EXCEL in statistics

VIKOM INSTITUTE

One stop solution for EXCEL Learning

www.vikominstitute.com

ABOUT THE AUTHOR

VIKAL JAIN

He is the founder of **VIKOM INSTITUTE**. He is an associate member of the Institute of Chartered Accountants of India. He is pursuing CFA level 1 from USA and holds a diploma in IFRS (India) from ICAI. He has also done B.Com (H).

EXCEL SKILLS AND COMPUTER EFFICIENCY

- Wrote a book on “Excel Finance” and Excel Basic to advances
- **More than 14, 00, 000 likes on “EXCEL TRICKS” page on Facebook.**
- **More than 20,000 group members on Facebook**
- Knowledge of More than 250 formula’s in Excel and Macro’s in Excel
- VAST experience of MS Excel Open Office, 2003, 2007, 2010, 2013 and 2016
- Teaching Excel at **VIKOM INSTITUTE** and responding Excel queries across the world.
- Well Versed with MS – Word, PowerPoint, Microsoft Publishers, other Microsoft utilities, ERP Packages (Tally ERP9, Tally 9, 7.2 ,6.3), Web Surfing .
- Working experience of SAP

ABOUT THE **VIKOM INSTITUTE**

VIKOM INSTITUTE is **one stop solution of Excel Training**. It offers you standardized and customized templates, Online Excel training, Corporate Seminars in Excel, Excel E-books, free excel shortcuts files, daily tricks and so on. It is joint venture of various professionals who are experts in Finance, International Taxation, Software Development, Business Modeling, Financial Modeling, costing and so on.

MISSION

VIKOM has only one mission that is to increase effectiveness and efficiency of doing work. “**Removal of Redundancy”**, “**Doing smart work”**, “**Presenting better”**, “**Beyond copy & paste”** are the four main keys of the **VIKOM.**

DEDICATION

This short booklet is dedicated to my mother, my father and my sister who have motivated me, stime to time and my FACEBOOK PAGE likers who have increased my excel experience.

ABOUT THE BOOKLET

The purpose of this booklet is to improve your calculation efficiency in statistics not to clarify you what is statistics. I have put some source name while writing the book you can go to the link and check the source for more detail.

Important Information

- ➔ You CANNOT take the print of booklet
- ➔ You CANNOT copy the contents of the booklet
- ➔ This book considers "**TREES ARE THE GREEN GOLD"**. So keep E-Reading.

FEEDBACK

For any suggested improvements in course content and booklet email us at info@vikominstitute.com

Contents

Business Statistics

A definition from "WIKIPEDIA"

Business statistics is the science of good decision making in the face of uncertainty and is used in many disciplines such as financial analysis, econometrics, auditing, production and operations including services improvement, and marketing research.

These sources feature regular repetitive publication of series of data. This makes the topic of time series especially important for business statistics. It is also a branch of applied statistics working mostly on data collected as a by-product of doing business or by government agencies. It provides knowledge and skills to interpret and use statistical techniques in a variety of business applications.

A typical business statistics course is intended for business majors, and covers statistical study, descriptive statistics (collection, description, analysis, and summary of data), probability, and the binomial and normal distributions, test of hypotheses and confidence intervals, linear regression, and correlation.

Excel in Business Statistics

We all know Business Statistics has lots of handy calculation like calculation of mean, median, standard deviation, kurtosis, skewness and so on. Whenever we do this calculation then, this takes around 10 times of our time.

Excel helps to save our time; it does calculation in less than a minute. Excel has variety of ADD IN to perform the statistics. We have used variety of examples to improve your efficiency in Excel in the booklet.

Useful Definitions

Population

A population is defined as all members of a specified field group.

Sample

A sample is subset of a population.

Sample Statistic

A sample statistic is a quantity computed from a used to describe a sample.

Basic Statistics

SUM function

SUM function allows you to add two or more values in excel. If you want to add the values in excel then this function will help you.

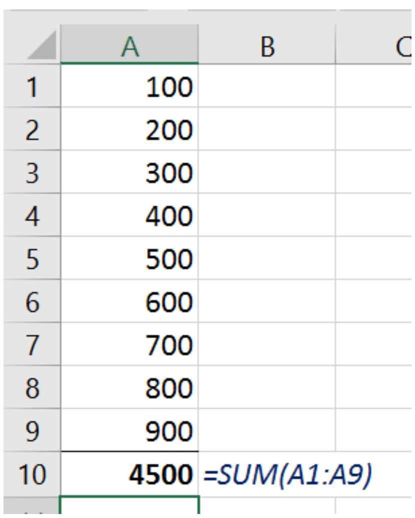
Mathematics

If you want to add 900+800+700+600+500+400+300+200+100 = 4500

Excel

Syntax: SUM(number1,[number2],...)

Let's try SUM function



	A	B	C
1	100		
2	200		
3	300		
4	400		
5	500		
6	600		
7	700		
8	800		
9	900		
10	**4500**	*=SUM(A1:A9)*	

See the above figure SUM function has automatically calculated the total of all values. You can try Alt + = to do the fast sum in excel.

COUNT function

COUNT function allows you to count the cells in a range in excel.

Mathematics

In math counting is so easy like 900, 800, 700, 600, 500, 400, 300, 200 and 100 are 9 numbers.

Excel

Syntax: COUNT(value1, [value2], ...)

Let's try COUNT function

	A	B	C
1	100		
2	200		
3	300		
4	400		
5	500		
6	600		
7	700		
8	800		
9	900		
10	**9**	*=COUNT(A1:A9)*	

See the above figure COUNT function has automatically counted the number of cells in a range.

MIN Function

MIN function in excel allows you to find the minimum value in range.

Mathematics

In math finding minimum value can be difficult if the data is huge in size. However it can calculated through observing the values if range. Like for example your range is 900, 800, 700, 600, 500, 300, 300, 200 and 100 so the minimum value is 100

Excel

Syntax: MIN(number1, [number2], ...)

Let's try MIN function

	A	B
1	100	
2	200	
3	300	
4	300	
5	500	
6	600	
7	700	
8	800	
9	900	
10	**100**	*=MIN(A1:A9)*

See the above figure MIN function has calculated the minimum value in less than a second. It can search the minimum values in any range, whether range has 100, 000 values, 10,000 values or any number of values.

MAX function

The function of MAX is as same as MIN function, However instead of finding lowest value it finds highest value in a range. Even the syntax of the MAX is also same.

Syntax: MAX(number1, [number2], ...)

	A	B	C
1	100		
2	200		
3	300		
4	400		
5	500		
6	600		
7	700		
8	800		
9	900		
10	**900**	*=MAX(A1:A9)*	

RANGE

Range is the difference between MAXIMUM value and the MINIMUM values of the range

Mathematics

In mathematics range can be found by this formula

MAXIMUM value in a range – MINIMUM value in a range

Excel

In excel you have to apply two formula to calculate the Range first is MAX function and second is the MIN function. Let’s calculate the range using Excel.

	A	B
1	100	
2	200	
3	300	
4	400	
5	500	
6	600	
7	700	
8	800	
9	900	
10	**100**	*=MIN(A1:A9)*

	A	B	C
1	100		
2	200		
3	300		
4	400		
5	500		
6	600		
7	700		
8	800		
9	900		
10	**900**	*=MAX(A1:A9)*	

In the above example we have taken the same range and calculate the Maximum value and the Minimum value of the range by applying MAX and MIN function.

Therefore the range in statistics would be 900- 100 = 800

MODE function

The most repetitive value in a range is called as MODE.

Mathematics

In math it can be calculate through observing the range suppose we have a range 100, 200, 300, 300, 500, 600, 700, 800, 900. So the MODE would be 300 since it frequency of repetition is two in the range.

Excel

Syntax : MODE(number1,[number2],...)

	A	B	C
1	100		
2	200		
3	300		
4	300		
5	500		
6	600		
7	700		
8	800		
9	900		
10	**300**	*=MODE(A1:A9)*	

See the calculation is too easy by using excel. Just apply MODE function and find the MODE.

AVERAGE function (Arithmetic Mean)

Arithmetic mean the simple average of the all values in a range.

Mathematics

In math the calculation of arithmetic mean is easy but sometimes it can take time when your data huge in size. Let' s take an example 100, 200, 300, 400, 500, 600, 700, 800, 900 so the arithmetic mean would be (100+200+300+400+500+600+700+800+900)/9 i.e.500

Excel

Syntax : Average(number1,[number2],...)

	A	B	C
1	100		
2	200		
3	300		
4	400		
5	500		
6	600		
7	700		
8	800		
9	900		
10	**500**	*=AVERAGE(A1:A9)*	

See the calculation is too easy by using excel. Just apply AVERAGE function and select you range it will calculate the arithmetic mean or simple average

MEDIAN function

The median is the value of the middle item of a set of items that has been sorted into ascending or descending order.

Mathematics

In an odd numbered sample of n items the median occupies (n+1)/2 position. In an even numbered sample we define the median as the mean of the Values of the means of the values of items occupying the n/2 and (n+2)/2 positions (the two middle items)

Example 1: Sample is 100,200,300,400,500,600,700,800,900 i.e 9 items or odd items

Solution: (9+1)/2 = 5^{th} item i.e 500

Example 2: Sample is 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000 i.e 10 items or even items

Solution: Median would be Mean of (10)/2 = 5^{th} item i.e 500 and (10+2)/2 i.e. 600. So the median is (500+600)/2 = 550

Excel

Syntax: Median(number1,[number2],...)

ODD items arranged in ascending order

	A	B	C
1	100		
2	200		
3	300		
4	300		
5	500		
6	600		
7	700		
8	800		
9	900		
10	**500**	*=MEDIAN(A1:A9)*	

EVEN items arranged in ascending order

	A	B	C
1	100		
2	200		
3	300		
4	400		
5	500		
6	600		
7	700		
8	800		
9	900		
10	1000		
11	**550**	*=MEDIAN(A1:A10)*	

DATA not arranged in ascending or descending order now median is

	A	B	C
1	400		
2	600		
3	800		
4	500		
5	100		
6	200		
7	700		
8	300		
9	900		
10	1000		
11	550	*=MEDIAN(A1:A10)*	

See in excel you don’t have to worry either about even or odd number or ascending or descending order. It can calculate median after applying all steps easily.

GEOMEAN function (Geometric Mean)

Geometric mean is the most frequently used to average rates of changes over time or to compute the growth rate of a variable

Mathematics

In math Geometric mean can be calculated with this formula

$$= \sqrt[n]{a_1 a_2 \cdots a_n}.$$

n = total number of the items

and a1, a2.... Is the items

Let's calculate the geometric mean of 100, 200, 300, 400, 500, 600, 700, 800, 900

(100*200*300*400*500*600*700*800*900)^(1/9)

417.717

Excel

Syntax: GEOMEAN(snumber1,[number2],...)

	A	B	C
1	400		
2	600		
3	800		
4	500		
5	100		
6	200		
7	700		
8	300		
9	900		
10	**414.717**	*=GEOMEAN(A1:A9)*	
11			

HARMEAN function (Harmonic Mean)

In mathematics, the harmonic mean (sometimes called the subcontrary mean) is one of several kinds of average, and in particular one of the Pythagorean means. Typically, it is appropriate for situations when the average of rates is desired.

Mathematics

The harmonic mean can be expressed as the reciprocal of the arithmetic mean of the reciprocals. As a simple example, the harmonic mean of 1, 2, and 4 is

$$\frac{3}{\frac{1}{1}+\frac{1}{2}+\frac{1}{4}} = \frac{1}{\frac{1}{3}(\frac{1}{1}+\frac{1}{2}+\frac{1}{4})} = \frac{12}{7}.$$

i.e 1.71429

Excel

Syntax: Harmean(number1,[number2],...)

	A	B	C
11			
12			
13	1		
14	2		
15	4		
16	**1.71429**	*=HARMEAN(A13:A15)*	
17			

STDEV function (Standard Deviation)

In statistics, the standard deviation (SD, also represented by the Greek letter sigma σ or s) is a measure that is used to quantify the amount of variation or dispersion of a set of data values

Mathematics

For a finite set of numbers, the standard deviation is found by taking the square root of the average of the squared deviations of the values from their average value. For example, the marks of a class of eight students (that is, a **population**) are the following eight values:

$$2,\ 4,\ 4,\ 4,\ 5,\ 5,\ 7,\ 9.$$

These eight data points have the mean (average) of 5:

$$\frac{2+4+4+4+5+5+7+9}{8} = 5.$$

First, calculate the deviations of each data point from the mean, and square the result of each:

$$\begin{array}{ll} (2-5)^2 = (-3)^2 = 9 & (5-5)^2 = 0^2 = 0 \\ (4-5)^2 = (-1)^2 = 1 & (5-5)^2 = 0^2 = 0 \\ (4-5)^2 = (-1)^2 = 1 & (7-5)^2 = 2^2 = 4 \\ (4-5)^2 = (-1)^2 = 1 & (9-5)^2 = 4^2 = 16. \end{array}$$

The variance is the mean of these values:

$$\frac{9+1+1+1+0+0+4+16}{8} = 4.$$

and the *population* standard deviation is equal to the square root of the variance:

$$\sqrt{4} = 2.$$

Excel

Syntax: STDEV(snumber1,[number2],...)

	A	B	C
19	2		
20	4		
21	4		
22	4		
23	5		
24	5		
25	7		
26	9		
27	**2.13809**	*=STDEV(A19:A26)*	
28			
29			

VAR function (VARIANCE)

Variance is defined as the average of the squared deviations around the mean.

Mathematics

Standard deviation is the square root of the Variance. So if we want to calculate the variance then we have to multiply standard deviation by standard deviation itself.

In above example VARINACE would be

=2*2 = 4

EXCEL

Syntax: VAR(snumber1,[number2],...)

	A	B	C
18			
19	2		
20	4		
21	4		
22	4		
23	5		
24	5		
25	7		
26	9		
27	**4.57143**	*=VAR(A19:A26)*	

STANDARD DEVIATION and VARIANCE of the POPULATION and SAMPLE

MATHS formula

When you calculate the variance and standard deviation of Population you take n as divisor while calculating the sample variance and sample standard deviation of sample you must have take n-1

EXCEL formula

Sample Variance

Syntax: VAR.S(snumber1,[number2],...)

Population Variance

Syntax: VAR.P(snumber1,[number2],...)

Sample standard deviation

Syntax: STDEV.S(snumber1,[number2],...)

Population standard deviation

Syntax: STDEV.P(snumber1,[number2],...)

Coefficient of Variation

CV is the ratio of the standard deviation of a set of observations to their mean value.

Mathematics

$$\widehat{c}_v = \frac{s}{\bar{x}}$$

Where CV = Coefficient of Variation

S = Standard Deviation

$\bar{x}$ = Mean

EXCEL

STDEV function/ AVERAGE function

SKEW function (Skewness)

If a return distribution is symmetrical about its mean, then each side of the distribution is a mirror image of the other.

One of the most important distributions is the normal distribution. The symmetrical is, bell shaped distribution plays an important role in the mean – variance model of sample selection.

The normal distribution has the following characteristics

1. Its mean and median are equal
2. It is completely described by two parameters – its mean and variance.

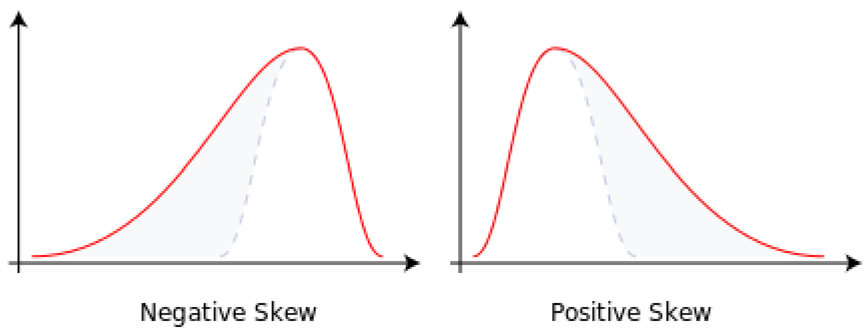
3. Roughly 68 % of its observation lie between plus and minus one standard deviation from the mean; 95% lie between plus and minus two standard deviations and 99 lie between plus and minus three standard deviations.

A distribution that is not symmetrical is called skewed.

Type of Skewness:-

A. Negative skew: The left tail is longer; the mass of the distribution is concentrated on the right of the figure. The distribution is said to be left-skewed, left-tailed, or skewed to the left.

B. Positive skew: The right tail is longer; the mass of the distribution is concentrated on the left of the figure. The distribution is said to be right-skewed, right-tailed, or skewed to the right.



Mathematics

When **n IS SMALL** then SKEWNESS can be calculated through this formula

$$S_K = \left[\frac{n}{(n-1)(n-2)}\right] \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}$$

When **n IS LARGE** then Skewness would be

$$S_K \approx \left(\frac{1}{n}\right) \frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}$$

EXCEL

Syntax: SKEW(snumber1,[number2],...)

	A	B	C
1	100		
2	200		
3	300		
4	300		
5	500		
6	600		
7	700		
8	800		
9	900		
10	**0.1019**	*=SKEW(A1:A9)*	

Analysis: In the above example SKEWNESS is 0.1019 then our curve would be positively skewed.

KURT function (Kurtosis)

KURTOSIS (from Greek: κυρτός, kyrtos or kurtos, meaning "curved, arching") is a measure of the "tailedness" of the probability distribution of a real-valued random variable

The kurtosis of any univariate normal distribution is 3. It is common to compare the kurtosis of a distribution to this value. Distribution with kurtosis **less than 3** are said to be **PLATYKURTIC**. An example of a platykurtic distribution is the uniform distribution, which does not have positive-valued tails. Distributions with kurtosis **greater than 3** are said to be **LEPTOKURTIC**. An example of a leptokurtic distribution is the Laplace distribution, which has tails that asymptotically approach zero more slowly than a Gaussian.

Mathematics

When **n IS SMALL** then KURTOSIS would be

$$K_E = \left(\frac{n(n+1)}{(n-1)(n-2)(n-3)} \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{s^4} \right) - \frac{3(n-1)^2}{(n-2)(n-3)}$$

When **n IS LARGE** then KURTOSIS would be

$$= \frac{1}{n} \frac{\sum (X - \bar{X})^4}{s^4} - 3$$

EXCEL

	A	B
1	100	
2	200	
3	300	
4	300	
5	500	
6	600	
7	700	
8	800	
9	900	
10	**-1.43089**	*=KURT(A1:A9)*

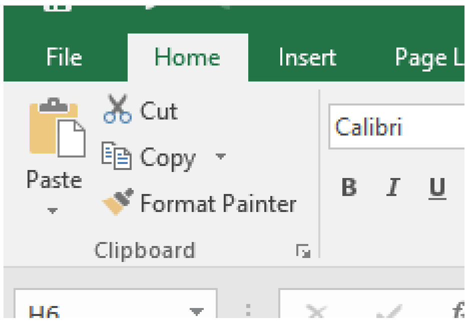
Analysis: Above calculation has shown that data has fairly less kurtosis

Descriptive Statistics

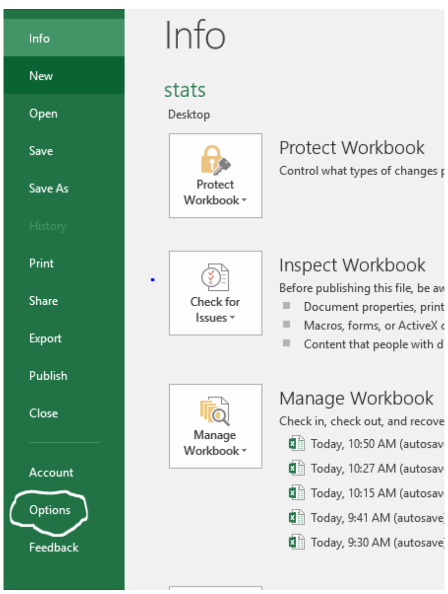
The Data Analysis ToolPak has a Descriptive Statistics tool that provides you with an easy way to calculate summary statistics for a set of sample data. Summary statistics includes Mean, Standard Error, Median, Mode, Standard Deviation, Variance, Kurtosis, Skewness, Range, Minimum, Maximum, Sum, and Count.

To work with Descriptive statistics we have follow these steps

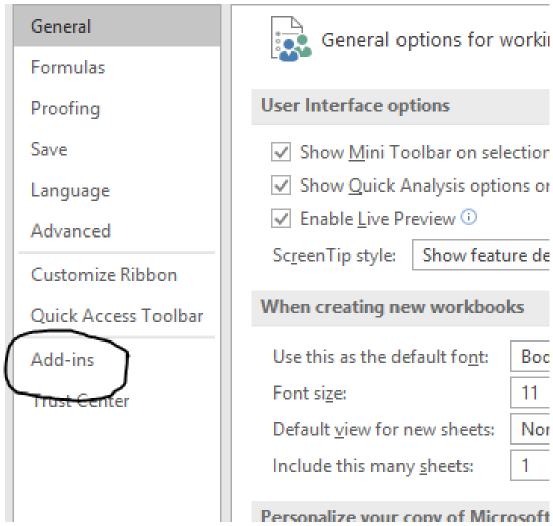
Step 1: Go to file menu



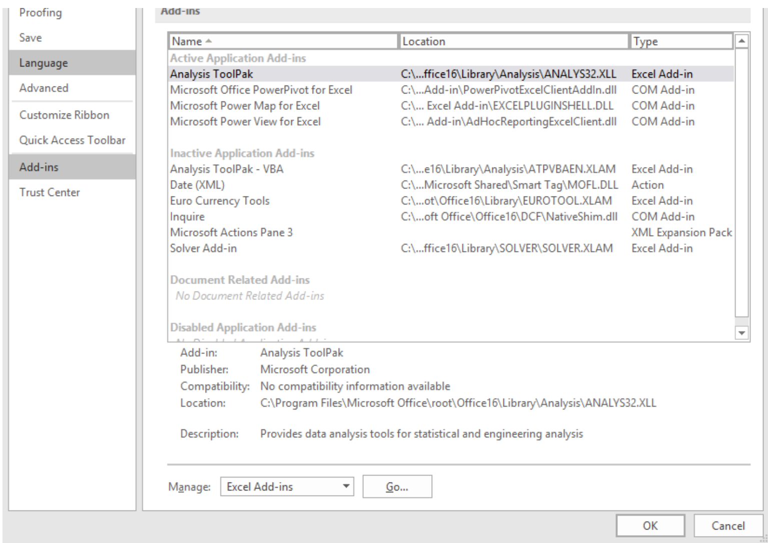
Step 2: Go to options



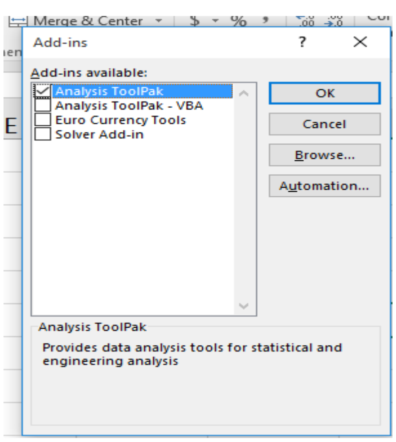
Step 3: Go to ADD IN



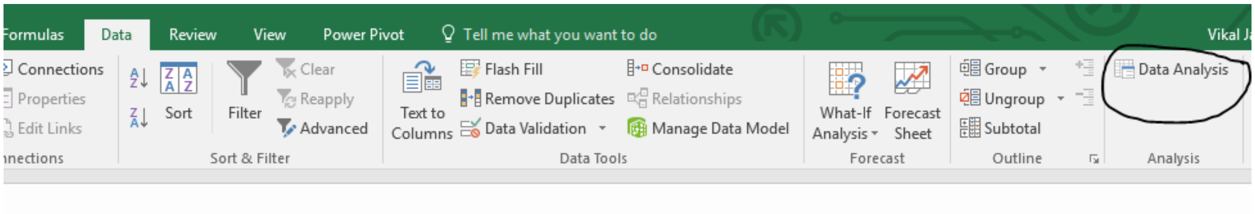
Step 4: Select Analysis ToolPak and press go



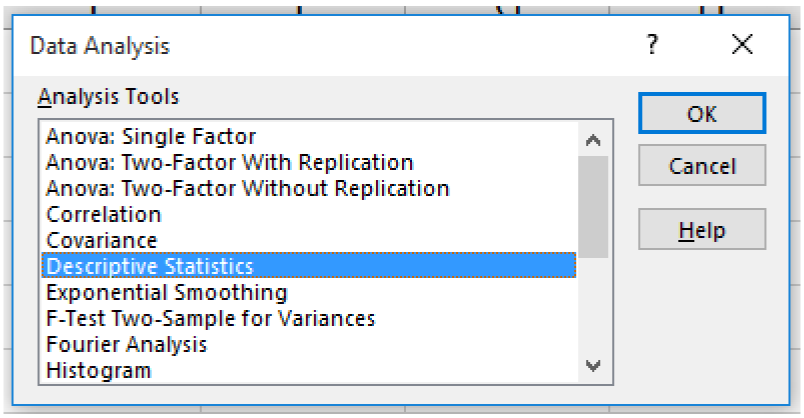
Step 5: Check the checkbox Analysis ToolPak



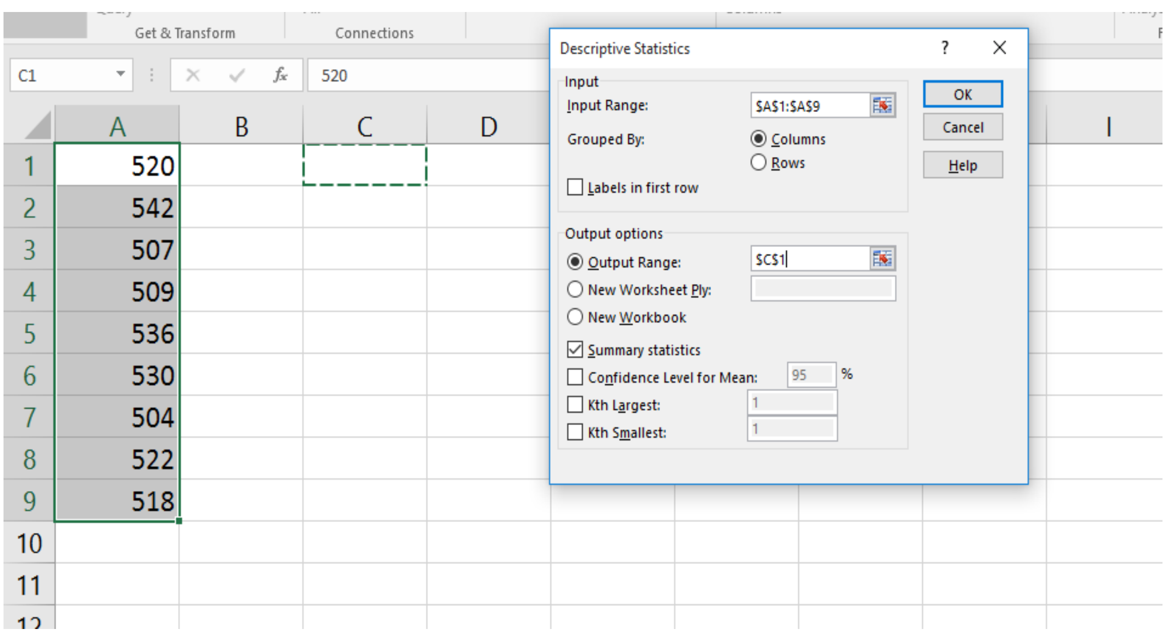
Step 6: Go to DATA menu and select DATA analysis



Step 7: Select Descriptive Statistics and press ok



Step 8: Now select the INPUT range , OUTPUT cell and summary statistics then press ok



Now your summary statistics table is ready. This table is showing error in MODE value since we don't have any mode in our range.

	A	B	C	D	E
1	520		*Column1*		
2	542				
3	507		Mean	520.8888889	
4	509		Standard Error	4.388888889	
5	536		Median	520	
6	530		Mode	#N/A	
7	504		Standard Deviation	13.16666667	
8	522		Sample Variance	173.3611111	
9	518		Kurtosis	-1.03955702	
10			Skewness	0.310876595	
11			Range	38	
12			Minimum	504	
13			Maximum	542	
14			Sum	4688	
15			Count	9	

Normal Distribution

The defining characteristics of a normal distribution are as follows:

- The normal distribution is completely described by two parameters—its mean, μ, and variance, σ2. We can also define a normal distribution in terms of the mean and the standard deviation, σ. As a consequence, we can answer any probability question about a normal random variable if we know its mean and variance (or standard deviation).
- The normal distribution has a skewness of 0 (it is symmetric). The normal distribution has a kurtosis (measure of peakedness) of 3; its excess kurtosis (kurtosis − 3.0) equals 0.17 As a consequence of symmetry, the mean, median, and the mode are all equal for a normal random variable.
- A linear combination of two or more normal random variables is also normally distributed.

Mathematics

To calculate Normal distribution we must have three things

1. MEAN
2. Standard deviation
3. Normal distribution table

Steps to calculate Normal Distribution

1. Calculate value of Z i.e. (X – Mean)/ σ
2. Calculate N(Z) using Normal distribution table

Consider the problem of finding the probability of getting less than a certain value under any normal probability distribution. As an illustrative example, let us suppose the SAT scores nationwide are normally distributed with a **mean and standard deviation of 500 and 100, respectively**. Answer the following questions based on the given information:

A. **What is the probability that a randomly selected student score will be less than 600 points?**
B. **What is the probability that a randomly selected student score will exceed 600 points?**
C. **What is the probability that a randomly selected student score will be between 400 and 600?**

Solution 1: Z = (600-500)/100 = 1

N(Z) = N(1) = 0.3413 (Probability from 600 to 500)

0.5 (probability from less than 500)

Therefore total probability of score less than 600 points is 0.3413 + 0.5 = 0.8413

EXCEL

NORM.DIST function

Syntax : NORM.DIST(x,mean,standard_dev,cumulative)

X	Required. The value for which you want the distribution.
Mean	Required. The arithmetic mean of the distribution.
Standard-dev	Required. The standard deviation of the distribution.
Cumulative	Required. A logical value that determines the form of the function. If cumulative is TRUE, NORMDIST returns the cumulative distribution function; if FALSE, it returns the probability mass function.

	A	B	C	D	E
1	Mean	500			
2	SD	100			
3	X	600			
4	Probaility	0.84134	=NORM.DIST(B3,B1,B2,TRUE)		
5					
6					

Solution 2: Using common sense we can answer part "b" by subtracting 0.84134474 from 1. So the part "b" answer is 1- 0.8413474 or 0.158653. This is the probability that a randomly selected student's score is greater than 600 points.

Solution 3: To answer part "c", use the same techniques to find the probabilities or area in the left sides of values 600 and 400. Since these areas or probabilities overlap each other to answer the question you should subtract the smaller probability from the larger probability. The answer equals 0.84134474 - 0.15865526 that is, 0.68269.

NORM.INV function

Syntax : NORM.INV(probability,mean,standard_dev)

Probability	Required. A probability corresponding to the normal distribution.
Mean	Required. The arithmetic mean of the distribution.
Standard-dev	Required. The standard deviation of the distribution.

	A	B	C	D
1	Mean	500		
2	SD	100		
3	Probaility	0.84134		
4	X	599.998	=NORM.INV(B3,B1,B2)	
5				
6				

Confidence Interval for the mean

Before understanding Confidence interval for the mean let's understand STANDARD ERROR

Standard Error

The standard error of the mean (SEM) is the standard deviation of the sample-mean's estimate of a population mean. (It can also be viewed as the standard deviation of the error in the sample mean with respect to the true mean, since the sample mean is an unbiased estimator.) SEM is usually estimated by the sample estimate of the population standard deviation (sample standard deviation) divided by the square root of the sample size (assuming statistical independence of the values in the sample):

Mathematics

$$\mathrm{SE}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where

s is the sample standard deviation (i.e., the sample-based estimate of the standard deviation of the population), and

n is the size (number of observations) of the sample.

EXCEL

For calculating standard error in excel we have to use mathematics through combination of the formulas

	A	B	C	D	E	F
1	528					
2	571					
3	599					
4	512					
5	558					
6	557					
7	518					
8	588					
9	558					
10	519					
11	9.64342	*=(STDEV.S(A1:A10))/(SQRT(COUNT(A1:A10)))*				

Confidence Interval for the mean

In statistics, a confidence interval (CI) is a type of interval estimate of a population parameter. It is an observed interval (i.e., it is calculated from the observations), in principle different from sample to sample, that frequently includes the value of an unobservable parameter of interest if the experiment is repeated. How frequently the observed interval contains the parameter is determined by the confidence level or confidence coefficient

Mathematics

$$\bar{x} \pm Z \bullet (S/\sqrt{n})$$

Where $\bar{x}$ = mean of the sample

Z = Interval coefficient

S = Standard Deviation of the sample

N = sample size

$(S/\sqrt{n})$ is also known as standard error.

Example: -

Sample Size = 100

$\bar{x}$ = 0.45

Confidence level = 90%

Standard deviation = 0.30

$$\overline{X} \pm z_{0.05}\frac{s}{\sqrt{n}} = 0.45 \pm 1.65\frac{0.30}{\sqrt{100}} = 0.45 \pm 1.65(0.03) = 0.45 \pm 0.0495$$

Therefore confidence interval spans is 0.45-0.0495 = 0.4005 to 0.45 + 0.0495 = 0.4995

EXCEL

Syntax : CONFIDENCE(alpha,standard_dev,size)

Alpha	Required. The significance level used to compute the confidence level. The confidence level equals 100*(1 - alpha)%, or in other words, an alpha of 0.05 indicates a 95 percent confidence level.
Standard_Dev	Required. The population standard deviation for the data range and is assumed to be known.
Size	Required the sample size

	A	B	C
1	Sample Size	100	
2	$\bar{x}$	0.45	
3	Confidence Level	90%	
4	SD	0.3	
5		0.049345609	=CONFIDENCE.NORM(1-B3,B4,B1)
6			
7	Confidence Interval spans		
8	Upper Level	0.400654391	=B2-B5
9	Lower Level	0.499345609	=B2+B5
10			
11			

Using Analysis ToolPak

Step 1. Enter data in cells A1 to A10 (on the spreadsheet)

Step 2. From the menus select Tools

Step 3. Click on Data Analysis then choose the Descriptive Statistics option then click OK.

On the descriptive statistics dialog, click on Summary Statistic. After you have done that, click on the confidence interval level and type 95% - or in other problems whatever confidence interval you desire. In the Output Range box enter B1 or whatever location you desire.

Now click on OK. The screen shot would look like the following:

	A	B	C
1	523	*Column1*	
2	571		
3	548	Mean	541.7
4	538	Standard Error	6.545651652
5	551	Median	545.5
6	501	Mode	#N/A
7	543	Standard Deviation	20.69916799
8	564	Sample Variance	428.4555556
9	552	Kurtosis	0.335291702
10	526	Skewness	-0.618378835
11		Range	70
12		Minimum	501
13		Maximum	571
14		Sum	5417
15		Count	10
16		Confidence Level(95.0%)	14.80729277
17			

So the confidence level spans would be (541.7 – 14.8079277) = 526.8927 to (541.7 + 14.8079277) =556.5073

Test of Hypothesis Concerning the population mean

Hypothesis

A statistical hypothesis is a statement about the distribution of X. It is a claim about the statement about a statement whether the statement is true or not.

Hypothesis Testing

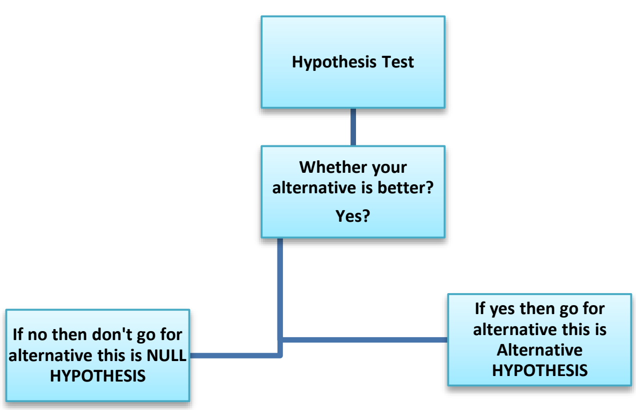
A hypothesis test is a statistical test that is used to determine whether there is enough evidence in a sample of data to infer that a certain condition is true for the entire population.

Type of Hypothesis Testing

A hypothesis test examines two opposing hypotheses about a population: the **NULL HYPOTHESIS** and the **ALTERNATIVE HYPOTHESIS**. The null hypothesis is the statement being tested. Usually the null hypothesis is a statement of "no effect" or "no difference". The alternative hypothesis is the statement you want to be able to conclude is true.

Examples of questions you can answer with a hypothesis test include:

- Does the mean height of undergraduate women differ from 66 inches?
    - If height is differ then your alternative is better therefore go for alternative otherwise don't go. If you will go with alternative then it will alternative hypothesis otherwise it will be null hypothesis.
- Is the standard deviation of their height equal less than 5 inches?
- Do male and female undergraduates differ in height?



Steps involved in Hypothesis

Source http://support.minitab.com/en-us/minitab/17/topic-library/basic-statistics-and-graphs/hypothesis-tests/basics/example-of-a-hypothesis-test/#

[Example of performing a basic hypothesis test

You can follow **SIX** basic steps to correctly set up and perform a hypothesis test. For example, the manager of a pipe manufacturing facility must ensure that the diameters of its pipes equal 5cm. The manager follows the basic steps for doing a hypothesis test.

Specify the hypotheses.

1. **First**, the manager formulates the hypotheses. The null hypothesis is: The population mean of all the pipes is equal to 5 cm. Formally, this is written as: $H0: \mu = 5$

   Then, the manager chooses from the following alternative hypotheses:

Condition to test	Alternative Hypothesis
The population mean is less than the target.	one sided: $\mu < 5$
The population mean is greater than the target.	one sided: $\mu > 5$
The population mean differs from the target.	two sided: $\mu \neq 5$

   Because they need to ensure that the pipes are not larger or smaller than 5 cm, the manager chooses the two-sided alternative hypothesis, which states that the population mean of all the pipes is not equal to 5 cm. formally, this is written as $H1: \mu \neq 5$

2. Determine the power and sample size for the test.

   The manager uses a power and sample size calculation to determine how many pipes they need to measure to have a good chance of detecting a difference of 0.1 cm or more from the target diameter.

3. Choose a significance level (also called alpha or α).

   The manager selects a significance level 0.05, which is the most commonly used significance level.

4. Collect the data.

   They collect a sample of pipes and measure their diameters.

5. Compare the p-value from the test to the significance level.

After they perform the hypothesis test, the manager obtains a p-value of 0.004. The p-value is less than the significance level of 0.05.

6. Decide whether to reject or fail to reject the null hypothesis.

   The manager rejects the null hypothesis and concludes that the mean pipe diameter of all pipes is not equal to 5cm.]

	Condition of null hypothesis	
Possible action	**True**	**False**
Fail to reject H_0 why fail to reject since our alternativ is not better	**Correct** **(1-α)**	**Type II error** **β**
Reject H_0 why accept alternative hypothesis since our alternative is better	**Type I error** **α**	**Correct** **(1-β)**

TYPE I ERROR – when based on our sample we reject a true null hypothesis

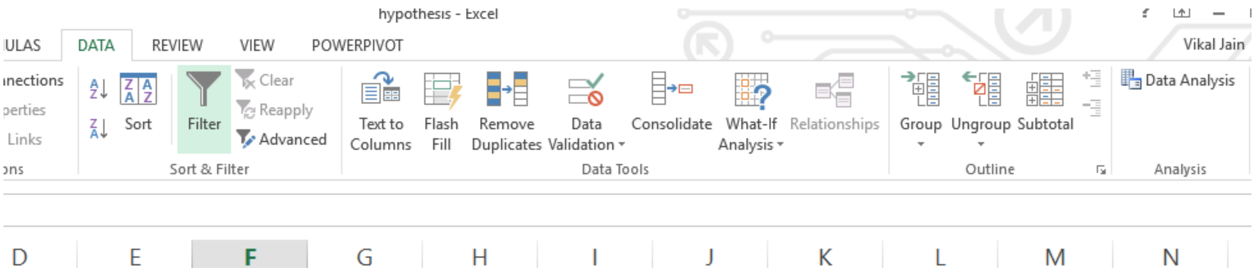
TYPE II ERROR – when based on our sample we cannot reject a true null hypothesis

Hypothesis Testing in Excel

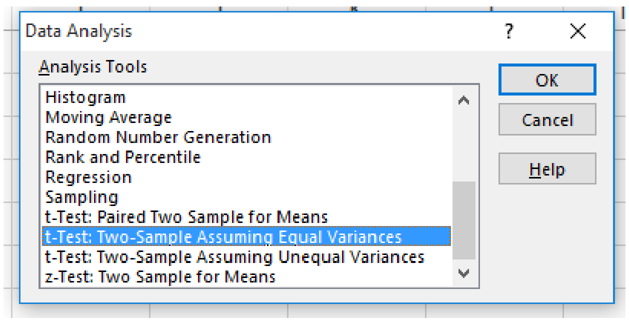
The first step is to get the data for the two groups you want organized so that the observations from one group are together and the observations for the second group are together. They don't have to be in separate columns, but this is one way to do it. They would just be sorted in one column so that the observations from the first group are all together and so that the observations from the second group are all together.

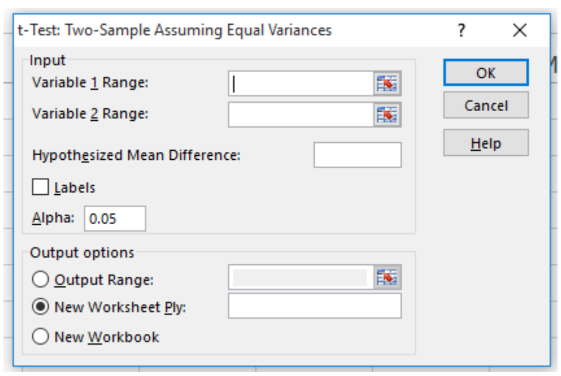
	A	B	C
1	A	B	
2	10.5795	6.57982	
3	12.6463	1.33274	
4	9.49893	3.88998	
5	9.62545	8.14074	
6	10.7885	13.5844	
7	2.50968	6.98592	
8	9.6693	0.77645	
9	1.589	8.98073	
10	1.25627	5.3386	
11	11.5408	2.55416	
12	13.6196	6.90455	
13	9.08767	4.07868	
14	3.97529	1.46596	
15	9.60774	14.8351	
16	7.49405	8.35698	

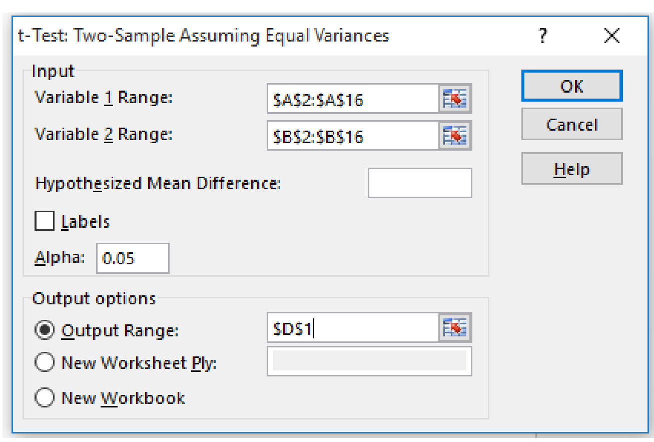
Now go to DATA analysis



Then, complete the t-Test box by highlighting the appropriate areas, indicating whether labels are included and specifying an output range.







The Hypothesized Mean Difference refers to the difference that you are testing for between the mean for the first group and the mean for the second group. If, as would usually be the case, you are testing whether or not the two groups have equal means, the Hypothesized Mean difference would be zero.

Also, the window asks for Alpha, or the probability of a Type 1 error. This is used to generate critical values in the output and isn't all that important as you will also be getting p-values with your output.

The output looks like this

t-Test: Two-Sample Assuming Equal Variances		
	Variable 1	*Variable 2*
Mean	8.232542298	6.253654271
Variance	15.99600692	17.56051156
Observations	15	15
Pooled Variance	16.77825924	
Hypothesized Mean Difference	0	
df	28	
t Stat	1.323056535	
P(T<=t) one-tail	0.098260094	
t Critical one-tail	1.701130934	
P(T<=t) two-tail	0.196520188	
t Critical two-tail	2.048407142	

The mean, variance and number of observations are given for each group.

The t-Stat is given.

The p-values and critical values are given for one and two tailed tests. If these are small, the means are significantly different. That is, you can reject the null hypothesis of equality in favor of the alternative hypothesis that:

1. the means are unequal, in the case of a two-tailed test

OR

2. the mean of B is greater than the mean of A, in the case of a one-tailed test.

Note that the both conditions are not satisfied because the sample mean for A was greater than the sample mean for B. Given the data above, you would reject the one sided null hypothesis that the mean of A was greater than or not equal to the mean of A.

Here’s another example where the null hypotheses would be rejected:

In this version, the mean are not so close that the null hypothesis that the mean for A is not equal to the mean for B can be rejected, as indicated by the p-values of 0.098260094 and 0.196520188 for the one and two-tailed tests, respectively.

HERE POINT TO BE NOTED IS T TEST CAN BE PEROFMED ONLY AND ONLY IF SIGMA (MEAN OF POPULATION) IS NOT GIVEN

BUT if sigma is given then Z TEST can be performed

ONLY basic of part of hypothesis so that we can link hypothesis to EXCEL.

ANOAVA ANAYLSIS (Analysis of Variances)

Source http://home.ubalt.edu/ntsbarsh/excel/excel.htm

[In this section the objective is to see whether or not means of three or more populations based on random samples taken from populations are equal or not. Assuming independents samples are taken from normally distributed populations with equal variances, Excel would do this analysis if you choose one way anova from the menus. We can also choose Anova: two way factor with or without replication option and see whether there is significant difference between means when different factors are involved.

Single-Factor ANOVA Test

In this case we were interested to see whether there a significant difference among hourly wages of student assistants in three different service departments here at the University of Baltimore. Six student assistants were randomly were selected from the three departments and their hourly wages were recorded as following:

ARC	CSI	TCC
10.00	6.50	9.00
8.00	7.00	7.00
7.50	7.00	7.00
8.00	7.50	7.00
7.00	7.00	6.50

Enter data in an Excel work sheet starting with cell A2 and ending with cell C8. The following steps should be taken to find the proper output for interpretation.

Step 1. From the menus select Tools and click on Data Analysis option.

Step 2. When data analysis dialog appears, choose Anova single-factor option; enter A2:C8 in the input range box. Select labels in first row.

Step3. Select any cell as output(in here we selected A11). Click OK.

The general form of ANOVA table looks like following:

Source of Variation	SS	Df	MS	F	P-value	F crit
Between Groups	SSTR	K-1	MSTR	MST/MSE	0.046725	3.682316674
Within Groups	SSE	n_t-K	MSE			
Total						

Suppose the test is done at level of significance a = 0.05, we reject the null hypothesis. This means there is a significant difference between means of hourly incomes of student assistants in these departments.

The Two-way ANOVA without Replication

In this section, the study involves six students who were offered different hourly wages in three different department services here at the University of Baltimore. The objective is to see whether the hourly incomes are the same. Therefore, we can consider the following:

Factor: Department

Treatment: Hourly payments in the three departments

Blocks: Each student is a block since each student has worked in the three different departments

Student	ARC	CSI	TCC
1	10.00	7.50	7.00
2	8.00	7.00	6.00
3	7.00	6.00	6.00
4	8.00	6.50	6.50
5	9.00	8.00	7.00
6	8.00	8.00	6.00

The general form of ANOVA table would look like:

Source of Variation	Sum of Squares	Degrees of freedom	Mean Squares	F
Treatment	SST	K-1	MST	F=MST/MSE
Blocks	SSB	b-1	MSB	
Error	SSE	(K-1)(b-1)	MSB	
Total	SST	nt-1		

To find the Excel output for the above data the following steps can be taken:

Step 1. From the menus select Tools and click on Data Analysis option.

Step2. When data analysis box appears: select Anova two-factor without replication then Enter A2: D8 in the input range. Select labels in first row.

Step3. Select an output range (in here we selected A11) then OK.

SUMMARY	COUNT	SUM	AVERAGE	VARIANCE
1	3	24.5	8.166667	2.583333
2	3	21	7	1
3	3	19.5	6.5	0.25
4	3	21.5	7.166667	0.583333
5	3	23	7.666667	2.333333
6	3	22	7.333333	1.333333
ARC	6	50	8.333333	1.066667
CSI	6	43	7.166667	0.666667

TCC	6	38.5	6.416667	0.241667

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Rows	4.902778	5	0.980556	1.972067	0.168792	3.325837
Columns	11.19444	2	5.597222	11.25698	0.002752	4.102816
Error	4.972222	10	0.497222			
Total	21.06944	17				

NOTE: F=MST/MSE =0.980556/0.497222 = 1.972067
F = 3.33 from table (5 numerator DF and 10 denominator DF)
Since 1.972067<3.33 we fail to reject the null.

Conclusion: There is not sufficient evidence to conclude that hourly rates differ for the three departments.

Two-Way ANOVA with Replication

Referring to the student assistant and the work study hourly wages here at the university of Baltimore the following data shows the hourly wages for the two categories in three different departments:

	ARC	CSI	TCC
	6.50	6.10	6.90
Work Study	6.80	6.00	7.20
	7.10	6.50	7.10
	7.40	6.80	7.50
Student Assistant	7.50	7.00	7.00
	8.00	6.60	7.10
Factors			

Factor A: Student job category (in here two different job categories exists)

Factor B: Departments (in here we have three departments)

Replication: The number of students in each experimental condition. In this case there are three replications.

Interaction:

	A	B	C	D	E
1		ARC	CSI	TCC	
2	Work Study	6.5	6.1	6.9	
3		6.8	6	7.2	
4		7.1	6.5	7.1	
5	Student Assistant	7.4	6.8	7.5	
6		7.5	7	7	
7		8	6.6	7.1	
8					

Anova: Two-Factor With Replication

SUMMARY	CSI	TCC	Total
6.5			
Count	3	3	6
Sum	18.6	21.2	39.8
Average	6.2	7.066666667	6.633333333
Variance	0.07	0.023333333	0.262666667
7.4			
Count	3	3	6
Sum	20.4	21.6	42
Average	6.8	7.2	7
Variance	0.04	0.07	0.092
Total			
Count	6	6	
Sum	39	42.8	
Average	6.5	7.133333333	
Variance	0.152	0.042666667	

Conclusion:
Mean hourly income differ by job category.
Mean hourly income differ by department.
Interaction is not significant.

Goodness-of-Fit Test for Discrete Random Variables

The **CHI-SQUARE** distribution can be used in a hypothesis test involving a population variance. However, in this section we would like to test and see how close a sample results are to the expected results.

Example: The Multinomial Random Variable

In this example the objective is to see whether or not based on a randomly selected sample information the standards set for a population is met. There are so many practical examples that can be used in this situation. For example it is assumed the guidelines for hiring people with different ethnic background for the US government is set at 70% (WHITE), 20%(African American) and 10%(others), respectively. A randomly selected sample of 1000 US employees shows the following results that is summarized in a table.

ETHNIC BACKGROUND	EXPECTED NUMBER OF EMPLOYEES	OBSERVED FROM SAMPLE
WHITE	700 =70%OF 1000	750
AFRICAN American	200 =20%OF 1000	170
OTHERS	100 =10%OF 1000	80

As you see the observed sample numbers for groups two and three are lower than their expected values unlike group one which has a higher expected value. Is this a clear sign of discrimination with respect to ethnic background? Well depends on how much lower the expected values are. The lower amount might not statistically be significant. To see whether these differences are significant we can use Excel and find the value of the CHI-SQUARE. If this value falls within the acceptance region we can assume that the guidelines are met otherwise they are not. Now lets enter these numbers into Excel spread-sheet. We used cells B7-B9 for the expected proportions, C7-C9 for the observed values and D7-D9 for the expected frequency. To calculate the expected frequency for a category, you can multiply the proportion of that category by the sample size (in here 1000). The formula for the first cell of the expected value column, D7 is 1000*B7. To find other entries in the expected value column, use the copy and the paste menu as shown in the following picture. These are important values for the chi-square test. The observed range in this case is C7: C9 while the expected range is D7: D9. The null and the alternative hypothesis for this test are as follows:

H_0 : $P_W = 0.70$, $P_A=0.20$ and $P_O=0.10$

H_A: The population proportions are not $P_W = 0.70$, $P_A= 0.20$ and $P_O = 0.10$

Now lets use Excel to calculate the p-value in a **CHI-SQUARE** test. **Step 1.**Select a cell in the work sheet, the location which you like the p value of the **CHI-SQUARE** to appear. We chose cell D12.

Step 2. From the menus, select **insert** then click on the **Function** option, **Paste Function** dialog box appears.

Step 3. Refer to function category box and choose **statistical**, from function name box select **CHITEST** and click on **OK**.

Step 4. When the **CHITEST** dialog appears: Enter C7: C9 in the **actual-range** box then enter D7: D9 in the **expected-range** box, and finally click on **OK**.

The p-value will appear in the selected cell, D12.

As you see the p value is 0.002392 which is less than the value of the level of significance (in this case the level of significance, a= 0.10). Hence the null hypothesis should be rejected. This means based on the sample information the guidelines are not met. Notice if you type "=CHITEST(C7:C9,D7:D9)" in the formula bar the p-value will show up in the designated cell.

NOTE: Excel can actually find the value of the CHI-SQUARE. To find this value first select an empty cell on the spread sheet then in the formula bar type "=CHIINV(D12,2)." D12 designates the p-Value found previously and 2 is the degrees of freedom (number of rows minus one). The CHI-SQUARE value in this case is 12.07121. If we refer to the CHI-SQUARE table we will see that the cut off is 4.60517 since 12.07121>4.60517 we reject the null. The following screen shot shows you how to the CHI-SQUARE value.

Test of Independence: Contingency Tables

The CHI-SQUARE distribution is also used to test and see whether two variables are independent or not. For example based on sample data you might want to see whether smoking and gender are independent events for a certain population. The variables of interest in this case are smoking and the gender of an individual. Another example in this situation could involve the age range of an individual and his or her smoking habit. Similar to case one data may appear in a table but unlike the case one this table may contains several columns in addition to rows. The initial table contains the observed values. To find expected values for this table we set up another table similar to this one. To find the value of each cell in the new table we should multiply the sum of the cell column by the sum of the cell row and divide the results by the grand total. The grand total is the total number of observations in a study. Now based on the following table test whether or not the smoking habit and gender of the population that the following sample taken from are independent. On the other hand is that true that males in this population smoke more than females?

You could use formula bar to calculate the expected values for the expected range. For example to find the expected value for the cell C5 which is replaced in c11 you could click on the formula bar and enter C6*D5/D6 then enter in cell C11.

Step 1. Observed Range b4:c5

Smoking and gender

	yes	no	total
male	31	69	100
female	45	122	167
total	76	191	267

Step 2. Expected Range b10:c11

28.46442	71.53558
47.53558	119.4644

So the observed range is b4:c5 and the expected range is b10:c11.

Step 3. Click on *fx*(paste function)

Step 4. When Paste Function dialog box appears, click on *Statistical* in function category and CHITEST in the function name then click OK.

When the CHITEST box appears, enter b4:c5 for the actual range, then b10:c11 for the expected range.

Step 5. Click on OK (the p-value appears). 0.477395

Conclusion: Since p-value is greater than the level of significance (0.05), fails to reject the null. This means smoking and gender are independent events. Based on sample information one can not assure females smoke more than males or the other way around.

Step 6. To find the chi-square value, use CHINV function, when Chinv box appears enter 0.477395 for probability part, then 1 for the degrees of freedom.

Degrees of freedom=(number of columns-1)X(number of rows-1)

CHI-SQUARE=0.504807

Test Hypothesis Concerning the Variance of Two Populations

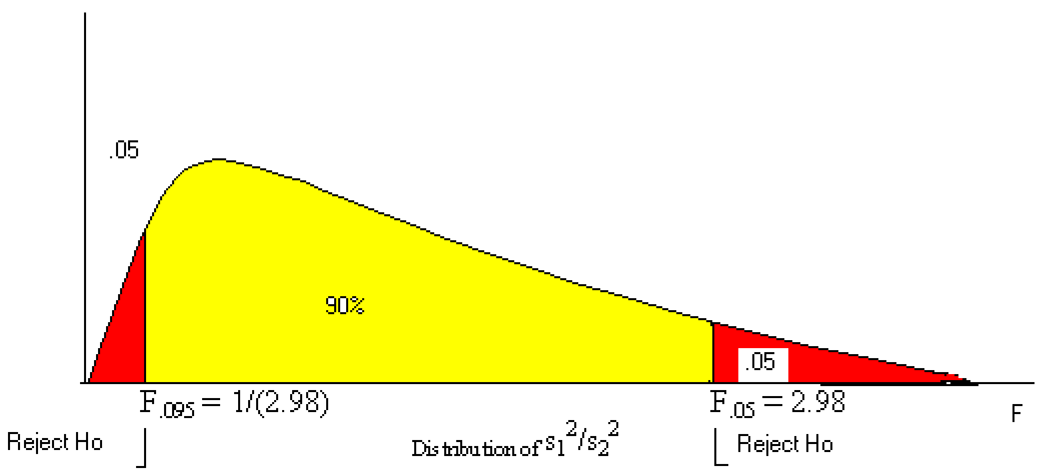
In this section we would like to examine whether or not the variances of two populations are equal. Whenever independent simple random samples of equal or different sizes such as n_1 and n_2 are taken from two normal distributions with equal variances, the sampling distribution of s_1^2/s_2^2 has F distribution with n_1- 1 degrees of freedom for the numerator and n_2 - 1 degrees of freedom for the denominator. In the ratio s_1^2/s_2^2 the numerator s_1^2 and the denominator s_2^2 are variances of the first and the second sample, respectively. The following figure shows the graph of an F distribution with 10 degrees of freedom for both the numerator and the denominator. Unlike the normal distribution as you see the F distribution is not symmetric. The shape of an F distribution is positively skewed and depends on the degrees of freedom for the numerator and the denominator. The value of F is always positive.

Now let see whether or not the variances of hourly income of student-assistant and work-study students based on samples taken from populations previously are equal. Assume that the hypothesis test in this case is conducted at a = 0.10. The null and the alternative are:

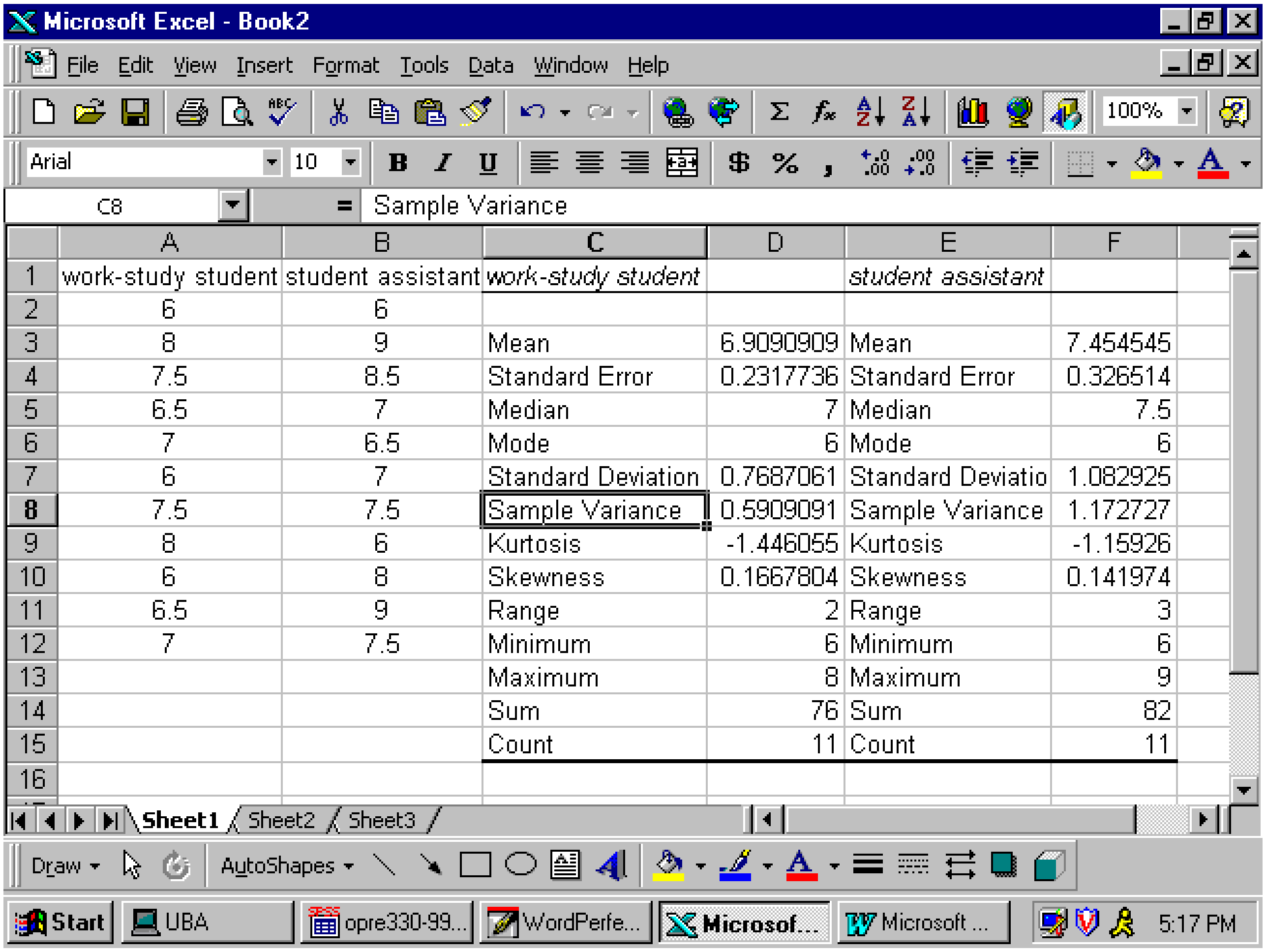
$$H_0: \delta_1^2 = \delta_2^2$$

$$H_a: \delta_1^2 \neq \delta_2^2$$

Rejection Rule: Reject the null hypothesis if F< $F_{0.095}$ or F> $F_{0.05}$ where F, the value of the test statistic is equal to s_1^2/s_2^2, with 10 degrees of freedom for both the numerator and the denominator. We can find the value of $F_{.05}$ from the F distribution table. If s_1^2/s_2^2, we do not need to know the value of $F_{0.095}$ otherwise, $F_{0.95}$ = 1/ $F_{0.05}$ for equal sample sizes.

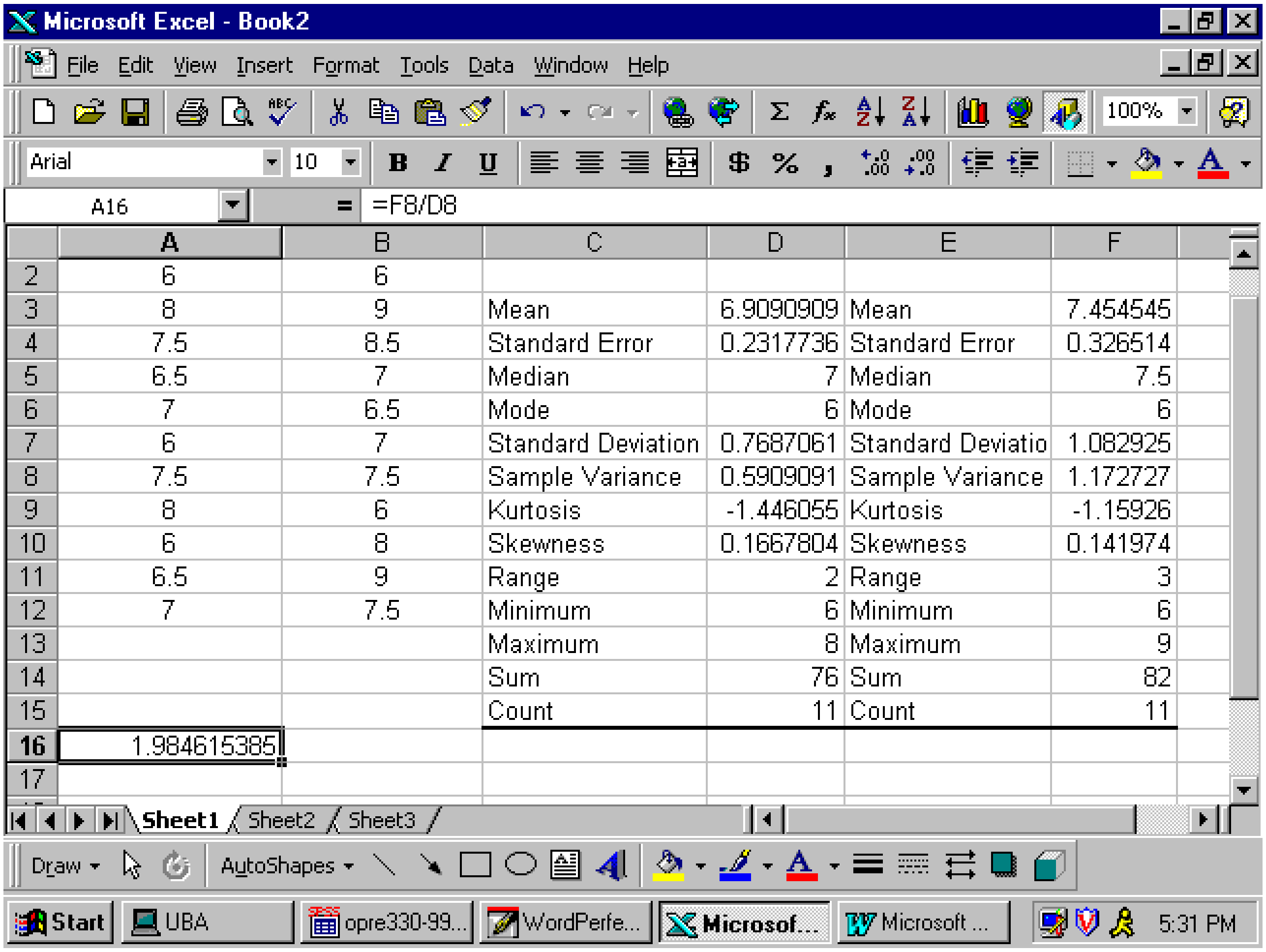


A survey of eleven student-assistant and eleven work-study students shows the following descriptive statistics. Our objective is to find the value of s_1^2/s_2^2, where s_1^2 is the value of the variance of student assistant sample and s_2^2 is the value of the variance of the work study students sample. As you see these values are in cells F8 and D8 of the descriptive statistic output.



	A	B	C	D	E	F
1	work-study student	student assistant	*work-study student*		*student assistant*	
2	6	6				
3	8	9	Mean	6.9090909	Mean	7.454545
4	7.5	8.5	Standard Error	0.2317736	Standard Error	0.326514
5	6.5	7	Median	7	Median	7.5
6	7	6.5	Mode	6	Mode	6
7	6	7	Standard Deviation	0.7687061	Standard Deviatio	1.082925
8	7.5	7.5	Sample Variance	0.5909091	Sample Variance	1.172727
9	8	6	Kurtosis	-1.446055	Kurtosis	-1.15926
10	6	8	Skewness	0.1667804	Skewness	0.141974
11	6.5	9	Range	2	Range	3
12	7	7.5	Minimum	6	Minimum	6
13			Maximum	8	Maximum	9
14			Sum	76	Sum	82
15			Count	11	Count	11
16						

To calculate the value of s_1^2/s_2^2, select a cell such as A16 and enter cell formula = F8/D8 and enter. This is the value of F in our problem. Since this value, F=1.984615385, falls in acceptance area we fail to reject the null hypothesis. Hence, the sample results do support the conclusion that student assistants hourly income variance is equal to the work study students hourly income variance. The following screen shoot shows how to find the F value. We can follow the same format for one tail test(s).]



	A	B	C	D	E	F
2	6	6				
3	8	9	Mean	6.9090909	Mean	7.454545
4	7.5	8.5	Standard Error	0.2317736	Standard Error	0.326514
5	6.5	7	Median	7	Median	7.5
6	7	6.5	Mode	6	Mode	6
7	6	7	Standard Deviation	0.7687061	Standard Deviatio	1.082925
8	7.5	7.5	Sample Variance	0.5909091	Sample Variance	1.172727
9	8	6	Kurtosis	-1.446055	Kurtosis	-1.15926
10	6	8	Skewness	0.1667804	Skewness	0.141974
11	6.5	9	Range	2	Range	3
12	7	7.5	Minimum	6	Minimum	6
13			Maximum	8	Maximum	9
14			Sum	76	Sum	82
15			Count	11	Count	11
16	1.984615385					
17						

Linear Correlation and Regression Analysis

Source http://home.ubalt.edu/ntsbarsh/excel/excel.htm#rtesthypo

[In this section the objective is to see whether there is a correlation between two variables and to find a model that predicts one variable in terms of the other variable. There are so many examples that we could mention but we will mention the popular ones in the world of business. Usually independent variable is presented by the letter x and the dependent variable is presented by the letter y. A business man would like to see whether there is a relationship between the number of cases of sold and the temperature in a hot summer day based on information taken from the past. He also would like to estimate the number cases of soda which will be sold in a particular hot summer day in a ball game. He clearly recorded temperatures and number of cases of soda sold on those particular days. The following table shows the recorded data from June 1 through June 13. The weatherman predicts a 94F degree temperature for June 14. The businessman would like to meet all demands for the cases of sodas ordered by customers on June 14.

DAY	Cases of Soda (Qty)	Temperature
01-Jun	57	56
02-Jun	59	58
03-Jun	65	63
04-Jun	67	66
05-Jun	75	73
06-Jun	81	78
07-Jun	86	85
08-Jun	88	85
09-Jun	88	87
10-Jun	84	84
11-Jun	82	88
12-Jun	80	84
13-Jun	83	89

Now let's use Excel to find the linear correlation coefficient and the regression line equation. The linear correlation coefficient is a quantity between -1 and +1. This quantity is denoted by R.

- The closer R to +1 the stronger positive (direct) correlation and
- Similarly the closer R to -1 the stronger negative (inverse) correlation exists between the two variables.]

Mathematics

Calculation of CORRELATION is totally dependent on calculation of COVARINACE.

COVARIANCE

The covariance between two securities A and B may be calculated using the following formula:

$$COV_{AB} = \frac{\sum [R_A - \overline{R_A}][R_B - \overline{R_B}]}{N}$$ At the beginning please add the summation sign in the numerator

where

COV_{AB} = Covariance between *x* and y.

R_A = Return of security *x*.

R_B = Return of security *y*.

$\overline{R_A}$ = Expected or mean return of security *x*.

$\overline{R_B}$ = Expected or mean return of security *y*.

N = Number of observations.

The calculation of covariance can be understood with the help of following table:

Calculation of Covariance

Year	R_X	Deviation $R_X - \overline{R_x}$	R_Y	Deviation $R_Y - \overline{R_Y}$	$[R_x - \overline{R_x}][R_y - \overline{R_y}]$
1	11	-4	18	5	-20
2	13	-2	14	1	-2
3	17	2	11	-2	-4
4	19	4	9	-4	-16
	$\overline{R_x}$ = 15		$\overline{R_y}$ =13		-42

$$Cov_{xy} = \frac{\sum_{i=1}^{n} [R_x - \overline{R_x}][R_y - \overline{R_y}]}{n} = \frac{-42}{4} = -10.5$$

Excel

Syntax: - COVAR(array1, array2)

Correlation calculation in Mathematics

	A	B	C	D
1	DAY	Cases of Soda	Temperature	
2	1-Jun	57	56	
3	2-Jun	59	58	
4	3-Jun	65	63	
5	4-Jun	67	66	
6	5-Jun	75	73	
7	6-Jun	81	78	
8	7-Jun	86	85	
9	8-Jun	88	85	
10	9-Jun	88	87	
11	10-Jun	84	84	
12	11-Jun	82	88	
13	12-Jun	80	84	
14	13-Jun	83	89	
15				
16	Covariance		**116.6686391**	*=COVAR(C2:C14,B2:B14)*
17				

CORRELATION

The coefficient of correlation is expressed as:

$$r_{AB} = \frac{Cov_{AB}}{\sigma_A \sigma_B}$$

where

r_{AB} = Coefficient of correlation between x and y.

Cov_{AB} = Covariance between A and B.

σ_A= Standard deviation of A.

σ_B = Standard deviation of B.

Excel

Syntax: - CORREL(array1, array2)

	A	B	C	D	E
1	DAY	Cases of Soda	Temperature		
2	1-Jun	57	56		
3	2-Jun	59	58		
4	3-Jun	65	63		
5	4-Jun	67	66		
6	5-Jun	75	73		
7	6-Jun	81	78		
8	7-Jun	86	85		
9	8-Jun	88	85		
10	9-Jun	88	87		
11	10-Jun	84	84		
12	11-Jun	82	88		
13	12-Jun	80	84		
14	13-Jun	83	89		
15					
16	Correlation		**0.966598577**	*=CORREL(C2:C14,B2:B14)*	
17					

Using DATA Analysis

Step 1. From the menus choose Tools and click on Data Analysis.

Step 2. When Data Analysis dialog box appears, click on correlation.

Step 3. When correlation dialog box appears, enter B2:C14 in the input range box. Click on Labels in first row and enter A18 in the output range box. Click on OK.

	A	B	C	D	E
1	DAY	Cases of Soda	Temperature		
2	1-Jun	57	56		
3	2-Jun	59	58		
4	3-Jun	65	63		
5	4-Jun	67	66		
6	5-Jun	75	73		
7	6-Jun	81	78		
8	7-Jun	86	85		
9	8-Jun	88	85		
10	9-Jun	88	87		
11	10-Jun	84	84		
12	11-Jun	82	88		
13	12-Jun	80	84		
14	13-Jun	83	89		
15					
16	Correlation		**0.966598577**	*=CORREL(C2:C14,B2:B14)*	
17					
18		*Column 1*	*Column 2*		
19	Column 1	1			
20	Column 2	0.966598577	1		
21					

As you see the correlation between the number of cases of soda demanded and the temperature is a very strong positive correlation. This means as the temperature increases the demand for cases of soda is also increasing. The linear correlation coefficient is 0.966598577 which is very close to +1.

Regression equation

Now let's follow same steps but a bit different to find the regression equation.

Step 1. From the menus choose Tools and click on Data Analysis

Step 2. When Data Analysis dialog box appears, click on regression.

Step 3. When Regression dialog box appears, enter B2:B14 in the y-range box and C2:C14 in the x-range box. Click on labels.

Step 4. Enter A10 in the output range box.

Note: The regression equation in general should look like Y=m X + b. In this equation m is the slope of the regression line and b is its y-intercept.

SUMMARY OUTPUT

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.966598577
R Square	0.934312809
Adjusted R Square	0.928341246
Standard Error	2.919383191
Observations	13

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1333.479989	1333.479989	156.4603497	7.58511E-08
Residual	11	93.75078034	8.522798213		
Total	12	1427.230769			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	9.17800767	5.445742836	1.685354587	0.120044803	-2.807991498	21.16400684	-2.807991498	21.16400684
X Variable 1	0.879202711	0.07028892	12.50841116	7.58511E-08	0.724497841	1.033907581	0.724497841	1.033907581

The relationship between the number of cans of soda and the temperature is: Y = 0.879202711 X + 9.17800767

The number of cans of soda = 0.879202711*(Temperature) + 9.17800767. Referring to this expression we can approximately predict the number of cases of soda needed on June 14. The weather forecast for this is 94 degrees, hence the number of cans of soda needed is equal to; The number of cases of soda=0.879202711*(94) + 9.17800767 = 91.82 or about 92 cases.

Intercept

You can also calculate intercept using INTERCEPT function in **EXCEL**

Syntax : INTERCEPT(Array1, Array2)

	A	B	C	D	E
1	DAY	Cases of Soda	Temperature		
2	1-Jun	57	56		
3	2-Jun	59	58		
4	3-Jun	65	63		
5	4-Jun	67	66		
6	5-Jun	75	73		
7	6-Jun	81	78		
8	7-Jun	86	85		
9	8-Jun	88	85		
10	9-Jun	88	87		
11	10-Jun	84	84		
12	11-Jun	82	88		
13	12-Jun	80	84		
14	13-Jun	83	89		
15					
16	Intercept		**9.17800767**	*=INTERCEPT(B2:B14,C2:C14)*	
17					

EXCEL 2016 Statistics functions List

S.No.	Function	Description	Year of Introduction
1	AVEDEV function	Returns the average of the absolute deviations of data points from their mean	
2	AVERAGE function	Returns the average of its arguments	
3	AVERAGEA function	Returns the average of its arguments, including numbers, text, and logical values	
4	AVERAGEIF function	Returns the average (arithmetic mean) of all the cells in a range that meet a given criteria	
5	AVERAGEIFS function	Returns the average (arithmetic mean) of all cells that meet multiple criteria	
6	BETA.DIST function	Returns the beta cumulative distribution function	2010
7	BETA.INV function	Returns the inverse of the cumulative distribution function for a specified beta distribution	2010
8	BINOM.DIST function	Returns the individual term binomial distribution probability	2010
9	BINOM.DIST.RANGE function	Returns the probability of a trial result using a binomial distribution	2013
10	BINOM.INV function	Returns the smallest value for which the cumulative binomial distribution is less than or equal to a criterion value	2010
11	CHISQ.DIST function	Returns the cumulative beta probability density function	2010
12	CHISQ.DIST.RT function	Returns the one-tailed probability of the chi-squared distribution	2010
13	CHISQ.INV function	Returns the cumulative beta probability density function	2010
14	CHISQ.INV.RT function	Returns the inverse of the one-tailed probability of the chi-squared distribution	2010
15	CHISQ.TEST function	Returns the test for independence	2010
16	CONFIDENCE.NORM function	Returns the confidence interval for a population mean	2010
17	CONFIDENCE.T function	Returns the confidence interval for a population mean, using a Student's t distribution	2010
18	CORREL function	Returns the correlation coefficient between two data sets	
19	COUNT function	Counts how many numbers are in the list of arguments	
20	COUNTA function	Counts how many values are in the list of arguments	
21	COUNTBLANK function	Counts the number of blank cells within a range	

22	COUNTIF function	Counts the number of cells within a range that meet the given criteria	
23	COUNTIFS function	Counts the number of cells within a range that meet multiple criteria	
24	COVARIANCE.P function	Returns covariance, the average of the products of paired deviations	2010
25	COVARIANCE.S function	Returns the sample covariance, the average of the products deviations for each data point pair in two data sets	2010
26	DEVSQ function	Returns the sum of squares of deviations	
27	EXPON.DIST function	Returns the exponential distribution	2010
28	F.DIST function	Returns the F probability distribution	2010
29	F.DIST.RT function	Returns the F probability distribution	2010
30	F.INV function	Returns the inverse of the F probability distribution	2010
31	F.INV.RT function	Returns the inverse of the F probability distribution	2010
32	F.TEST function	Returns the result of an F-test	2010
33	FISHER function	Returns the Fisher transformation	
34	FISHERINV function	Returns the inverse of the Fisher transformation	
35	FORECAST function	Returns a value along a linear trend NOTE In Excel 2016, this function is replaced with FORECAST.LINEAR as part of the new Forecasting functions, but it's still available for compatibility with earlier versions.	
36	FORECAST.ETS function	Returns a future value based on existing (historical) values by using the AAA version of the Exponential Smoothing (ETS) algorithm NOTE This function isn't available in Excel 2016 for Mac.	2016
37	FORECAST.ETS.CONFINT function	Returns a confidence interval for the forecast value at the specified target date NOTE This function isn't available in Excel 2016 for Mac.	2016
38	FORECAST.ETS.SEASONALITY function	Returns the length of the repetitive pattern Excel detects for the specified time series NOTE This function isn't available in Excel 2016 for Mac.	2016
39	FORECAST.ETS.STAT function	Returns a statistical value as a result of time series forecasting NOTE This function isn't available in Excel 2016 for Mac.	2016
40	FORECAST.LINEAR function	Returns a future value based on existing values NOTE This function isn't available in Excel 2016 for Mac.	2016
41	FREQUENCY function	Returns a frequency distribution as a vertical array	

42	GAMMA function	Returns the Gamma function value	2013
43	GAMMA.DIST function	Returns the gamma distribution	2010
44	GAMMA.INV function	Returns the inverse of the gamma cumulative distribution	2010
45	GAMMALN function	Returns the natural logarithm of the gamma function, Γ(x)	
46	GAMMALN.PRECISE function	Returns the natural logarithm of the gamma function, Γ(x)	2010
47	GAUSS function	Returns 0.5 less than the standard normal cumulative distribution	2013
48	GEOMEAN function	Returns the geometric mean	
49	GROWTH function	Returns values along an exponential trend	
50	HARMEAN function	Returns the harmonic mean	
51	HYPGEOM.DIST function	Returns the hypergeometric distribution	
52	INTERCEPT function	Returns the intercept of the linear regression line	
53	KURT function	Returns the kurtosis of a data set	
54	LARGE function	Returns the k-th largest value in a data set	
55	LINEST function	Returns the parameters of a linear trend	
56	LOGEST function	Returns the parameters of an exponential trend	
57	LOGNORM.DIST function	Returns the cumulative lognormal distribution	2010
58	LOGNORM.INV function	Returns the inverse of the lognormal cumulative distribution	2010
59	MAX function	Returns the maximum value in a list of arguments	
60	MAXA function	Returns the maximum value in a list of arguments, including numbers, text, and logical values	
61	MEDIAN function	Returns the median of the given numbers	
62	MIN function	Returns the minimum value in a list of arguments	
63	MINA function	Returns the smallest value in a list of arguments, including numbers, text, and logical values	
64	MODE.MULT function	Returns a vertical array of the most frequently occurring, or repetitive values in an array or range of data	2010
65	MODE.SNGL function	Returns the most common value in a data set	2010
66	NEGBINOM.DIST function	Returns the negative binomial distribution	2010
67	NORM.DIST function	Returns the normal cumulative distribution	2010
68	NORM.INV function	Returns the inverse of the normal cumulative distribution	2010
69	NORM.S.DIST function	Returns the standard normal cumulative distribution	2010
70	NORM.S.INV function	Returns the inverse of the standard normal cumulative distribution	2010

71	PEARSON function	Returns the Pearson product moment correlation coefficient	
72	PERCENTILE.EXC function	Returns the k-th percentile of values in a range, where k is in the range 0..1, exclusive	2010
73	PERCENTILE.INC function	Returns the k-th percentile of values in a range	2010
74	PERCENTRANK.EXC function	Returns the rank of a value in a data set as a percentage (0..1, exclusive) of the data set	2010
75	PERCENTRANK.INC function	Returns the percentage rank of a value in a data set	2010
76	PERMUT function	Returns the number of permutations for a given number of objects	
77	PERMUTATIONA function	Returns the number of permutations for a given number of objects (with repetitions) that can be selected from the total objects	2013
78	PHI function	Returns the value of the density function for a standard normal distribution	2013
79	POISSON.DIST function	Returns the Poisson distribution	2010
80	PROB function	Returns the probability that values in a range are between two limits	
81	QUARTILE.EXC function	Returns the quartile of the data set, based on percentile values from 0..1, exclusive	2010
82	QUARTILE.INC function	Returns the quartile of a data set	2010
83	RANK.AVG function	Returns the rank of a number in a list of numbers	2010
84	RANK.EQ function	Returns the rank of a number in a list of numbers	2010
85	RSQ function	Returns the square of the Pearson product moment correlation coefficient	
86	SKEW function	Returns the skewness of a distribution	
87	SKEW.P function	Returns the skewness of a distribution based on a population: a characterization of the degree of asymmetry of a distribution around its mean	2013
88	SLOPE function	Returns the slope of the linear regression line	
89	SMALL function	Returns the k-th smallest value in a data set	
90	STANDARDIZE function	Returns a normalized value	
91	STDEV.P function	Calculates standard deviation based on the entire population	2010
92	STDEV.S function	Estimates standard deviation based on a sample	2010
93	STDEVA function	Estimates standard deviation based on a sample, including numbers, text, and logical values	
94	STDEVPA function	Calculates standard deviation based on the entire population, including numbers, text, and logical values	
95	STEYX function	Returns the standard error of the predicted y-value for each x in the regression	

96	T.DIST function	Returns the Percentage Points (probability) for the Student t-distribution	2010
97	T.DIST.2T function	Returns the Percentage Points (probability) for the Student t-distribution	2010
98	T.DIST.RT function	Returns the Student's t-distribution	2010
99	T.INV function	Returns the t-value of the Student's t-distribution as a function of the probability and the degrees of freedom	2010
100	T.INV.2T function	Returns the inverse of the Student's t-distribution	2010
101	T.TEST function	Returns the probability associated with a Student's t-test	2010
102	TREND function	Returns values along a linear trend	
103	TRIMMEAN function	Returns the mean of the interior of a data set	
104	VAR.P function	Calculates variance based on the entire population	2010
105	VAR.S function	Estimates variance based on a sample	2010
106	VARA function	Estimates variance based on a sample, including numbers, text, and logical values	
107	VARPA function	Calculates variance based on the entire population, including numbers, text, and logical values	
108	WEIBULL.DIST function	Returns the Weibull distribution	2010
109	Z.TEST function	Returns the one-tailed probability-value of a z-test	2010

Appendix – 1

NORMAL DISTRIBUTION TABLE

	0.00	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Appendix – 2

Confidence Level table

Confidence Level	Z
0.70	1.040
0.75	1.150
0.80	1.280
0.85	1.440
0.90	1.645
0.92	1.750
0.95	1.960
0.96	2.050
0.98	2.330
0.99	2.580

Appendix – 3

Hypothesis Testing Examples

1. Suppose we would like to determine if the typical amount spent per customer for dinner at a new restaurant in town is more than $20.00. A sample of 49 customers over a three-week period was randomly selected and the average amount spent was $22.60. Assume that the standard deviation is known to be $2.50. Using a 0.02 level of significance, would we conclude the typical amount spent per customer is more than $20.00?

2. Suppose an editor of a publishing company claims that the mean time to write a textbook is at most 15 months. A sample of 16 textbook authors is randomly selected and it is found that the mean time taken by them to write a textbook was 12.5. Assume also that the standard deviation is known to be 3.6 months. Assuming the time to write a textbook is normally distributed and using a 0.025 level of significance, would you conclude the editor's claim is true?

3. Suppose, according to a 1990 demographic report, the average U. S. household spends $90 per day. Suppose you recently took a random sample of 30 households in Huntsville and the results revealed a mean of $84.50. Suppose the standard deviation is known to be $14.50. Using a 0.05 level of significance, can it be concluded that the average amount spent per day by U.S. households has decreased?

4. Suppose the mean salary for full professors in the United States is believed to be $61,650. A sample of 36 full professors revealed a mean salary of $69,800. Assuming the standard deviation is $5,000, can it be concluded that the average salary has increased using a 0.02 level of significance?

5. Historically, evening long-distance calls from a particular city have averaged 15.2 minutes per call. In a random sample of 35 calls, the sample mean time was 14.3 minutes. Assume the standard deviation is known to be 5 minutes. Using a 0.05 level of significance, is there sufficient evidence to conclude that the average evening long-distance call has decreased?

6. Suppose a production line operates with a mean filling weight of 16 ounces per container. Since over- or under-filling can be dangerous, a quality control inspector samples 30 items to determine whether or not the filling weight has to be adjusted. The sample revealed a mean of 16.32 ounces. From past data, the standard deviation is known to be .8 ounces. Using a 0.10 level of significance, can it be concluded that the process is out of control (not equal to 16 ounces)?

Solutions

1. Ho: $\mu = 20$

   Ha: $\mu > 20$

   $$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{22.60 - 20}{\frac{2.50}{\sqrt{49}}} = 7.28$$

   Reject Ho if Z > 2.06

   Reject Ho

   There is sufficient evidence to conclude the typical amount spent per customer is more than \$20.00, $\alpha = 0.02$.

2. Ho: $\mu = 15$

   Ha: $\mu < 15$

   $$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{12.5 - 15}{\frac{3.6}{\sqrt{16}}} = -2.78$$

   Reject Ho if Z < -1.96

   Reject Ho

   There is sufficient evidence to conclude the editor's claim is true, $\alpha = 0.025$.

3. Ho: $\mu = 90$

   Ha: $\mu < 90$

   $$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{84.50 - 90}{\frac{14.50}{\sqrt{30}}} = -2.078$$

   Reject Ho if Z < -1.65

   Reject Ho

There is sufficient evidence to conclude the average amount spent per day by U.S. households has decreased, $\alpha = 0.05$.

4. Ho: $\mu = 61{,}650$

   Ha: $\mu > 61{,}650$

   $$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{69{,}800 - 61{,}650}{\frac{5{,}000}{\sqrt{36}}} = 9.78$$

   Reject Ho if Z > 2.06

   Reject Ho

   There is sufficient evidence to conclude the average salary has increased, $\alpha = 0.02$.

5. Ho: $\mu = 15.2$

   Ha: $\mu < 15.2$

   $$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{14.3 - 15.2}{\frac{5}{\sqrt{30}}} = 1.065$$

   Reject Ho if Z < -1.65

   Fail to Reject Ho

   There is insufficient evidence to conclude the average evening long-distance call has decreased, $\alpha = 0.05$.

6. Ho: $\mu = 16$

   Ha: $\mu \neq 16$

   $$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{16.32 - 16}{\frac{0.8}{\sqrt{30}}} = 2.19$$

   Reject Ho if Z < - 1.65 OR Z > 1.65

   Reject Ho

   There is sufficient evidence to conclude the process is out of control, $\alpha = 0.10$.

IMPORTANT LINKS

1. Registration of the online Course

   www.vikominstitute.com

2. Excel query group on FACEBOOK or become our FACEBOOK family part

   https://www.facebook.com/groups/Tricksexcel/

3. Want cool excel tricks then like us on face book and get daily updates

   https://www.facebook.com/exceltricks

CONTACT DETAILS

Website	:	www.vikominstitute.com
Email I'd	:	info@vikominstitute.com

WANNA COURSE BEYOND EXCEL BASIC TO ADVANCE

ADVANCES EXCEL SKILLS AND ADVANCE FORMULAS (10 HOURS COURSE)

1. Data Validation
2. Conditional Formatting
3. Printing with Excel
4. Hyperlinks
5. Advance Filter
6. VLOOKUP VS INDEX MATCH
7. AND, OR AND NOT functions
8. ISNA and ISERROR functions
9. Consolidate
10. Flash Fill

---THE END---