3 DESCRIPTIVE STATISTICS II: SUMMARIZING THE DATA

This is the second in a series of two chapters on descriptive statistics. In the last chapter we covered data visualization. In this chapter, we will move on to look at the data in a more quantitative manner using summary statistics. Summary statistics of a dataset are specific numbers that give us a 'big picture' of the overall data. The most commonly used summary statistics include:

- A measure of location or central tendency like mean, mode, or median
- A measure of dispersion or spread like variance, or standard deviation

Let's look at each one in more detail.

3.1 MEASURES OF CENTRAL TENDENCY

A measure of central tendency is a summary statistic which tries to describe the entire dataset with a single central or middle value (and hence the name!). The centre of a data set can be measured in different ways, and the method chosen can lead to different conclusions.

Three measures of central tendency are mean, median and mode. Each of these measures follows a different methodology to identify the central value.

Mean

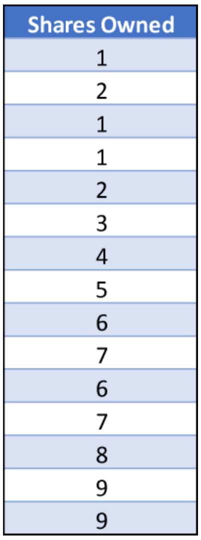
Simply put, mean is the average i.e. the sum of all observations divided by the number of observations.

Thus, if mean is represented by μ, it can be calculated using the following formula:

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

Where,
N = Number of observations in population
x_i = value of i^{th} observation of the data

For example, if Sophie wants to find the average number of shares held by her over the last fifteen days, she needs to find the mean of the "Shares Owned" column/variable.



Shares Owned
1
2
1
1
2
3
4
5
6
7
6
7
8
9
9

Thus,

$$\textbf{mean number of shares} = (1+2+1+1+2+3+4+5+6+7+6+7+8+9+9)\,/\,15 = 4.73$$

Weighted Mean

Sometimes, data can be presented in the form of a frequency distribution or weights.

Weighted average or weighted mean is the mean resulting from the sum of multiplication of each value of data with a weight reflecting its importance divided by the total weight of all observations.

$$weighted\ mean = \frac{\sum x_i w_i}{\sum w_i}$$

$$E(X) = \frac{\sum x_i f_i}{N}$$

$$E(X) = \sum_{1}^{N} x_i p_i$$

$$V(X) = \sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

$$V(X) = \sigma^2 = \sum pi(x_i - \mu)^2$$

$$SD(X) = \sqrt{V(X)} = \sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

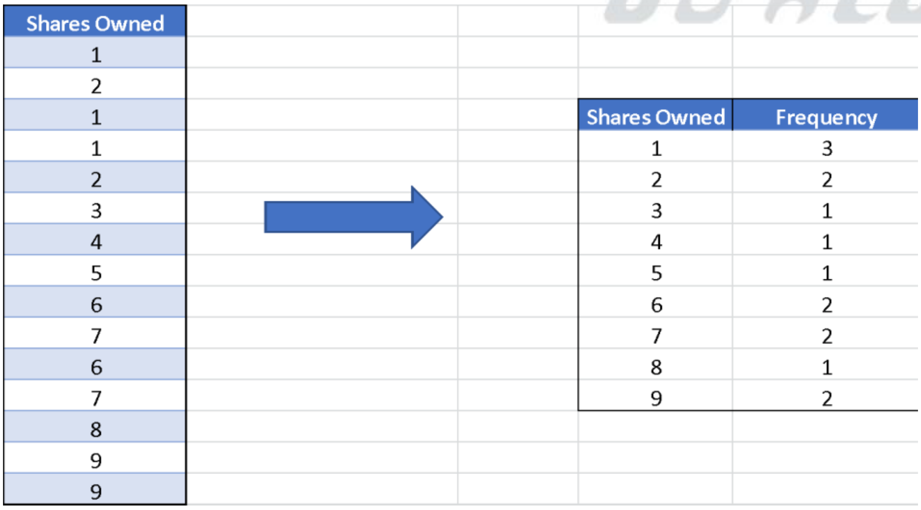
$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

$$S_n = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$$

$$S = \sqrt{\frac{\sum(x_i - \underline{x})^2}{n - 1}}$$

Where,
w_i = weight of the i^th value of data
x_i = value of the i^th observation

For example, the “Shares Owned” variable can be presented as:



Shares Owned
1
2
1
1
2
3
4
5
6
7
6
7
8
9
9

Shares Owned	Frequency
1	3
2	2
3	1
4	1
5	1
6	2
7	2
8	1
9	2

The weight in this case is just the frequency.

Thus, we can also find the mean as:

Weighted mean = (1*3 + 2*2 + 3*1 + 4*1 + 5*1 + 6*2 + 7*2 + 8*1 + 9*2) / 15
= 4.73

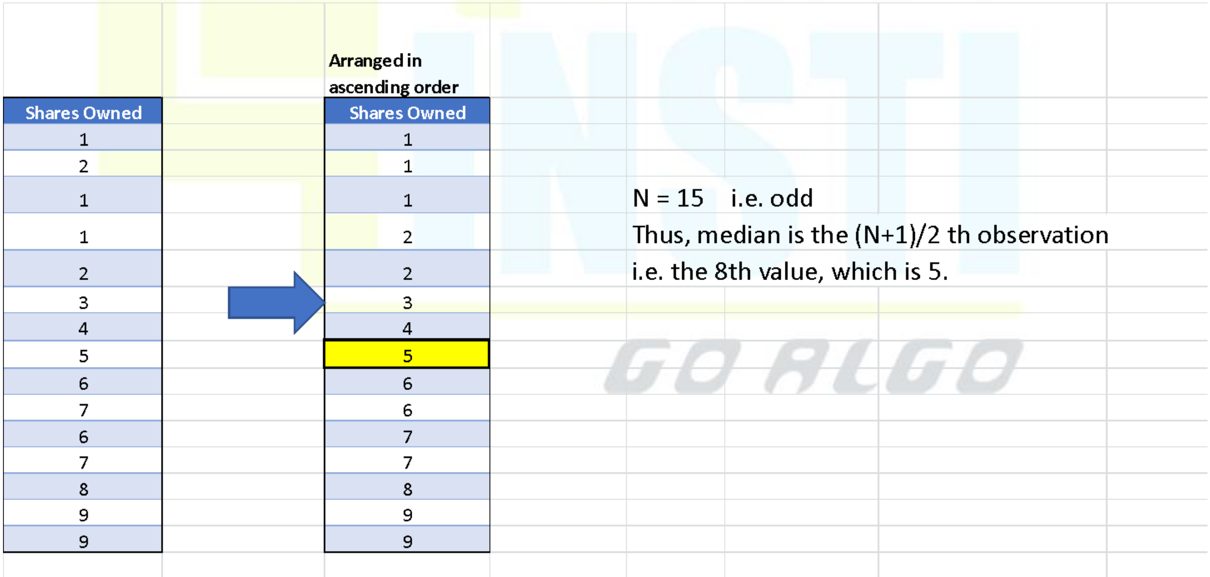
Median

Median is the middle value for the dataset once it is arranged in ascending order. One feature of median is that outliers or extreme values have very less influence on its value. Median is calculated on the basis of number of observations (N):

For **odd** number of observations,
Median = Central observation or $(N+1)/2$ th observation

For **even** number of observations,
Median = Midway between the two central observations = $\{[N/2]^{th} + [(N/2) + 1]^{th}\}/2$

For instance, consider the "Shares Owned" variable again. Its median can be calculated by first arranging it in increasing order as follows:



Shares Owned
1
2
1
1
2
3
4
5
6
7
6
7
8
9
9

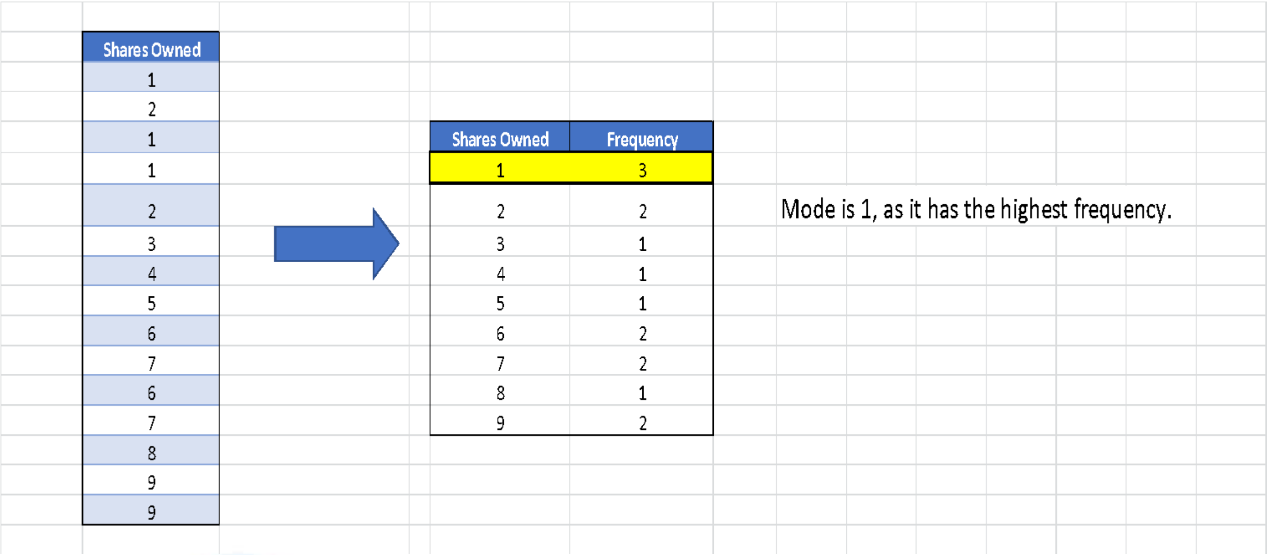
Arranged in ascending order

Shares Owned
1
1
1
2
2
3
4
5
6
6
7
7
8
9
9

N = 15 i.e. odd
Thus, median is the (N+1)/2 th observation
i.e. the 8th value, which is 5.

Mode

Mode is the value that occurs most often in a set of observations or dataset. It is not very useful for continuous data but is informative for discrete measurements. For instance, we can use the frequency distribution of the "Shares Owned" variable again to spot its mode as follows:



Shares Owned
1
2
1
1
2
3
4
5
6
7
6
7
8
9
9

Shares Owned	Frequency
1	3
2	2
3	1
4	1
5	1
6	2
7	2
8	1
9	2

Mode is 1, as it has the highest frequency.

When there are two or more values with the highest frequency, then mode does not give a useful representation of central or typical value of distribution. This is one of its limitations.

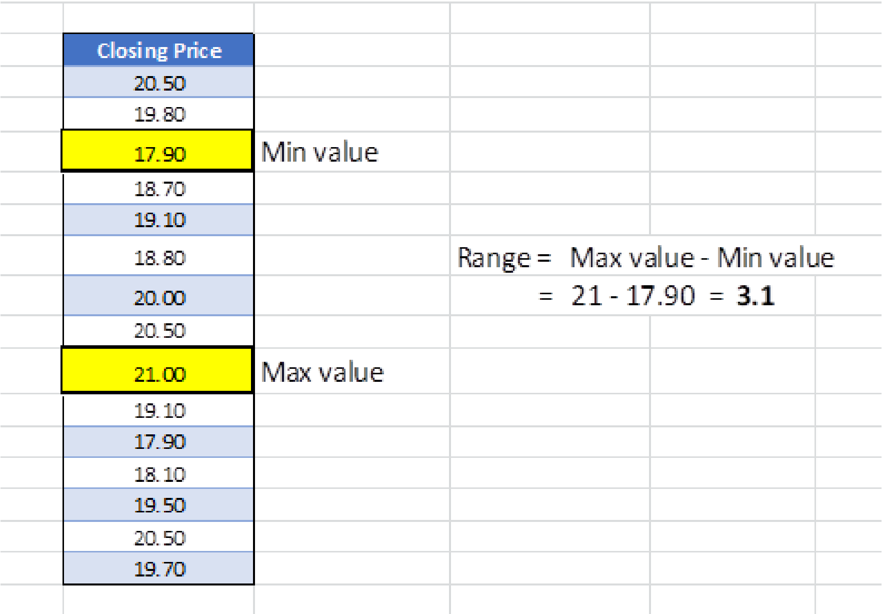
3.2 Measures of Dispersion

Measures of dispersion provide an idea about how stretched or squished the spread of values in a dataset is. Two datasets can have the same mean but considerably different dispersions. Thus, summarizing the data with only a measure of central tendency can be misleading. In order to adequately describe the dataset, a measure of dispersion is used along with measures of central tendency.
The most commonly used measures of dispersion are range, standard deviation and variance.

Range

Range is the simplest measure of calculating variability and provides the difference between the largest and smallest value of a dataset.

For example, consider the variable "Closing Price" again. Its range can be calculated as follows:



Closing Price	
20.50	
19.80	
17.90	Min value
18.70	
19.10	
18.80	
20.00	
20.50	
21.00	Max value
19.10	
17.90	
18.10	
19.50	
20.50	
19.70	

Range = Max value - Min value
= 21 - 17.90 = **3.1**

One of the limitations of using range is that it provides information only about the extreme values and does not give any idea about how the dataset is distributed in between those values.

Standard Deviation

Standard deviation is a measure of spread of data relative to its mean. High deviation indicates that data points are quite far away from the mean while less deviation indicates that data points are very close to the mean.

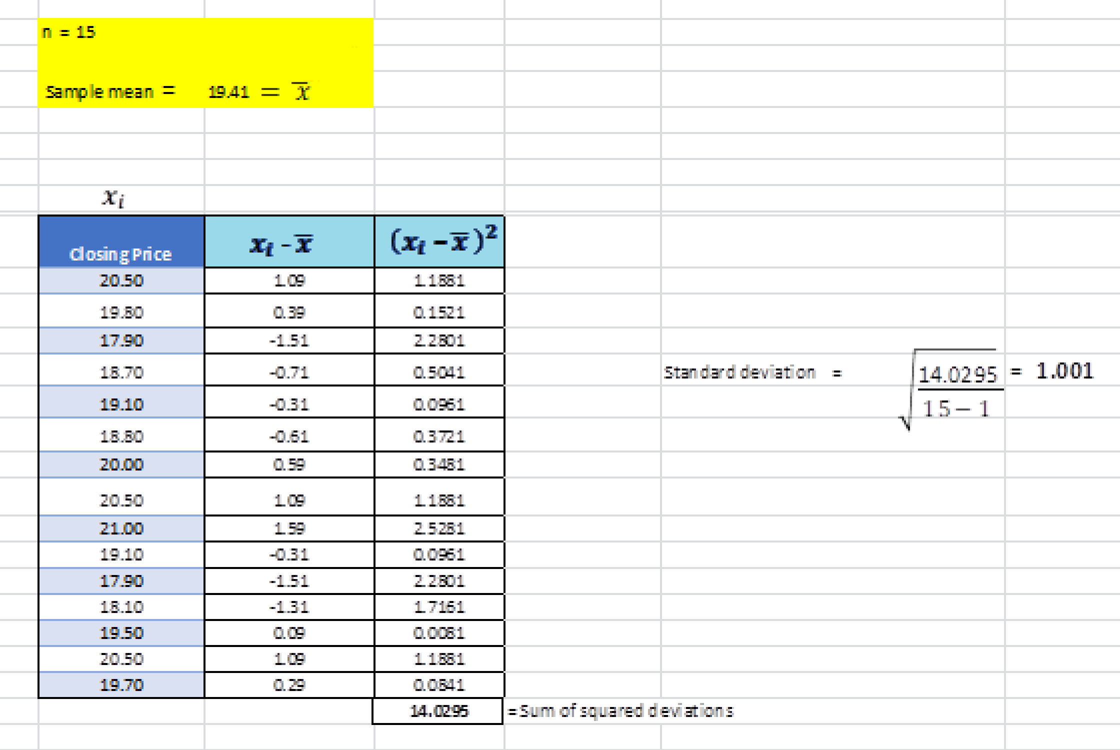
Standard deviation is widely used as a proxy for volatility in the field of quantitative finance. While conducting studies and research, we often encounter a subset of overall values of data, also called a sample dataset.
A sample standard deviation is calculated using the following formula –

$$\boxed{\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}}$$

Where,
σ= Sample standard deviation
x_i = value of i^{th} observation
n = number of observations
$\underline{x}$ is the mean of the sample values (x_is)
$\sum_{i=1}^{n}(x_i - \bar{x})^2$ is the sum of all squared deviations from the sample mean $\bar{x}$

Standard deviation has the same unit as the dataset of observations and can only be a non-negative value.

For example, if Sophie wants to get a value which represents the volatility in the price of stock XYZ for the previous fifteen days that she's traded, she can calculate the standard deviation of the variable "Closing Price" as follows:



n = 15

Sample mean = 19.41 = $\overline{x}$

x_i

Closing Price	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
20.50	1.09	1.1881
19.80	0.39	0.1521
17.90	-1.51	2.2801
18.70	-0.71	0.5041
19.10	-0.31	0.0961
18.80	-0.61	0.3721
20.00	0.59	0.3481
20.50	1.09	1.1881
21.00	1.59	2.5281
19.10	-0.31	0.0961
17.90	-1.51	2.2801
18.10	-1.31	1.7161
19.50	0.09	0.0081
20.50	1.09	1.1881
19.70	0.29	0.0841
		14.0295

=Sum of squared deviations

Standard deviation = $\sqrt{\dfrac{14.0295}{15-1}}$ = 1.001

Thus, based on the data, the **standard deviation of Closing Price is $ 1.001.**

Variance

Variance is another measure of the spread of data that is linked closely with the standard deviation. Variance is calculated by averaging the squared differences from the mean. A small variance means the data points are close to the mean while a large variance value signifies highly dispersed data.

Mathematically, standard deviation is the square root of the variance or variance is just the square of the standard deviation. Both essentially convey the same information.

Variance is calculated using the following formula –

$$\boxed{\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

For example, the variance of the variable "Closing Price" is nothing but the square of the value of the standard deviation:

Variance $= (Standard\ deviation)^2 = (1.001 \times 1.001) = 1.002$