Session 7

Using the Statistical Add-Ins (Part I)

SESSION 7: Statistical Add-Ins Part I

Random Number Generation Tool

Statistics deals with data, and many times we want to simulate results by using fictitious data. This means that we would want to generate data that follows a particular distribution. **Excel** has a tool that allows us to create this type of data.

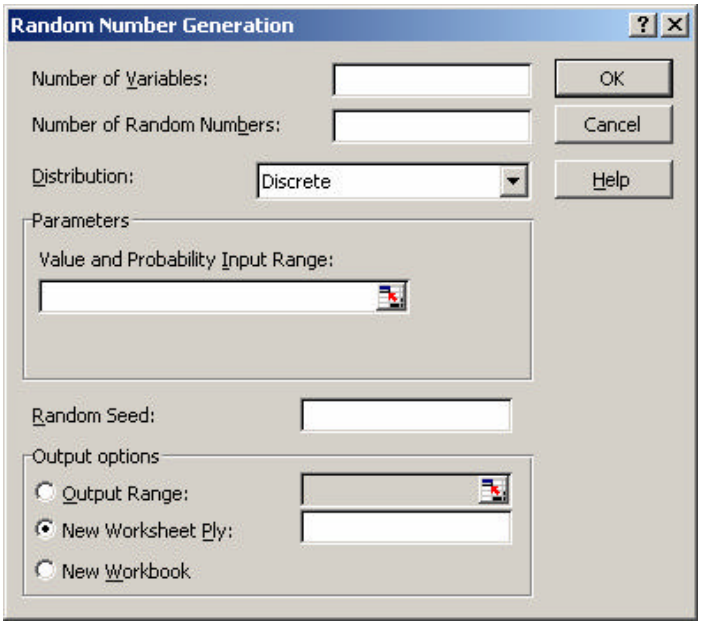
This analysis tool fills a range with independent random numbers drawn from one of several distributions (like Normal, Bernoulli, Poisson, etc.).

Let us open a new datasheet in Excel.

To use the **Random Number Generation** tool, use the following steps:

Click on **Tools** ➢ **Data Analysis** ➢ **Random Number Generation**➢ **OK**.

Excel displays the following dialog box.



Then make your selections from the options provided.

The following table describes the options in the **Random Number Generation** dialog box:

Option	Description
Number of Variables	Enter the number of columns of values you want in the output table. If you do not enter a number, Excel fills all

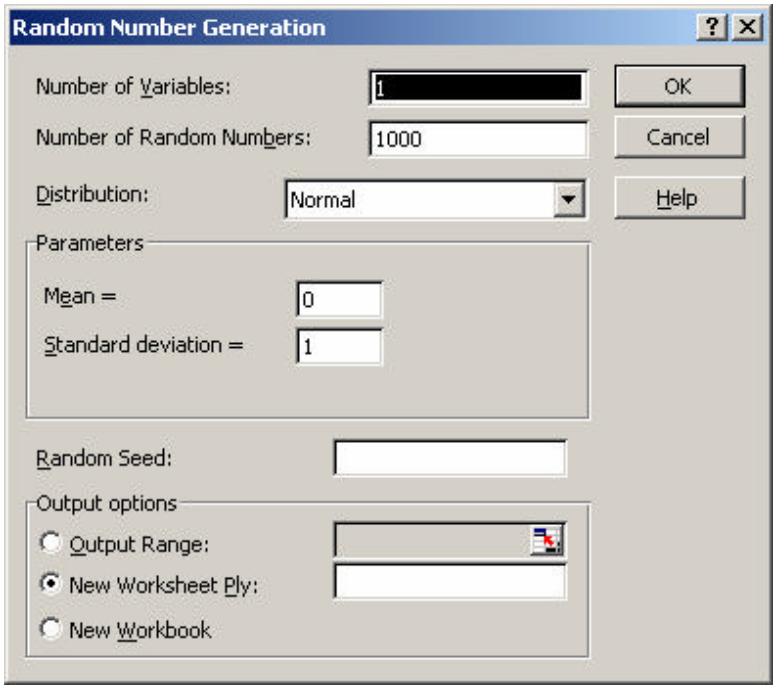
	columns in the output range you specify.
No of random numbers	Enter the number of data points you want to see. Each data point appears in a row of the output table. If you do not enter a number, Excel fills all rows in the output range you specify.
Distribution	Click the distribution method you want to use to create random values.
Parameters	Enter a value or values to characterize the distribution selected. (e.g. normal needs mean and standard deviation)
Random Seed	Enter an optional value from which to generate random numbers. You can reuse this value later to produce the same random numbers.

A distribution can be chosen in the following way:

Option	Description	Parameters
Uniform	Variables are drawn with equal probability from all values in the range. A common application uses a uniform distribution in the range 0...1.	Lower and upper bound.
Normal	A common application uses a mean of 0 and a standard deviation of 1 for the standard normal distribution.	Mean and standard deviation.
Bernoulli	Bernoulli variables have the value 0 or 1. If the variable is less than or equal to the probability of success, the Bernoulli random variable is assigned the value 1; otherwise, it is assigned the value 0.	Probability of success (*p*).
Binomial	The sum of Bernoulli trials gives a binomial random variable.	Probability of success (*p*), and number of trials (*n*).
Poisson	Poisson distribution is often used to characterize the number of events that occur per unit of time (e.g. no of telephone calls in 1 hour).	Lambda (1 over the average).
Patterned	Creating a patterned sequence	Characterized by a lower and upper

		bound, a step, repetition rate for values, and repetition rate for the sequence.
Discrete	The range must contain two columns: The left column contains values, and the right column contains probabilities associated with the value in that row. The sum of the probabilities must be 1.	A value and the associated probability.

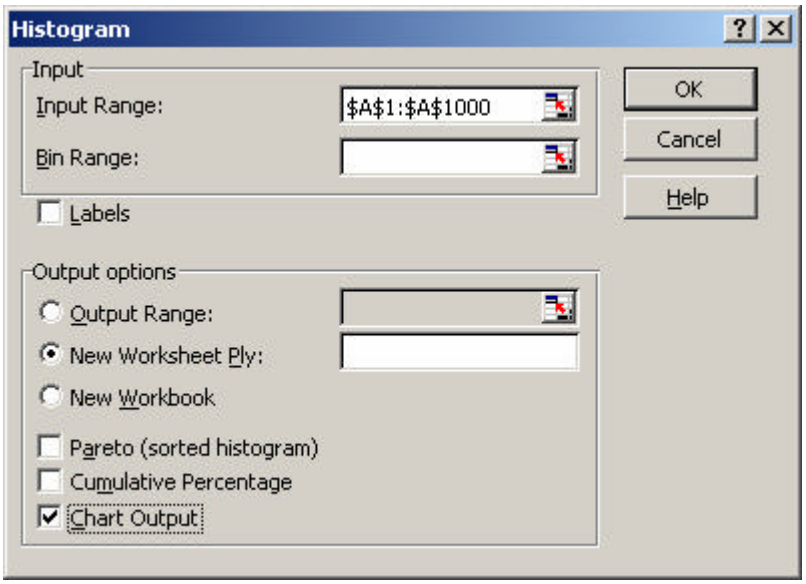
So as an example, let us try to obtain a normal variable having 1000 rows and check its normality by using a histogram. Therefore, we must specify 1 in the **Number of variables,** 1000 in the **Number of Random Numbers** and select **Normal** as the distribution. You notice that the form changes to



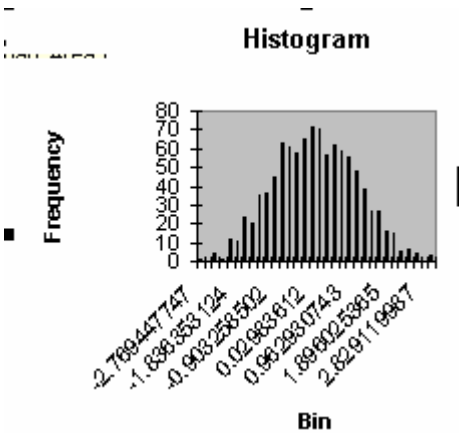
We can now enter the values for the mean and the standard deviation. If you leave these unaltered, you would have created a standard normal distribution. Press **OK**.

Excel executes the **Random Number Generation** macro according to your specifications.

To obtain a histogram, click on **Tools** ➢ **Data Analysis** ➢ **Histogram**. Fill in the options as shown.



The histogram obtained follows, which shows that **Excel** has in fact generated numbers from a standard normal distribution.



z-Test: Two Sample for Means

The ***z*-Test** tool performs a two-sample z-test for means with known variances. This tool is used to test hypotheses about the difference between two population means. Use the *z*-Test when both the two samples have a size greater than 30, and when you know the variances of the two populations.

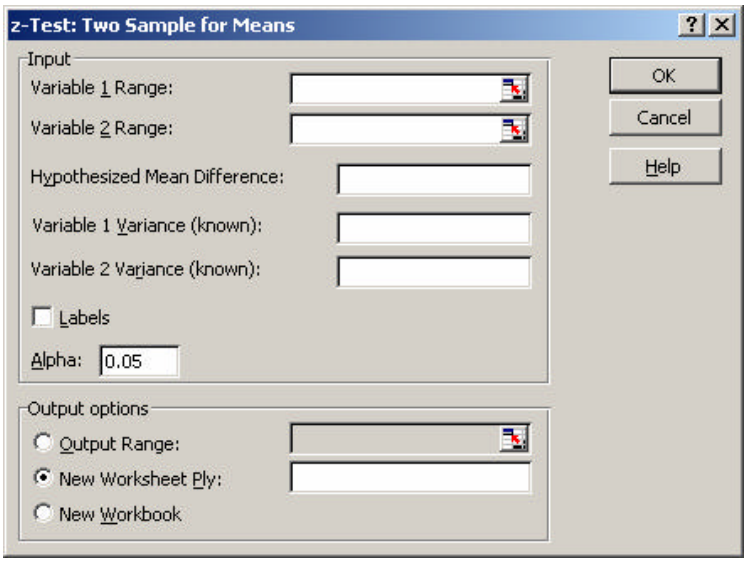
In hypothesis testing, we are always testing for the null hypothesis (H_0), against the alternative hypothesis (H_1). In this case

H_0 : the two means are equal

H_1 : the two means are not equal

To demonstrate the **z-Test**, we are going to use the data file **gss91t.xls**. In the data file there are two variables, **paeduc** and **maeduc** which give the number of years of education of the father and the mother respectively. It is also estimated that the variances of the variables are 1579 and 1160 respectively.

To use the ***z*-Test Analysis** tool, click on **Tools** ➢ **Data Analysis** ➢ **z-Test: Two Sample for Means**. **Excel** will display the following box

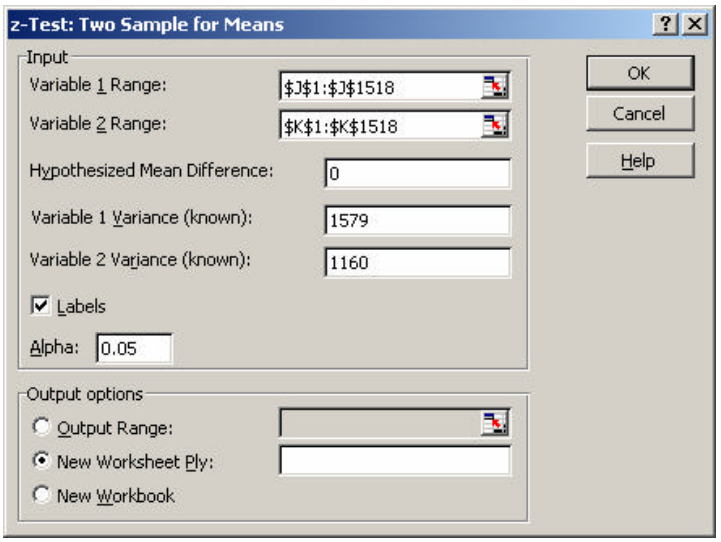


You can then make selections from the options provided. The following table describes the options in the ***z*-Test** dialog box:

Option	Description
Variable 1 Range	Enter the cell reference for the first range of data you want to analyse. The range must consist of a single column or row of data.
Variable 2 Range	Enter the cell reference for the second range of data you want to analyse. The range must consist of a single column or row of data.
Hypothesized Mean Difference	Enter the number you want for the shift in sample means. A value of 0 (zero) indicates that the sample means are hypothesized to be equal.
Variable 1 Variance (known)	Enter the known population variance for the Variable 1 input range.
Variable 2 Variance (known)	Enter the known population variance for the Variable 2 input range.

Alpha	Enter the confidence level for the test. This value must be in the range 0...1. The alpha level is a significance level related to the probability of having a type I error (rejecting a true hypothesis).

For our example, the form must be filled in the following way:



Note that common values for alpha are 0.05 (to indicate a confidence level of 95%) and 0.01 (to indicate a confidence level of 99%).

Click **OK** and **Excel** executes the ***z*-Test** macro according to your specifications.

The following output is obtained.

	A	B	C
1	z-Test: Two Sample for Means		
2			
3		PAEDUC	MAEDUC
4	Mean	36.49505603	27.124588
5	Known Variance	1579	1160
6	Observations	1517	1517
7	Hypothesized Mean Difference	0	
8	z	6.973616451	
9	P(Z<=z) one-tail	1.55442E-12	
10	z Critical one-tail	1.644853476	
11	P(Z<=z) two-tail	3.10885E-12	
12	z Critical two-tail	1.959962787	

The following table provides a brief description of the elements that Excel calculates for each row/column in the input range when you execute the ***z*-Test** macro:

Option	Description
Mean	The arithmetic average of both variables, arrived at by adding the values and dividing by the number of values.
Known Variance	The two variances inputted in the dialog form.
Observations	No of observations in both variables
Hypothesised Mean Difference	The difference that we are trying to test, inputted in the dialog form.
z	The *z*-score, worked out by a suitable formula $(\frac{\overline{x_1} - \overline{x_2} - m_{\overline{x_1} - \overline{x_2}}}{s_{\overline{x_1} - \overline{x_2}}})$
$P(z \le z)$ **one-tail**	This gives the one-tail probability (either less than or greater than) that the difference is equal to the hypothesised mean. If this probability is less than alpha, then we reject H_0 and accept H_1, if the probability is greater than alpha, we accept H_0.
z Critical one-tail	The Critical value (corresponding to the standard normal distribution) for a 1-tailed test.
$P(z \le z)$ **two-tail**	This gives the two-tail probability (both less than and greater than) that the difference is equal to the hypothesised mean.
z Critical two-tail	The Critical value (corresponding to the standard normal distribution) for a 2-tailed test. If we are testing a two-tailed hypothesis, then if *z* is greater than the critical point, we reject H_0 and accept H_1, if *z* is less than the critical point, we accept H_0.

The probability of both 1-tail and 2-tail is almost 0, indicating that the 2 samples have different means. In fact, from the data we can see that the number of years of education of the father is higher than that of the mother.

This might be correct, but if you look closely at the data, you will note that we did not remove the 98 and 99 values (the missing data). So this might be affecting our results. Therefore it is best to modify the data and then try to re-work the **z-test**.

F-test: Two Sample for Variances

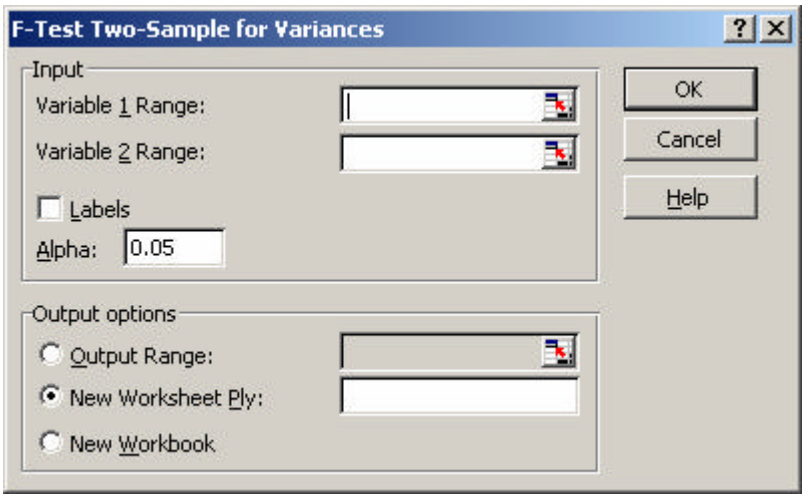
The ***F*-Test** tool performs a two-sample ***F*-test** to compare two population variances. This is useful when the two population variances are unknown, and therefore we cannot use the previous ***z*-test** for means. To be able to use the ***t*-test**, we need to determine whether the two populations have the same variance or not. In such a case, use the ***F*-test**. The ***F*-test** compares the ***F*-score** to the ***F* distribution**.

In this case, the null hypothesis (H_0) and the alternative hypothesis (H_1) are:

> H_0 : the two populations have the same variance
>
> H_1 : the two populations do not have the same variance

We will use the **statlaba.xls** file to work out this statistical analysis. In this data set, the children were weighed and measured (among other things) at the age of ten. We want to know whether there is any difference in the average heights of boys and girls (CTH) at this age. We do not know the variance and hence we cannot perform a **z-test**. Therefore we have to use a **t-test**. Before doing a ttest, we want to check which **t-test** we need, i.e. we have to perform an **F-test** to check on the variance of the 2 populations.

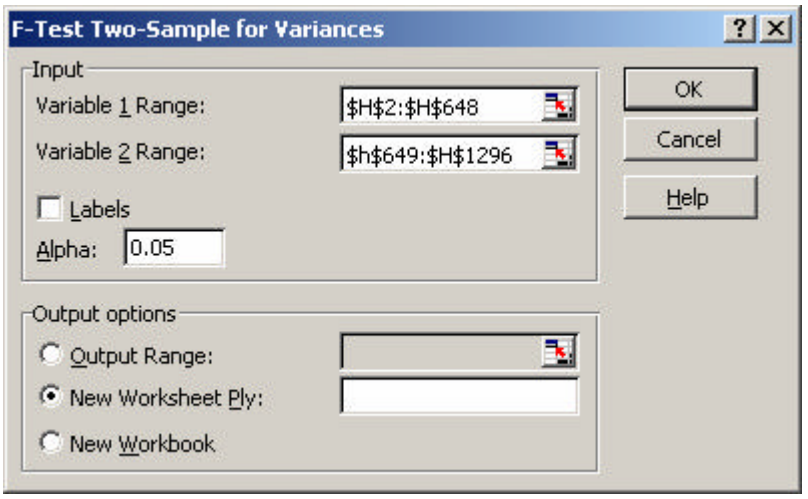
To use the ***F*-Test Analysis** tool, click on **Tools** ➢ **Data Analysis** ➢ **F-Test: Two Sample for Variances**. **Excel** will display the following box.



The selections are quite similar to previous forms in **Excel**. You need to move the 1st part of the column in **Variable 1 Range**, and the 2nd part in **Variable 2 Range**. Remember that in **Excel** we have no way how to use a grouping variable, so I would suggest that you 1st sort your data

according to the grouping variable (**SEX**) and then you open the above form. In this case the 1st range is from the 2nd cell to the 648th cell, while the 2nd range runs from the 649th cell to the 1296th cell.

The completed form is



Press **OK**. Excel executes the ***F*-Test** macro according to your specifications. The following output is obtained.

	A	B	C
1	F-Test Two-Sample for Variances		
2			
3		*Variable 1*	*Variable 2*
4	Mean	53.28686244	53.6441358
5	Variance	6.716251286	6.382469422
6	Observations	647	648
7	df	646	647
8	F	1.052296665	
9	P(F<=f) one-tail	0.258559031	
10	F Critical one-tail	1.138228622	

The following table provides a brief description of the elements that Excel calculates for each row/column in the input range when you execute the ***F*-Test** macro:

Option	Description
Mean	The arithmetic average of both variables, arrived at by adding the values and dividing by the number of values.
Variance	The two variances calculated from the data points.
Observations	No of observations in both variables
df	The degrees of freedom. This is calculated by (no of observations –1). This is a parameter used by the *F*-distribution.

F	The *F*-score. Note that the shape of the distribution only permits one-tailed analysis.
$P(F \le f)$ **one-tail**	This gives the one-tail probability that the two populations have the same variance. If this probability is less than alpha, then we reject H_0 and accept H_1, if the probability is greater than alpha, we accept H_0.
***F* Critical one-tail**	The Critical value (corresponding to the *F* distribution) for a 1-tailed test.

In this example, the probability is 0.25 which is greater than alpha (0.05), indicating that the two samples have the same variance, and therefore we have to use the t-test assuming equal variances.

t-Test: Two Sample Assuming Equal Variance

The ***t*-Test** tool performs a two-sample student's ***t*-test**. This ***t*-test** form assumes that the means of both data sets are equal; it is referred to as a **homoscedastic *t*-test**. Use this ***t*-test** to determine whether two sample means are equal. You can use this test after you have found out that the two populations have equal variances. If any one of your sample sizes is less than 30, or any one of the population variances is unknown, use this test rather than the z-test.

In this case, the null hypothesis (H_0) and the alternative hypothesis (H_1) are:

> H_0 : the two populations have the same mean (assuming pooled variance)
>
> H_1 : the two populations do not have the same mean (assuming pooled variance)

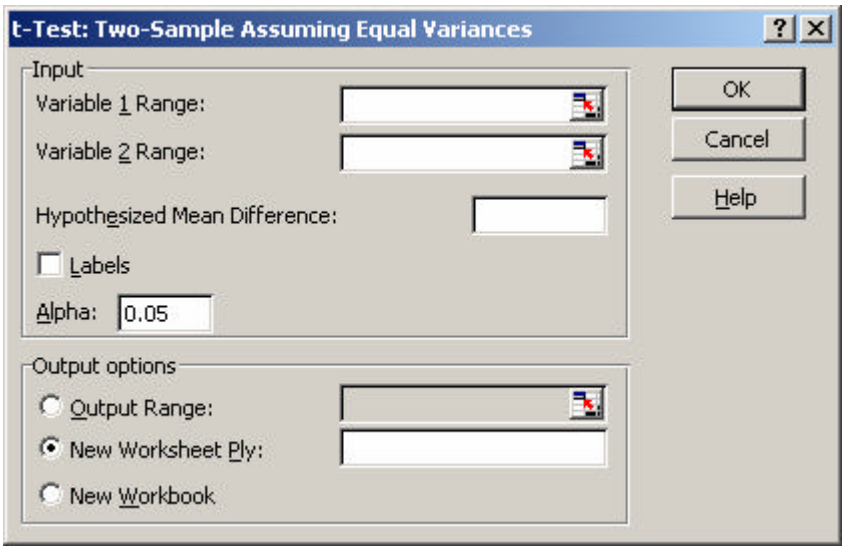
We will again use the **statlaba.xls** data set. In this data set, the children were weighed and measured (among other things) at the age of ten. We want to know whether there is any difference in the average heights of boys and girls at this age. We do this by performing a **T-Test**.

We start by stating our **Null Hypothesis**:

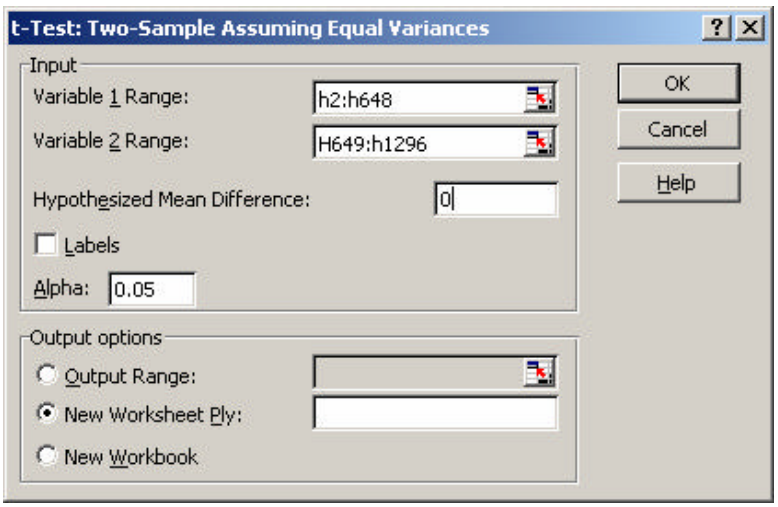
We assume there is no difference between boys and girls in terms of their height.

The **Alternative Hypothesis**, the one used if the Null Hypothesis is rejected, is therefore that there **is** some difference in the average heights of boys and girls at this age.

To perform the **T-Test**, click on **Tools** ➢ **Data Analysis** ➢ ***t*-Test: Two Sample Assuming Equal Variances**. **Excel** will display the following box.



As before, we have to sort the data before using this form since **Excel** does not have any grouping variable like other packages. If you are testing that the 2 means are the same, then remember to put a **0** in the **Hypothesized Mean Difference** box. The completed form should look like this.



Excel executes the ***t*-Test** macro according to your specifications. The following output is obtained. Note that the output is very similar to the output obtained whilst doing a ***z*-test**.

	A	B	C
1	t-Test: Two-Sample Assuming Equal Variances		
2			
3		*Variable 1*	*Variable 2*
4	Mean	53.28686244	53.6441358
5	Variance	6.716251286	6.382469422
6	Observations	647	648
7	Pooled Variance	6.549231281	
8	Hypothesized Mean Difference	0	
9	df	1293	
10	t Stat	-2.511945602	
11	P(T<=t) one-tail	0.006063891	
12	t Critical one-tail	1.64603307	
13	P(T<=t) two-tail	0.012127783	
14	t Critical two-tail	1.961798262	
15			

The following table provides a brief description of the elements that Excel calculates for each row/column in the input range when you execute the ***t*-Test** macro:

Option	Description
Mean	The arithmetic average of both variables, arrived at by adding the values and dividing by the number of values.
Variance	The two variances calculated from the data points.
Observations	No of observations in both variables
Pooled Variance	A common variance for the two populations.
df	The degrees of freedom. This is calculated by (total no of observations –2). This is a parameter used by the *t*-distribution.
***t* Stat**	The *t*-score. The *t*-distribution is very similar in shape to the normal distribution.
$P(T \le f)$ **one-tail**	This gives the one-tail probability that the two populations have the same means. If this probability is less than alpha, then we reject H_0 and accept H_1, if the probability is greater than alpha, we accept H_0.
***t* Critical one-tail**	The Critical value (corresponding to the *t* distribution) for a 1-tailed test.
$P(T \le f)$ **two-tail**	This gives the two-tail probability that the two populations have the same means. If this probability is less than alpha, then we reject H_0 and accept H_1, if the probability is greater than alpha, we accept H_0.
***t* Critical two-tail**	The Critical value (corresponding to the *t* distribution) for a 2-tailed test.

Note that if the ***t*-test** is used for samples with observations more than 30, and the variances are known, this will give the same results as a ***z*-test**.

The first part of the output gives some summary statistics; the numbers in each group, and the mean, standard deviation and standard error of the mean for the height.

Our Null Hypothesis says that there is no difference between the boys and girls in terms of their heights; in other words, we are testing whether their mean difference, (-0.357), is **significantly different from zero**. If it is, we must reject the Null Hypothesis, and instead take the Alternative.

Excel calculates the **t-value**, the **degrees of freedom** and the **Significance Level** (for both 1 tailed and 2 tailed) and displays them in the columns headed **t Stat**, **df** and P (T<=t) **(two-tail)**. Just as with other tests, we can make our decision quickly based on the displayed Significance Level.

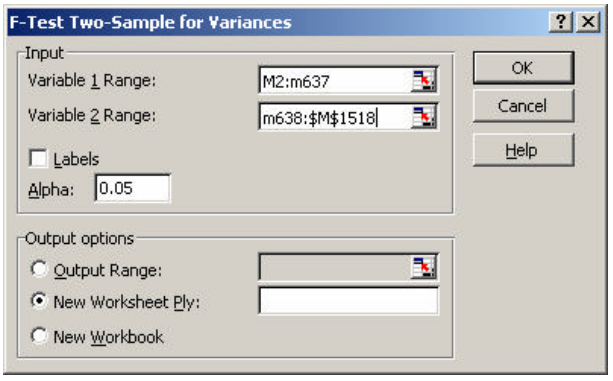
If the Significance Level is less than 0.05, we reject the Null Hypothesis and take the Alternative Hypothesis instead.

In this case, with a Significance Level of 0.012, we say that there is evidence, at the 5% level, to suggest that there **is** a difference between the heights of boys and girls at age ten (the Alternative Hypothesis).

t-Test: Two Sample Assuming Unequal Variance

Open **gss91t.xls**. We are going to look at the variable **PRESTG80** – 'R's Occupational Prestige Score (1980)', which has a scale of 0 to 100. We want to see whether gender affects this score. Before using the ***t*-test**, we need to make sure that the variances of the 2 samples are equal (pooled), otherwise we have to use the unequal variances ***t*-test**.

Click on **Tools ➢ Data Analysis ➢ F-Test: Two Sample for Variances**. If the ***F*-Test** returned a significant result, then we have to use the ***t*-test** with unequal variances. This ***t*-test** form assumes that the variances of both ranges of data are unequal; it is referred to as a ***heteroscedastic t*-test**.
The following is the completed form for the **F-test**.



This is the output obtained.

	A	B	C
1	F-Test Two-Sample for Variances		
2			
3		*Variable 1*	*Variable 2*
4	Mean	42.81918239	38.19182747
5	Variance	216.491663	303.5938422
6	Observations	636	881
7	df	635	880
8	F	0.713096357	
9	P(F<=f) one-tail	2.8826E-06	
10	F Critical one-tail	0.885246099	

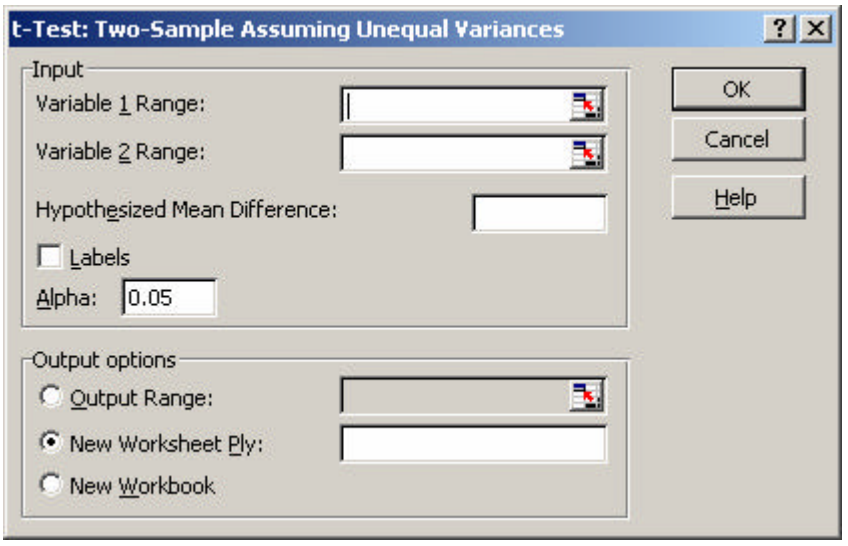
(Remember to sort your data set according to **Sex**)

You can see that the probability value obtained is almost 0, indicating that the 2 samples do not have the same variance. Therefore we cannot use the **t-test assuming equal variances**, but the **t-test assuming unequal variances**.

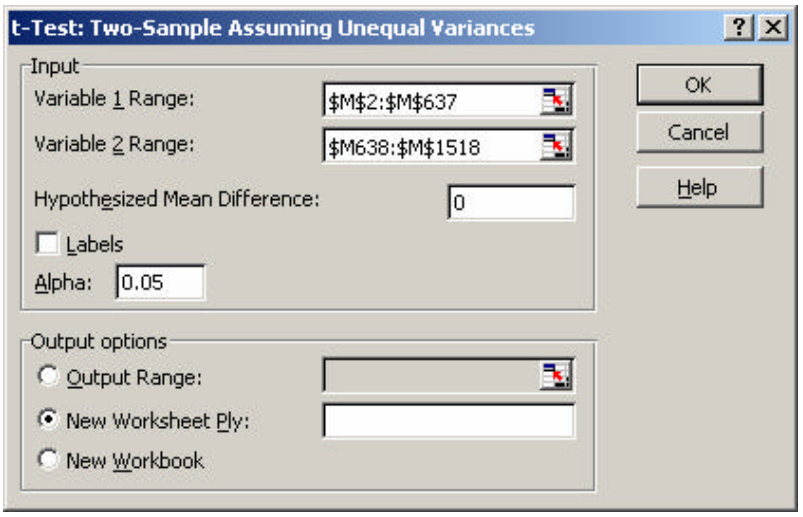
In this case, the null hypothesis (H_0) and the alternative hypothesis (H_1) are:

> H_0 : the two populations have the same mean (assuming unpooled variance)
>
> H_1 : the two populations do not have the same mean (assuming unpooled variance)

Click on **Tools ➢ Data Analysis ➢ t-Test: Two Sample Assuming Unequal Variances. Excel** displays the following dialog form.



This is identical to the other **t-test**, however **Excel** will use different calculations to work out the result. The completed form looks like this.



Click **OK**, and the following output is obtained.

t-Test: Two-Sample Assuming Unequal Variances		
	Variable 1	*Variable 2*
Mean	42.81918239	38.19182747
Variance	216.491663	303.5938422
Observations	636	881
Hypothesized Mean Difference	0	
df	1478	
t Stat	5.590986492	
P(T<=t) one-tail	1.34255E-08	
t Critical one-tail	1.645885277	
P(T<=t) two-tail	2.68511E-08	
t Critical two-tail	1.961570888	

The following table provides a brief description of the elements that Excel calculates for each row/column in the input range when you execute the ***t*-Test** macro:

Option	Description
Mean	The arithmetic average of both variables, arrived at by adding the values and dividing by the number of values.
Variance	The two variances calculated from the data points.
Observations	No of observations in both variables
Hypothesized Mean Difference	The test difference between the two data sets.

df	Worked out by a complex formula. This is a parameter used by the *t*-distribution.
***t* Stat**	The *t*-score. The *t*-distribution is very similar in shape to the normal distribution.
$P(T \le f)$ **one-tail**	This gives the one-tail probability that the two populations have the same means. If this probability is less than alpha, then we reject H_0 and accept H_1, if the probability is greater than alpha, we accept H_0.
***t* Critical one-tail**	The Critical value (corresponding to the *t* distribution) for a 1-tailed test.
$P(T \le f)$ **two-tail**	This gives the two-tail probability that the two populations have the same means. If this probability is less than alpha, then we reject H_0 and accept H_1, if the probability is greater than alpha, we accept H_0.
***t* Critical two-tail**	The Critical value (corresponding to the *t* distribution) for a 2-tailed test.

In this example, the probability is less than alpha (0.05) indicating that gender does affect the prestige score.

t-Test: Paired Two Sample for Mean Analysis

Imagine you want to compare two groups that are somehow **paired**; for example, husbands and wives, or mothers and daughters. Knowing about this pairing structure gives extra information, and you should take account of this when performing the T-Test.

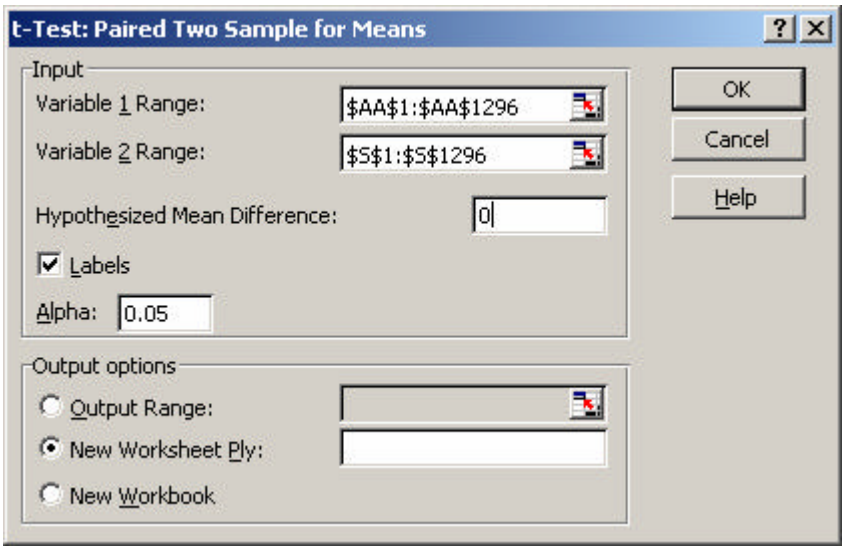
In the **statlaba.xls** data, we have the weights of the parents when their child was aged 10 in **FTW** and **MTW**. If we want to know if there is a difference between males and females in terms of weight, we can perform a **Paired Samples T-Test** on these two variables. Our **Null Hypothesis** is that there is no difference.

In this case, the null hypothesis (H_0) and the alternative hypothesis (H_1) are:

> H_0 : there is no change in the paired data
>
> H_1 : there is a change in the paired data

Click on **Tools** ➢ **Data Analysis** ➢ **t-Test: Paired Two-Sample For Means**.

Fill in the range for **FTW** as **Variable 1** and the range for **MTW** as **Variable 2**. The **Hypothesized Mean Difference** should again be set to **0**. The following is the completed form.



Excel executes the ***t*-Test** macro according to your specifications. The following is the output obtained.

	A	B	C
1	t-Test: Paired Two Sample for Means		
2			
3		*FTW*	*MTW*
4	Mean	177.2826255	143.1969112
5	Variance	678.9355171	787.3777339
6	Observations	1295	1295
7	Pearson Correlation	0.233860448	
8	Hypothesized Mean Difference	0	
9	df	1294	
10	t Stat	36.58124849	
11	P(T<=t) one-tail	4.6548E-202	
12	t Critical one-tail	1.646030796	
13	P(T<=t) two-tail	9.3096E-202	
14	t Critical two-tail	1.961798262	

The following table provides a brief description of the elements that Excel calculates for each row/column in the input range when you execute the ***t*-Test** macro:

Option	Description
Pearson Correlation	It is a measure of how the two variables are related. A relation close to 0 indicates almost no relation. We can see also that we have a negative relation.

As with the **Independent Samples T-Test**, we are first given some summary statistics; we are also given the **Correlation** between the two variables (we will deal with **Correlation** in a later session). We can that the difference between the weights of the males and females is -34.09 – is this significantly different from zero?

We use this table just as we did in the **Independent Samples T-Test**, and since the **P(T<=t) (two-tail)** column shows a value of less than 0.05, we can say that there is evidence, at the 5% level, to reject the **Null Hypothesis** that there is no difference between the mothers and fathers in terms of their weight.

Practical Session 7

1. Open a new data sheet. Create a new variable **normal** with 1000 cases. This variable should be normally distributed with mean 2 and standard deviation 1. Create another variable **uniform** that is uniformly distributed. Draw a histogram of the 2 variables to check their shape.

2. Using the data file **gss91t.xls**, modify the data set so that you remove the missing values from **paeduc** and **maeduc**. Obtain a *z*-test for the means using variances of 1579 and 1160 respectively. Check whether the parents have the same level of education.

Open **statlaba.xls**. At the age of ten, the children in the sample were given two tests; the **Peabody Picture Vocabulary Test** and the **Raven Progressive Matrices Test**. Their scores are stored in the variables **CTP** and **CTR**. Create a new variable called **TESTS** which is the sum of the two tests; this new variable will be used in Questions 3 and 4.

In each of the questions below, state your **Null** and **Alternative Hypotheses**, which of the two you accept on the evidence of the relevant test, and the **Significance Level**.

3. Use an **Independent Samples T-Test** to decide whether there is any difference between boys and girls in terms of their scores.

4. By pairing the parents of the child, decide whether there is any difference between fathers and mothers in terms of the heights. (Use **FTH** and **MTH**).