Excel
Data Analysis

Course objectives:

- Import data
- Use statistical functions in Excel
- Create histograms
- Gain insights from your data

Student Training and Support

Phone: (07) 3346 4312

Email: askus@library.uq.edu.au

Web: https://web.library.uq.edu.au/library-services/training/

Staff Training (Bookings)

Phone   (07) 3365 2666
Email   staffdev@uq.edu.au
Web   http://www.uq.edu.au/staffdevelopment

Staff may contact their trainer with enquiries and feedback related to training content. Please contact Staff Development for booking enquiries or your local I.T. support for general technical enquiries.

Table of Contents

Exercise document:

Go to https://web.library.uq.edu.au/library-services/training/training-resources and click on Data Analysis (ZIP,40.9 KB) to download. Save these files on your H:/ drive or to your local machine or a USB drive.

Statistical Function definitions can be found at:
https://support.office.com/en-us/article/statistical-functions-reference-624dac86-a375-4435-bc25-76d659719ffd

Importing External Data

Data located in compatible external files can be imported into excel without the need to retype all the information again. Depending on the format of the data you would like to import, different methods can be used, including opening and saving in Excel, linking to data, importing data and copying and pasting data into excel.

Exercise 1.	*Importing External Data*

Open the spreadsheet **Data Analysis_Exercises.xlsx** (which can be found under the Excel section on the Library Training Resources page. The **External Data Link** sheet is selected.
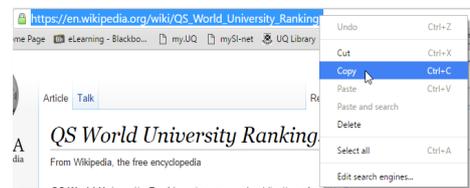
Importing Data from websites
Data from websites and other sources can be imported into Excel if it is in an appropriate format.

1. Copy the URL of the web page with the data you want to import.

    e.g. World University Rankings on Wikipedia (which can be found in cell A1 of the **External Data Link** sheet)

    https://en.wikipedia.org/wiki/QS_World_University_Rankings

Note: For this exercise ignore From Web in the Get External Data group. It will bring in the entire web page and not just a selected table
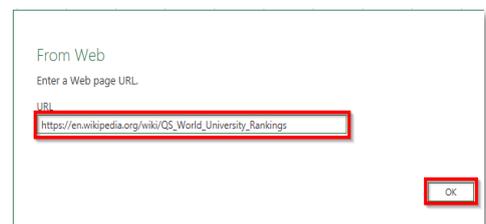
2. Navigate to the **Data** tab

3. Click on **New Query** (in the Get & Transform group)

4. From the drop down menu, select **From Other Sources** > **From Web**

This opens the dialogue box for you to enter the URL of the web page with the data you want to import
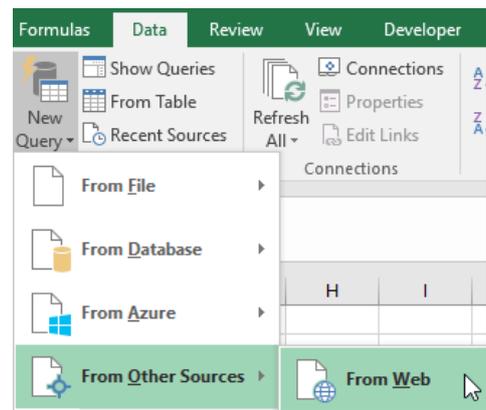
5. **Paste** the **URL** in the From Web dialogue box and click **OK**
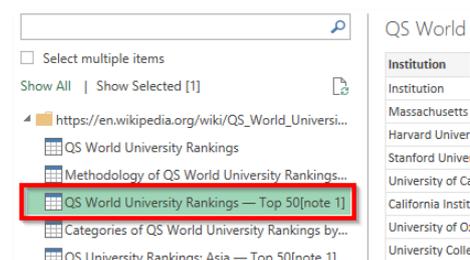
The Navigator Pane will open with a list of data that can be imported into excel

6. Select the required data set (*QS World University Rankings – Top 50)* on the left pane of the Navigator to preview it

NB: You can use the edit button to clean the data before importing

7. Select *QS World University Rankings – Top 50*

8. Click on **Load**

THE UNIVERSITY
OF QUEENSLAND
AUSTRALIA

A connection will be created to the data on the website. This will ensure that refreshing your excel file will update the data to the latest version. Excel will then open a new worksheet with the imported data.
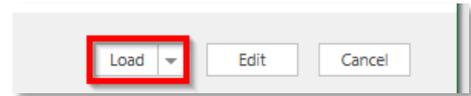
Refresh Linked Data

9. Click on any cell within the data table

10. Click on the **Data** tab

11. Select **Refresh All**

NB: Refresh all will refresh all connections in the workbook. If you want to refresh data on a single sheet click Refresh
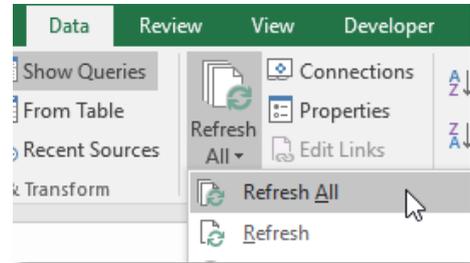
NB: You may get a Microsoft Excel Security Notice about connections to external data sources. You can safely click OK here but see the section on **Considerations when importing data into Excel** below for further information.

Considerations when importing data into Excel

Malware / Macros – Unfortunately there are ways to hide malware inside Excel files. This is usually done via "macros" which are little programs that are typically created to do complex or repetitive tasks. Because hackers have exploited these tools, Microsoft has disabled macros by default in Excel. In fact, when you open an Excel file from an untrusted source, you will get a security warning like this one. If you are working on data from an unknown or untrusted source, use caution before "Enabling Editing"

Some hackers have even learned to use social engineering techniques to try and trick users into turning macros back on. For example there may be an image in the file that appears blurred with a note that it is for security reasons. The goal is to get you to enable macros so that you can 'see' the image when, in reality, enabling the macro allows the virus to run.   Of course if you have good anti-virus / anti-malware programs installed, they will go a long way towards mitigating that threat.

## References within a file or sheet to external data

You can refer to the contents of cells in another Excel workbook by creating an external reference. An external reference (also called a link) is a reference to a cell or range on a worksheet in another Excel workbook, or a reference to a defined name in another workbook.   If your data is coming from a source beyond your immediate control, you may find that these 'links' are broken.   If you don't have access to the workbooks/worksheets where the underlying data lives, you won't be able to use it via the link in the spreadsheet you are currently working on.

Microsoft Excel: *Data Analysis*

Exercise 2.   *Importing data from a file*
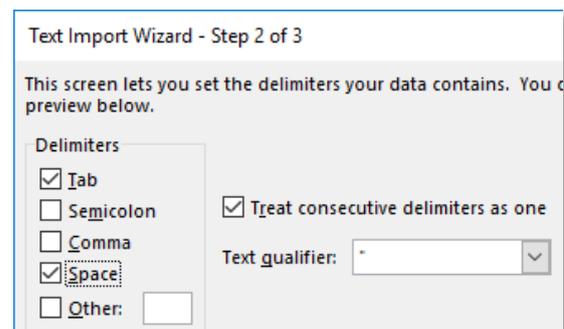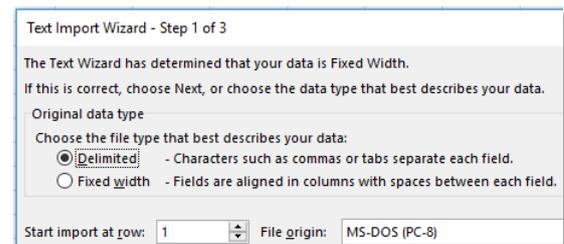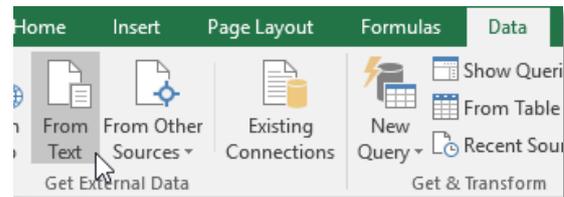
Open exercise files and enable content

1. Open the exercise file **Data Analysis_Exercises.xlsx** and select the **Importing Data & Histograms** worksheet.

2. Click on the **Enable Content** button on the Security Warning (if necessary)

3. If you get a Security Warning dialog box. Click on **Yes**
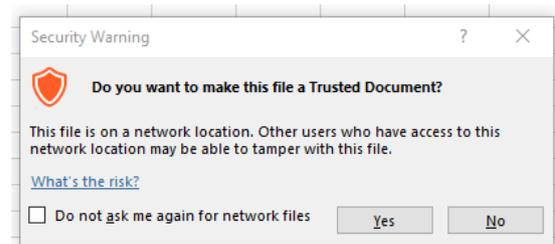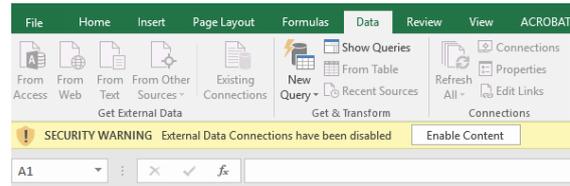
Note: In Office 365 (Windows version) Microsoft removed the Text Import Wizard as an option when using steps below. They force you to use the Power Query window which does not have the "Treat consecutive delimiters as one" option. You can get around this by opening the text file directly in Excel which will launch the wizard below
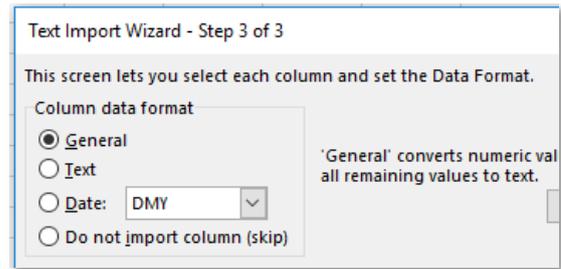
Import data from text file:

4. Click the **Data tab**

5. Click **From Text** (in the Get External Data group)

6. Locate **data_analysis.txt**

7. Click on **Import** (in **Mac** – **Get Data**)

8. Click on **Delimited** option

9. Click **Next**

10. **Tick** the following options:

    Tab

    Space

    Treat consecutive delimiters as one

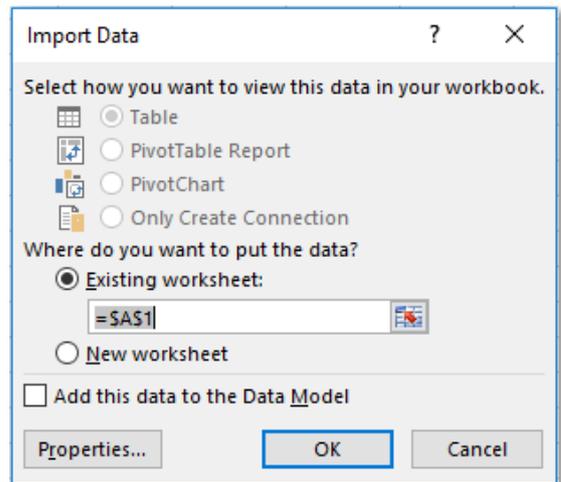11. Click **Next**

12. Ensure **General** option is selected

13. Click **Finish**

Text Import Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

Column data format
- ● General
- ○ Text
- ○ Date: DMY
- ○ Do not import column (skip)

'General' converts numeric val
all remaining values to text.

14. Assign data to **$A$1** in existing worksheet

15. Click **OK**

Import Data   ?   ✕

Select how you want to view this data in your workbook.
- ○ Table
- ○ PivotTable Report
- ○ PivotChart
- ○ Only Create Connection

Where do you want to put the data?
- ● Existing worksheet:
  =$A$1
- ○ New worksheet

☐ Add this data to the Data Model

Properties...   OK   Cancel

Microsoft Excel: *Data Analysis*
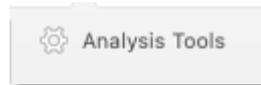
Descriptive Statistics

Descriptive statistics is the discipline of quantitatively (expressed as numbers) describing the main features of a collection of data. Excel's **Analysis Toolpak** add-in offers a variety of features to undertake statistical computations and graphing. Descriptive Statistics is included to provide statistical averages (mean, mode, median), standard error, standard deviation, sample variance, kurtosis and confidence levels of sample data.
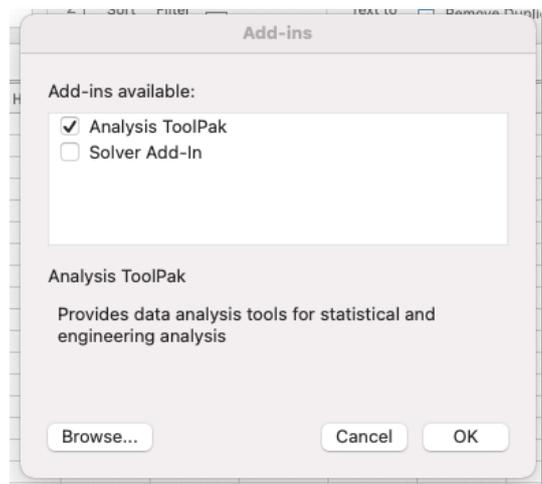
Exercise 3.	*Using Descriptive Statistics*

Mac users may need to add the Analysis Tool Pack

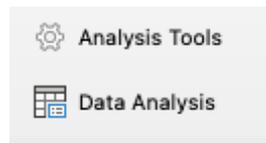**Data Tab** – Far right hand side – click **Analysis Tools** button
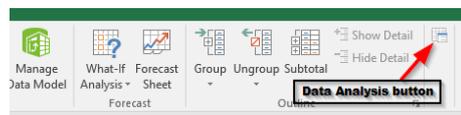
Click next to **Analysis ToolPak**

Choose **OK**

The **Data Analysis button** will now be visible

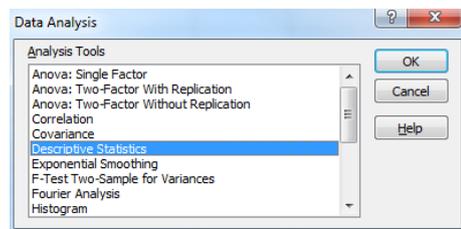1. Click **Data Analysis** (at the far right of ribbon) on the **Data** tab

2. Click **Descriptive Statistics**
3. Click **OK**

4. Highlight cells **$A$1:$D$201** for **Input Range**

5. Select **Grouped by columns**

6. Click **Labels in first row** box

7. Click **Output Range**

8. Highlight cell **$G$1** for **Output Range**

9. Select **Summary statistics**

10. Click **OK**

NB: To obtain descriptive statistics for one group ensure that only one column is selected.
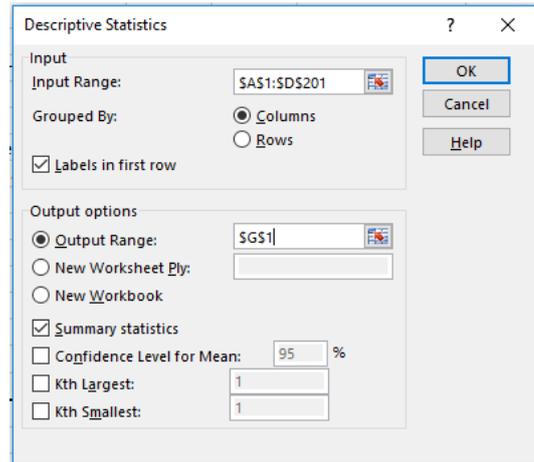
Statistical Functions

Exercise 4.	*Using basic statistical functions in Excel*

To use Basic Statistical Functions

1. Ensure you are on the **Basic Statistics** worksheet

2. Select the Home tab

3. Click in cell **C14**

4. Click **AutoSum**
   Check the range is (C5:C11)

5. Press **Enter**

6. Use Autofill to calculate sum for remaining weeks)

7. Calculate with statistical functions

   Sample size = COUNT
   Mean = AVERAGE
   Minimum value = MIN
   Maximum value = MAX

Note: Mean and Average are different terms for the same thing when dealing with Statistics

8. Select cells **C14** to **C18**

9. Autofill across to fill cells in
   remaining weeks

14	sum	1981	
15	count	7	
16	average	283	
17	min	123	
18	max	787	

NB: For quick statistical reference refer to status bar after highlighting a selection of values. Adjust options on status bar by right clicking on it and selecting items.

✓	Average
✓	Count
	Numerical Count
	Minimum
	Maximum
✓	Sum

Count: 12

Using Variance and Standard Deviation in Excel

Variance is a measure of the average of the squared difference from the mean.
Here is how it is defined manually:

- Subtract the mean from each value in the data. This gives you a measure of the distance of each value from the mean.
- Square each of these distances (so that they are all positive values), and add all of the squares together.
- Divide the sum of the squares by the number of values in the data set.
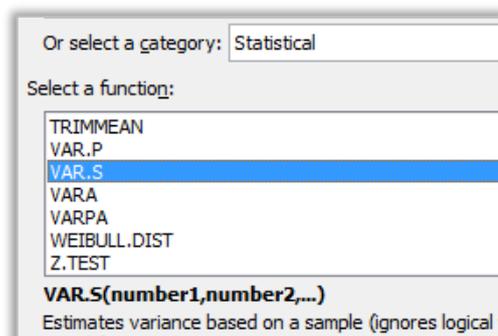 (if calculating variance for a sample subtract 1 from the number of values)

Mean of the Values	283	1. Subtract the mean	2. Square the result	3. Add (Sum) the squares	4. Divide by the number of values
Mon	154	-129	16641	379000	54142.85714
Tue	178	-105	11025		
Wed	458	175	30625		
Thu	156	-127	16129		5. Using Variance function
Fri	125	-158	24964		54142.85714
Sat	787	504	254016		
Sun	123	-160	25600		

The standard deviation (σ) is simply a measure of how close the values are to the average. A smaller number means the values are bunched whilst a larger number indicates values that are spread out.
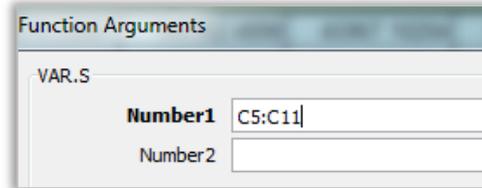
Exercise 5.	*Variance and Standard deviation*

To use Variance Function on a sample

1. Click in cell **C21**

2. Click f_x button in formula bar

3. Change category to Statistical

4. Click on **VAR.S** function

Or select a category: Statistical

Select a function:

TRIMMEAN
VAR.P
VAR.S
VARA
VARPA
WEIBULL.DIST
Z.TEST

VAR.S(number1,number2,...)
Estimates variance based on a sample (ignores logical

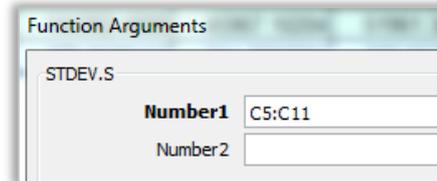5. Select range (C5:C11)

6. Click on **OK**

To use Standard Deviation Function on a sample

1. Click in cell **C22**

2. Click f_x button in formula bar

3. Change category to Statistical

4. Click on **STDEV.S** function
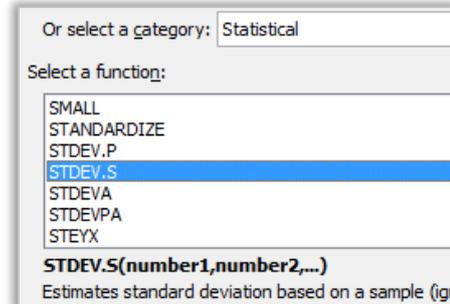
5. Select range (C5:C11)

6. Click on **OK**

Repeat steps above for **entire population** using range **(C5:I11)**
- Click cell C25: Overall Average: =AVERAGE**(C5:I11)**
- Click cell C26: Overall Variance: =VAR.P**(C5:I11)**
- Click cell C27: Overall Std Deviation = STDEV.P**(C5:I11)**
- Click cell C33: Overall Sum = SUM**(C5:I11)**

To find WeeklyTotal as a percentage of the Overall Total
1. Go to cell C34
2. Enter =C14/C33 in the formula bar
3. Press function key F4

Note: This will change cell reference C33 to absolute reference $C$33

4. Press enter
5. Autofill across (D34:I34)
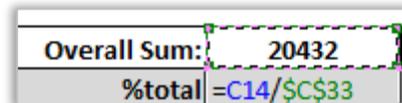
Histograms and Frequency

A histogram is used to display tabulated frequencies of data in graphical form. It is able to show the proportion of data that fits into specific categories or bins. For example, we may want to find out how many items were of a particular length, e.g. 100mm. Excel provides a Histogram tool which is available via the **Analysis ToolPak** add-in. With the latest versions of Excel there is now a **Histogram chart** available in the **Statistics chart** options.

Exercise 6.	*Creating histograms*

Use worksheet "**Importing Data & Histograms**"
Using the tool in **Data Analysis**
  Prepare data for a histogram of weights

| Bin |
|---|
| 0 |
| 50 |
| 100 |
| 150 |
| 200 |
| 250 |
| 300 |
| 350 |
| 400 |
| 450 |
| 500 |

1. Go to cell F19
2. Type "**Bin**"
3. Go to cell F20
4. Type **0**
5. Go to cell F21
6. Type **50**
7. Select F20 and F21
8. **Autofill** to display a value of 500 in cell F30

Input Range: This is the data that you want to analyse by using the Histogram tool.
Bin Range: This represents the intervals that you want the Histogram tool to use for measuring the input data in the data analysis.
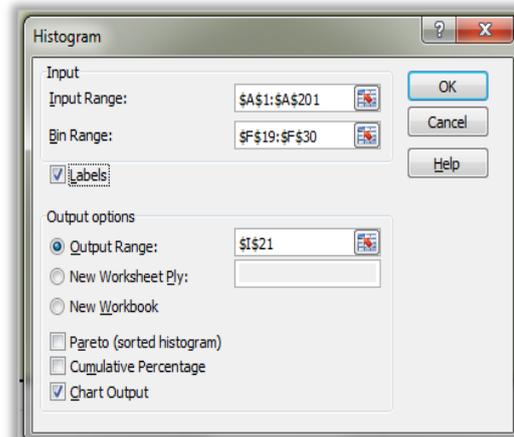
9. Click **Data Analysis** (at the far right of the ribbon) on **Data** tab
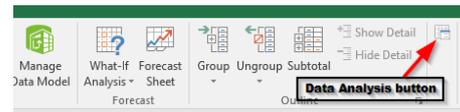
10. Click on **Histogram**
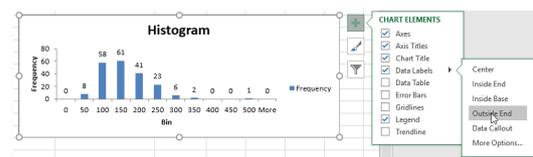11. **Click OK**

Complete the dialog box as follows:
  • **Input Range** = $A1:$A201
  • **Bin Range** = $F$19:$F$30
  • Tick **Labels**
  • **Output Range**: $I$21
  • Tick **Chart Output**

12. Click **OK**

To display the frequencies in Histogram:
1. Click on Histogram in worksheet
2. Click **Data Labels** on Add Chart Element button
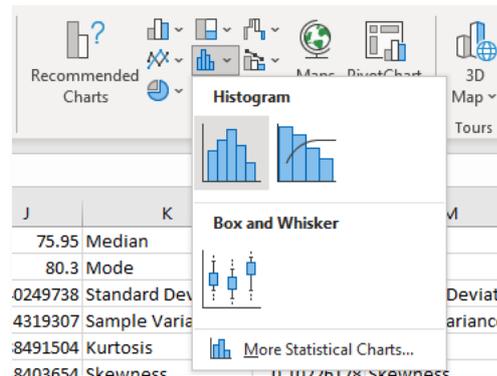3. Select **Outside End**

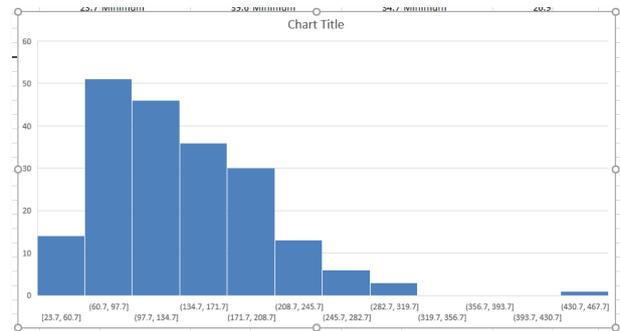**NB**: Table with Bin and Frequency headings will appear along with Histogram graph.
Resize graph as required.

Using the **Statistics Chart - Histogram** option

Select the data range **A1:A201**
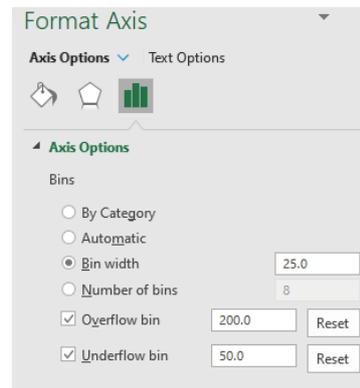**Insert tab** – **Charts - Statistics Chart - Histogram**

A Histogram will appear

Windows:
Single click the **X axis** – **Double click** the **X axis** to launch the Format Axis panel on the right of the screen.
Choose the **Axis Option** and **expand** the **Axis Options**
Set the **Bin Width** to 25
Set the **Overflow bin** to 200
Set the **Underflow bin** to 50

Mac:
Right mouse click the blue **data series columns**
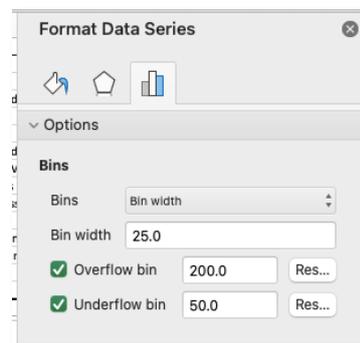**Choose Format Data Series**…
Expand the Data Series Options (if necessary)
Change Bins – Auto to Bin Width
Set the **Bin Width** to 25
Set the **Overflow bin** to 200
Set the **Underflow bin** to 50

Microsoft Excel: *Data Analysis*

Correlation and Linear Regression

A **correlation** is a number between -1 and +1 that summarizes the relationship between two variables. A correlation close to +1 is strong and positive, whereas a correlation close to -1 is strong but negative. A zero correlation means there is no relationship between variables.

Linear regression is a statistical approach to modelling the relationship between a scalar variable *y* and one or more explanatory variables denoted *X*. It can be used for predication or forecasting.
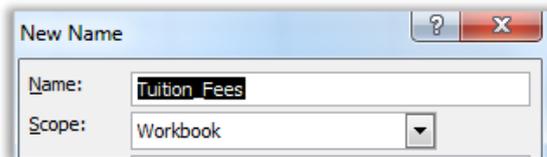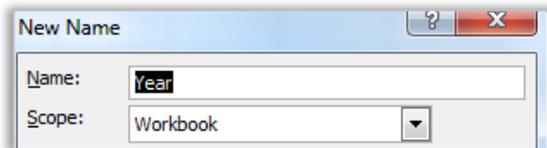
Exercise 7.	*Calculate Correlation Co-efficient*

Select worksheet "**Correlation & Linear Regression**"

**Name cells to find correlation:**
1. Select cells**(B4:B14)**
2. Click Define Name (near middle of ribbon) on Formulas Tab
3. Check name is "**Year**"
4. Click on **OK**

5. Select cells **(C4:C14)**
6. Click Define Name on Formulas Tab
7. Check name is "**Tuition_Fees**"
8. Click on **OK**

To calculate correlation co-efficient
1. Go to cell **B17**
2. Click f_x button in formula bar
3. Select Correl function
4. In Array 1, type **Year** (or press F3 for the Paste Name dialog box; Choose the name Year and press OK)
5. In Array 2, type **Tuition_Fees**
6. Click on **OK**
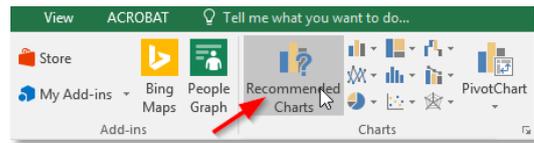7. Format cell B17 to 2 decimal places
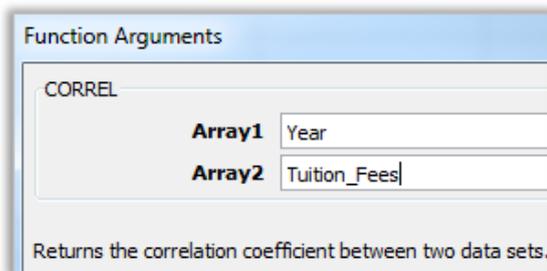
Note: You will be presented with a strong positive correlation of +0.99 between Year and Tuition Fee increases
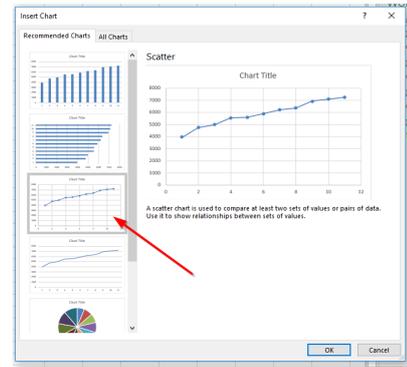
Exercise 8.   *Create Chart and Linear Regression*

Create a chart

1. Select cells**(B4:C14)**
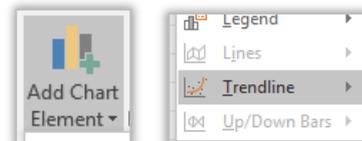2. Insert Tab > Charts group > Recommended Charts

3. Select **Scatter**

Add the regression line

1. Click **Add Chart Element** button – **Trendline** – **Linear Trendline**

2. The Trendline will appear on the chart

3. **Right click** the Trendline
4. Choose **Format Trendline**

5. Within Trendline Options….
6. Select Checkbox to "**Display Equation on Chart**"
   Select Checkbox to **"Display R-squared value on chart"**
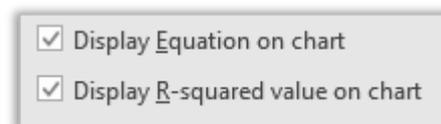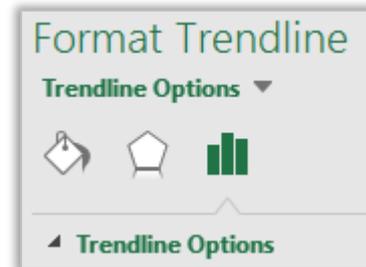
Note: The equation and R squared value will appear towards the top right of the chart. If the formulas are obscured by the Trendline, you can move them by selecting the text box with the formulas and then drag it to where you want.

To Find Regression Summary

1. Click on **Data Analysis** on **Data** tab (far right on ribbon)
2. Select **Regression**
3. Click on **OK**

4. Input Y range, Select **C4:C14**
5. Input X range, Select **B4:B14**
6. Output Range, Select **A22**
7. Click on **OK**

Note: You will be presented with **Summary Output** which includes regression analysis
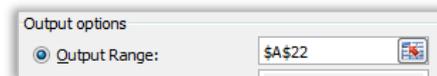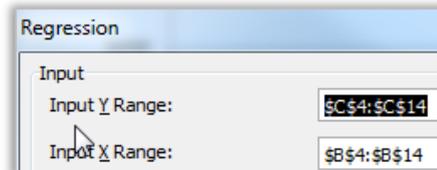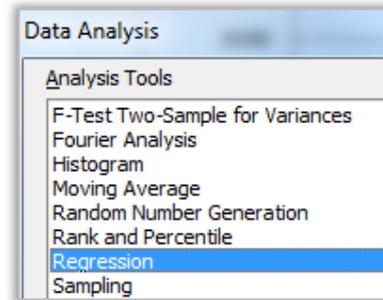
Interpreting results: A demonstrated strong positive correlation:

Equation (*Y=mx+c*) **Y = 308.63x + 4018.1** Matches the coefficients in regression summary

Intercept indicates the predicted cost of tuition in the Year 2000. This is the line of best fit value not the actual value*(the line of best fit value for Y if X=0)*

X Variable indicates the average increase in $ in tuition fees year to year approximately $308.63

Forecasting

Forecasting is estimating the likelihood of an event taking place in the future, based on available data. Statistical forecasting concentrates on using the past to predict the future by identifying trends, patterns and business drives within the data to develop a forecast.

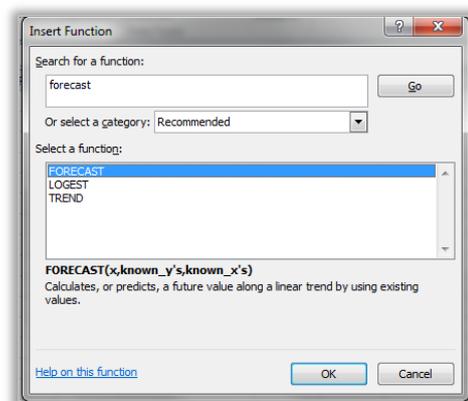| Exercise 9.	*Forecasting*

Use worksheet "**Correlation & Linear Regression**"

In Excel the **FORECAST** function takes raw trendline data, an input (independent variable) and returns the dependent variable

1. Click in **$C$20**

2. Click the **Insert Function** button

3. Select **Forecast** from the list of functions (search for Forecast in the search box if you cannot see it)

4. X, select **B20**

5. Known_y's, select C4:C14 (the range name Tuition_Fees will appear)

6. Known_x's, select B4:B14 (the range name Year will appear)

7. Note how the indicated answer matches the Intercept value of the regression analysis

8. Click **OK**

9. In cell B20 type 20 to forecast the cost of tuition fees in year 20
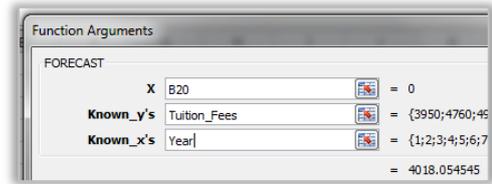
	Year	Fees	
Forecast	20	10190.6	

T Tests

TTests are performed when you have two sets of measurements or results from given populations and you would like to compare them to see if they are significantly different.

For example you may have two lists of measurements from the same set of people. The first set of measurements may have been taken in the morning and the second set in the afternoon. This type of TTest is known as a related TTest or a paired TTest because you have tested the same population twice.

Alternatively if you had two sets of measurements taken from two sets of people with one set being in the morning and the other in the afternoon you would have an unpaired or independent TTest. This is because you have tested two different populations.

If you are sure about the direction of differences, for example that the morning measurements are faster than the afternoon then you perform a one tail t test.
If you are unsure about the difference between the values perform a two tail t test.

A result is called "statistically significant" if the result of the t test comes in at below .05. This is often referred to as the P Value.

| Exercise 10.	*Significance tests*

On the T-Test spreadsheet are two series of measurements.

These measurements are paired as they are from the same population but taken at different times.

|	A	B	C	D
1		Reaction	Times	
2		Morning	Afternoon	
3	Person 1	935	978	
4	Person 2	955	980	
5	Person 3	967	1017	
6	Person 4	1002	973	
7	Person 5	1000	1006	
8	Person 6	964	1017	
9	Person 7	952	995	
10	Person 8	933	1048	
11				
12				
13				

1. Select cell B12

Using the Insert Function button search for and locate the T.Test function.
Note: The TTest function is still available for compatibility purposes with Excel 2007 and below.
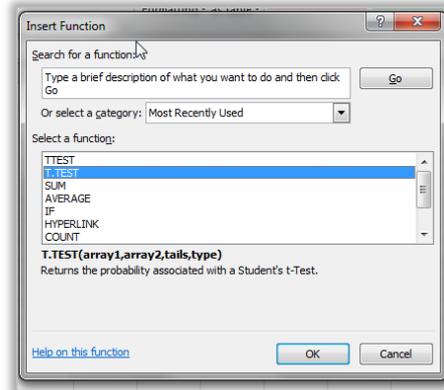
In the T.Test Function Arguments dialog box Array1 and Array2 are the cell ranges containing the two columns of measurements.
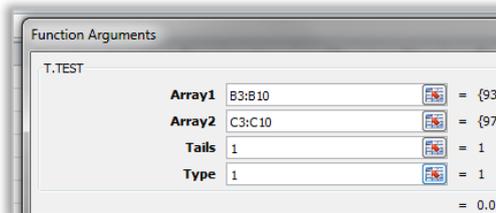In this case B3:B10 and C3:C10
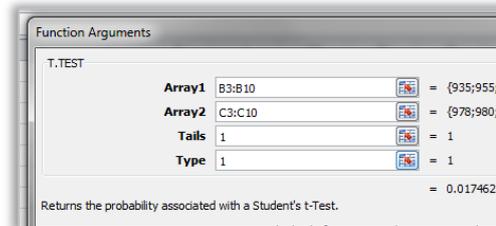
Tails can be either a 1 or a 2
Use 1 if you are sure about the direction of the differences.
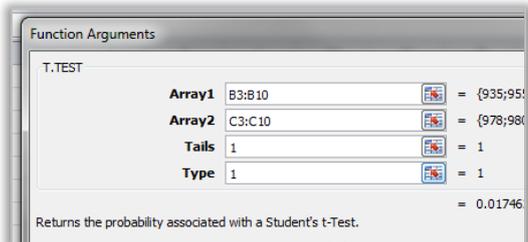Use 2 if you are unsure about the direction of the differences.

Type can either be a 1, 2 or 3
Use 1 if your data is from a paired population.
Use 2 if your data is from an unpaired population with an equal variance.
Use 3 if your data is from an unpaired population with an unequal variance.

ANOVA: Analysis of Variance

In its simplest form, ANOVA provides a statistical test of whether or not the means of several groups are all equal. The ANOVA test is the initial step in identifying factors that are influencing a given data set. Anova should be performed on 3 or more groups of data.
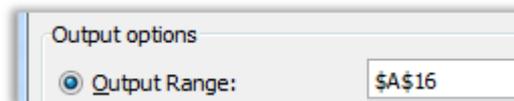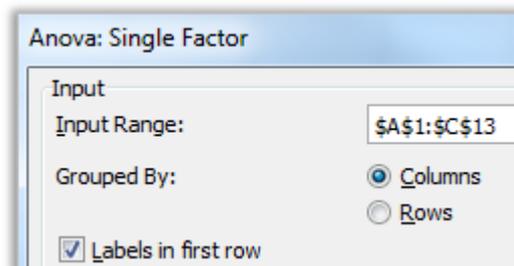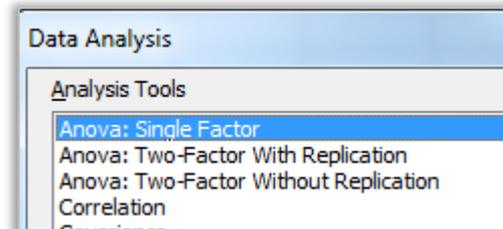
Exercise 11.	ANOVA: Analysis of Variance

Use worksheet "**ANOVA - Rank & percentile**"
To conduct the one-way ANOVA

2. Click on **Data Analysis** on the **Data** Tab (far right on ribbon)

3. Select **Anova: Single Factor**

4. Click **OK**.

5. Select the **input** range (A1:C13) (automatically absolute references)

6. Click "**Labels in first row**" option

7. Select **Output** Range (A16)

8. Click **OK**.

Note: Descriptive statistics and ANOVA summary table are displayed on screen

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Lecture	12	773	64.41667	84.62879
Online	12	846	70.5	164.6364
Video	12	844	70.33333	215.8788

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	288.1667	2	144.0833	0.929282	0.404924	3.284918
Within Groups	5116.583	33	155.048			
Total	5404.75	35				

Interpreting results: In the summary section we can see the mean exam results for each class, But are these differences statistically significant?

There are two types of hypotheses. Null (negative) or Alternative (positive). It is best practice to use null hypotheses so no personal opinions creep in to the testing statement.

A null hypothesis is a default position and can never be proven. Statistically results can only reject or fail to reject the null hypotheses.

Null hypotheses are always phrased as a negative statement e.g. There is no real difference between the effectiveness of lectures, online delivery and video delivery.

The test result shows F **=0.93** With a critical **P-value** of **.4**, the critical **F = 3.285**. Therefore, since the F statistic is smaller than the critical value, we fail to reject the null hypothesis. Remember from before the P value is statistically significant if it is below .05. This value of .4 shows there is some connection in the data though. So, we fail to reject that there is no difference between the effectiveness of lectures, online delivery and video delivery. These values may be explained by the small sample size. A larger sample of data may give more statistically significant results. Apparently, the differences we saw in this sample were simply due to random sampling error.

Rank and Percentiles

Percentile rank means the percentage of scores that fall "at or below" a certain number. Percentiles are most often used for determining the relative standing of an individual in a population or the rank position of the individual. Percentiles measure position from the bottom.

Exercise 12.	*Obtaining your Rank*

Use worksheet "**ANOVA - Rank & percentile**"
1. Click **Data Analysis on the Data Tab** (far right on ribbon)
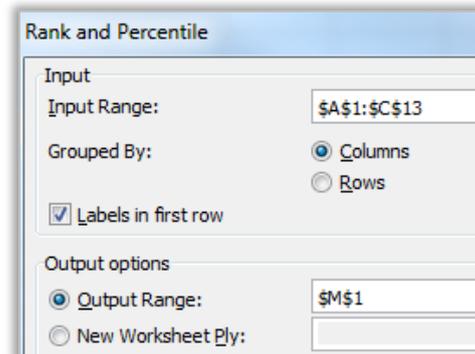2. Click **Rank and Percentile**
3. Click **OK**
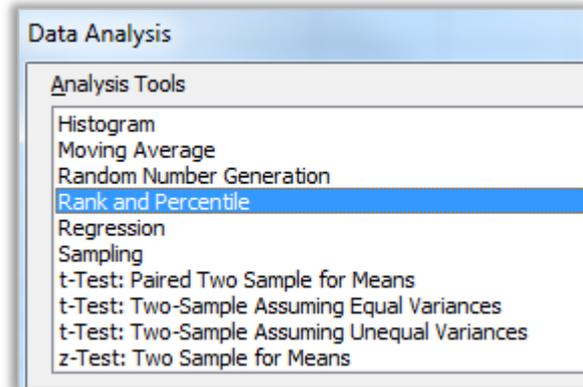
Complete dialog box:
4. Highlight cells $A$1:$C$13 for **Input Range**

NB: In this instance, do not merely click on column A header as the program will process every row in the spreadsheet.

1. In **Grouped By**, select Columns
2. Click **Labels in first row**
3. Select **Output Range** as **$M$1**
4. Click **OK**

Interpreting results:

Point - The location of the value within the original list. This can be used to quickly sort the output table into the same order of the original list.

Original - This is the column containing the original values. This column has the same column name as the original list since we used labels in the first row.

Rank - This is the rank of the corresponding number in the list.

Percent - This is the numbers percentage rank within the list. This percentage indicates the proportion of the list which are <u>below</u> this given number.