Glossary

The ISI glossary of statistical terms provides definitions in a number of different languages: http://isi.cbs.nl/glossary/index.htm

Adjusted r^2 Adjusted R squared measures the proportion of the variation in the dependent variable accounted for by the explanatory variables.

Aggregate price index A measure of the value of money based on a collection (a basket) of items and compared to the same collection of items at some base date or a period of time.

Alpha, α Alpha refers to the probability that the true population parameter lies outside the confidence interval. Not to be confused with the symbol alpha in a time series context i.e. exponential smoothing, where alpha is the smoothing constant.

Alternative hypothesis (H_1) The alternative hypothesis, H_1, is a statement of what a statistical hypothesis test is set up to establish.

Analysis of variance (ANOVA) Analysis of variance is a method for testing hypotheses about means.

Arithmetic mean The sum of a list of numbers divided by the number of numbers.

Autocorrelation Autocorrelation is the correlation between members of a time series of observations and the same values shifted at a fixed time interval.

Base index A value of a variable relative to its previous value at some fixed base.

Beta, β Beta refers to the probability that a false population parameter lies inside the confidence interval.

Binomial distribution A Binomial distribution can be used to model a range of discrete random data variables.

Bonferroni t test The Bonferroni test is a statistical procedure that adjusts the alpha level to allow multiple t tests to be used following the ANOVA.

Box plot A box plot is a way of summarizing a set of data measured on an interval scale.

Box-and-whisker plot A box-and-whisker plot is a way of summarizing a set of data measured on an interval scale.

Categorical variable A set of data is said to be categorical if the values or observations belonging to it can be sorted according to category.

Causal forecasting methods Methods that forecast one variable on the basis of relating it to another variable.

Central Limit Theorem The Central Limit Theorem states that whenever a random sample is taken from any distribution (μ, σ^2), then the sample mean will be approximately normally distributed with mean μ and variance σ^2/n.

Central tendency Measures the location of the middle or the centre of a distribution.

Chi square distribution The chi square distribution is a mathematical distribution that is used directly or indirectly in many tests of significance.

Chi square test Apply the chi square distribution to test for homogeneity, independence, or goodness of fit.

Classical additive time series model One of the models in classical time series analysis that assumes that components (trend, cyclical, seasonal, and random component) need to be added to compose the time series.

Classical time series analysis Approach to forecasting that decomposes a time series into certain constituent components (trend, cyclical, seasonal and, random component), makes estimates of each component and then re-composes the time series and extrapolates into the future.

Classical time series mixed model One of the models in classical time series analysis that assumes that components (trend, cyclical, seasonal, and random component) need to be added and multiplied to compose the time series.

Classical time series multiplicative model One of the models in classical time series analysis that assumes that components (trend, cyclical, seasonal, and random component) need to be multiplied to compose the time series.

Coefficient of determination (COD) The proportion of the variance in the dependent variable that is predicted from the independent variable.

Coefficient of variation The coefficient of variation measures the spread of a set of data as a proportion of its mean.

Confidence interval ($1 - \alpha$) A confidence interval gives an estimated range of values which is likely to include an unknown population parameter.

Contingency table A contingency table is a table of frequencies classified according to the values of the variables in question.

Continuous probability distribution If a random variable is a continuous variable, its probability distribution is called a continuous probability distribution.

Continuous variable A set of data is said to be continuous if the values belong to a continuous interval of real values.

Covariance Covariance is a measure of how much two variables change together.

Critical test statistic The critical value for a hypothesis test is a limit at which the value of the sample test statistic is judged to be such that the null hypothesis may be rejected.

Cumulative frequency distribution The cumulative frequency for a value x is the total number of scores that are less than or equal to x.

Cyclical component A component in the classical time series analysis approach to forecasting that covers cyclical movements of the time series, usually taking place over a number of years.

Deflating values Converting current prices into constant prices by using one of the standard indices, such as CPI (Consumer Price Index).

Degrees of freedom Refers to the number of independent observations in a sample minus the number of population parameters that must be estimated from sample data.

Differencing A method of transforming a time series, usually to achieve stationarity. Differencing means that every current value in the time series is subtracted from the previous value.

Directional test Implies a direction for the implied hypothesis (one tailed test).

Discrete probability distribution If a random variable is a discrete variable, its probability distribution is called a discrete probability distribution.

Discrete variable A set of data is said to be discrete if the values belonging to it can be counted as 1, 2, 3, ...

Dispersion The variation between data values is called dispersion.

Error measurement A method of validating the quality of forecasts. Involves calculating the mean error, the mean squared error, the percentage error, etc.

Estimate An estimate is an indication of the value of an unknown quantity based on observed data.

Event An event is any collection of outcomes of an experiment.

Expected frequency In a contingency table the expected frequencies are the frequencies that you would predict in each cell of the table, if you knew only the row and column totals, and if you assumed that the variables under comparison were independent.

Expected value The expected value of a random data variable indicates its population average value.

Exponential smoothing One of the methods of forecasting that uses a constant (or several constants) to predict future values by 'smoothing' the past values in the series. The effect of this constant decreases exponentially as the older observations are taken into calculation.

Exponential trend An underlying time series trend that follows the movements of an exponential curve.

Ex-post forecasts Values produced from a forecasting model that are fitted to historical data.

Factor A factor of an experiment is a controlled independent variable; a variable whose levels are set by the experimenter.

Five-number summary A five-number summary is especially useful when we have so many data that it is sufficient to present a summary of the data rather than the whole data set.

Forecasting A method of predicting the future values of a variable, usually represented as the time series values.

Forecasting errors A difference between the actual and the forecasted value in the time series.

Forecasting horizon A number of the future time units until which the forecasts will be extended.

Frequency distributions Systematic method of showing the number of occurrences of observational data in order from least to greatest.

Frequency polygon A graph made by joining the middle-top points of the columns of a frequency histogram.

Friedman's test for > 2 medians The Friedman rank test is primarily used to test whether c sample groups have been selected from populations having equal medians.

F test for variances Tests whether two population variances are the same based upon sample values.

Grouped frequency distributions Data arranged in intervals to show the frequency with which the possible values of a variable occur.

Histogram A histogram is a way of summarizing data that are measured on an interval scale (either discrete or continuous).

Homogeneity of variance Population variances are equal.

Hypothesis test procedure A series of steps to determine whether to accept or reject a null hypothesis, based on sample data.

Independent events Two events are independent if the occurrence of one of the events has no influence on the occurrence of the other event.

Index number A value of a variable relative to its previous value at some base.

Interaction Two independent variables interact if the effect of one of the variables differs depending on the level of the other variable.

Intercept Value of the regression equation (y) when the x value = 0.

Interquartile range The interquartile range is a measure of the spread of or dispersion within a data set.

Interval scale An interval scale is a scale of measurement where the distance between any two adjacent units of measurement (or 'intervals') is the same but the zero point is arbitrary.

Irregular component A component in the classical time series analysis approach to forecasting that is uncovered by other components. It has to be random in shape.

Kruskal-Wallis test for > 2 medians The Kruskal-Wallis test compares the medians of three or more independent groups.

Kurtosis Kurtosis is a measure of the 'peakedness' or the distribution.

Least squares The method of least squares is a criterion for fitting a specified model to observed data. If refers to finding the smallest (least) sum of squared differences between fitted and actual values.

Level The number of levels of a factor or independent variable is equal to the number of variations of that factor that were used in the experiment.

Level of confidence The confidence level is the probability value $(1 - \alpha)$ associated with a confidence interval.

Linear relationship Simple linear regression aims to find a linear relationship between a response variable and a possible predictor variable by the method of least squares.

Linear trend model A model that uses the straight line equation to approximate the time series.

Logarithmic trend A model that uses the logarithmic equation to approximate the time series.

Main effect This is the simple effect of a factor on a dependent variable.

Mann-Whitney U test The Mann-Whitney U test is used to test the null hypothesis that two populations have identical distribution functions against the alternative hypothesis that the two distribution functions differ only with respect to location (median), if at all.

McNemar's test for matched pairs McNemar's test is a non-parametric method used on nominal data to determine whether the row and column marginal frequencies are equal.

Mean The mean is a measure of the average data value for a data set.

Mean absolute deviation (MAD) The mean value of all the differences between the actual and forecasted values in the time series. The differences between these values are represented as absolute values, i.e. the effects of the sign are ignored.

Mean absolute percentage error (MAPE) The mean value of all the differences between the actual and forecasted values in the time series. The differences between these values are represented as absolute percentage values, i.e. the effects of the sign are ignored.

Mean error (ME) The mean value of all the differences between the actual and forecasted values in the time series.

Mean percentage error (MPE) The mean value of all the differences between the actual and forecasted values in the time series. The differences between these values are represented as percentage values.

Mean square error (MSE) The mean value of all the differences between the actual and forecasted values in the time series. The differences between these values are squared to avoid positive and negative differences cancelling each other.

Median The median is the value halfway through the ordered data set.

Mode The mode is the most frequently occurring value in a set of discrete data.

Moving averages Averages calculated for a limited number of periods in a time series. Every subsequent period excludes the first observation from the previous period and includes the one following the previous period. This becomes a series of moving averages.

Multiple comparisons Multiple comparisons problem occurs when one considers a set, or family, of statistical inferences simultaneously.

Multiple regression Multiple linear regression aims to find a linear relationship between a response variable and several possible predictor variables.

Multivariate methods Methods that use more than one variable and try to predict the future values of one of the variables by using the values of other variables.

Nominal scale A set of data is said to be nominal if the values belonging to it can be assigned a label rather than a number.

Non-parametric Non-parametric tests are often used in place of their parametric counterparts when certain assumptions about the underlying population are questionable.

Non-stationary time series A time series that does not have a constant mean and oscillates around this moving mean.

Normal distribution The normal distribution is a symmetrical, bell-shaped curve, centred at its expected value.

Normal probability plot Graphical technique to assess whether the data is normally distributed.

Null hypothesis (H_0) The null hypothesis, H_0, represents a theory that has been put forward but has not been proved.

Observed frequency In a contingency table the observed frequencies are the frequencies actually obtained in each cell of the table, from our random sample.

Ogive (or cumulative frequency polygon) A distribution curve in which the frequencies are cumulative.

One tail test A one tail test is a statistical hypothesis test in which the values for which we can reject the null hypothesis, H_0, are located entirely in one tail of the probability distribution.

Ordinal variable A set of data is said to be ordinal if the values belonging to it can be ranked.

Outlier An outlier is an observation in a data set which is far removed in value from the others in the data set.

Parametric Any statistic computed by procedures that assume the data were drawn from a particular distribution.

Pearson's coefficient of correlation Pearson's correlation coefficient measures the linear association between two variables that have been measured on interval or ratio scales.

Point estimate A point estimate (or estimator) is any quantity calculated from the sample data which is used to provide information about the population.

Poisson distribution Poisson distributions model a range of discrete random data variables.

Polynomial trend A model that uses an equation of any polynomial curve (parabola, cubic curve, etc.) to approximate the time series.

Population mean The population mean is the mean value of all possible values.

Population standard deviation The population standard deviation is the standard deviation of all possible values.

Population variance The population variance is the variance of all possible values.

Power trend A model that uses an equation of a power curve (a parabola) to approximate the time series.

Probability A probability provides a number value to the likely occurrence of a particular event.

P-value The p-value is the probability of getting a value of the test statistic as extreme as or more extreme than that observed by chance alone, if the null hypothesis is true.

Qualitative variable Variables can be classified as descriptive or categorical.

Quantitative variable Variables can be classified using numbers.

Quartiles Quartiles are values that divide a sample of data into four groups containing an equal number of observations.

Random component A component in time series analysis that has to act as a random variable, i.e. have some constant mean and the variance, as well as to exhibit no pattern.

Random sample A random sample is a sampling technique where we select a sample from a population of values.

Rank coefficient of correlation Spearman's rank correlation coefficient is applied to data sets when it is not convenient to give actual values to variables but one can assign a rank order to instances of each variable.

Ranks List data in order of size.

Range The range of a data set is a measure of the dispersion of the observations.

Ratio variable Ratio data are continuous data where both differences and ratios are interpretable and have a natural zero.

Region of rejection The range of values that leads to rejection of the null hypothesis.

Residual The residual represents the unexplained variation (or error) after fitting a regression model.

Residuals The differences between the actual and predicted values. Sometimes called forecasting errors. Their behaviour and pattern has to be random.

Sample space The sample space is an exhaustive list of all the possible outcomes of an experiment.

Sampling distribution The sampling distribution describes probabilities associated with a statistic when a random sample is drawn from a population.

Sampling error Sampling error refers to the error that results from taking one sample rather than taking a census of the entire population.

Scatter plot A scatter plot is a plot of one variable against another variable.

Seasonal component A component in the classical time series analysis approach to forecasting that covers seasonal movements of the time series, usually taking place inside one year's horizon.

Seasonal correlation A correlation between the observations given in the corresponding units of time within which the seasonality repeats itself.

Seasonal time series A time series, represented in the units of time smaller than a year, that shows regular pattern in repeating itself over a number of these units of time.

Significance level, α The significance level of a statistical hypothesis test is a fixed probability of wrongly rejecting the null hypothesis, H_0, if it is in fact true.

Sign test The sign test is designed to test a hypothesis about the location of a population distribution.

Simple price index A value of a price for one item relative to the previous price for the same item at some base.

Simple regression analysis Simple linear regression aims to find a linear relationship between a response variable and a possible predictor variable by the method of least squares.

Skewness Skewness is defined as asymmetry in the distribution of the data values.

Slope Gradient of the fitted regression line.

Standard deviation Measure of the dispersion of the observations (A square root value of the variance)

Standard error of forecast The square root of the variance of all forecasting errors.

Stated limits The lower and upper limits of a class interval.

Statistic A statistic is a quantity that is calculated from a sample of data.

Statistical independence Two events are independent if the occurrence of one of the events gives us no information about whether or not the other event will occur.

Stationary time series A time series that does have a constant mean and oscillates around this mean.

Student's t distribution The t distribution is the sampling distribution of the t statistic.

Symmetrical A data set is symmetrical when the data values are distributed in the same way above and below the middle value.

Test of association The chi square test of association allows the comparison of two attributes in a sample of data to determine if there is any relationship between them.

Test statistic A test statistic is a quantity calculated from our sample of data.

Tied ranks Two or more data values share a rank value.

Time period An unit of time by which the variable is defined (an hour, day, month, year, etc.).

Time series A variable measured and represented per units of time.

Time series plot A chart of a change in variable against time.

Transformations A method of changing the time series, usually to make it stationary. Most common method for transforming the time series is differencing or sometimes taking differences of every observation from the mean value.

Trend component A component in the classical time series analysis approach to forecasting that covers underlying directional movements of the time series.

Two tail test A two tail test is a statistical hypothesis test in which the values for which we can reject the null hypothesis, H_0, are located in both tails of the probability distribution.

Type I error, α A type I error occurs when the null hypothesis is rejected when it is in fact true.

Type II error, β A type II error occurs when the null hypothesis, H_0, is not rejected when it is in fact false.

Unbiased When the mean of the sampling distribution of a statistic is equal to a population parameter, that statistic is said to be an unbiased estimator of the parameter.

Univariate methods Methods that use only one variable and try to predict its future values through some pre-defined method.

Variable A variable is a symbol that can take on any of a specified set of values.

Variance Measure of the dispersion of the observations.

Wilcoxon signed rank t test The Wilcoxon signed ranks test is designed to test a hypothesis about the location of the population median (one or two matched pairs).