 7's, while the other values occur no more than twice. However, the MODE() function would return 4, because 4 appears before 7.

## Quartiles and percentiles

The  $k^{\text{th}}$  **percentile** is a value  $x$  such that  $k\%$  of the data values are less than or equal to  $x$ , and the rest are greater than or equal to  $x$ . The 50<sup>th</sup> percentile is the same as the median. The position of the  $k^{\text{th}}$  percentile in a sorted list of  $N$  values is given by the equation

$$p = 1 + (N - 1) k / 100.$$

In Excel, percentiles of a list of values are computed with the PERCENTILE() function.

For example, in a sorted list of 37 values, the 25<sup>th</sup> percentile occurs in position 10.

$$p = 1 + (37 - 1) \times 25 / 100 = 1 + 36 / 4 = 10.$$

In Excel, PERCENTILE(A1:A37, 0.25) would return the 10<sup>th</sup> smallest value in the list.

If there were only 36 values, then we would require the value in position 9.75. This value should lie somewhere between the 9<sup>th</sup> and 10<sup>th</sup> values, but closer to the 10<sup>th</sup> value than the 9<sup>th</sup>. One way to calculate the (9.75)<sup>th</sup> value is to add 75% of the 10<sup>th</sup> value to 25% of the 9<sup>th</sup> value. Similarly, the value in position 14.3 can be calculated by adding 30% of the 15<sup>th</sup> value to 70% of the 14<sup>th</sup> value. This is not a universal practice, but it is the one that Excel follows.

The 25<sup>th</sup> percentile is also the **first quartile**, and the 75<sup>th</sup> percentile is called the **third quartile**. The 50<sup>th</sup> percentile could be called the second quartile, but we call it the median instead.

**Example:** Find the mean, the median, the mode, the third quartile, and the 30<sup>th</sup> percentile for the following data set.

6, 10, 10, 12, 13, 15, 19, 20, 25

### Solution

The number of observations is  $N = 9$ . The mean is  $(6+10+10+12+13+15+19+20+25)/9 = 14.44$ . The median is the number in position  $p = (9+1)/2 = 5$ , which is 13. The mode is 10, because it is the only value that is repeated. The third quartile is the value in position  $p = 1 + (9 - 1) \cdot (.75) = 7$ , which is 19.

The 20<sup>th</sup> percentile is the number in position  $p = 1 + (9 - 1)(.30) = 3.4$ . We add together 40% of the 4<sup>th</sup> number and 60% of the 3<sup>rd</sup> number:  $(.40)(12) + (.60)(10) = 10.8$ .

## Variance and standard deviation

The **deviation** of a data value is difference between the value and the mean value. If a value is above average then it has a positive deviation, and if it is below average then it has a negative deviation. For example, if Anthony scores 88 on an examination, and the average score is 81, then the deviation of Anthony's score is 7. On the other hand, a score of 68 on this exam would represent a deviation of  $-13$ .

The **variance** of a data set is the mean squared deviation. To compute the variance, we find the mean, and subtract it from each data value to find the deviations. We square each deviation, add these numbers together, and divide by the number of data values. The **standard deviation** is the square root of the variance. We write  $\text{Var}(X)$  for the variance of  $X$ , and  $\sigma(X)$  for the standard deviation of  $X$ .

The Excel functions for variance and standard deviation are  $\text{VARP}()$  and  $\text{STDEVP}()$ .

**Example:** Find the mean, variance and standard deviation for the following data set.  
5, 7, 9, 9, 10

### Solution

The mean is  $\bar{X} = (5 + 7 + 9 + 9 + 10) / 5 = 40/5 = 8$ , the variance is 3.2, and the standard deviation is  $\sqrt{3.2} = 1.79$ .

	X	$X - \bar{X}$	$(X - \bar{X})^2$
	5	-3	9
	7	-1	1
	9	1	1
	9	1	1
	10	2	4
Sum:	40	0	16
Average:	8	0	3.2

## Frequency tables

The **frequency** of a value is the number of times that it occurs in the data set. The **cumulative frequency** of a value is the number of terms that are less than or equal to that value – it is found by adding up the frequencies of all values up to the given value. A frequency table contains information about the frequency of each value in the data set.

**Example:** Give a frequency table for the following data set:

1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 6, 6, 6, 7, 8, 8, 9, 9, 9

**Solution:**

Value	Frequency	Cumulative Frequency
1	2	2
2	3	5
3	6	11
4	3	14
5	3	17
6	3	20
7	1	21
8	2	23
9	3	26

Notice that the last number in the third column (cumulative frequency) gives the total number of observations. In the example above,  $N = 26$ . We can also use the cumulative frequencies to find the value in position  $p$ . To do this, we scan the cumulative frequencies until we find the first number that is greater than or equal to  $p$ , then we read off the number at the beginning of the row. For example, the 13<sup>th</sup> smallest number is **4**, and the 18<sup>th</sup> smallest number is **6**. Thus, a frequency table makes it easy to find the median and the percentiles of a data set.

**Example:** Using the frequency table below, calculate the cumulative frequencies, then find the median, the first quartile and the third quartile.

Value	Frequency
0	2
1	2
2	5
3	2
4	3
5	3
6	1
7	2
8	5

**Solution:**

Value	Frequency	Cumulative Frequency
-------	-----------	-------------------------

0	2	2
1	2	4
<b>2 1<sup>st</sup> Quartile</b>	5	<b>9</b>
3	2	11
<b>4 Median</b>	3	<b>14</b>
5	3	17
6	1	18
<b>7 3<sup>rd</sup> Quartile</b>	2	<b>20</b>
8	5	<b>25</b>

The cumulative frequencies (running totals) are shown above. From the last entry in this row we see that  $N = 25$ . The median is the value in position  $p = (25+1)/2 = 13$ , which is **4**. The first quartile is the value in position  $p = 1 + (25-1)/4 = 7$ , which is **2**. The third quartile is the value in position  $p = 1 + (25-1)*3/4 = 19$ , which is **7**.

### Calculating the mean from a frequency table

To find the sum of the data values, multiply each distinct value by its frequency, then add the products together. To find the average, divide the sum by  $N$ . We illustrate the procedure using the frequency table from the previous two examples.

Value	Frequency	Product
0	2	0
1	2	2
2	5	10
3	2	6
4	3	12
5	3	15
6	1	6
7	2	14
8	5	40
<b>Total:</b>	<b>24</b>	<b>105</b>
<b>Mean = 105/24 = 4.375</b>		

### Grouped Frequency Tables

If the data set contains a large number of different values, then it is helpful to separate the values into **class intervals** or **bins** when making a frequency table. The number of observations that fall into a given bin is called the class frequency. Here are some guidelines for making a grouped frequency table.

1. Each class interval should have the same width, in order to enable comparisons between different intervals.
2. The class intervals should be listed in order (ascending or descending).
3. There should be no gaps between class intervals.

**Example:** Create a grouped frequency table for the following set of exam scores.

60	90	76	81	59	46
48	61	57	78	86	65
63	54	68	93	71	78
79	67	75	87	76	74
86	57	62	95	80	70

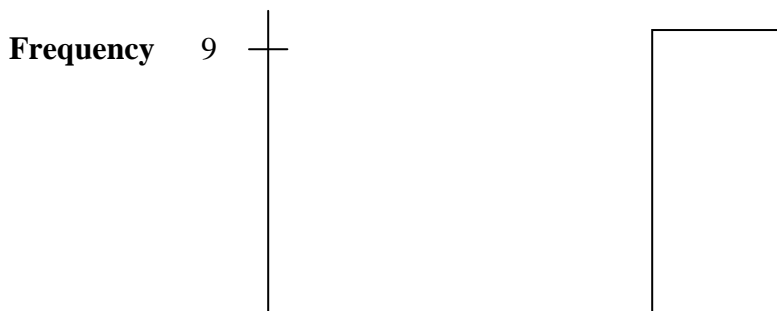
**Solution:**

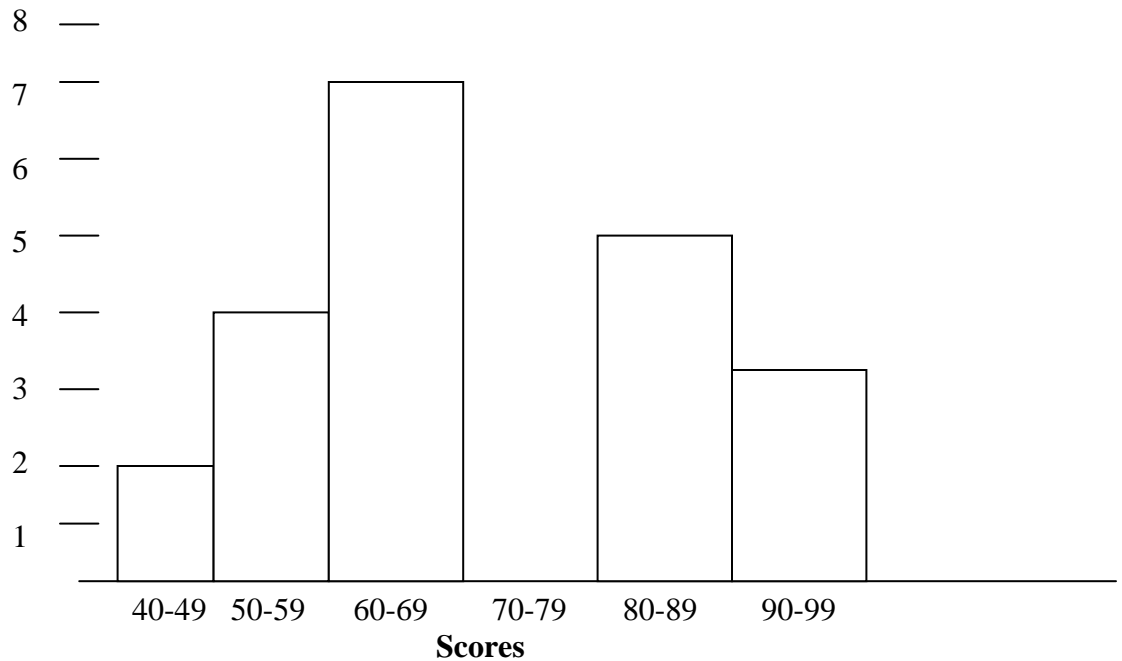
The lowest score is 46 and the highest score is 95. We will group the data into class intervals of width 10. The first interval (40 – 49) has two scores, 46 and 49; so it has a class frequency of 2. The second interval (50 – 59) has four scores: 54, 57, 57, 59. We continue to count the scores that fall in each interval, and obtain this group frequency table.

Class interval	Class Frequency
40 – 49	2
50 – 59	4
60 – 69	7
70 – 79	9
80 – 89	5
90 – 99	3

## ***Histograms***

A histogram is a bar chart that graphically displays the information in a grouped or ungrouped frequency table. It has one bar for each class interval, and the height of a bar corresponds to its frequency. It is important to use class intervals of equal width when creating a histogram, or else the histogram will give a misleading picture of the data. The histogram corresponding to our grouped frequency table is shown below.





## Stem-Leaf Diagrams

Creating a **stem-leaf diagram** is an efficient way to summarize your data set. Each data value is said to consist of a **stem** and a **leaf**. The leaf is the rightmost digit, and the stem is formed from the other digits. For example, if the value is 625, then the stem is **62** and the leaf is **5**.

The stems are listed in the first column, in ascending or descending order. To the right of each stem, we list all of the leaves with that stem. It is preferable that the leaves be in ascending order as well, but this is not strictly necessary. In our previous example, we had four scores with the stem 5 – they were 54, 57, 57, and 59. Thus, in our stem-leaf diagram, we write 4779 in the row corresponding to that stem. The full stem-leaf diagram is shown below.

Stem	Leaves
4	68
5	4779
6	0123578
7	014566889
8	01667
9	035

A stem-leaf diagram can be viewed as a kind of histogram on its side. Each stem corresponds to a class interval, and the length of the list of leaves is the class frequency. But a stem-leaf diagram offers a number of advantages.

1. Creating a stem-leaf diagram by hand is quick and easy.
2. It is fairly easy to calculate the median and other percentiles from a stem-leaf diagram.
3. The stem-leaf diagram contains all of the information that was present in the original data set. In contrast, a grouped frequency table contains less information than the original data set. We lose information by grouping the data into bins.

**Assignment:**

1. Suppose that your data set consists of the following 10 values.

7, 15, 13, 21, 7, 49, 16, 10, 21, 7

Using a hand-held calculator:

- a. Calculate the mean, median, and mode.
  - b. Calculate the first quartile and the third quartile.
  - c. Calculate the variance and the standard deviation.
2. Repeat the previous exercise using the following data set.

25.21, 24.79, 16.73, 24.88, 30.15, 29.76, 28.88, 23.29, 21.00
  3. The second worksheet of the attached Excel workbook lists daily high and low temperatures for March 2006 in the Twin Cities. Perform the following calculations for both data sets using Excel.
    - a. Calculate the mean, median, and mode.
    - b. Calculate the 40<sup>th</sup> percentile.
    - c. Calculate the variance and the standard deviation.
  4. Using the list of daily high temperatures from the previous exercise, create a grouped frequency table, a histogram, and a stem-leaf diagram.
  5. The third worksheet of the attached Excel workbook simulates rolling three dice 100 times. It has a frequency table showing how many times each total from 3 to 18 was obtained.
    - a. Calculate the cumulative frequencies.
    - b. Find the mean, the median, and the mode.
    - c. Find the first and third quartiles.