Lab 06 Microsoft Excel

Types of Data

Data can appear in one of four types:

- **Nominal –** the data is categorical, names, e.g. eye colour – green, brown or numbers on the back of soccer jerseys. You cannot do arithmetic on data.
- **Ordinal –**the data is categorical, it has order but the intervals between the values may not be equal, e.g. poor, satisfactory, good
- **Interval**- the data is numerical data. Differences between numbers start to take meaning. Celsuis and Fahrenheit are good examples. The difference between 20 degrees and 30 degrees is the same as the difference between 50 degrees to 60 degrees. Most statistic functions can be done but ratio statements don't make sense. 50 degrees is not twice as hot as 25 degrees.
- **Ratio –** the data is numerical data that can make statements like "twice as much" of "half as much". It makes sense to say that 4 inches is twice as much as 2 inches.

Descriptive Statistics part I

Descriptive Statistics is a term given to the analysis of 'describing' data

There are two main categories:

- Central tendency – some 'central' aspect of the data
- Measures of dispersion – how 'spread-out' or dispersed' the data is

There are 3 important Measures of Central Tendency:

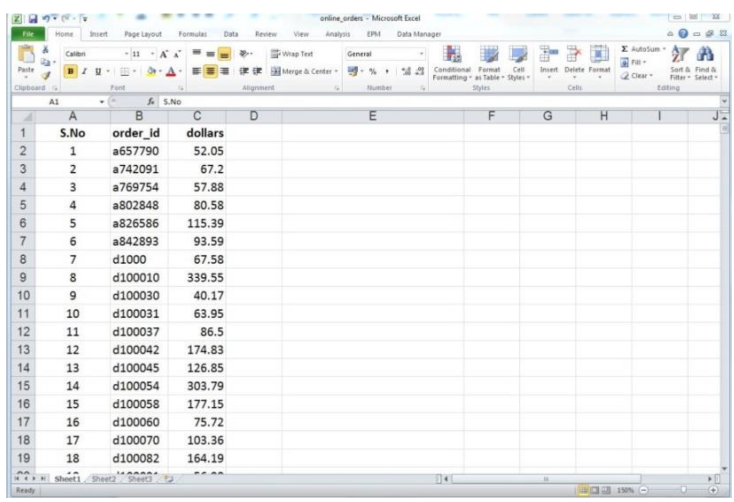
- Mean
- Median
- Mode

Download the ***online_orders.xls***. This data consists of orders in dollar value at an online grocery store in the US in a particular month. We will use this file for various descriptive statistics.

- The very first column of this data set is a running serial number, so we have a total of 11,121 observations.

- The second column is some internal identification mark given to these orders.

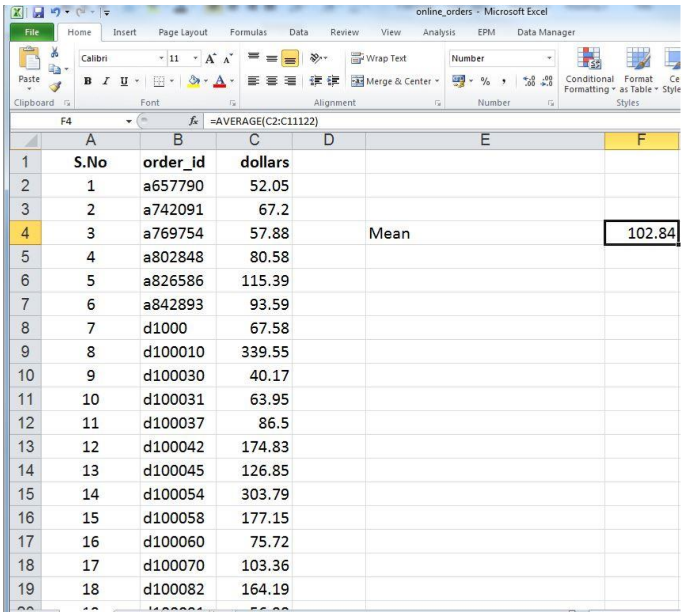
- The third column is the actual dollar value of the grocery order.



S.No	order_id	dollars
1	a657790	52.05
2	a742091	67.2
3	a769754	57.88
4	a802848	80.58
5	a826586	115.39
6	a842893	93.59
7	d1000	67.58
8	d100010	339.55
9	d100030	40.17
10	d100031	63.95
11	d100037	86.5
12	d100042	174.83
13	d100045	126.85
14	d100054	303.79
15	d100058	177.15
16	d100060	75.72
17	d100070	103.36
18	d100082	164.19

Mean

- So for example, the very first order of that month was $52.05. The tenth order in that month was worth $63.95, and so on
- Calculate mean. Use average C2 to C11122. The mean value is 102.84

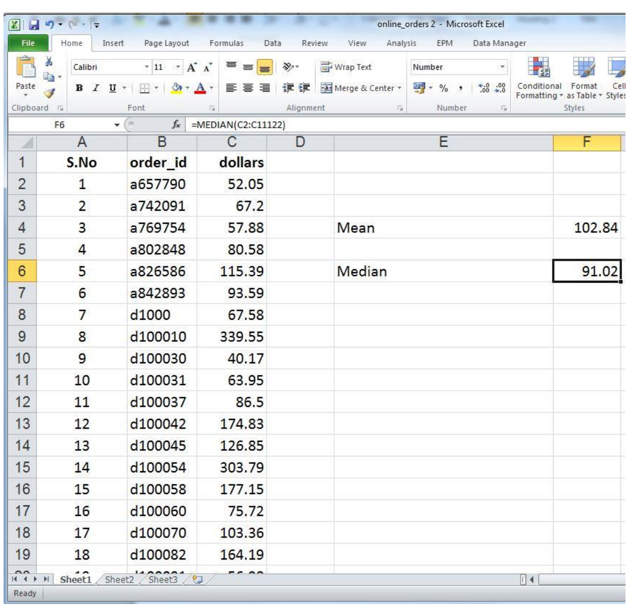


	A	B	C	D	E	F
1	S.No	order_id	dollars			
2	1	a657790	52.05			
3	2	a742091	67.2			
4	3	a769754	57.88		Mean	102.84
5	4	a802848	80.58			
6	5	a826586	115.39			
7	6	a842893	93.59			
8	7	d1000	67.58			
9	8	d100010	339.55			
10	9	d100030	40.17			
11	10	d100031	63.95			
12	11	d100037	86.5			
13	12	d100042	174.83			
14	13	d100045	126.85			
15	14	d100054	303.79			
16	15	d100058	177.15			
17	16	d100060	75.72			
18	17	d100070	103.36			
19	18	d100082	164.19			

Median

The Median of a set of ordered observation is a middle number that divides the data into two parts.

- The Excel command for median is =Median(number1, number2…)

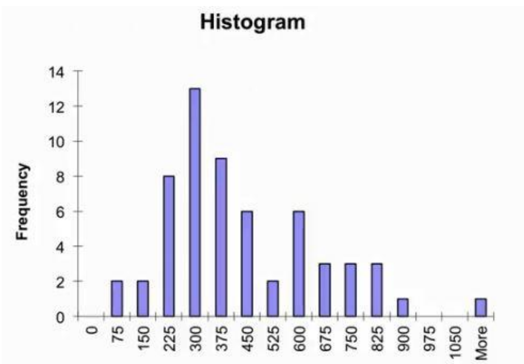


	A	B	C	D	E	F
1	S.No	order_id	dollars			
2	1	a657790	52.05			
3	2	a742091	67.2			
4	3	a769754	57.88		Mean	102.84
5	4	a802848	80.58			
6	5	a826586	115.39		Median	91.02
7	6	a842893	93.59			
8	7	d1000	67.58			
9	8	d100010	339.55			
10	9	d100030	40.17			
11	10	d100031	63.95			
12	11	d100037	86.5			
13	12	d100042	174.83			
14	13	d100045	126.85			
15	14	d100054	303.79			
16	15	d100058	177.15			
17	16	d100060	75.72			
18	17	d100070	103.36			
19	18	d100082	164.19			

Mean versus Median

We have a Histogram plot of CEO salaries of small firms, we saw this in a previous lab, this plot was skewed to the right.

There were fewer CEO with really large salaries – mean was greater median. This is what makes the data distribution skewed to the right.



(CEO Salaries in 1000's)

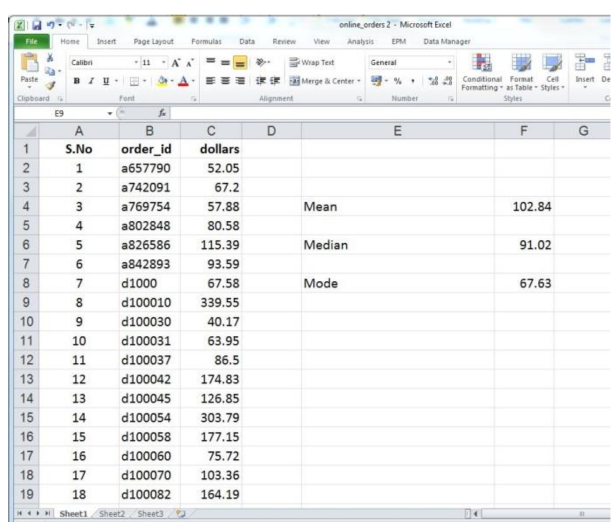
- If you calculate the mean and median for this data, it turns out to be $404,170 and $350,000 respectively.
- The mean is greater than the median, and this is what makes the data distribution skewed to the right.

Mode

- The mode is the most frequently occurring value in a set of data

MODE.SNGL(number1, number2

- The SNGL in the command stands for single mode
- A mode is not a very relevant statistic when the data is essentially continuous.
- For example, consider the daily exchange rate between the US dollar and Euro in a particular month.
- The mode is not very relevant because the nature of data is such that no value occurs more than once. Even if it occurs more than once, the likelihood of such occurrence would be low. Thus, little information is gained knowing the mode.



	A	B	C	D	E	F
1	S.No	order_id	dollars			
2	1	a657790	52.05			
3	2	a742091	67.2			
4	3	a769754	57.88		Mean	102.84
5	4	a802848	80.58			
6	5	a826586	115.39		Median	91.02
7	6	a842893	93.59			
8	7	d1000	67.58		Mode	67.63
9	8	d100010	339.55			
10	9	d100030	40.17			
11	10	d100031	63.95			
12	11	d100037	86.5			
13	12	d100042	174.83			
14	13	d100045	126.85			
15	14	d100054	303.79			
16	15	d100058	177.15			
17	16	d100060	75.72			
18	17	d100070	103.36			
19	18	d100082	164.19			

Measures of Dispersion/Spread

Let us consider two sets of salaries in two different small firms of seven employees each.

Firm 1	Firm 2
$34,500	$35,800
$30,700	$25,500
$32,900	$31,600
$36,000	$41,700
$34,100	$35,300
$33,800	$33,800
$32,500	$30,800
Mean = $33,500	**Mean = $33,500**
Median = $33,800	**Median = $33,800**

- The mean and median salaries in both firms are exactly the same.
- If you look closer at the salaries, the spread of salaries in Firm 2 is much more than in Firm 1.
- The visual shows these two sets of salaries plotted together:

- It is clearly seen that the spread of salaries in Firm 2 is greater than that in Firm 1.
- So how does one translate this difference in spread into some meaningful descriptive statistic?
- One way to do so is to calculate the range of data, which simply is the difference between maximum and minimum values in the data.

Calculate the range of data

The ‘Range’ measure

=Maximum of data – Minimum of data

<u>Range of salaries in Firm 1</u>

= Maximum Salary - Minimum Salary

= $36,000 - $30,700

= **$5,300**

<u>Range of salaries in Firm 2</u>

= Maximum Salary - Minimum Salary

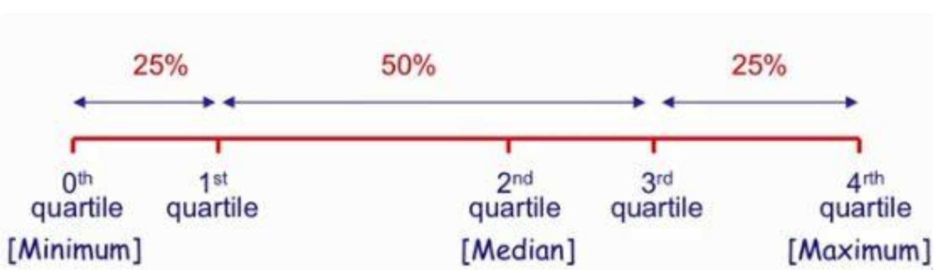
= $41,700 - $25,500

= **$16,200**

- The maximum salary in Firm 1 is $36,000 and the minimum is $30,700, giving us a range of $5,300.
- Similarly, the range of salaries in Firm 2 is $16,200. A higher range, implying greater dispersion or spread in the data.

The range measure of dispersion leads into yet another similar measure. Namely, the inter quartile range, or IQR.

The ‘Inter Quartile Range’ measure (IQR)



- This defines the middle 50% of data, leaving 25% of the data to the right, and 25% to the left.
- The 1st quartile is a number such that 25% of the observations are less than or equal to this number.
- Similarly, the 3rd quartile is a number such that 75% of the observations are less than or equal to that number.

- The median, incidentally, is the second quartile.
- The minimum number in the range is the zeroth quartile.
- And the maximum number in the range is the fourth quartile.
- The interquartile range or IQR is:

IQR = 3rd quartile – 1st quartile

Why do we prefer the inter quartile range over the range measure?

- As you will see most analysis specialization, almost always, we'll be dealing with a sample of data.
- Our effort will be to make some inferences about the population from which the sample comes.
- In that context, the range of a sample is not indicative of the range of population from where it comes.

In Excel, we use the QUARTILE.INC() function:

=QUARTILE.INC(array, quart)

- This function takes in two inputs, your data array and the particular quartile, zeroth, first, second, third, or the fourth you wish to calculate.
- The inter quartile range can then be calculated in Excel as the difference between the third and the fourth quartiles.

IQR = QUARTILE.INC(array,3) – QUARTILE.INC(array,1)

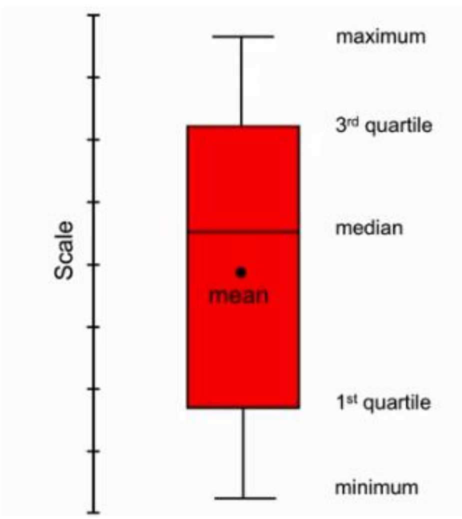
We wish to calculate the inter quartile range of the dollar value of all the orders received at this grocery store during the month.
=QUARTILE.INC(C2:C11122,3)QUARTILE.INC(C2:C11122,1)

	A	B	C	D	E	F
1	S.No	order_id	dollars			
2	1	a657790	52.05			
3	2	a742091	67.2			
4	3	a769754	57.88		Mean	102.84
5	4	a802848	80.58			
6	5	a826586	115.39		Median	91.02
7	6	a842893	93.59			
8	7	d1000	67.58		Mode	67.63
9	8	d100010	339.55			
10	9	d100030	40.17		Inter Quartile Range	63.32
11	10	d100031	63.95			
12	11	d100037	86.5			
13	12	d100042	174.83			
14	13	d100045	126.85			
15	14	d100054	303.79			
16	15	d100058	177.15			
17	16	d100060	75.72			
18	17	d100070	103.36			

Box Plot and Standard Deviation

In this next section, we will see a useful plot to visualize the various descriptive statistics, the box plot.

- Standard Deviation is the most commonly used measure of dispersion in data
- We will conceptually understand the measure, and see Excel commands to calculate it.
- A box plot is a nice visual representation of these various descriptive measures that we have studied until now. A box plot is also known as a box and whisker plot.



- The two whiskers on the top and bottom are the maximum and minimum of your data.
- The vertical line on the left-hand side is the scale of measurement.
- The rectangular box is your interquartile range, bounded by the first and third quartiles.
- The horizontal line inside the box is the median, and the dot represents the mean.
- In one visual you get a good summarization of data.
- Such a visual becomes particularly useful when you are comparing two data sets.

Excel in it's native form is not adept at producing a box plot. Nevertheless, learning to interpret data in terms of a box plot is a very useful skill to have. So we have covered two measures of dispersion in data.

The Standard Deviation measure

The standard deviation measure is much more commonly used in descriptive statistics for dispersion or spread in data.

How does it differ from the range and interquartile range measures?

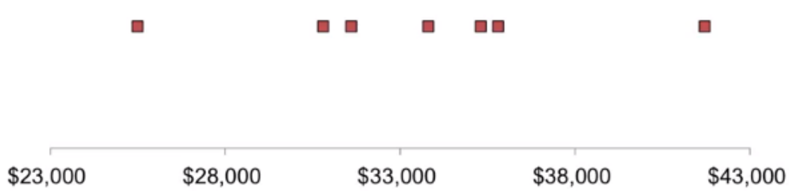
Consider the 7 salaries in a small firm:

- 35,800
- 25,500
- 31,600
- 41,700
- 35,300
- 33,800
- 30,800

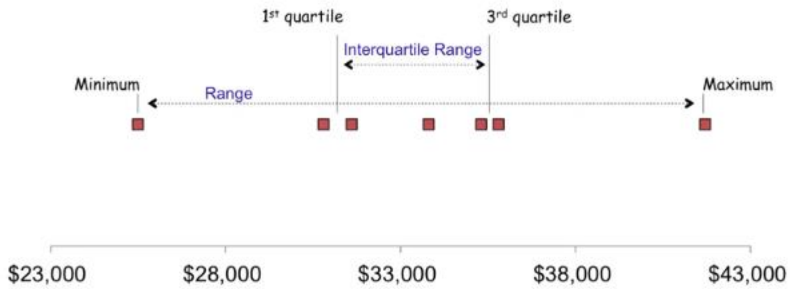
Mean = 33,500

Median = 33,800

The same information is visually displayed:



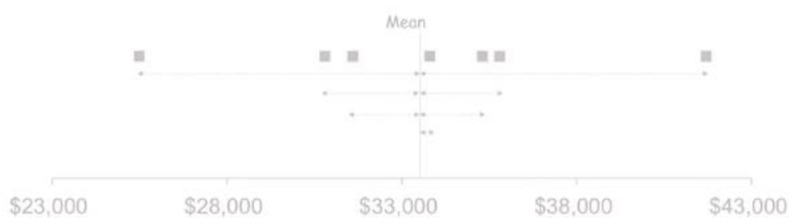
The range and interquartile range measures describe dispersion, or spread, in data in terms of difference between some high value and some low value. For example, the range is the difference between some maximum value and minimum value in data. The interquartile range is the difference between a high value, the third quartile and a low value, which is the first quartile.



On the other hand, the standard deviation first calculates the mean of the data. And then computes differences between each of the data points and the mean.

- It then combines these differences to give the standard deviation measure.
- N in this formula is the total number of observations.
- The standard deviation formula sums the square of differences, divides it by N, and then takes the square root of the result.

$$\text{Standard Deviation} \quad = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(difference_i)^2}$$



Anyway, you need not remember the formula as we will be using Excel to calculate the standard deviation. As we are considering the entire set of employees in the firm, or in other words, the population of employees in the firm. We use the Excel command STDEV.P to calculate standard deviation

The Excel Command (population standard deviation)

=STDEV.P(number1, number2, …)

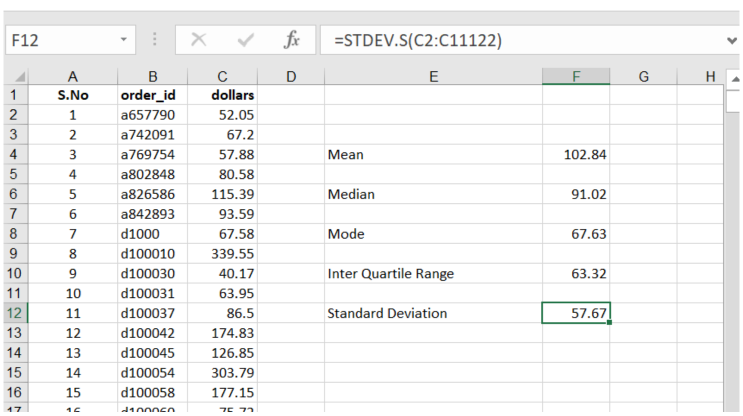
If we had data which was a sample from some larger population of data, which, by the way, typically would be the case in a majority of business applications. We would use the Excel command STDEV.S to calculate the standard deviation rather than STDEV.P. The P and S standing for population and sample.

Excel Command (sample standard deviation)

=STDEV.S(number1, number2, …)

The STDEV.S is the command that we will be using. Let us calculate the standard deviation of dollar orders in our grocery data file.

We wish to calculate the standard deviation of orders received at this grocery store. Or in other words, we wish to assess the dispersion, or spread, in this data.



F12 | =STDEV.S(C2:C11122)

	A	B	C	D	E	F	G	H
1	S.No	order_id	dollars					
2	1	a657790	52.05					
3	2	a742091	67.2					
4	3	a769754	57.88		Mean	102.84		
5	4	a802848	80.58					
6	5	a826586	115.39		Median	91.02		
7	6	a842893	93.59					
8	7	d1000	67.58		Mode	67.63		
9	8	d100010	339.55					
10	9	d100030	40.17		Inter Quartile Range	63.32		
11	10	d100031	63.95					
12	11	d100037	86.5		Standard Deviation	57.67		
13	12	d100042	174.83					
14	13	d100045	126.85					
15	14	d100054	303.79					
16	15	d100058	177.15					

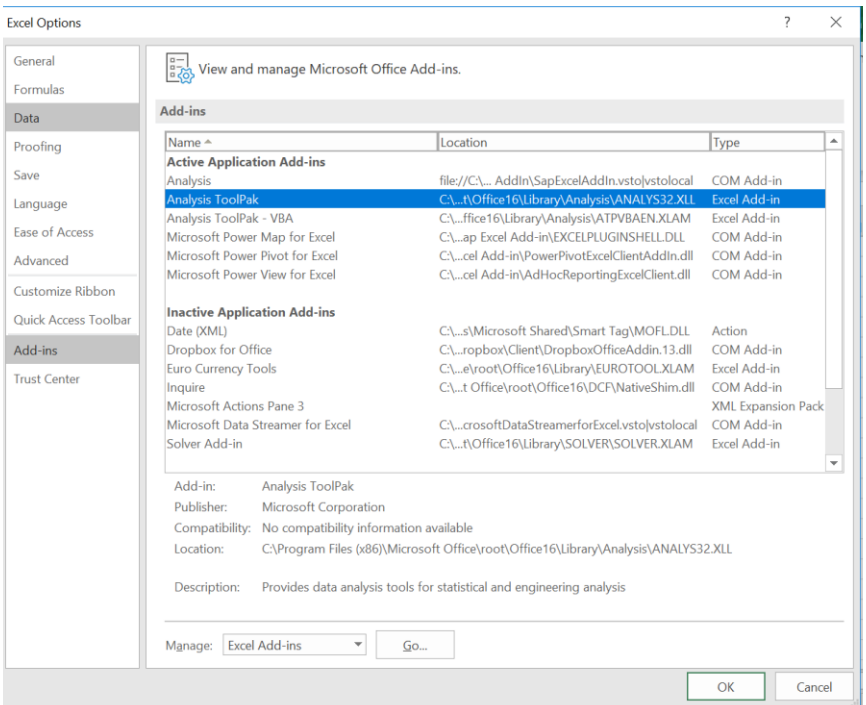
This gives us our standard deviation measure, 57.67. The units of standard deviation are the same as the data, so in this case it would be $57.67. This is the spread of my data.

Excel's Analysis ToolPak.

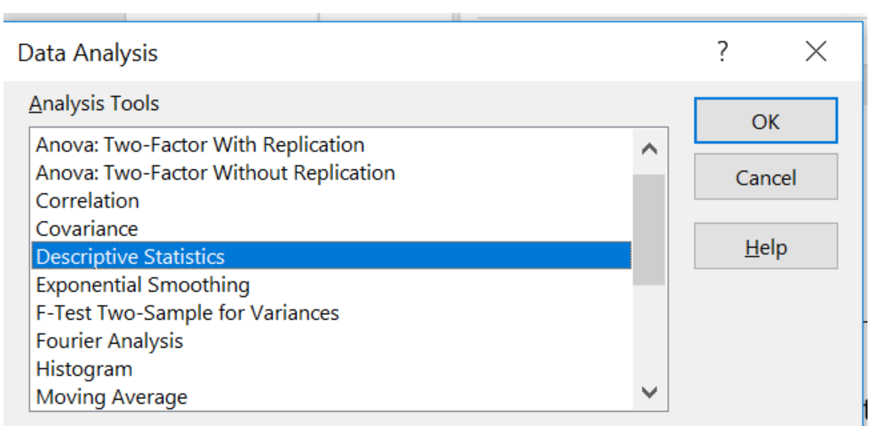
This is an extensive set of statistical analysis tools. It's an add-in that you have to install. It’s available on Excel 2016.

Open the spreadsheet, toolpak.xls.

Click on the File tab, click Options near the bottom of the left panel. In the dialog box, choose Add-ins near the bottom of the left panel. In the Add-ins list, pick Analysis Toolpak. Go.



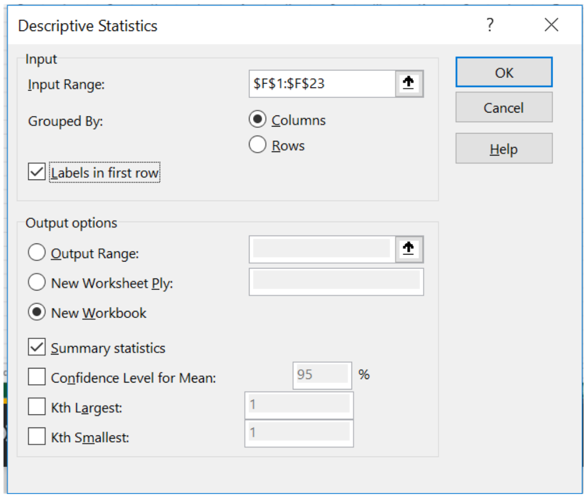
Data ribbon, data analysis



Scroll up and down and take a look at the list of tools that this Data Analysis Add-in provides.

Select Descriptive statistics.

We'll use it to calculate a number of statistics about the rainfall data. With the Input Range box active, enter the column that holds the data, starting with the label at the top, in cell F1, to the last cell in the column, cell F23. See screenshot:



Average Annual Rainfall (inches)	
Mean	65.59090909
Standard Error	3.159990062
Median	67
Mode	63
Standard Deviation	14.82166719
Sample Variance	219.6818182
Kurtosis	-0.324747374
Skewness	-0.503526479
Range	56
Minimum	32
Maximum	88
Sum	1443
Count	22

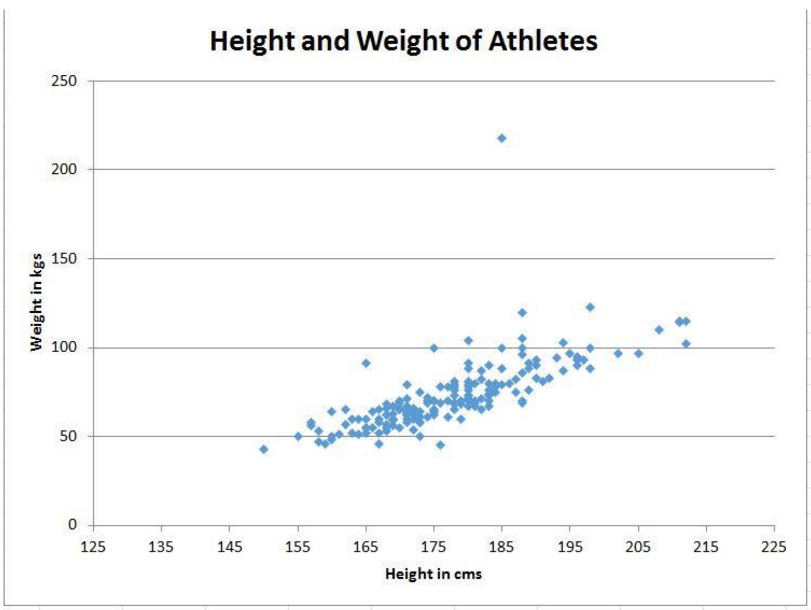
Our previous examples focused on two categories of descriptive statistics, measures of central tendency and measures of dispersion on a single variable. To recap, descriptive statistics are measures, a summary set of numbers that describe the multiple observations in a data.

The descriptive statistics on central tendency that we covered were the mean, median, and the mode. While the descriptive measures of dispersion or spreading the data were the range, interquartile range, standard deviation, and variance.

This section will introduce descriptive measures of association between two variables, the covariance and the correlation measures through using their Excel commands. To describe the co-variation among two variables, means how do the two variables vary together or co-vary?

Scatter plot

The scatter plot was introduced in lab 04. The scatter chart is known as the XY plot. A scatter chart is a visual representation of the relationship between two variables. We analysed the relationship between the height and weight of certain Olympic athletes from a recent Olympic games. Heights being represented on the horizontal axis and weights on the vertical axis. The pattern of the scatter plot visually indicates that there is a positive relationship between the height and weight of an athlete. That is as height increases, so does weight.



A descriptive measure of association attempts to numerically quantify this relationship between height and weight, that is in a set of numbers to describe this relationship, - the covariance and correlation.

Covariance

Covariance is a measure of how changes in one variable are associated with changes in a second variable. Specifically, covariance measures the degree to which two variables are linearly associated. A positive number indicates a positive relationship, e.g height and weight. The covariance measure is affected by change in units of variables in the variables as it is susceptible to the unit of measurement. We can arbitrarily inflate or deflate the covariance by choice of units. So as a result, the covariance measure is not very appropriate if one needs to assess the strength of relationship between two variables. The measure is fine as long as we need to know the direction of relationship. That is, when one variable increases or decreases, what happens to the other variable? The covariance cannot be directly interpreted in terms of how strong the relationship is.

Correlation

Both **covariance and correlation** indicate whether variables are positively or inversely **related**. **Correlation** also tells you the degree to which the variables tend to move together

We will use correlation to establish a relationship or connection between two or more things. The function in Excel is: =CORREL, the inputs are the two data ranges.

Excel Command (correlation)

=CORREL(*range1*, *range2*)

The correlation is not affected by change of units. It is always bound between -1 and +1, with the positive value of correlation indicating a positive relationship and negative value indicating a negative relationship. Further, closer the correlation is to a +1 or -1, stronger is the positive or negative relationship between the two variables. If its excess of positive 0.5 (> +0.5) it is considered a strong positive relationship. And a correlation less than negative 0.5 (>-0.5), is considered a strong negative relationship between the two variables.

- Range: −1 to +1
- Not affected by the units of measurement.

Using the **height-and-weight.xls** file we will calculate the correlation between height of the athlete in centimetres, and the weight in kilograms to confirm the correlation in height and weight.
Using the function CORREL =CORREL(C2:C1588,E2:E1588)

The following number 0.774625 is shown which indicates a strong positive relationship between heights and weights of athletes. The number is in excess of >+0.5.

K11 =CORREL(C2:C1588,D2:D1588)

	A	B	C	D	E	F	G	H	I	J	K
1	**Second name**	**First name**	**Height(cm)**	**Weight(kg)**	**Country/Team**	**Gender**					
2	Abbadi	Ilyas	170	69	Algeria	M					
3	Abbate	Simona	171	64	Italy	W					
4	Abdelaal	Hesham	167	52	Egypt	M					
5	Abdulrahman	Amer	168	68	United Arab Emirates	M					
6	Abian	Pablo	177	78	Spain	M					
7	Abrantes	Arnaldo	184	78	Portugal	M					
8	Abril	Erika	158	47	Colombia	W					
9	Abshero	Ayele	168	62	Ethiopia	M					
10	Abu Drais	Methkal	168	62	Jordan	M					
11	Achara	Kieron	208	110	Team GB	M		Correlation (height, weight)			0.774625
12	Achour	Dallal Merwa	176	45	Algeria	W					
13	Adam	Idrissa	178	79	Cameroon	M					
14	Adams	Antoine	180	79	Saint Kitts and Nevis	M					
15	Adams	Lyukman	194	87	Russian Federation	M					
16	Adams	Nicola	164	51	Team GB	W					
17	Adcock	Chris	183	80	Team GB	M					
18	Adeoye	Margaret	162	65	Team GB	W					
19	Adlington	Rebecca	179	70	Team GB	W					
20	Afroudakis	Georgios	194	103	Greece	M					
21	Agamennoni	Luca	188	96	Italy	M					

www.coursera.org. Basic Data Descriptors, Statistical Distributions, and Application to Business Decisions