a guide to

Probability and Statistics in Microsoft Excel™

Resources to support the learning of mathematics, statistics and OR in higher education.

www.mathstore.ac.uk

The Statistical Education through Problem Solving (STEPS) glossary

www.stats.gla.ac/steps/glossary

Probability and Statistics in Microsoft Excel™

Excel provides more than 100 functions relating to probability and statistics. It also has a facility for constructing a wide range of charts and graphs for displaying data. This leaflet provides a quick reference guide to assist you in harnessing Excel's statistical capability. Except where indicated, the features included here are available in Excel Versions 4.0 and above. Almost all the instructions here also apply to the spreadsheet facility in OpenOffice (http://openoffice.org-suite.com/); any slight variations in commands should be obvious to the user.

Excel is not designed for statistical computing. If you require statistical analysis beyond data validation and manipulation, tabulation, presentation and calculation of summary statistics, you are advised to use a bespoke statistical package such as Minitab or SPSS.

Excel has an Analysis Toolpak optional "add-in" facility that includes macros for carrying out many elementary statistical analyses. The instructions for installation of this add-in vary with the version of Excel — use the **Help** facility in Excel for further information on this. This add-in facility is not used in this leaflet.

There are two reasons why this add-in should be used with care:

- Unlike other spreadsheet functionality, which ensures that calculations automatically update in the light of changes elsewhere in the workbook, the output from the add-in is not dynamically linked to the source data. Hence if any of the data change the add-in must be run again to obtain updated output.
- Output from the add-in can be misleading (see http://support.microsoft.com/kb/829252 for example).

There are other commercially available add-ins that make use of Excel's familiar user interface but supplement its statistical functionality. Examples include:

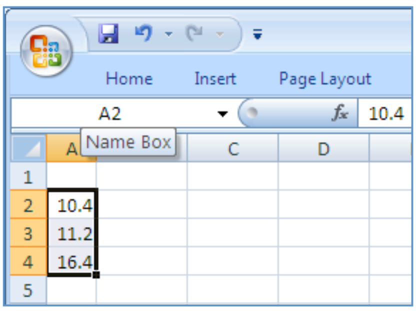
Analyse-it®	http://www.analyse-it.com/
R-Excel	http://rcom.univie.ac.at/
Unistat	http://www.unistat.com/
XLSTAT	http://www.xlstat.com/en/home/
StatTools	http://www.palisade.com/stattools/

Using this leaflet

Suppose you have a sample of three data, 10.4, 11.2 and 16.4, that you have entered into cells A2:A4 on a worksheet. In Excel a function, e.g. SUM, can be applied to these data in one of four ways:

=SUM(10.4, 11.2, 16.4)
=SUM(A2, A3, A4)
=SUM(A2:A4)
=SUM(x)   where x is the name attached to range A2:A4.

In this leaflet, for simplicity, we have chosen to refer to named ranges. To name a range, simply highlight the range of cells, click in the **Name Box** on the far left of the Formula Bar, type in the required name, e.g. x, then press Enter. In Excel 2007 names can be managed via **Formulas > Name Manager**.



If you prefer not to use names then in what follows simply replace the name of the range, e.g. x, by the range address, e.g. A2:A4.

Descriptive Statistics

Assuming a sample of data in range x

Sample total, Σx	=SUM(x)
Sample size, n	=COUNT(x)
Sample mean, $\Sigma x/n$	=AVERAGE(x)
Sample variance, s^2	=VAR(x)
Sample standard deviation, s	=STDEV(x)
Mean squared deviation	=VARP(x)
Root mean squared deviation	=STDEVP(x)
Corrected sum of squares, S_{xx}	=DEVSQ(x)
Raw sum of squares, Σx^2	=SUMSQ(x)
Minimum value	=MIN(x)
Maximum value	=MAX(x)
Range	=MAX(x)-MIN(x)
Lower Quartile, Q_1*	=QUARTILE(x, 1)
Median, Q_2	=MEDIAN(x)
Upper Quartile, Q_3*	=QUARTILE(x, 3)
Interquartile range, IQR	=QUARTILE(x, 3) - QUARTILE(x, 1)
K^{th} Percentile	=PERCENTILE(x, K%) where K is a number between 0 and 100
Mode	=MODE(x)

*Note: There are several different definitions for the upper and lower quartiles, so the values calculated by Excel may not agree with your textbook or other statistical calculation tools.

Boxplot See http://www.coventry.ac.uk/ec/~nhunt/boxplot.htm

Grouped Frequency Data

Assuming a frequency distribution with class midpoints stored in range x and frequencies in range f:

Sample size, n	=SUM(f)
Sample total, Σfx	=SUMPRODUCT(f, x)
Sample mean, $\Sigma fx/n$	=SUMPRODUCT(f, x)/SUM(f)
Corrected sum of squares, S_{xx}	=SUMPRODUCT(f, x, x)-SUMPRODUCT(f, x)^2/SUM(f)
Sample variance, s^2	=(SUMPRODUCT(f, x, x)-SUMPRODUCT(f, x)^2/SUM(f))/(SUM(f)-1)
Sample standard deviation, s	=SQRT(Sample variance)

Graphical Representations

Excel offers a wide range of chart types for displaying data. Many of these are over-elaborate. In particular, 3-D effects can be misleading and should be avoided.

In Excel 2007 to construct a chart for your data:

1. **Select** the range containing your data, including any row or column labels.
2. On the main ribbon, click on the **Insert** tab.
3. Under the **Charts** group of icons, select the chart type required, then the preferred chart subtype.
4. Under **Chart Tools** on the main ribbon, use the **Design**, **Layout** and **Format** tabs to customise the chart.

In earlier versions of Excel, select the data range and then **Insert > Chart** to invoke the Chart Wizard.

Permutations and Combinations

Number of different combinations of m objects selected from n objects

^nC_m =COMBIN(n, m)

Number of different permutations of m objects selected from n objects

^nP_m =PERMUT(n, m)

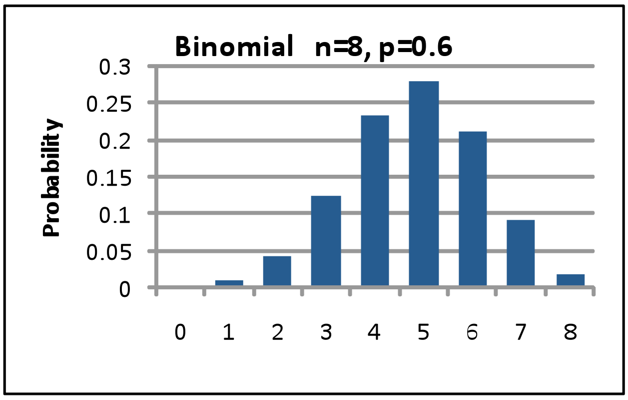
Standard Probability Distributions

Assuming a random variable X and constants a and b

Binomial **Bin(n, p)**

P(X=a) =BINOMDIST(a, n, p, FALSE)

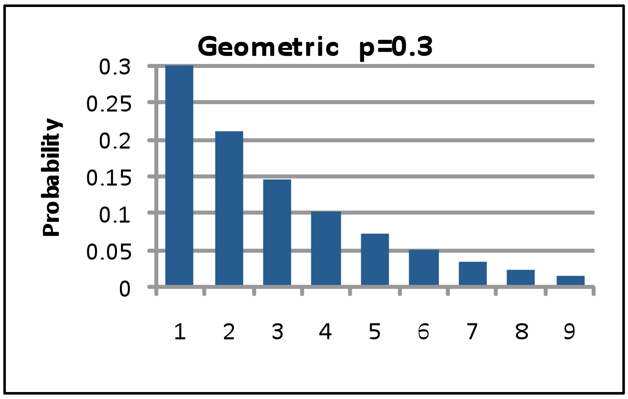
P(X≤a) =BINOMDIST(a, n, p, TRUE)



Geometric **Geom(p)**

P(X=a) =BINOMDIST(1, a, p, FALSE)/a

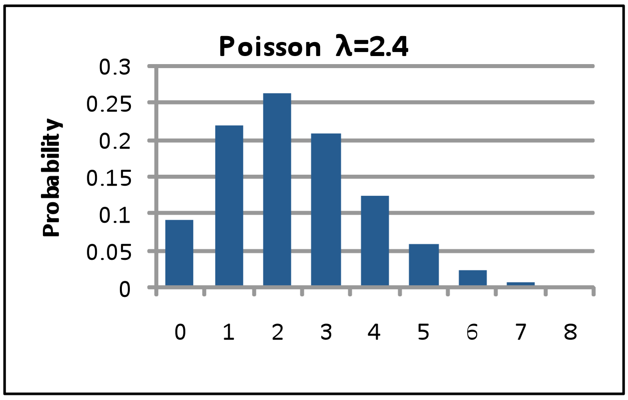
P(X≤a) =1-BINOMDIST(0, a, p, FALSE)



Poisson **Po(λ)**

P(X=a) =POISSON(a, lambda, FALSE)

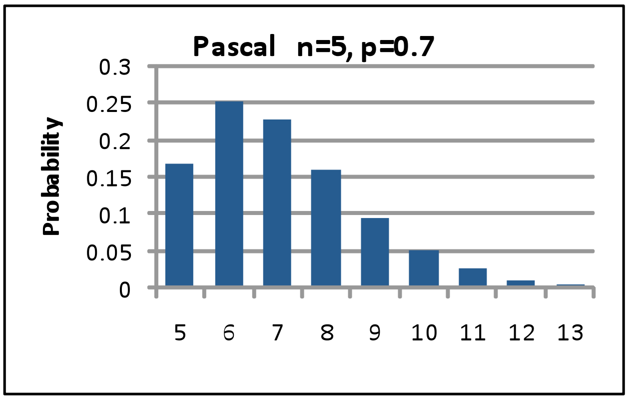
P(X≤a) =POISSON(a, lambda, TRUE)



Pascal **Pasc(n, p)**

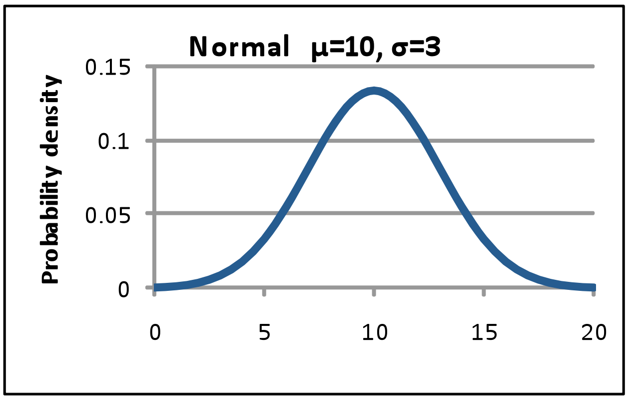
P(X=a) =NEGBINOMDIST(a-n, n, p)

P(X≤a) =BETADIST(p, n, a-n+1)/BETADIST(1, n, a-n+1)



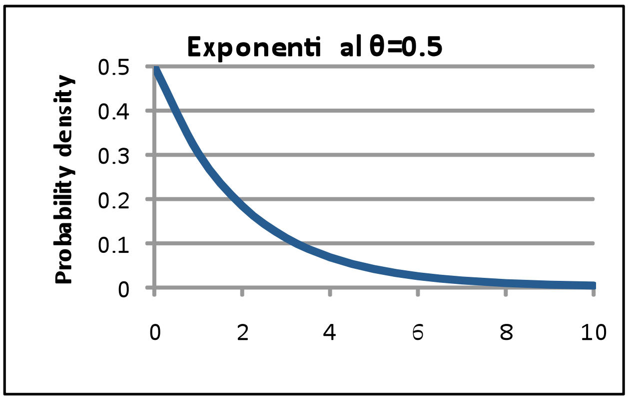
Normal **N(μ, σ²)**

f(a)	=NORMDIST(a, mu, sigma, FALSE)
P(X≤a)	=NORMDIST(a, mu, sigma, TRUE)
P(a≤X≤b)	=NORMDIST(b, mu, sigma, TRUE) - NORMDIST(a, mu, sigma, TRUE)
P(X≥b)	=1-NORMDIST(b, mu, sigma, TRUE)



Exponential **Expon(θ)**

f(a)	=EXPONDIST(a, theta, FALSE)
P(X≤a)	=EXPONDIST(a, theta, TRUE)
P(a≤X≤b)	=EXP(-a*theta)-EXP(-b*theta)
P(X≥b)	=EXP(-b*theta)



Gamma **Ga(α, β)**

f(a) =GAMMADIST(a, alpha, beta,
FALSE)

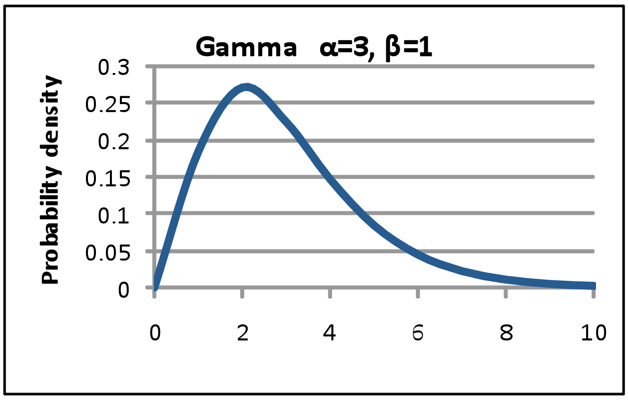
P(X≤a) =GAMMADIST(a, alpha, beta, TRUE)

P(a≤X≤b) = GAMMADIST(b, alpha, beta,
TRUE)

- GAMMADIST(a, alpha, beta,

TRUE)

P(X≥b) =1- GAMMADIST(b, alpha, beta,
TRUE)



Test Statistics for Popular Significance Tests

One sample test of a mean

Assuming a sample of data in range x, drawn from a population with mean μ and standard deviation σ:

H_0: $\mu=\mu_0$ H_1: $\mu\neq\mu_0$

Test statistic, z =(AVERAGE(x)-mu0)/(sigma/SQRT(COUNT(x))) assuming σ known

Test statistic, t =(AVERAGE(x)-mu0)/(STDEV(x)/SQRT(COUNT(x))) assuming σ unknown

One sample test of a variance

Assuming a sample of data in range x, drawn from a population with mean μ and standard deviation σ:

H_0: $\sigma^2=\sigma_0^2$ H_1: $\sigma^2>\sigma_0^2$

Test statistic, χ^2 =DEVSQ(x)/sigma0^2

Two sample test of difference between means

Assuming two samples of data in ranges x and y, drawn from populations with means μ_1 and μ_2 and equal variances:

H_0: $\mu_1 - \mu_2 = c$ H_1: $\mu_1 - \mu_2 \neq c$

Estimate the unknown common standard deviation by the pooled estimate:

s =SQRT((DEVSQ(x)+DEVSQ(y))/(COUNT(x)+COUNT(y)-2))

Test statistic, t =(AVERAGE(x)-AVERAGE(y)-c)/(s*SQRT(1/COUNT(x)+1/COUNT(y)))

Two sample test of ratio of variances

Assuming two samples of data in ranges x and y, drawn from populations with variances σ_1^2 and σ_2^2 :

H_0: $\sigma_1^2 = \sigma_2^2$ H_1: $\sigma_1^2 > \sigma_2^2$

Test statistic, F =VAR(x)/VAR(y)

Chi-squared test of association

Assuming a two-way contingency table of observed frequencies.

H_0: row factor independent of column factor

H_1: some association between row and column factors

The suggested layout below for a 4x2 table can easily be modified for tables of other sizes.

	A	B	C	D	E	F	G	H
1	90	Total	36	54				
2	Total	Observed	Col1	Col2		Expected	Col1	Col2
3	30	Row1	6	24		Row1	12	18
4	15	Row2	5	10		Row2	6	9
5	20	Row3	11	9		Row3	8	12
6	25	Row4	14	11		Row4	10	15
7								
8		P-value	0.020					
9		Deg. freedom	3					
10		Chi-squared	9.82					

A1: =SUM(C3:D6)

A3: =SUM(C3:D3) copy down to A6

C1: =SUM(C3:C6) copy across to D1

G3: =$A3*C$1/$A$1 copy into G3:H6

C8: =CHITEST(C3:D6,G3:H6)

C9: =(COUNT(A3:A6)-1)*(COUNT(C1:D1)-1)

C10: =CHIINV(C8,C9)

Critical Values and P-values for Statistical Tests

There are two approaches to conducting significance tests. Some analysts like to compare the test statistic with the critical value for a given significance level; others prefer to calculate the P-value corresponding to the test statistic. Excel can be used for either method.

Assuming significance level α, (typically α = 5% or 0.05):

Two-tailed z-test

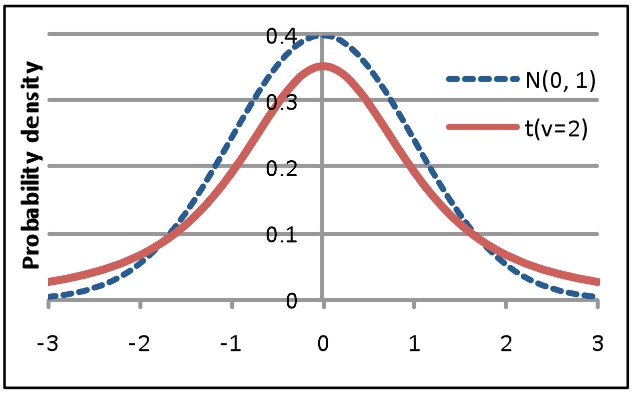
Upper tail critical value =NORMSINV(1-alpha/2)

P-value for given z =2*(1-NORMSDIST(ABS(z)))

Two-tailed t-test with v degrees of freedom

Upper tail critical value =TINV(alpha, v)

P-value for given t =TDIST(ABS(t), v, 2)



One-tailed χ^2-test with v degrees of freedom

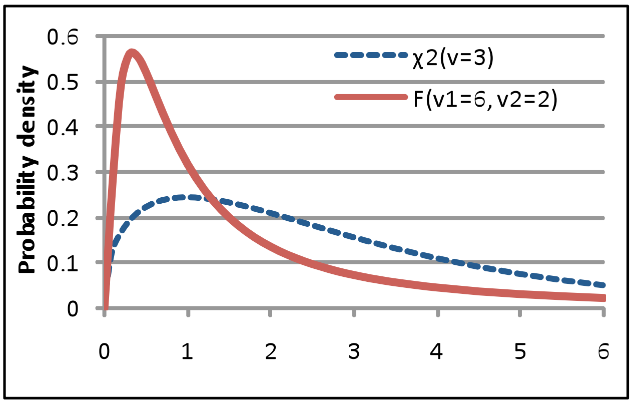
Upper tail critical value =CHIINV(alpha, v)

P-value for given chisquared =CHIDIST(chisquared, v)

One-tailed F-test with v_1 degrees of freedom in the numerator and v_2 in the denominator

Upper tail critical value =FINV(alpha, v1, v2)

P-value for given F =FDIST(F, v1, v2)



Confidence Limits

Assuming degree of confidence 100(1-α)% (e.g. for 95% confidence α =0.05):

<u>One-sample statistics, with data in range x</u>

For μ (σ known) Lower limit =AVERAGE(x)-NORMSINV(1-alpha/2)*sigma/SQRT(COUNT(x))

or =AVERAGE(x)-CONFIDENCE(alpha, sigma, COUNT(x))

Upper limit =AVERAGE(x)+NORMSINV(1-alpha/2)*sigma/SQRT(COUNT(x))

or =AVERAGE(x)+CONFIDENCE(alpha, sigma, COUNT(x))

For μ (σ unknown) Lower limit =AVERAGE(x)-TINV(alpha, COUNT(x)-1)*STDEV(x)/SQRT(COUNT(x))

Upper limit =AVERAGE(x)+TINV(alpha, COUNT(x)-1)*STDEV(x)/SQRT(COUNT(x))

For σ^2 Lower limit =(DEVSQ(x)/CHIINV(alpha/2,COUNT(x))-1)

Upper limit =(DEVSQ(x)/CHIINV(1-alpha/2,COUNT(x))-1)

Two-sample statistics, with data for the first sample in range x, and the second sample in range y

For $\mu_x - \mu_y$ (σ_x known, σ_y known)

Lower limit
=AVERAGE(x)-AVERAGE(y)-NORMSINV(1-alpha/2)*SQRT(sigmax^2/COUNT(x)+ sigmay^2/COUNT(y))

Upper limit
=AVERAGE(x)-AVERAGE(y)+NORMSINV(1-alpha/2)* SQRT(sigmax^2/COUNT(x)+ sigmay^2/COUNT(y))

For $\mu_x - \mu_y$ (σ_x and σ_y unknown but assumed equal)

Estimate the unknown common standard deviation by the pooled estimate:

s =SQRT((DEVSQ(x)+DEVSQ(y))/(COUNT(x)+COUNT(y)-2))

Lower limit
=AVERAGE(x)-AVERAGE(y)-TINV(alpha,COUNT(x)+COUNT(y)-2)*s*SQRT(1/COUNT(x)+ 1/COUNT(y))

Upper limit
=AVERAGE(x)-AVERAGE(y)+TINV(alpha,COUNT(x)+COUNT(y)-2)* s*SQRT(1/COUNT(x)+ 1/COUNT(y))

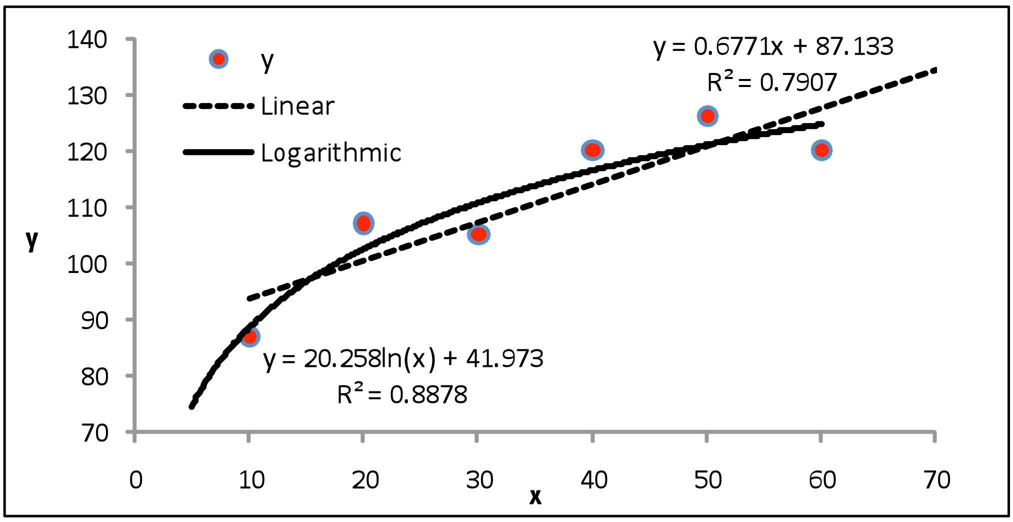
For σ_x^2 / σ_y^2

Lower limit =DEVSQ(x)/DEVSQ(y)/FINV(alpha/2, COUNT(x)-1, COUNT(y)-1)

Upper limit (DEVSQ(x)/DEVSQ(y)/FINV(1-alpha/2, COUNT(x)-1, COUNT(y)-1)

Simple Linear Regression

In Excel Versions 5 and above, a regression line (or trendline) can be added to a scatterplot by right-clicking on one of the plotted points and selecting **Add Trendline** from the shortcut menu. Both linear and a variety of non-linear models may be fitted to the data. The equation of the fitted model may be displayed, together with the value of the coefficient of determination, R^2. There are also options to extrapolate the trendline in either direction, or to force the trendline to have a specific intercept.



The trendline approach is purely graphical. To calculate predictions, regression functions must be used.

Assuming a sample of values of the independent variable in range x, and corresponding values of the dependent variable in range y:

Least squares estimate of intercept, a	=INTERCEPT(y, x)
Least squares estimate of slope, b	=SLOPE(y, x)
S_{xy}	=SUMPRODUCT(x, y)-COUNT(x)*AVERAGE(x)*AVERAGE(y)
S_{xx}	=DEVSQ(x)
S_{yy}	=DEVSQ(y)
Sample covariance, Cov(x,y)	=COVAR(x, y)*COUNT(x)/(COUNT(x)-1)
Estimate of σ, s	=STEYX(y, x)
Prediction of y at $x=x_0$, $\hat{y}=a + bx_0$	=FORECAST(x0, y, x)

Estimated standard error of individual predicted y at $x=x_0$

=STEYX(y, x)*SQRT(1+1/COUNT(x)+(x0-AVERAGE(x))^2/DEVSQ(x))

Estimated standard error of mean predicted y at $x=x_0$

=STEYX(y, x)*SQRT(1/COUNT(x)+(x0-AVERAGE(x))^2/DEVSQ(x))

Correlation

Assuming two samples of paired data in ranges x and y:

Pearson product moment correlation coefficient, r =CORREL(x, y)

Rank Correlation

Assuming two samples of paired data in ranges x and y with no ties:

Rank of i^{th} value in range x =RANK(INDEX(x, i), x, 1)

Assuming two samples of paired data in ranges x and y with some tied values:

Rank of i^{th} value in range x =(RANK(INDEX(x, i), x, 1)- RANK(INDEX(x, i), x, 0)+COUNT(x)+1)/2

Assuming that the ranges rx and ry contain the ranks of the data in x and y respectively:

Spearman rank correlation coefficient, r_S = CORREL(rx, ry)

	A	B	C	D	E	F
1	x	y	rx(Ascend)	ry(Ascend)	ry(Descend)	ry(Correct)
2	10	87	1	1	6	1
3	20	107	2	3	4	3
4	30	105	3	2	5	2
5	40	120	4	4	2	4.5
6	50	126	5	6	1	6
7	60	120	6	4	2	4.5
8						
9					Rank correlation	0.8407

In the example above:

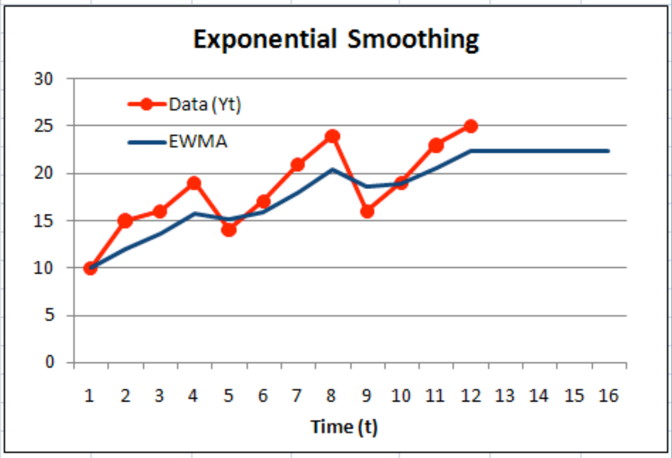
D2:	=RANK(B2, \$B\$2:\$B\$7, 1)	copy down to D7
E2:	=RANK(B2, \$B\$2:\$B\$7, 0)	copy down to E7
F2:	=(D2-E2+COUNT(\$B\$2:\$B\$7)+1)/2	copy down to F7
F9:	=CORREL(C2:C7, F2:F7)	adjusted for ties

Time Series

The examples below refer to three years of observed quarterly data.
Forecasts are made for a further four quarters (one extra year).

Level only

	A	B	C	D	E	F
1	Time (t)	Data (Y_t)	MA(5)	MA(4)	EWMA	alpha
2	1	10			10.00	0.4
3	2	15			12.00	
4	3	16	14.80	15.50	13.60	
5	4	19	16.20	16.25	15.76	
6	5	14	17.40	17.13	15.06	
7	6	17	19.00	18.38	15.83	
8	7	21	18.40	19.25	17.90	
9	8	24	19.40	19.75	20.34	
10	9	16	20.60	20.25	18.60	
11	10	19	21.40	20.63	18.76	
12	11	23			20.46	
13	12	25			22.27	
14	13		21.40	20.63	22.27	
15	14		21.40	20.63	22.27	
16	15		21.40	20.63	22.27	
17	16		21.40	20.63	22.27	



Simple moving average period 5
C4: =AVERAGE(B2:B6) copy down to C11
C14: =C$11 copy down to C17

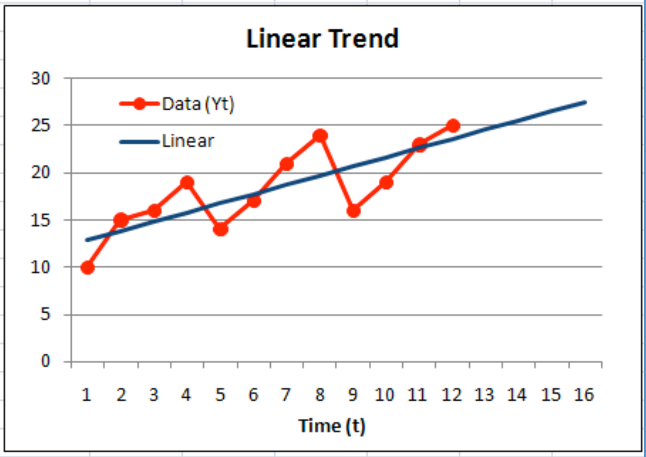
Centred moving average period 4
D4: =(AVERAGE(B2:B5)+AVERAGE(B3:B6))/2 copy down to D11
D14: =D$11 copy down to D17

Exponentially weighted moving average
E2: =B2 initial level estimate
E3: =$G2*B3+(1-$G2)*E2 copy down to E13
E14: =E$13 copy down to E17

The chart was drawn by highlighting B1:B17 and E1:E17 then using Insert > Charts > Line> 2-D Line.

Level and constant trend

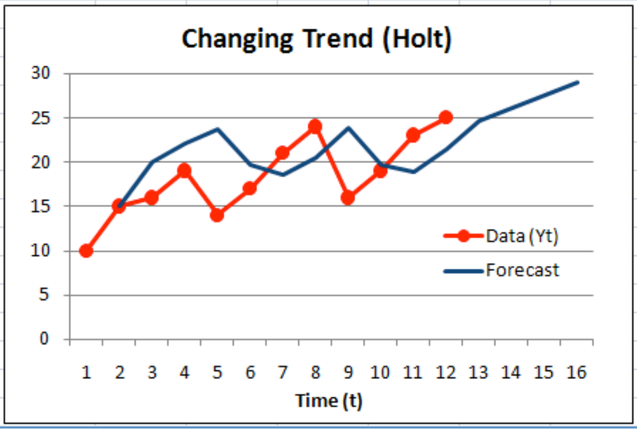
	A	B	C
1	Time (t)	Data (Y_t)	Linear
2	1	10	12.92
3	2	15	13.89
4	3	16	14.86
5	4	19	15.83
6	5	14	16.80
7	6	17	17.77
8	7	21	18.73
9	8	24	19.70
10	9	16	20.67
11	10	19	21.64
12	11	23	22.61
13	12	25	23.58
14	13		24.55
15	14		25.51
16	15		26.48
17	16		27.45



C2: =FORECAST(A2,$B$2:$B$13,$A$2:$A$13) copy down to C17

Level and changing trend

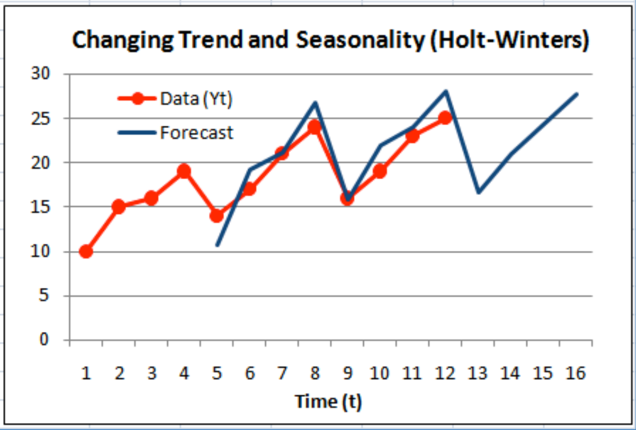
	A	B	C	D	E	F	G
1	Time (t)	Data (Y_t)	Level (μ_t)	Trend (R_t)	Forecast	alpha	beta
2	1	10	10.00	5.00		0.5	0.5
3	2	15	15.00	5.00	15.00		
4	3	16	18.00	4.00	20.00		
5	4	19	20.50	3.25	22.00		
6	5	14	18.88	0.81	23.75		
7	6	17	18.34	0.14	19.69		
8	7	21	19.74	0.77	18.48		
9	8	24	22.26	1.64	20.51		
10	9	16	19.95	-0.33	23.90		
11	10	19	19.31	-0.49	19.62		
12	11	23	20.91	0.56	18.82		
13	12	25	23.23	1.44	21.47		
14	13				24.68		
15	14				26.12		
16	15				27.56		
17	16				29.00		



C2: =B2 — initial level estimate
C3: =$F2*B3+(1-$F2)*(C2+D2) — copy down to C13
D2: =B3-B2 — initial trend estimate
D3: =$G2*(C3-C2)+(1-$G2)*D2 — copy down to D13
E3: =C2+D2 — copy down to E13
E14: =C$13+(A14-A$13)*D$13 — copy down to E17

Level, changing trend and seasonality

	A	B	C	D	E	F	G	H	I
1	Time (t)	Data (Y_t)	Level (μ_t)	Trend (R_t)	Season (S_t)	Forecast	alpha	beta	gamma
2	1	10			0.67		0.3	0.4	0.5
3	2	15			1.00				
4	3	16			1.07				
5	4	19	15.00	1.00	1.27				
6	5	14	17.50	1.60	0.73	10.67			
7	6	17	18.47	1.35	0.96	19.10			
8	7	21	19.78	1.33	1.06	21.14			
9	8	24	20.46	1.07	1.22	26.74			
10	9	16	21.62	1.11	0.74	15.79			
11	10	19	21.84	0.75	0.91	21.82			
12	11	23	22.30	0.64	1.05	24.05			
13	12	25	22.21	0.34	1.17	27.98			
14	13					16.61			
15	14					20.94			
16	15					24.34			
17	16					27.65			



C5: =AVERAGE(B2:B5) — initial level estimate
C6: =G$2*B6/E2+(1-G$2)*(C5+D5) — copy down to C13
D5: =(AVERAGE(B6:B9)-C5)/4 — initial trend estimate
D6: =H$2*(C6-C5)+(1-H$2)*D5 — copy down to D13
E2: =B2/C$5 — copy down to E5, initial seasonal estimates
E6: =I$2*B6/C6+(1-I$2)*E2 — copy down to E13
F6: =(C5+D5)*E2 — copy down to F13
F14: =(C$13+(A14-A$13)*D$13)*E10 — copy down to F17