

Using Spreadsheets for Probability and Statistics

by
Prof. Richard B. Goldstein

PROBABILITY

Multiplications $a*b*c$	@MULT(list)	@MULT(3,6,2) = 36
Combinations ${}_n C_r$	@COMB(R,N)	@COMB(2,5) = 10
Permutations ${}_n P_r$	@PERMUT(N,R)	@PERMUT(5,2) = 20
Factorial $n!$	@PERMUT(N,N)	@PERMUT(5,5) = 120

PROBABILITY DISTRIBUTIONS

Beta	@BETADIST(X,Z,W,<A>,)	Cumulative Prob.
	@BETAINV(Prob,Z,W,<A>,)	Inverse Prob.

$$\beta(z,w) = \beta(w,z) = \int_0^1 t^{z-1} (1-t)^{w-1} dt$$

$$f(x) = x^{z-1} (1-x)^{w-1} / \beta(z,w)$$

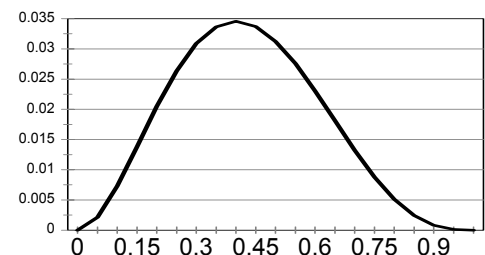
- X = Value at which to evaluate the function A # X # B
- Prob = Cumulative probability value; 0 # Prob # 1.
- Z = α parameter to the Beta distribution; must be > 0.
- W = β parameter to the Beta distribution; must be > 0.
- A = Optional lower bound to the interval (default is 0)
- B = Optional upper bound to the interval (default is 0)

@BETADIST returns the cumulative beta probability density function. The cumulative beta probability density function is a bounded distribution that is useful for studying variables such as percentages that may only take on values within a restricted range.

@BETAINV computes the inverse of the cumulative beta distribution function. If Prob = @BETADIST(X...), then @BETAINV(Prob...) = X.

Examples:

- @BETADIST(0.5,3,4,0,1) = 0.65625
- @BETADIST(0.4,3,4,0,1) = 0.45568
- @BETAINV(0.65625,3,4,0,1) = 0.5
- @BETAINV(0.45568,3,4,0,1) = 0.4



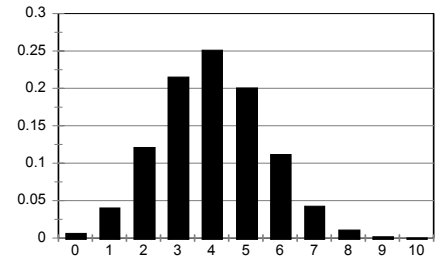
Binomial

$$@\text{BINOMDIST}(X,N,P,C) = \frac{N!}{X!(N-X)!} P^X (1-P)^{N-X} \text{ if } C = 0$$

- X = number of successes
- N = number of trials
- P = probability of success on each trial
- C = $\begin{cases} 0 & \text{for prob. of } X \text{ successes} \\ 1 & \text{for cumulative probability} \end{cases}$

Examples

$$\begin{aligned} @\text{BINOMDIST}(3,10,0.4,0) &= 0.214991 \\ @\text{BINOMDIST}(3,10,0.4,1) &= 0.382281 \end{aligned}$$



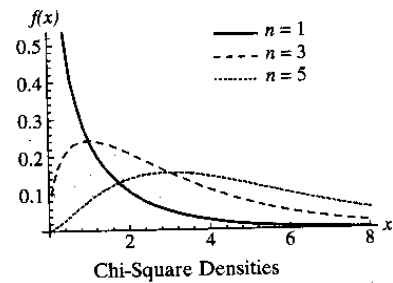
Chi-Square

$$@\text{CHIDIST}(X,N) = \int_0^x \frac{1}{2^{n/2} \Gamma(n/2)} t^{n/2-1} e^{-t/2} dt$$

- X = independent variable
- N = number of degrees of freedom

EXAMPLES

$$\begin{aligned} @\text{CHIDIST}(36.41503,24) &= 0.05 \\ @\text{CHIINV}(0.05,24) &= 36.41503 \end{aligned}$$



Exponential

$$@\text{EXPONDIST}(X,\lambda,C) = \lambda e^{-\lambda x} \text{ if } C = 0 \text{ and cumulative probability if } C = 1$$

- X = independent variable
- λ = parameter = 1/mean

F

$$@\text{FDIST}(X,N_1,N_2) = \frac{N_1^{N_1/2} N_2^{N_2/2}}{\beta(N_1, N_2)} \int_0^x t^{(N_1-2)/2} (N_2 + N_1 t)^{-(N_1+N_2)/2} dt$$

- X = independent variable
- N_1 = numerator degree of freedom
- N_2 = denominator degree of freedom

$$\begin{aligned} @\text{FDIST}(6.256057,5,4) &= 0.05 \\ @\text{FINV}(0.05,5,4) &= 6.256057 \end{aligned}$$

Gamma $@GAMMADIST(X,\alpha,\lambda,C) = \frac{\lambda^\alpha}{\Gamma(\alpha)} t^{\alpha-1} e^{-\lambda t}$ if C = 0

- X = independent variable
- α = Parameter to the gamma distribution; must be > 0.
- λ = Parameter to the gamma distribution; must be > 0.
- Cum = 1 to return the cumulative gamma distribution function;
0 to return the probability density function.

Note: when $\alpha = 1$ @GAMMADIST returns the exponential distribution

Examples

@GAMMADIST(18,8,2,1) = 0.676103

@GAMMADIST(18,8,2,0) = 0.058558

@GAMMAINV(0.676103,8,2) = 18

Hypergeometric $@HYPGEOMDIST(x,n,M,N) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}$

- x = successes in the sample.
- n = sample size
- M = successes in the population
- N = Population size.

Examples

Five cards are drawn from a deck of 52 playing cards. This formula calculates the probability that one of the five cards drawn is an ace (assuming there are only four aces in the deck):

@HYPGEOMDIST(1,5,4,52) = 0.299474

Log-Normal @LOGNORMDIST(X, μ , σ)

returns the cumulative log-normal distribution: $\frac{1}{t\sigma\sqrt{2\pi}} e^{-\frac{(\ln(t)-\mu)^2}{2\sigma^2}}$

- X = Value to evaluate the function; must be > 0
- μ = Mean of $\ln(x)$
- σ = Standard deviation of $\ln(x)$; must be > 0.

Example

@LOGNORMDIST(3,2.5,0.8) = 0.03991

Negative Binomial $@NEGBINOMDIST(n, s, p) = \binom{n + s - 1}{s - 1} p^s (1 - p)^n$

- n = Number of failures (n = 0, 1, 2, ...)
- s = Threshold of successes (s ≥ 1)
- p = Probability of a success

@NEGBINOMDIST returns the negative binomial distribution. Use it to determine the distribution of the number of failures you experience before achieving a given number of successes.

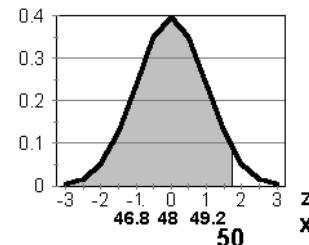
Example

A polling organization asks a sampling of voters if they favor Candidate A for reelection. Given that 55% of the city's voters favor Candidate A, this formula calculates the probability that the polling organization will contact 10 voters who do not favor her for reelection before contacting 1 voter who does favor her:

@NEGBINOMDIST(10,1,0.55) = 0.000187

Normal

@NORMDIST(x, μ, σ, C) = $\int_{-\infty}^x \frac{e^{-(t-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma}} dt$



- x = Value at which to evaluate function
- μ = Mean of the normal distribution
- σ = Standard deviation of the normal distribution
- C = 1 to return the cumulative normal distribution function
0 (the default) to return the probability density function

Examples

@NORMDIST(50,48,1.2,1) = 0.95221
@NORMDIST(50,48,1.2,0) = 0.082898
@NORMINV(0.95221,48,1.2) = 50

Poisson

@POISSON(n,μ,C) = $\frac{e^{-\mu} \mu^n}{n!}$

- n = Number of events
- μ = Expected numeric value for the mean over the distribution
- C = 1 to return the cumulative Poisson probability distribution that the

number of random events will be in the range from zero to N;
 0 to return the Poisson probability mass function that the number
 of events will be N.

Example

On average, Company Z receives 30 customer service phone calls per hour.
 What is the probability that Company Z will receive 35 calls in one hour?

@POISSON(35,30,0) = 0.045308

Student's t

$$@TDIST(x, n, Tails) = 1 - \frac{1}{\sqrt{n} \beta \left(\frac{1}{2}, \frac{n}{2}\right)} \int_{-x}^x \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} dt \text{ for Tails} = 2$$

X = Value at which to evaluate the distribution.

n = Integer number of degrees of freedom

Tails = 1 to return the area in a one-tailed distribution

2 to return the area in a two-tailed distribution

Example

@TDIST(2.228139,10,2) = 0.05 (area in each tail is 0.025)

@TINV(0.05,10) = 2.228

Weibull

$$@WEIBULL(x, \alpha, \beta, C) = \text{Mean Time to Failure} = \alpha \beta^\alpha t^{\alpha-1} e^{-(\beta t)^\alpha}$$

X = Function parameter to evaluate.

α = Parameter to the distribution; must be > 0.

β = Parameter to the distribution; must be > 0.

C = A numeric value (0 or 1) indicating whether to use the cumulative
 distribution function (1) or the probability density function (0).

@WEIBULL returns the Weibull distribution, which is used to calculate the
 mean time to failure of a device. If $\alpha = 1$, @WEIBULL returns the
 same value as @EXPONDIST with Lambda = $1/\beta$.

Examples

@WEIBULL(20,2,15,1) = 0.830987

@WEIBULL(20,2,15,0) = 0.030047

DESCRIPTIVE STATISTICS

SIMPLE MEASURES

Count	@COUNT(list)	- returns the number of non-blank cells in a list
Maximum	@MAX(list)	- returns the largest number or last date in a list
Minimum	@MIN(list)	- returns the smallest number or earliest date in a list
Largest	@LARGEST(list,N)	- returns the N th largest number value in a list
Smallest	@SMALLEST(list,N)	- returns the N th smallest number value in a list
Sum	@SUM(list)	- returns the total of all numeric values in a list

Note: To ignore labels when evaluating a list use @PUREAVG, @PURECOUNT, @PUREMAX, @PUREMIN, @PURESTD, @PURESTDS, @PUREVAR, or @PUREVARS

MEASURES OF LOCATION

Mean or Simple Average	@AVG(list)	- returns the arithmetic mean of numeric values in a list
Geometric Mean	@GEOMEAN(list)	- returns the geometric mean of numeric values in a list: Geometric Mean = $\sqrt[n]{x_1 x_2 \dots x_n}$
Harmonic Mean	@HARMEAN(list)	- return the harmonic mean of numeric values in a list: Harmonic Mean = $\frac{1}{\frac{1}{n} \left(\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_n} \right)}$ $@HARMEAN(3,4,5,6,7) = 4.575163$
Weighted Mean	@WEIGHTAVG(DataBlock, WeightsBlock, <Type>)	
	DataBlock	= Block reference or name where values to be averaged are stored.
	WeightsBlock	= Block reference or name where data affecting the weighting are stored; WeightsBlock must have the same dimensions as DataBlock.

Type = Optional value that tells Corel Quattro Pro how to calculate the weighted average:
 0 = divide by sum of values in WeightsBlock; default if you omit the argument
 1 = divide by number of values in DataBlock

returns a weighted average of the values in a block.

@WEIGHTAVG returns ERR if DataBlock and WeightsBlock are not the same dimensions.

Examples

Your son's school places heavy emphasis on math and science courses, and weights them accordingly in comparison with humanities courses. His fall semester grades are in the following table, along with the school's weighting of the courses:

	A	B	C
1	Course	Weight	Grade
2	English	2	70.0%
3	Geography	2	65.0%
4	Math	3	95.0%
5	Chemistry	3	91.0%

@WEIGHTAVG(C2..C5,B2..B5) = 82.8%

@AVG(C2..C5) = 80.3%

@MEDIAN @MEDIAN(list) - returns the middle value in a range of values in a data set arranged in ascending or descending order. If the number of values in the data set is even, the median is the mean of the two middle values.

@MODE @MODE(list)- returns the most frequent value in a list

@PERCENTILE @PERCENTILE(list, X) - returns a number from the list at the percentile indicated by X.

list = A numeric array or a block of values.

X = A percentile value between 0 and 1, inclusive.

Examples

@PERCENTILE({4,5,7,9,10,12,13,16},0) = 4
@PERCENTILE({4,5,7,9,10,12,13,16},0.25) = 6.5
@PERCENTILE({4,5,7,9,10,12,13,16},1/7)=5
@PERCENTILE({4,5,7,9,10,12,13,16},0.2)=5.8

@RANK

@RANK(X, list, Order)

X = a number in the list
list = one or more numeric or block values.
Order = Flag indicating how to sort the list of numbers:
any nonzero value = ascending order;
0 = descending order.

@RANK(4,{2,5,4,8},0) = 3
@RANK(4,{2,5,4,8},1) = 2

@PERCENTRANK @PERCENTRANK(list, X , digits)

X = a number in the list
list = one or more numeric or block values
digits = number of decimal digits (default is 3)

@PERCENTRANK({2,8,5,1,7},1,2)=0
@PERCENTRANK({2,8,5,1,7},2,2)=0.25
@PERCENTRANK({2,8,5,1,7},5,2)=0.5

@QUARTILE

@QUARTILE(list, X)

list = a numeric array or a block of values.
X = Number signifying what quartile value to return:

0 = minimum value in Array
1 = 25th percentile
2 = 50th percentile (median)
3 = 75th percentile
4 = maximum value in Array

If the quartile falls between two discrete values in the list, a fractional value is determined using linear interpolation.

Examples

@QUARTILE({4,5,7,9,10,12,13,16},0) = 4
@QUARTILE({4,5,7,9,10,12,13,16},1) = 6.5
@QUARTILE({4,5,7,9,10,12,13,16},2) = 9.5

MEASURES OF SPREAD AND HIGHER MOMENTS

@VAR

@VAR(List)

List = One or more numeric or string values, cell addresses, and block references or names, separated by commas.

@VAR calculates the population variance of all nonblank, numeric cells in

List, using the n method (biased):
$$\frac{\sum (x_i - \bar{x})^2}{n}$$

Examples

@VAR(23,24,25) = 0.666666667

@VAR("Adam",53) = 702.25, the same as for @VAR(0,53)

@VARS

@VARS(List) calculates the sample variance of all nonblank, numeric

cells in List, using the n - 1 method (unbiased):
$$\frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Examples

@VARS(23,24,25) = 1

@VARS("Adam",53) = 1404.5 , the same as for @VARS(0,53))

@STD

@STD(list) = @SQRT(@VAR(list)) population variance

@STDS

@STDS(list) = @SQRT(@VARS(list)) sample variance

@SKEW

@SKEW(List) List = One or more numeric or block values.

@SKEW returns the skewness of a distribution. Skewness characterizes the degree of asymmetry of a distribution around its mean value. Use @SKEW when you want a non-dimensional quantity to measure the "shape" of a distribution rather than its moment, which is a measure in the same units as the elements of the distribution. @SKEW finds the skewness coefficient using this formula:

$$\frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^3$$

A positive result means that the distribution is skewed to the right (the median is less than the mean). A negative result means that the distribution is skewed to the left (the median is greater than the mean).

@SKEW returns 0 when the distribution is symmetrical around its mean. If there are less than three data points in List, or if the standard deviation is zero, @SKEW returns ERR.

Example: @SKEW(4,5,8,5,7,12,6,9,2,5) = 0.685055

@KURT @KURT(List)

List = One or more numeric or block values.

@KURT returns the kurtosis of List. The kurtosis of a data set measures a distribution's closeness to normality, indicating relative peakedness or flatness. A kurtosis greater than zero is referred to as leptokurtic. A kurtosis less than zero is referred to as platykurtic.

List must have four or more values. The standard deviation of List must not be 0.

@KURT uses this formula:

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

where s is the sample standard deviation.

Examples

@KURT(5,7,9,12,14,15,4,9,5,6) = -1.11117

@KURT(9.7,10,9.5,9.3,10.2,10,9.5,11) = 1.780277

@KURT(20,25,27,22,35,28) = 0.876754

@CORREL @CORREL(List1,List2) returns the correlation coefficient r for two lists of size n

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

@PEARSON @PEARSON(List1,List2) same as @CORREL

@SEMEAN @SEMEAN(list) returns the standard error of the sample mean
for values in a given block = $\sqrt{\frac{s}{n}}$

Example: @SEMEAN({5,3,7,8}) = 1.108678

STATISTICAL TESTS

@CHITEST @CHITEST(Actual, Expected)

Actual = Block containing actual values.
 Expected = Block containing expected values.

@CHITEST computes the probability that the actual and expected frequencies are similar by chance. @CHITEST returns the probability for a chi-square test distribution with $(r - 1)(c - 1)$ degrees of freedom, where r = number of rows, and c = number of columns.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Actual and Expected must have the same number of values and must contain multiple rows or columns of data.

Example

This example refers to cells in the next figure. The chi-square statistic for the data in the next figure is 16.25813 and the degrees of freedom is 4.

@CHITEST(C3..E5,C7..E9) = 0.002692

	A	B	C	D	E
1	Soft Drink Flavors				
2	Age	Ranges	Cola	Orange	Lemon-lime
3	Actual	Under 25	120	65	55
4		26-50	100	45	85
5		Over 50	75	35	70
6					
7	Expected	Under 25	108.93	53.53	77.54
8		26-50	104.38	51.31	74.31
9		Over 50	81.69	40.16	58.15

@FTEST @FTEST(List1, List2)

List1 = First array of numeric values.
 List2 = Second array of numeric values.

@FTEST returns the results of an F-test run against the samples in List1 and List2. An F-test is a one-tailed probability that the differences in the sample variances in List1 and List2 are different. Use @FTEST to determine if two samples have significantly different variances (that is, if data sets were drawn from different parent populations).

List1 and List2 must have more than two values. The variance of List1 or

List2 must not be zero.

$$F = \frac{SS_1 / (n_1 - 1)}{SS_2 / (n_2 - 1)}$$

Example

@FTEST({75,82,83,85,85,90},{80,86,92,93,95,96}) = 0.637248

@TTEST

@TTEST(List1, List2, Tails, Type)

List1 = First array of numeric values.

List2 = Second array of numeric values.

Tails = 1 to return a one-tailed test;
2 to return a two-tailed test.

Type = A discrete variable specifying the type of test to conduct:
1 = a paired test;
2 = a two-sample equal variance test;
3 = a two-sample unequal variance test.

@TTEST returns the probability associated with the Student's t-Test. Use @TTEST to test the means of two small samples.

If List1 and List2 have a different number of data points, @TTEST returns ERR.

Example

@TTEST({62,77,73,69,54,67,59,76},{64,80,72,53,69,63,76,74},2,2) = 0.68502

@ZTEST

@ZTEST(List, X, <S>)

List = A numeric array or a block of values.

X = A value to test against the mean of the values in Array.

S = Population standard deviation; if omitted, @ZTEST uses the sample standard deviation.

@ZTEST returns the two-tailed probability of a z-test. @ZTEST calculates a z-score, which is the distance between X and the mean for Array, and then returns the two-tailed probability of the z-score for a normal distribution. Use @ZTEST to test whether a value is drawn from a large sample population.

$$z = \frac{(\bar{x} - \mu)}{(\sigma / \sqrt{n})}$$

Example: @ZTEST({10,12,14,17,19,21,22,25},15) = 0.087352

DATA FREQUENCY DISTRIBUTION

Use Tools | Numerical Tools | Frequency

Value Cells: A1.J6

Bin Cells: A8.B17

	A	B	C	D	E	F	G	H	I	J
1	13	47	10	3	16	20	17	40	4	2
2	7	25	8	21	19	15	3	17	14	6
3	12	45	1	8	4	16	11	18	23	12
4	6	2	14	13	7	15	46	12	9	18
5	34	13	41	28	36	17	24	27	29	9
6	14	26	10	24	37	31	8	16	12	16
7										
8	5	7								
9	10	11								
10	15	13								
11	20	11								
12	25	5								
13	30	4								
14	35	2								
15	40	3								
16	45	2								
17	50	2								
18		0								

MULTIPLE OR SIMPLE LINEAR REGRESSION

To perform regression analysis: $Y = \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k + \beta_0$

- 1 Choose the variable data you want to analyze. You can have one dependent variable that you believe is influenced by independent variables; for example, quarterly revenues.
- 2 Make sure the data cells you want to use in the regression analysis have the same number of rows.
- 3 Click Tools | Numeric Tools | Regression.
- 4 Specify the column of dependent data.
- 5 Specify the column(s) of independent data; they may be in noncontiguous columns.
- 6 Specify the upper left cell of the output cells where you want Corel Quattro Pro to write the regression information.
- 7 If you want to force the y-intercept value to zero, set Y Intercept to Zero.

Tip:

Regression tables are not automatically updated. If you alter the values in the independent and dependent cells, use Tools Numeric Tools Regression again to see the new results.

The Regression Output Cells

The regression output is nine rows deep and three columns wider than the number of columns in the independent cells. Make sure to leave enough blank space; any underlying data will be overwritten. The output cells contain this information:

Constant

The y-intercept value, zero if Y Intercept is set to Zero instead of Compute.

Std Err of Y Est

The estimated standard error of y values; the degree of deviation of observed y values from predicted values.

R Squared

The variance; the degree of relationship between independent and dependent variables. With one independent variable, R Squared is the square of the correlation between the two variables.

No. of Observations

The number of values for each variable; the number of rows in the regression table.

Degrees of Freedom

The number of observations minus the number of independent values being computed by the regression. With Y Intercept set to Zero, Degrees of Freedom = (No. of Observations) - (number of independent variables); with Y Intercept computed, Degrees of Freedom = (No. of Observations) - (number of independent variables + 1).

X Coefficient(s)

The regression coefficients of the independent (x) variables; the slope of the regression line representing the relationship between each independent variable and the dependent variable.

Std Err of Coef.

An error estimate of the X Coefficient above it. Interpret each coefficient as the X Coefficient value plus or minus the Standard Error of Coefficient.

	A	B	C	D	E
1		Bedrooms	Baths	Acres	Cost
2	1	3	2	0.5	\$120,000
3	2	4	2.5	0.5	\$180,000
4	3	2	1	0.25	\$80,000
5	4	5	3	2	\$300,000
6	5	4	3	1	\$250,000
7					
8		Regression Output:			
9	Constant			-9024.390	
10	Std Err of Y Est			30603.682	
11	R Squared			0.971376	
12	No. of Observations			5	
13	Degrees of Freedom			1	
14					
15	X Coefficient(s)		243.902	60975.610	63414.634
16	Std Err of Coef.		55016.087	58925.537	44151.058

Independent: A2.C6

Dependent: E2.E6

Output: A8

MISCELLANEOUS

@SUMSQ(list) returns $\sum_{i=1}^n (x_i)^2$

@SUMXMY2(list1,list2) returns $\sum (x_i - y_i)^2$

@SUMXY(list1,list2) returns $\sum (x_i y_i)$ can be used for expected values

@RAND returns a random number between 0 and 1

@RAND*9+1 = a random number between 1 and 10

@RAND*1000 = a random number between 0 and 1000

@RANDBETWEEN(M,N) returns a random integer between N and M

Normally distributed or other random numbers can be generated by a method such as the following: @NORMINV(@RAND,70,10) will return a normally distributed random number with mean of 70 and standard deviation of 10.