A Brief Review of Statistics and Microsoft Excel Statistical Functions

This document provides a brief review of basic statistics used in the CAS Limited Attendance Seminar on Reserve Variability. This review also describes implementation of the statistics discussed in Microsoft Excel. Four overall categories are discussed within this document:

1. Measures of central tendency
2. Measures of dispersion
3. Regression functions
4. Probability distributions and simulation of random variables

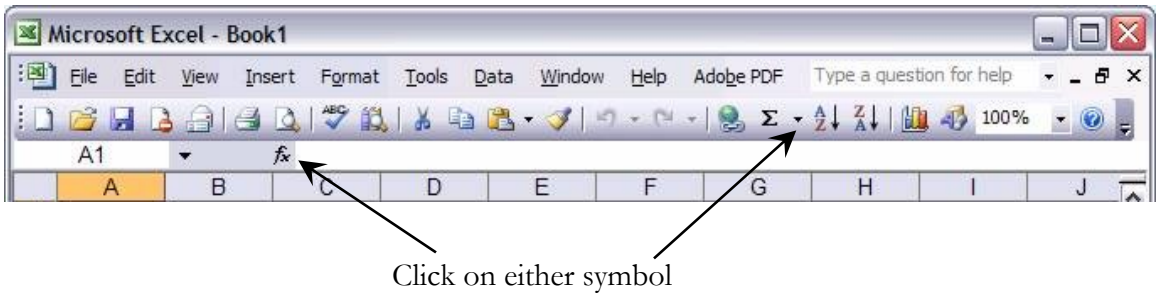
Sample Data for Exercises

Two sample databases are provided for exercises used in this document. They are:

1. A sample of Texas workers compensation claims.[1] The original claims database was downloaded from the Texas department of insurance web site. Only claims in excess of $25,000 are included.
2. Mack Excess General Liability loss triangle. The data is excess casualty automatic facultative general liability data excluding environmental and asbestos compiled by the Reinsurance Association of America in 1991. This data was used by Mack in his papers on modeling loss reserves. The data includes cumulative losses, incremental losses and age-to-age factors. A spreadsheet containing the logs of the data is also provided.

Help with Microsoft Excel Statistical Functions

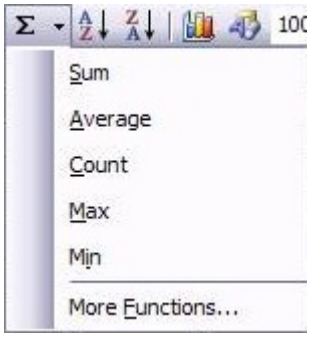
While this document provides an overview of statistical functions in Excel that will be used in the Limited Attendance Seminar, an excellent source of information on the statistical functions are Excel's function and Help menu. Note that the menus displayed and functions described in this document relate to Excel 2003, although other versions of Excel should be similar. In this section, a brief introduction to utilizing Excel's function and Help menus for learning more about specific statistical functions is explained. On the top row of icons on the spreadsheet toolbar, click on the downward tab next to the summation symbol (**Σ**) or the function symbol (***fx***), after moving to a cell on the worksheet where you want to enter a function.



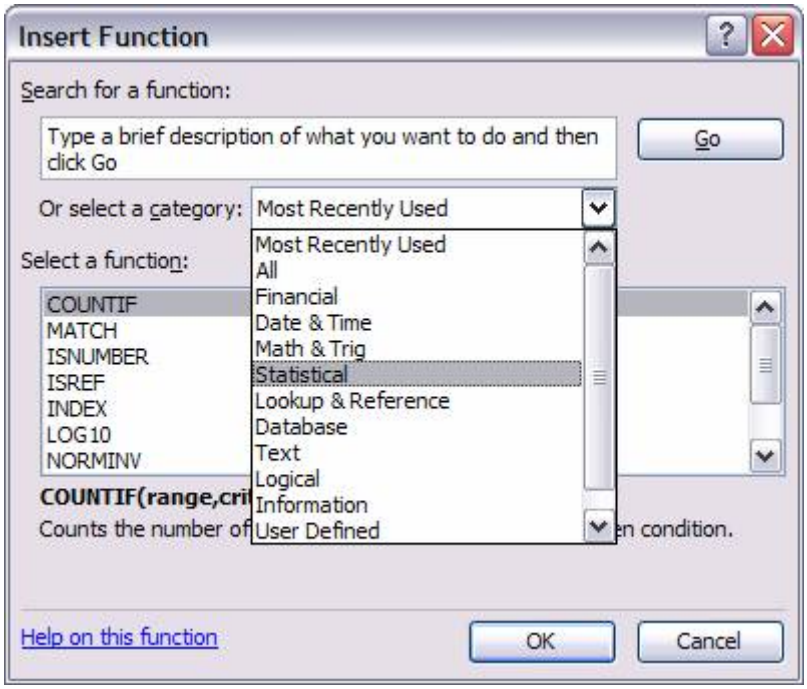
[1] The claims were identified in a claim data base as work related.

A Brief Review of Statistics and Microsoft Excel Statistical Functions

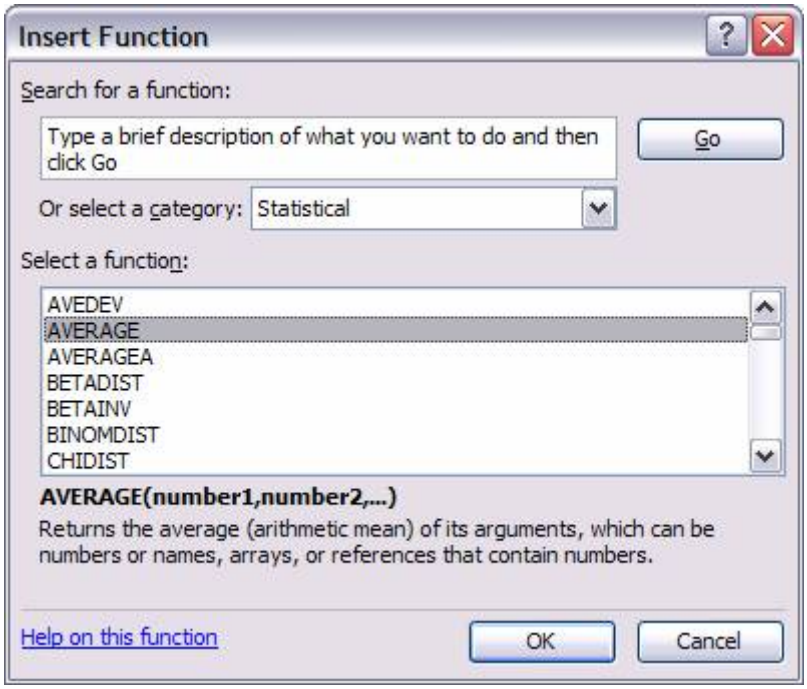
After clicking the ▼ or "down triangle" symbol next to the summation symbol, a list of recently used functions appears. Below the list is a line and then "More Functions". Clicking on "More Functions" causes the "Insert Function" dialogue box to appear that lists the many categories of functions. (Clicking on the function symbol causes the "Insert Function" dialogue box to appear.)



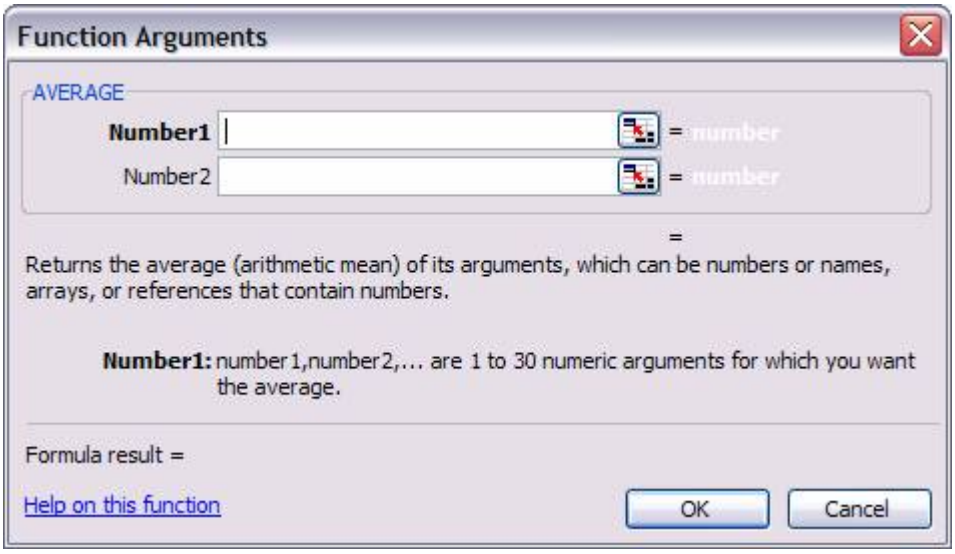
The list of function categories includes statistical functions, the functions we are primarily interested in during the seminar. Highlight the "Statistical" category and click on it.



Next, find a specific statistical function you want information about after clicking on "Statistical". For instance, suppose you want to know how to use the function AVERAGE to compute the mean of a variable. Find AVERAGE in the list and click on it.

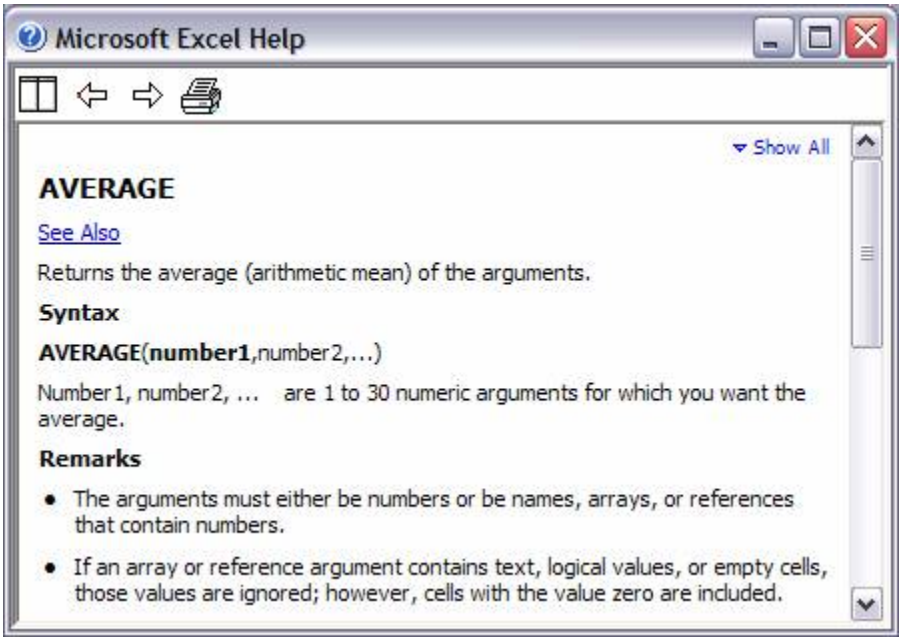


After you have selected the function you want to use (in this case the Average function), clicking on the OK button will enter the function in the cell in the worksheet and a "Function Arguments" dialogue box is displayed which prompts you to enter the data you want to use as function arguments. Below the input box(es) is a short definition of what the function does, as well as a description of the argument box your cursor is in. To learn more about the function and how to use, click on the "Help on this function" link in the bottom left corner. For instance, suppose you wish to find out how to supply a range of data as an argument to the average function.



A Brief Review of Statistics and Microsoft Excel Statistical Functions

After clicking on Help, a dialogue box appears which defines the function and how to use it. The "Microsoft Excel Help" dialogue box may also display one or more examples of its use and possibly some related functions which can be accessed using the "See Also" link.



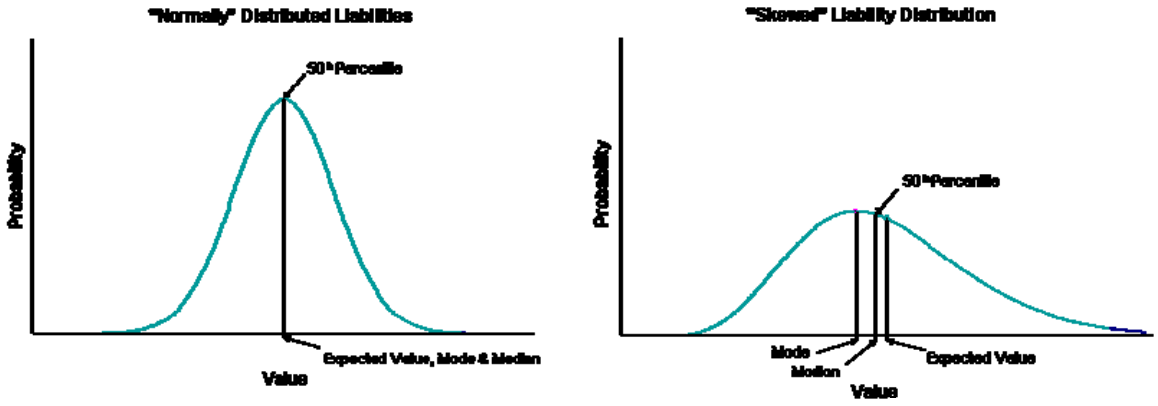
Other help is available from Microsoft's web site: www.microsoft.com. From the main website menu, you can navigate to "Product Families: Office" or "Resources: Support" or "Resources: Knowledge Base". From any of these main navigation routes, you can get more specific help or run a search.

I. Measures of Central Tendency

Measures of central tendency provide a single numeric summary for a probability distribution. Actuaries are most familiar with the mean or average of a distribution, but other measures of central tendency can also be used. For instance, when actuaries select age-to-age factors from a loss development triangle, they are choosing a value (which could represent the average, maximum likelihood, 5-year excluding high/low or some other value) for the future development that will be observed when the claims underlying the triangle mature to their ultimate value.

For symmetric distributions, the mean, median and mode are equal, but for skewed distributions (those which most commonly characterize insurance data) this is not the case and the actuary must choose which statistic best represents the distribution. Note that if the actuary chooses the median for skewed distributions as his/her best estimate, on average, over a large number of analyses, the estimate (of reserves) will usually be below what is actually experienced when extreme values are factored in (see graph below).

A Brief Review of Statistics and Microsoft Excel Statistical Functions



Microsoft Excel Functions for Central Tendency

AVERAGE

Provides an estimate of the mean or expected value of the data. To use it, supply the range of data you want the average of "=AVERAGE(data range)." For example, if you wanted the average of the observations of the data in cells A1 through A10, you might enter =AVERAGE (A1:A10) into cell A12.

MEDIAN

Returns the median or 50th percentile of the data. To use it, supply the range of data you want the median of "=MEDIAN (data range)."

TRIMMEAN

Computing a trimmed mean is a way to temper the influence of extreme values in the estimate. A trimmed mean excludes k% of the data from the calculation, where k is a judgmentally selected proportion of the data to exclude. The calculation excludes the $k/2$% highest and lowest values. To use, supply the range of data you want the trimmed mean of and specify the proportion to exclude. The proportion must be between zero and one "=TRIMMEAN(data range, proportion)."

GEOMEAN

The geometric mean is a measure that is often used for data that are expressed as rates of change (such as the return on stocks or other investments). In a sample of size n, it returns the nth root of the product of all n sample items and is particularly relevant for data that follow the logNormal distribution. "=GEOMEAN(data range)".

WEIGHTED AVERAGES

Actuaries often used weighted averages rather than arithmetic averages. For instance, it is common to compute weighted average loss development factors. The rational behind a weighted average is to produce a more stable estimate than can be obtained from an un-weighted average. When using a weighted average, a common practice is to select a weight

for an observation that is believed to be inversely proportional to its variance so that observations contributing more variability to the estimate are given less weight.

$$\bar{x} = \sum_{i=1}^{N} w_i x_i, \quad w_i \propto \frac{1}{\sigma^2}$$

For instance when computing average severities for a given accident period and development age, those severities based on cells with higher claim volumes will be more stable than the severities based on cells with lower claim volumes. It therefore makes sense to give more weight to the cells with larger claim volumes when fitting a severity model using the data. This can be accomplished by using claim count as the weight variable.

Excel does not have a function that can be used to compute weighted averages. However certain functions are helpful when computing weighted averages:

SUMPRODUCT

Computes the product of two columns or rows. If one of the columns is the weight associated with each observation and the other is the sample values, the weighted average can be computed. "=SUMPRODUCT(weight variable range, sample value variable range)" gives the same result as creating a column that is the product of the column with the first variable times the column with the second variable, and then summing the result. Dividing the SUMPRODUCT by the SUM of the weight variable range results in the weighted average.

SUMIF

Calculates a conditional sum for the values in sum range which met the specified criteria using "=SUMIF(criterion range, criterion, sum range)." For instance, if you wish to sum only the losses at 12 months of maturity for which there also losses at 24 months of maturity (*i.e.*, drop the most recent period from the sum), using the criterion ">0", along with specification of the 24 month values as the criterion range and the 12 month values as the sum range will accomplish this.

Exercises, Section I – Measures of Central Tendency

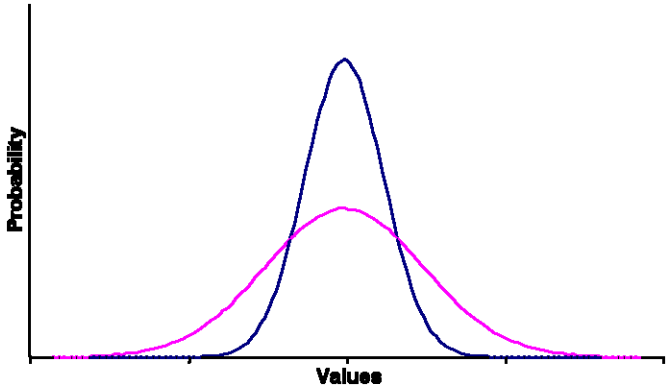
1. Using the Mack data and age-to-age factors from the Mack data, compute the following:
   a. arithmetic mean of age-to-age factors
   b. arithmetic mean of incremental losses
   c. trimmed mean (at 25%) of incremental losses
   d. trimmed mean (at 25%) of age-to-age factors
   e. weighted average of age-to-age factors where the weight is the cumulative incurred losses for the prior development age
   f. weighted average of incremental losses where the weight is the cumulative incurred losses for the prior development age
   g. median age-to-age factors
   h. geometric mean of age-to-age factors

2. Using the WC claims sample data, compute the following:
   a. mean severity
   b. median severity
   c. 10% trimmed mean of severity
   d. mean of logs of severity (compute the mean only on the log scale. do not transform it)
   e. median of logs of severity (compute the median only on the log scale. do not transform it)

Additional resources on measures of central tendency:
The web site listed below provides a brief tutorial on measures of central tendency:
simon.cs.vt.edu/SoSci/Site/MMM/mmm.html

II. Measures of Dispersion

Measures of dispersion supply a summary statistic describing how dispersed the values are in a sample of data. That is, how far from the center of the data do the observations tend to be? For the normal distribution about two thirds of the data lies within one standard deviation of the mean. Thus, if two distributions have the same mean but a different standard deviation, the one with the higher standard deviation displays more dispersion.



STDEV

Supplies the standard deviation of a sample using the formula:

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}},\ N\text{=sample size, } \bar{x} \text{ is the mean}$$

The standard deviation is the square root of the variance. To use it, supply the range of data you want the standard deviation of "=STDEV(data range)." This statistic is also known as the sample standard deviation.

A Brief Review of Statistics and Microsoft Excel Statistical Functions

VAR

Supplies the variance or second moment about the mean of the sample. It is also the square of the standard deviation. To use supply the range of data you want the variance of "=VAR(data range)."

STDEVP

The population standard deviation. That is, the data you apply the function to represents the entire population of a distribution not a sample from it. Do not confuse this with STDEV. In general, the function you will be using is STDEV not STDEVP, as you will be dealing with samples and STDEV supplies the unbiased estimate for the standard deviation of a sample. The formula for STDEVP is:

$$\sigma_p = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}, N = \text{sample size}, \bar{x} \text{ is the mean}$$

AVEDEV

The average deviation is the average of the absolute deviations from the sample mean. The average deviation is a more robust measure than standard deviation as it more resistant to the influence of outliers. To use it, supply the range of data you want the average deviation of "=AVEDEV(data range)."

COEFFICIENT OF VARIATION

This is an important scale free measure of dispersion. That is, when the values of one data series are higher in magnitude than the values of another data series, the standard deviations will be higher because of the effect of scale. For instance, when money is on a dollar scale it is 1/100 the scale of money measured in cents and therefore has a much smaller standard deviation. A way to remove this effect and compute relative dispersion is to divide the standard deviation by the mean. If your estimate of the mean of the distribution is the average, the Excel formula for the computation is "=STDEV(data range)/AVERAGE(data range)."

SKEWNESS

Measures how far the data departs from a symmetric shape. It is the third moment about the distribution's mean divided by the cube of the standard deviation. A symmetric distribution such as the Normal has a skewness of zero. Most insurance distributions are positively skewed. Note that age-to-age factors, because they cannot be less than zero, are necessarily skewed. Some asset distributions are negatively skewed. Use the function "=SKEW(data range)" to compute the skewness of your data.

KURTOSIS

Measures the heavy "tailedness" (or light "tailedness") of a distribution. It is the normalized fourth moment about the mean. Insurance data tend to be heavily tailed. Use the Excel

function "=KURT(data range)" to compute the kurtosis. The Normal distribution has a kurtosis of 3. Note than some statistical software packages (not Excel) subtract 3 from the kurtosis before reporting it.

Exercises, Section II – Measures of Dispersion

3. Using the age-to age factors from the Mack data, compute the following:
   a. sample standard deviation of age-to-age factors
   b. coefficient of variation of age-to-age factors
   c. population standard deviation of age-to-age factors
   d. sample standard deviation of incremental losses at each age
   e. population standard deviation of incremental losses at each age
   f. average deviation of incremental losses at each age
   g. which is larger, the sample standard deviation or the average deviation?

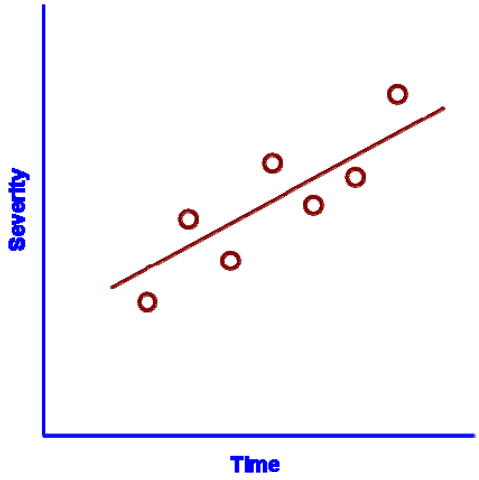
4. Using the WC claims sample data, compute the following:
   a. sample standard deviation of severity
   b. population standard deviation of severity
   c. coefficient of variation of severity
   d. skewness of severity
   e. kurtosis of severity

Additional resources on measures of dispersion:
The concepts of dispersion reviewed in this section are also covered in any introductory statistics text. Additional material is available for free from many educational web sites. The web site listed below provides a brief tutorial on measures of dispersion:
http://147.134.144.30/knudsen/DISPERS/sld001.htm;
http://pse.cs.vt.edu/SoSci/converted/Dispersion_I/activity.html

Hayne in "Estimation of Statistical Variation in Development Factor Method", *Proceedings of The Casualty Actuarial Society*, 1985, pp 25 – 43, describes a procedure for modeling the variability of loss development factors, utilizing the variances of the logs of loss development factors. This is an excellent reference for procedures that will be covered in the course.

A Brief Review of Statistics and Microsoft Excel Statistical Functions

III. Regression



Linear models are the backbone of much of statistical analysis. An example of a linear regression model is fitting a line to claim severities observed over a number of years in order to estimate the severity trend. Linear models are of the form:

$$Y = \alpha + \beta X + \varepsilon$$

where X is a predictor or independent variable, such as accident year, Y is a dependent variable such as claim severity and ε denotes a random error from an assumed probability distribution (although distribution free methods of regression exist, they will not be covered here). The parameters α and β are the constant and coefficient of the regression line.

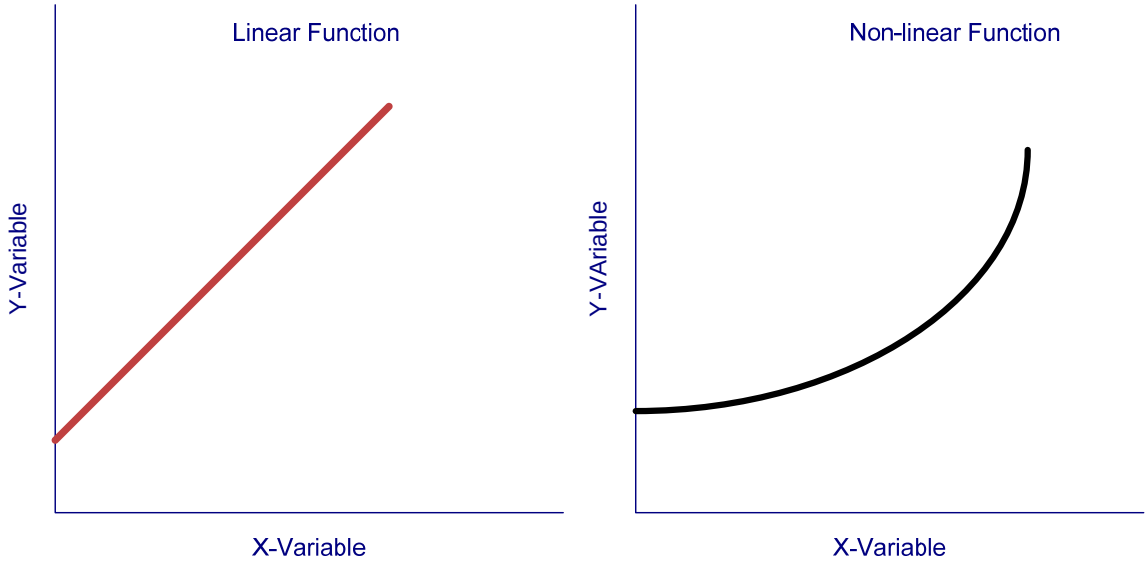
If X is categorical rather than numeric, the linear model is referred to as an analysis of variance (ANOVA) and X is coded as a series of binary dummy variables. If for instance the dummy variable denoted gender, the binary dummy variable might be coded as 1 for females and 0 for males. With categorical variables that have more than two categories, there is one less dummy variable than the number of categories. An example of an ANOVA model in reserving might be the additive model for loss development (see A Simulation Test of Prediction Errors of Loss Reserve Estimation Techniques, James Stanard, *Proceedings of the Casualty Actuarial Society*, 1985) where the expected development for each development age is the average of the observed incremental amount in the historic data for that age. Thus, each development age constitutes a category of the independent variable in the model.

A regression can have more than one predictor variable, in which case it is called a multiple regression model.

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

A Brief Review of Statistics and Microsoft Excel Statistical Functions

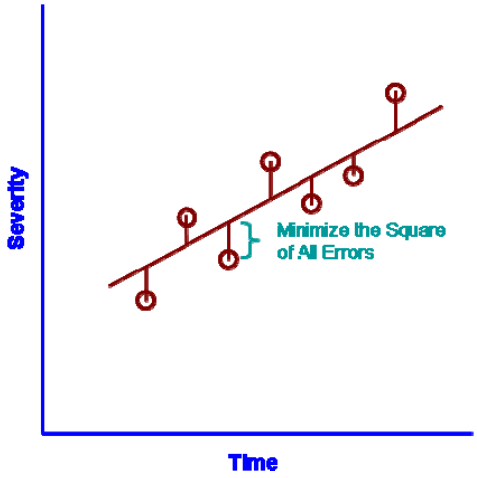
The word "linear" indicates that the relationship between the explanatory variables and the target variable is linear.



If there is a single explanatory variable, the Expected (or Fitted) values will all be on a straight line. Sometimes, simple transformations, such as the logarithmic transformation can be used to create a new explanatory variable (*e.g.*, the log of X) that has a linear relationship with the dependent variable. In practice, linear models are immensely rich and have been used in nearly every field of scientific work. The challenge is to select a data set that is appropriate, or adjust the data set so that the linear assumption is reasonable.

How is a Regression Fit?

Ordinary least squares regression (the most common approach) minimizes the sum of the squared deviations between the actual and fitted values. This is also referred to as the sum squared error (SSE). In the case of a regression with one independent variable, the least squares formulas are simple enough to be programmed into a spreadsheet, adding only a few extra columns to the computational steps. Unless you are interested in gaining experience with the formulas and insight into how they fit your data, you will use the functions built into Excel or into the statistical software you are using. The formulas for multiple regressions are more complicated and in general, require matrix mathematics for their solution. These formulas can also be accessed through Excel functions and the Data Analysis ToolPak of Excel.



The Excel regression functions are:

FORECAST

The formula "=FORECAST(target *x*, known *y*'s, known *x*'s)" is for forecasting linear tends. Typically, the known *x*'s are time periods. The known *y*'s are the data you are forecasting and target *x* is the value of the independent variable you are forecasting *y* for. For instance, an actuary may have a time series of annual claim severities (the *y*'s) by accident year for 1995 through 2005 (the *x*'s) and want to forecast the severity for 2007 (the target *x*).

RSQ

Using "=RSQ(known *y*'s, known *x*'s)" returns the coefficient of determination which is also the square of the linear correlation coefficient between the variables *y* and *x*. The linear correlation coefficient is also known as the Pearson correlation coefficient and it measures the magnitude of the linear co-movement between two variables. Non-parametric measures of correlation which are useful when a relationship that is not necessarily linear between two variables also exist, but will not be discussed here.

The square of the correlation coefficient is used as a rough guide for the goodness of fit for a linear regression function. A rule of thumb is that the percentage of variance in the dependent variable explained by the regression equals $\hat{\rho}^2$.[2] Its formula is:

$$R^2 = \hat{\rho}^2 = \left(\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2 \Sigma(y-\bar{y})^2}} \right)^2$$

When assessing the overall fit of a regression with multiple variables, the "*x*" variable in the R^2 formula is the fitted value from the regression.

[2] It is common in statistics to denote an estimate by placing a "hat" symbol over the parameter. Thus the estimate of ρ is denoted $\hat{\rho}$.

A Brief Review of Statistics and Microsoft Excel Statistical Functions

CORREL

Using "=CORREL(*y*-range, *x*-range)" provides the Pearson correlation coefficient between two random variables *y* and *x*. The correlation measures linear co-movement between the two variables. The correlation coefficient is also equal to the square root of the R^2 statistic. It is possible the two variables are strongly related, but the relationship is not linear. Such relationships may yield low correlation coefficients.

INTERCEPT

The formula "=INTERCEPT(known *y*'s, known *x*'s)" is also known as the constant. It equals the parameter α in the regression equation specification.

SLOPE

The formula "=SLOPE(known *y*'s, known *x*'s)" equals the parameter β in the regression equation specification. Unlike, the TREND or LINEST functions there is no option to exclude a constant or intercept.

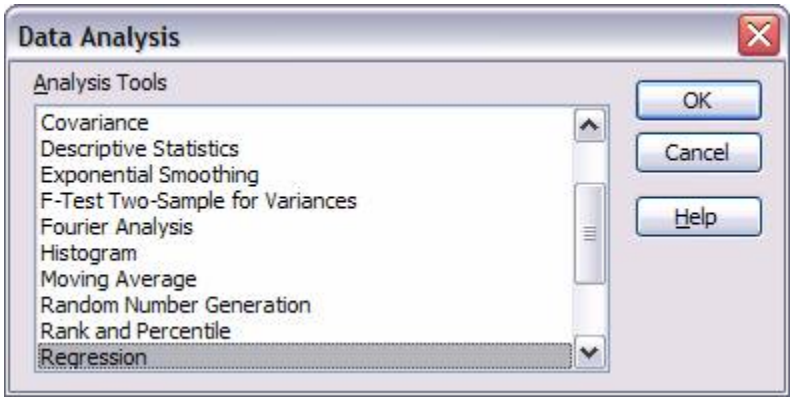
TREND

Using "=TREND(known *y*'s, known *x*'s, new *x*'s, constant)" is very much like the FORECAST function. It will fit a linear trend line to a series of independent and dependent variables specified separately and then forecast one or more new values (the new *x*'s). You will need to specify whether the trend line includes a constant (TRUE, or omitted, for yes on constant).

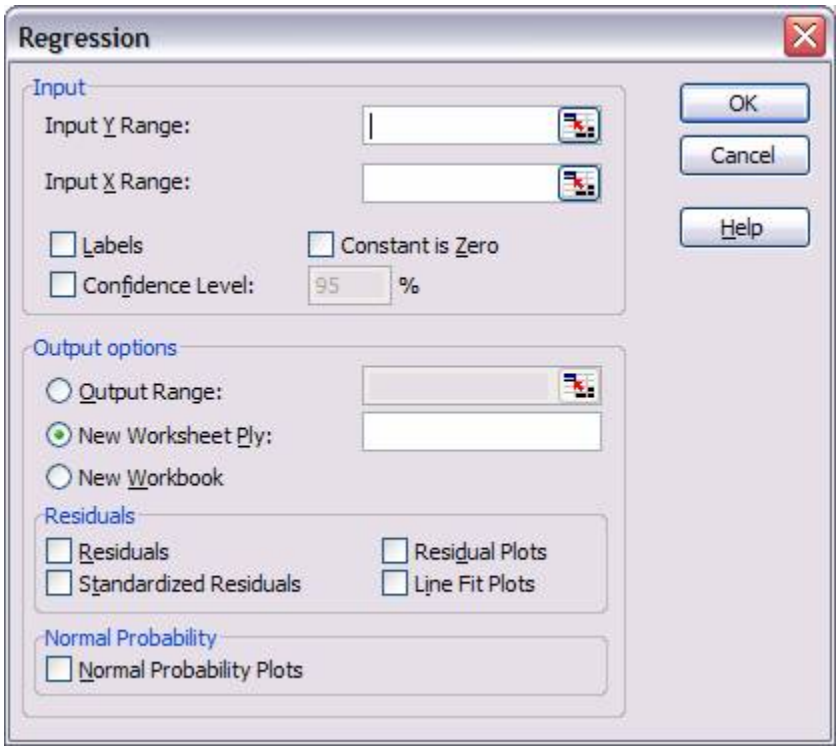
LINEST

The formula "=LINEST(known *y*'s, known *x*'s, constant, stats)" is the general linear regression function with a number of parameters that determine its output. The know *y*'s specifies the range of the dependent variable. The known *x*'s specifies the range of the independent variables, which must all be located in a single contiguous range. Constant specifies whether the linear regression has a constant (TRUE means "yes" and FALSE means "no" constant). Stats is a logical value specifying whether you want to return additional statistics. If it is TRUE, LINEST will return additional statistics. The additional statistics include the standard error for the coefficients R^2, the standard error of the *y* estimate and the F-statistic for the regression. To compute only the slope of a single variable regression use "=LINEST(known *y*'s, known *x*'s)".

To access Regression from the Data Analysis ToolPak, click on "Tools" from the Main Menu, find "Data Analysis…" and click on it. A "Data Analysis" dialogue box listing a number of statistical procedures will appear:



Find the regression option and click on it. Another dialogue box appears:



In the "Regression" dialogue box, there are two boxes which must be filled in; the "Input Y Range" and the "Input X Range." Thus the range of data points with the dependent variable will be entered in the Input Y Range. The range of data points for the independent variable(s) will be entered in the Input X Range. Since there may be two or more columns containing independent variables, all variables must be located next to each other. That is, all the data for the independent variable must be in a contiguous range in Excel. Also, there must be the same number of rows for the independent variables as for the dependent variable.

By filling in an output range you can specify an output range. By clicking the box for Residuals, Standardized Residuals, Residual Plots and Line Fit Plots and Normal Probability Plots you can obtain goodness of fit statistics that help you assess the quality of the regression model.

Certain techniques used in loss reserving can be formulated as regression models. For instance Venter (in Testing the Assumptions of Age-To-Age Factors, *Proceedings of the Casualty Actuarial Society*, 1998) pointed out that the chain ladder method used frequently in loss reserving is equivalent to a linear regression with no constant.

$$Y = \beta X + \varepsilon$$

where Y denotes the incremental losses for a given age, X denotes the cumulative losses (for the same accident year) for the previous development age, and β is the age to age factor (minus 1) for a specific development period. Thus, the development factor can be estimated by fitting a regression with no constant. Other models underlie many other reserving approaches.

Exercises, Section III – Regression

1. Use the Trend function to compute the 1991 value for the Mack incremental losses at ages 1, 2, and 3. Use a constant in your trend function.

2. Compute the correlations up to age 7 between each set of cumulative losses and the prior development cumulative losses for the Mack data

3. Compute the correlations up to age 7 between each set of incremental losses and the prior development cumulative losses for the Mack data

4. Compute the correlations up to age 6 between each set of age-to-age factors and the prior development period age-to-age factors for the Mack data

5. Extra credit question: Are the correlations in 4, significant?

Additional resources on regression:
The regression procedure used by the Analysis Toolpak of Excel produces a number of statistics such as T-statistics for the coefficients, an F-statistic, R^2 and an adjusted R^2. A short discussion of these and several other regression statistics can be found in the following source: A Note Regarding the Evaluation of Multiple Regression Models, Gregg Alf, *Proceedings of the Casualty Actuarial Society*, 1984.

IV. Probability Distributions and Simulation of Random Variables

Because models of reserve variability are stochastic models, that is, based upon probabilistic assumptions about the distribution of key underlying components of the model, probability

distributions play a key role. The following are probability distributions we will be using in the Limited Attendance Seminar.

A. Continuous Distributions

UNIFORM Distribution

A common distribution in introductory statistics courses is the Uniform distribution. For instance if data is uniformly distributed between 0.0 and 1.0, it is equally likely that any number in this range will occur and its probability density function (PDF) is 1 and its cumulative distribution function (CDF) is x. If a number is uniformly distributed between two constants a and b, its PDF is $\frac{1}{b-a}$ and its CDF at x is $\frac{x-a}{b-a}$.

UNIFORM Random Number Generation

A fundamental component of simulation is the generation of numbers which behave like samples from a probability distribution. In Excel the function for generating Uniform(0,1) numbers is RAND. To use it, type "=RAND()". No argument is needed by the function.

To generate a random number between a and b, use the formula "=RAND()*(b-a)+a". Note: Every time you recalculate in Excel (which is every time you type a number or formula into Excel if recalculation is set on automatic), the random numbers will change. If you want to freeze them, range value your random numbers by first copying them and then using Edit, Paste Special, Values, while leaving the cursor at the position of the first random number. The RAND function is found in the "Math & Trig" category rather than the "Statistics" category of the "Insert Function" dialogue box.

Using "=RANDBETWEEN(a,b)" will generate a random number between a and b, but you may need to have the Analysis ToolPak add-in installed.

NORMAL Distribution

If you want to know the cumulative probability of a Normal random variable, use the NORMDIST function. A variant of this function, NORMSDIST, provides the cumulative probability for a standard Normal variable. Thus "=NORMSDIST(1.645)" returns 0.95. The PDF for the Normal distribution is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

For "=NORMSDIST(x)", x is a standard Normal variable value. Alternatively, for "=NORMDIST(x, mean, standard deviation, cumulative)" x is a value which depends on the mean and standard deviation used for the Normal variable. If you want the probability density at x, rather than the cumulative probability, supply FALSE for cumulative.

A Brief Review of Statistics and Microsoft Excel Statistical Functions

Using "=NORMINV(probability, mean, standard deviation)" generates the inverse of a Normal variable at the cumulative probability specified. The mean and standard deviations are typically the values that would be obtained from the AVERAGE and STDEV functions applied to a sample of data, but they could be theoretical parameters selected *a priori*. Thus, if you wanted the value of a standard Normal (mean=0, standard deviation = 1) variable at its 95th percentile use "=NORMINV(.95, 0, 1)" and 1.645 (when rounded to three places) will be returned. Thus, the NORMINV function allows us to find percentiles of the Normal probability distribution.

One of the ways this is helpful is in computing confidence intervals around estimates. For instance, suppose we estimated the mean of Normally distributed data and wanted the 95% confidence interval for our estimate. The lower end of the confidence interval is found as "=NORMINV(0.025, mean, standard deviation)". The upper end of the confidence interval is found as "=NORMINV(0.975, mean, standard deviation)". Please note, however, that property and casualty insurance data typically does not follow a Normal or symmetrical distribution.

The easiest way to generate a random Normal variate is to first generate a uniform random number and then use the NORMINV function – *e.g.*, "=(NORMINV(RAND(), mean, standard deviation)". Thus, a common way to generate a random variable from a probability distribution is to generate its percentile from a uniform random distribution using RAND and then apply the inverse function for the probability distribution of interest to compute the value of a random variable.

LOGNORMAL Distribution

Using "=LOGNORMDIST(*x*, mean log scale, standard deviation log scale)" gives the cumulative probability of a logNormally distributed variable with mean and standard deviation being the moments of the log of the random variate. That is the actual mean is $e^{(\mu+\frac{1}{2}\sigma^2)}$. The probability density function for the logNormal distribution is:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

Using "=LOGINV(probability, mean log scale, standard deviation log scale)" gives the inverse of a logNormally distributed variable at the specified cumulative probability with mean and standard deviation being the moments of the log of the random variate.

GAMMA Distribution

Using "=GAMMADIST(*x*, *k*, *theta*, cumulative)" returns the cumulative probability of a Gamma random variable with parameters *k* and *theta*. To compute the density value at *x* rather than the cumulative probability, supply FALSE as the last argument. The PDF for the Gamma distribution is:

$$f(x) = \frac{1}{\theta^k \Gamma(k)} x^{k-1} e^{-\frac{x}{\theta}},\ x > 0$$

Or, alternatively: $f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},\ x > 0$ with $\alpha = k$ and $\beta = 1/\theta$.

Using "=GAMMAINV(probability, *k*, *theta*)" returns the inverse of a Gamma random variable with parameters *k* and *theta*. Note that for the parameterization of the Gamma used by Excel[3], the mean of the distribution equals *k* * *theta*.

PARETO Distribution

Many commonly used distributions do a relatively poor job of approximating the "tails" of distributions. The Pareto distribution, a relatively "heavy-tailed" distribution, is frequently used to model large claims, or claims larger than some threshold value, k. The cumulative and density distribution functions for the Pareto are:

$$F(x) = 1 - \left(\frac{K}{x}\right)^q;\ \ f(x) = \frac{q}{K}\left(\frac{K}{x}\right)^{q-1},\ K < x < \infty$$

Microsoft Excel does not have a function for the Pareto distribution, but a random variable from the Pareto is easily simulated using the inverse of the cumulative distribution, along with a uniform random variable. The formula is "=*K*/(1-RAND())$^{(1/q)}$." See A Practical Guide to Single Parameter Pareto Distribution, Stephen Philbrick, *Proceedings of the Casualty Actuarial Society*, 1985, for an introduction to the Pareto distribution and a simple method for estimating its parameters.

T Distribution

Using "=TDIST(*x*, degrees of freedom, tails)" returns the cumulative probability of a Students T random variable at *x* with the specified degrees of freedom. You must specify whether you want the one or two tailed probability where 1= one tailed and 2 = two tailed. TDIST does not take a value below zero. The cumulative probability of a negative T-variate is 1.0 minus the cumulative probability of the absolute value of the variate. The probability density function for the T distribution is:

$$f(x) = \frac{\Gamma((v+1)/2)}{\sqrt{v\pi}\,\Gamma(v/2)(1 + x^2/v)^{(v+1)/2}},\ \ -\infty < x < \infty$$

[3] Note that the Gamma functions in Excel only use single precision in their calculations which can return errors in some situations. For a wonderful example of double precision Gamma functions (as well as many other functions) see: members.aol.com/iandjmsmith/examples.xls.

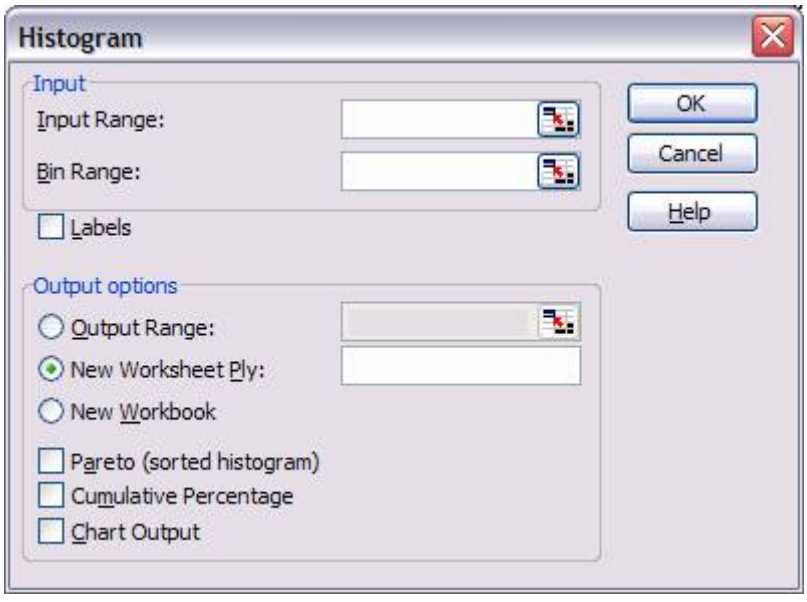
Using "=TINV(probability, degrees of freedom, tails)" returns the inverse of a Students T random variable with the specified degrees of freedom. You can only obtain the two tailed probability percentile from this function. This reflects the common usage of the Student's T for hypothesis testing and constructing confidence intervals.

EMPIRICAL Distribution

A continuous distribution can often be approximated by a discreet Empirical distribution. Given a sample from a continuous distribution, the empirical cumulative distribution is computed as follows:

$$F_n(x) = \frac{\text{count of records} \le x}{n + 0.5}$$

Note that a denominator of n+0.5 is often used in statistics rather than n.[4] This is a correction that allows for an observation in the population to be higher than any sample value. When simulating from an empirical distribution, it will be necessary to select a maximum value for the distribution, where the empirical CDF is 1.0. It is common to use frequency tables and histograms to group the distribution information into intervals and create a grouped empirical probability distribution. It is easiest to use to histogram feature of the Analysis Toolpak to compute the frequencies for the binned data. It will be necessary to select the upper endpoints of the intervals used for binning when using the Histogram procedure:



The endpoints of size intervals are supplied in the box "Bin Range" and the data to be grouped is supplied to the histogram procedure in "Input Range". If the box "Chart

[4] Other adjustments are also used.

Output" on the bottom is not clicked, only a table displaying the frequencies in each bin is displayed.

To simulate values from an empirical sample of data use the percentile function. "=PERCENTILE(data range, *p*)" will return the p^{th} percentile of the sample specified in the data range. The parameter *p* can be a random uniform number generated with the RAND function.

Exercises, Section IV.A. – Continuous Distributions

1. Simulate 100 random variables from a Uniform[0, 1] distribution and compute the mean of:
   a. the first 10 simulated observations
   b. all 100 simulated observations

2. Using the mean and standard deviation of the age 1-2 Mack development factors computed from Exercises I and II, simulate 100 observations from a Normal distribution

3. Simulate 100 random variables from a Pareto distribution with parameter *K* equal to $100,000 and Parameter *q* equal to 1.2. Calculate the mean, standard deviation and coefficient of variation.
   a. Limit the random Pareto variables to $250,000 and compute the mean, standard deviation and coefficient of variation of the limited simulated variables.
   b. Use the random Pareto variables in excess of $250,000 and compute the mean, standard deviation and coefficient of variation of the limited simulated variables.

4. Create a histogram of the data in the WC sample

5. Simulate 100 random variables from an Empirical distribution derived from the sample workers compensation severities. Compute the mean of the sample.

6. Create a histogram of the simulated WC data
   a. compare it to the histogram in #4

B. Claim (Discrete) Random Variables

POISSON Distribution

Using "=POISSON(*x*, λ, cumulative)" returns the cumulative probability of a Poisson random variable with parameters *x* and λ for the number of events and expected number of events, respectively. To compute the density value at *x* rather than the cumulative probability, supply FALSE as the last argument. The formulas for the Poisson are:

A Brief Review of Statistics and Microsoft Excel Statistical Functions

PDF: $f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$; CDF: $F(x) = \sum_{k=0}^{x} \frac{e^{-\lambda}\lambda^k}{k!}$, where $x = 0,1,2,\ldots$

NEGATIVE BINOMIAL Distribution

Using "=NEGBINOMDIST(*x*, *r*, *p*)" returns the probability of a Negative Binomial random variable at *x* with parameters *r* (number of failures) and *p* (probability of success). The PDF formula for the Negative Binomial distribution is:

$$f(x) = \binom{x+r-1}{r-1} p^r (1-p)^x,\ 0 < p < 1 \text{ and } r > 0$$

EMPIRICAL Discrete Distributions

When you want to find the cumulative probability for an empirical discrete distribution you will need to create a table in Excel with the distributions values in the first column and the cumulative distribution in the second column. You will then need to use the VLOOKUP function (alternatively HLOOKUP can also be used if you create a horizontal rather than a vertical table).

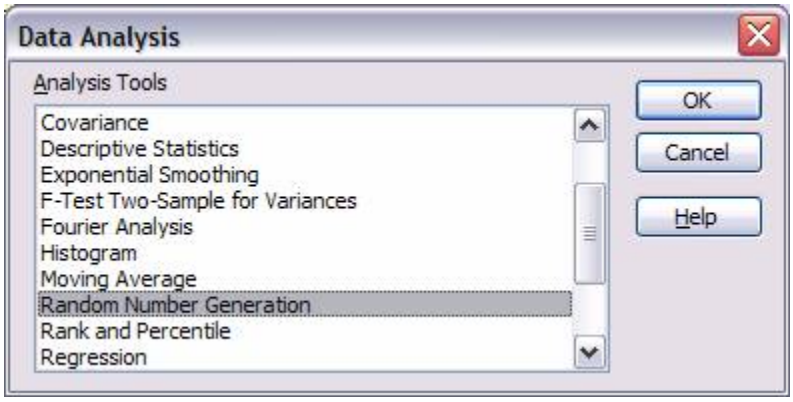
Using "=VLOOKUP(lookup value, table range, column number)" will return the value from a table associated with the lookup value specified. For example, let the first column of the table contain the cumulative probability of each of the claim amounts and the second column values of a discrete claims distribution, such as 0, 1, 2, *etc.*:

Cumulative Probability	Value
0.0	0
0.1	1
0.2	2
0.3	3
0.4	4
0.5	5
0.6	6

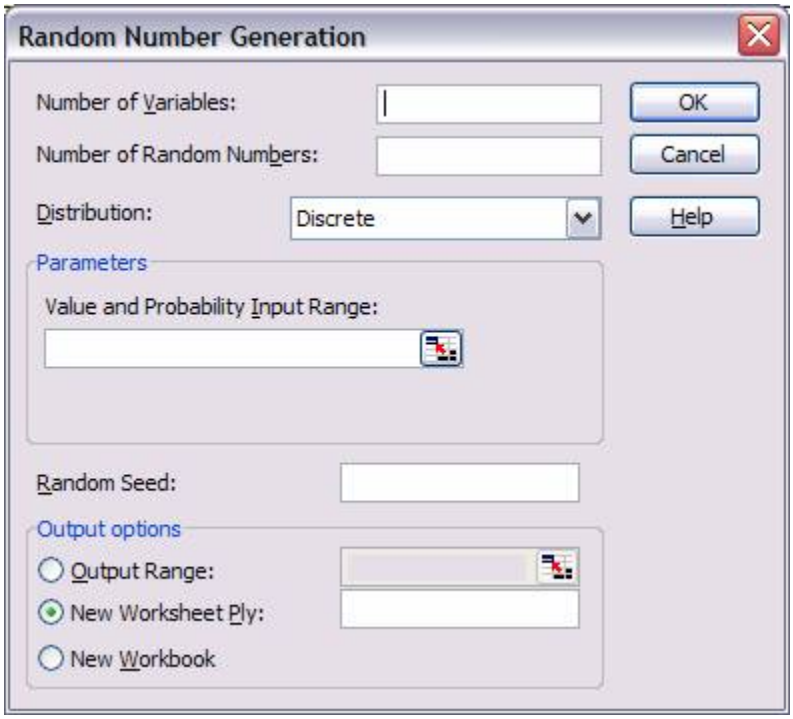
Then the formula "=VLOOKUP(RAND(), table range, 2)" would return the value of 3 when the random value was greater than or equal to 0.3 and less than 0.4. In this example, each of the claim values has a PDF of 0.1, except for 6 which has a PDF of 0.4.

Random Variable Generation: ANALYSIS TOOLPAK

A second way to generate random variables is to use the Analysis ToolPak.[5] To access the Analysis ToolPak, click on "Tools" on the main menu, and then click on the "Data Analysis…" option. A "Data Analysis" dialogue box displaying a list of the analytical procedures available appears, click on "Random Number Generation."

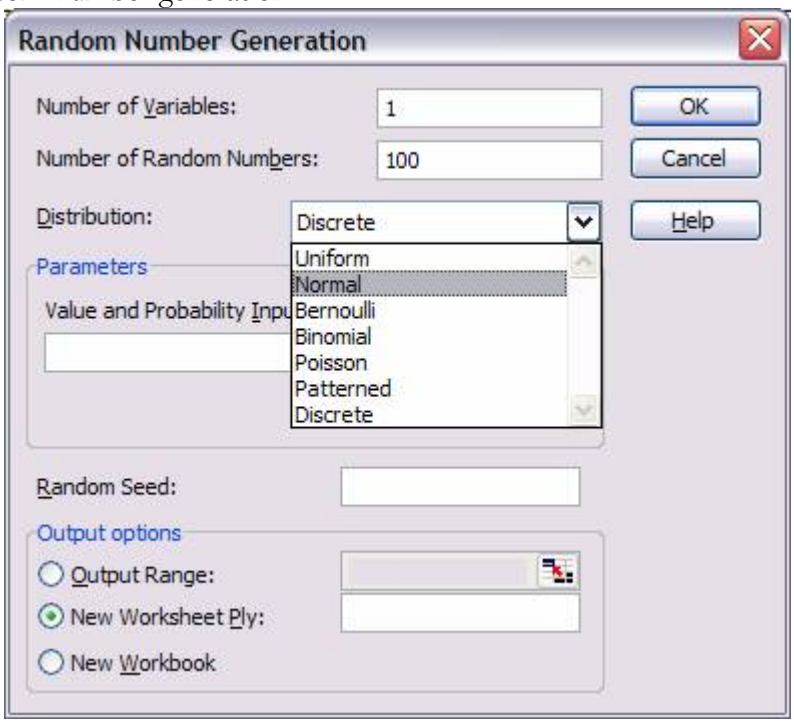


A "Random Number Generation" dialogue box displaying the arguments you need to supply the procedure appears.

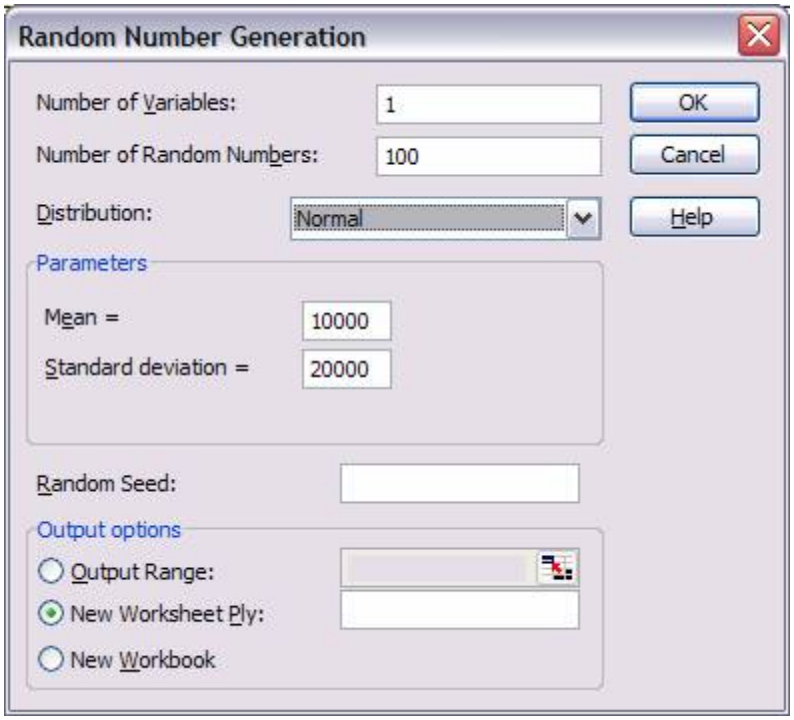


[5] While the Analysis ToolPak is provided with Excel, it is not always installed when Microsoft Office is initially installed on your computer. Some readers might need their "Office" CD's and passwords at hand to be able to install the ToolPaks.

You must fill in the boxes that specify the number of random variables you are generating, the number of randomly generated values for the variables and the distribution that you wish to use for random number generation.



Once you have selected a distribution, an additional box will appear where you will enter the parameters for the distribution.



A Brief Review of Statistics and Microsoft Excel Statistical Functions

COLLECTIVE RISK MODEL

This is a well known approach to constructing a probability distribution for insurance loss amounts. It assumes losses are the result of 1) a random process that generates the number of claims and 2) a random process that generates random severities for each of the random claims. Thus, total losses are the convolution of a random claim count N, and N random claim severities. Where L is a variable denoting total aggregate losses, the process can be denoted as follows:

$$L = X_1 + X_2 + X_3 + \ldots + X_N$$

Where N, the claim count is a random variable from a discrete distribution such as the Poisson or Negative Binomial and each X is an random variable, typically from a continuous distribution such as the logNormal or Pareto.

An algorithm for the collective risk model is as follows:

1. Generate N random claim counts from a claim probability distribution
2. For each random claim, generate a random claim severity from a claim severity distribution
3. Apply any per claim limits to each random claim severity
4. Sum the limited claim severities
5. Apply any aggregate limits to the summed limited losses

Repeat steps 1-5 many times (typically 1,000 or more) and tabulate the results. Compute means, standard deviations, percentiles and other statistics for the total loss distribution. The algorithm is readily implemented using simulation.

Exercises, Section IV.B. – Claim (Discrete) Random Variables

1. Simulate 100 claims (records) from a Poisson with mean 3.

2. For each of the 100 simulated Poisson claim counts, N, simulate N logNormal severities with location parameter, μ equal to 9.5 and scale parameter σ equal to 2.

3. For each of the simulated Poisson-logNormal combinations, sum the random severities to derive total aggregate losses.

4. Compute the mean and standard deviation for the 100 count/severity combinations (i.e., the aggregate loss simulations).

5. Prepare a histogram of the simulated observations in step 4.

A Brief Review of Statistics and Microsoft Excel Statistical Functions

V. Parameter Risk

Because parameter risk plays an important role in the total variability of unpaid claim reserves, this section provides a brief introduction to the topic. In this section, parameter risk is treated from a Bayesian perspective. In the actuarial literature, the universe of risk in outstanding claim estimates is usually divided into three major types:

- *Process Risk* – The inherent variability of the process; for example the variability of the ultimate losses for a line of business, for one year around its "true" mean. ASOP 43 defines it as "The risk that actual results of a stochastic process differ from projections, when the parameters are known with certainty".

- *Parameter Risk* – The risk that the model's parameter has been under or overestimated, i.e., that it is not known with certainty. For instance if the analyst estimated the mean and standard deviation of an ultimate loss distribution, these estimates are unlikely to be the "true" parameter. Rather the estimates themselves are random variables with a probability distribution. ASOP 43 defines it as "The risk associated with the estimation of the parameters that underlie methods and models".

- *Model Risk* – The risk that the model used to develop estimates is a poor approximation of the actual experience. For instance, the lognormal distribution may have been used to model claim severity, but the "true" distribution might be the Pareto distribution. ASOP 43 defines model risk as "The risk that the methods or models are not representative of how the claim amounts will emerge and develop".

A classic example:
A classic example of parameter risk is provided by Dropkin (1959) and Mayerson (1964). The simple example relates to claim frequency in automobile insurance. The example assumes that automobile claim frequency follows a Poisson distribution. The Poisson is a distribution with one parameter, which we denote λ. In addition, the parameter, λ is assumed to vary between policyholders. In Dropkin and Mayerson's example, λ is distributed according to the gamma distribution. The gamma is also referred to as the prior distribution for the parameter. The Poisson is a conditional distribution, as the particular distribution is dependent on the value of λ. To compute the unconditional distribution of claim frequency for the data, one must integrate over all possible values of the random parameter using the prior distribution:

$$\int_{-\infty}^{\infty} f(n/\lambda)\, f(\lambda)\partial\lambda =$$

$$\int_{-\infty}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} f(\lambda)\partial\lambda = \int_{-\infty}^{\infty} \frac{\lambda^n}{n!} e^{-\lambda} \left(\frac{\Gamma(\alpha)}{\beta^\alpha}\lambda^{\alpha-1}\right)\partial\lambda = \frac{1/\beta}{1+1/\beta}\binom{-\partial}{n}\left(\frac{-1}{1+1/\beta}\right)^{-\alpha}$$

A Brief Review of Statistics and Microsoft Excel Statistical Functions

The result is that the unconditional distribution (that is, the distribution that reflects parameter variance) of claim frequency follows a negative binomial distribution. This distribution is more dispersed than the Poisson. A straightforward way to use simulation to approximate the unconditional distribution is as follows:

- Simulate the parameter λ' from a gamma distribution, $\Gamma(\alpha, \beta)$ where the mean of the distribution = λ
- Simulate a random claim frequency, *n*, from the distribution Poisson(λ')
- Repeat many times

The resulting simulation outcomes are an approximation to the unconditional distribution and reflects both process and parameter risk under the assumed prior distribution for the parameter.

Exercises, Section V – Parameter Risk

1. Assume claim counts follow the Poisson distribution with a mean of 100 and can be approximated by a normal distribution. Assume that the Poisson parameter λ follows a gamma distribution with parameter α =5. Simulate 100 random claim counts from the unconditional distribution.

2. Create a histogram of the distribution you simulated.

Additional resources on simulation:
For further assistance on generating random variables, go to the Microsoft office website by clicking Help, then Office on the Web. Then type in "generate random variables".

A number of training options appear. Select "Introduction to Monte Carlo simulation".

Other training sponsored by the CAS:
The Committee on the Theory of Risk has prepared a one day course introducing actuaries to statistical procedures which can be used to test the assumptions of loss development models. This class does not cover simulation or other methods of modeling variability but focuses on key statistical assumptions underlying reserving procedures. The class is available to the CAS's Regional Affiliates and as a Limited Attendance Seminar. The class is an excellent compliment to the Limited Attendance Seminar on Reserve Variability.

A Brief Review of Statistics and Microsoft Excel Statistical Functions

References:

All references are noted in a separate Bibliography for the Limited Attendance Seminar on Reserve Variability and are generally available on the CAS web site, www.casact.org.