COMMON CONCEPTS IN STATISTICS

M.Tevfik Dorak, MD, PhD

***Please use this address next time*: http://www.dorak.info/mtd/glosstat.html**

***See also* Common Terms in Mathematics; Statistical Analysis in HLA-Disease Association Studies;**

Epidemiology (incl. **Genetic Epidemiology Glossary**)

[Please note that the best way to find an entry is to use the Find option from the Edit menu, or CTRL + F]

Absolute risk: Probability of an event over a period of time; expressed as a cumulative incidence like 10-year risk of 10% (meaning 10% of individuals in the group of interest will develop the condition in the next 10 year period). It shows the actual likelihood of contracting the disease and provides more realistic and comprehensible risk than **relative risk/odds ratio**.

Addition rule: The probability of any of one of several mutually exclusive events occurring is equal to the sum of their individual probabilities. A typical example is the probability of a baby to be homozygous or heterozygous for a Mendelian recessive disorder when both parents are carriers. This equals to 1/4 + 1/2 = 3/4. A baby can be either homozygous or heterozygous but not both of them at the same time; thus, these are mutually exclusive events (see also **multiplication rule**).

Adjusted odds ratio: In a multiple logistic regression model where the response variable is the presence or absence of a disease, an odds ratio for a binomial exposure variable is an adjusted odds ratio for the levels of all other risk factors included in a **multivariable model**. It is also possible to calculate the adjusted odds ratio for a continuous exposure variable. An adjusted odds ratio results from the comparison of two strata similar at all variables except exposure (or the marker of interest). It can be calculated when stratified data are available as contingency tables by **Mantel-Haenszel test**.

Affected Family-Based Controls (AFBAC) Method: One of several family-based association study designs (**Thomson, 1995**). This one uses affected siblings as controls and examines the sharing between two affected family members. The parental marker alleles not transmitted to an affected child or never transmitted to an affected sib pair form the so-called affected family-based controls (AFBAC) population. See also **HRR** and **TDT** and **Genetic Epidemiology**.

Age-standardized rate: An age-standardized rate is a weighted average of the age-specific rates, where the weights are the proportions of a standard population in the corresponding age groups. The potential confounding effect of age is removed when comparing age-standardized rates computed using the same standard population (from the **Glossary** of **Disease Control Priorities Project**.)

Alternative hypothesis: In practice, this is the hypothesis that is being tested in an experiment. It is the conclusion that is reached when a null hypothesis is rejected. It is the opposite of null hypothesis, which states that there is a difference between the groups or something to that effect.

Analysis of molecular variance (AMOVA): A statistical (analysis of variance) method for analysis of molecular genetic data. It is used for partitioning diversity within and among populations using nucleotide sequence or other molecular data. AMOVA produces estimates of variance components and F-statistic analogs (designated as phi-statistics). The significance of the variance components and phi-statistics is tested using a permutational approach, eliminating the normality assumption that is inappropriate for molecular data (**Excoffier, 1992**). AMOVA can be performed on **Arlequin**. For examples, see **Roewer, 1996**; **Stead, 2003**; **Watkins, 2003**); see also **AMOVA Lecture Note** (**EEB348**).

ANCOVA: See **covariance models**.

ANOVA (analysis of variance): A test for significant differences between multiple means by comparing variances. It concerns a normally distributed response (outcome) variable and a single categorical explanatory (predictor) variable, which represents treatments or groups. ANOVA is a special case of multiple regression where indicator variables (or orthogonal polynomials) are used to describe the discrete levels of factor variables. The term analysis of variance refers not to the model but to the method of determining which effects are statistically significant. Major assumptions of ANOVA are the homogeneity of variances (it is assumed that the variances in the different groups of the design are similar) and normal distribution of the data within each treatment group. Under the null hypothesis (that there are no mean differences between groups or treatments in the population), the variance estimated from the within-group (treatment) random variability (**residual sum of squares** = RSS) should be about the same as the variance estimated from between-groups (treatments) variability (**explained sum of squares** = ESS). If the null hypothesis is true, mean ESS / mean RSS (variance ratio) should be equal to 1. This is known as the **F test** or variance ratio test (see also **one-way** and **two-way ANOVA**). The ANOVA approach is based on the partitioning of sums of squares and degrees of freedom associated with the response variable. ANOVA interpretations of main effects and interactions are not so obvious in other regression models. An accumulated ANOVA table reports the results from fitting a succession of regression models to data from a factorial experiment. Each main effect is added on to the constant term followed by the interaction(s). At each level an F test result is also reported showing the extra effect of adding each variable so it can be worked out which model would fit best. In a two-way ANOVA with equal replications, the order of terms added to the model does not matter, whereas, this is not the case if there are unequal replications. When the assumptions of ANOVA are not met, its non-parametric equivalent **Kruskal-Wallis test** may be used (a tutorial on **ANOVA**, and **ANOVA posttest**). See also **MANOVA**.

Arithmetic mean: M = (x_1 + x_2 + ... x_n) / n (n = sample size).

Association: A statistically significant correlation between an environmental exposure or a biochemical/genetic marker and a disease or condition. An association may be an artifact (random error-chance, bias, confounding) or a real one. In population genetics, an association may be due to **population stratification**, **linkage disequilibrium**, or direct causation. A significant association should be presented together with a measure of the strength of association (**odds ratio**, **relative risk** or **relative hazard** and its 95% confidence interval) and when appropriate a measure of potential impact (**attributable risk, prevented fraction, attributable fraction/etiologic fraction**).

Assumptions: Certain conditions of the data that are required to be met for validity of a statistical test. **ANOVA** generally assumes normal distribution of the data within each treatment group, homogeneity of the variances in the treatment groups, and independence of the observations. In **regression analysis**, main assumptions are the normal distribution of the response variable, constant variance across fitted values, independence of **error terms**, and the consistency of underlying hazard rate over time (**proportionality assumption**) in **Cox Proportional Hazard Models**. See **StatNotes**: **Testing of Assumptions**.

Asymptotic: Refers to a curve that continually approaches either the x or y axis but does not actually reach it until x or y equals infinity. The axis so approached is the asymptote. An example is the **normal distribution curve**.

Asymptotically unbiased: In point estimation, the property that the bias approaches zero as the sample size (N) increases. Therefore, estimators with this property improve as N increases. See also **bias**.

Attributable risk (AR): Also called excess risk or risk difference. A measure of potential impact of an association. It quantifies the additional risk of disease following exposure over and above that experienced by individuals who are not exposed. It shows how much of the disease is eliminated if no one had the risk factor (unrealistic). The information contained in AR combines the **relative risk** and the risk factor prevalence. The larger the AR, the greater the effect of the risk factor on the exposed group. See also **prevented fraction**, **Walter, 1978**; **PowerPoint presentation on AR**; and **Attributable Risk Applications in Epidemiology**. For online calculation, see **EpiMax Table Calculator**.

Attributable fraction (etiologic fraction): It shows what proportion of disease in the exposed individuals is due to the exposure.

Balanced design: An experimental design in which the same number of observations is taken for each combination of the experimental factors.

Bartlett's test: A test for homogeneity of variance.

Bayesian inference: An inference method radically different from the classical frequentist approach which takes into account the prior probability for an event. Established as a new method by **Reverend Thomas Bayes**. See a slide presentation on **Introduction to Bayesian Statistics**; **+Plus**: **Bayesian Statistics Explained**; **Bayesian Revolution in Genetics**; **BMJ Collection on Bayesian Statistics**; **Books**; **Partition Software for Online Bayesian Analysis**; **Bayesian Calculator**; **A Bayesian Perspective on Interpreting Statistical Significance (InStat)** & **Bayesian Calculator**.

Bayes' method in genetic counseling: This method uses available additional information to modify risks calculated purely by Mendelian probabilities. It combines prior and conditional probabilities to give joint and posterior probabilities of unknown events. See **Bayesian Methods in Genetic Risk Calculation**, **Bayesian Analysis and Risk Assessment in**

Genetic Counseling and Testing and **A Unified Approach to Bayesian Risk Calculations**.

Bernoulli distribution models the behavior of data taking just two distinct values (0 and 1).

Bias: An estimator for a parameter is unbiased if its expected value is the true value of the parameter. Otherwise, the estimator is biased. It is the quantity E = (q -hat) - q. If the estimate of q is the same as actual but unknown q, the estimate is unbiased (as in estimating the mean of normal, binomial and Poisson distributions). If bias tends to decrease as n gets larger, this is called **asymptotic unbiasedness**. See reviews on epidemiologic meaning of bias: **Bias in Clinical Trials**, **Bias & Confounding in Molecular Epidemiology**, **Bias of Ascertainment** in **Complex Disease Genetics**, and **Bias and Confounding Lecture Note**.

Binary (dichotomous) variable: A discrete random variable that can only take two possible values (success or failure).

Binomial distribution: The binomial distribution gives the probability of obtaining exactly *r* successes in *n* independent trials, where there are two possible outcomes one of which is conventionally called success (**Binomial Distribution**; **Online Binomial Test** for observed vs expected value; **Binomial Probability Calculator**).

Blocks: Homogeneous grouping of experimental units (subjects) in experimental design. Groups of experimental units that are relatively homogeneous in that the responses on two units within the same block are likely to be more similar than the responses on two units in different blocks. Dividing the experimental units into blocks is a way of improving the accuracy between treatments. Blocking will minimize variation between subjects that is not attributable to the factors under study. Blocking is similar to matching in two-sample tests or stratification to control for confounding.

Blocking: When the available experimental units are not homogeneous, grouping them into blocks of homogeneous units (stratification) will reduce the experimental error variance. This is called blocking where differences between experimental units other than those caused by treatment factors are taken into account. This is like comparing age-matched groups (blocks) of a control group with the corresponding blocks in the patients group in an investigation of the side effects of a drug as age itself may cause differences in the experiment. Block effects soak up the extra, uninteresting and already known variability in a model. Blocking is preferable to randomization when the factors that might affect the outcome are known.

Bonferroni Correction: This is a multiple comparison technique used to adjust the a error level. See also **HLA and Disease Association Studies**, **Online Bonferroni Correction**, **GraphPad Guide to Multiple Comparisons & a commentary by Perneger, 1998**).

Bootstrap: An application of resampling statistics. It is a data-based simulation method used to estimate variance and bias of an estimator and provide confidence intervals for parameters where it would be difficult to do so in the usual way (**Online Resampling Book**).

Canonical: Something that has been reduced to its simplest form.

Carryover effect: Any effect of a drug that lasts beyond the period of treatment. This is a worry in drug trials with **crossover design** and the reason for the washout period between treatments.

Case-control study: A design preferred over cohort studies for relatively rare diseases in which cases with a disease or exposure are compared with controls randomly selected from the same study base. This design yields **odds ratio** as opposed to **relative risk** from cohort studies. See **Case-control Studies Chapter** in **Epidemiology for the Uninitiated**.

Causal relationship: It does not matter how small it is, a *P* value does not signify causality. To establish a causal relationship, the following non-statistical evidence is required: consistency (reproducibility), biological plausibility, dose-response, temporality (when applicable) and strength of the relationship (as measured by odds ratio/relative risk/hazard ratio). See **Hills's criteria of causality**; **Seven Common Errors in Statistics**; and Causality by DR Cox, JR Stat Soc A 1992;155:291 (**JSTOR-UK link**). The original reference for Hill's criteria is Hill AB: The environment and disease: association or causation. Proc R Soc Medicine 1965;58:295-300.

Categorical (nominal) variable: A variable that can be assigned to categories. A non-numerical (**qualitative**) variable measured on a (discrete) **nominal** scale such as gender, drug treatments, disease subtypes; or on an **ordinal** scale such as low, median or high dosage. A variable may alternatively be **quantitative** (**continuous** or **discrete**). See **GraphPad QuickCalc**: **Categorical Data Analysis**.

Censored observation: Observations that survived to a certain point in time before dropping out from the study for a reason other than having the outcome of interest (lost to follow up or not enough time to have the event). Thus, censoring is simply an incomplete observation that has ended before time-to-event. These observations are still useful in **survival analysis**.

Central limit theorem: The means of a relatively large (>30) number of random samples from any population (not necessarily a normal distribution) will be approximately normally distributed with the population mean being their mean and variance being the (population variance / n). This approximation will improve as the sample size (the number of samples) increases. See **Mathematical Basis**; **QuickTime Demonstration**; **JAVA Demonstration**; **Simulation**.

Chi-squared distribution: A distribution derived from the **normal distribution**. Chi-squared (C^2) is distributed with v degrees of freedom with mean = v and variance = 2v. (**Chi-squared Distribution**, **a Lecture on Chi-Squared Significance Tests**).

Chi-squared test: The most commonly used test for frequency data and goodness-of-fit. In theory, it is nonparametric but because it has no parametric equivalent, it is not classified as such. It is not an exact test and with the current level of computing facilities, there is not much excuse not to use Fisher's exact test for 2x2 contingency table analysis instead of Chi-squared test. Also for larger contingency tables, the G-test (log-likelihood ratio test) may be a better choice. The Chi-square value is obtained by summing up the values (residual^2/fit) for each cell in a contingency. In this formula, residual is the difference between the observed value and its expected counterpart and fit is the expected value. See **Statistical Analysis in HLA and Disease Association Studies** for assumptions and restrictions of the Chi-squared test. (**Tables of critical values of t, F and Chi-square**).

Cochran's Q Test: A nonparametric test examining change in a dichotomous variable across more than two observations. If there are two observations, **McNemar's test** should be used.

Coefficient of determination (R^2): See **Multiple regression correlation coefficient**.

Coefficient of variation: It is a measure of spread for a set of data. It is a measure of variation in relation to the mean. Calculated as standard deviation divided by the mean (x100). (**Online Calculator for Coefficient of Variation and Other Descriptive Statistics**).

Cohort effect: The tendency for persons born in certain years to carry a relatively higher or lower risk of a given disease. This may have to be taken into account in case-control studies.

Concomitant variable: See **covariance models**.

Conditional (fixed-effects) logistic regression: The conditional logistic regression (CLR) model is used in studies where cases and controls can be matched (as pairs) with regard to variables believed to be associated with the outcome of interest. The model creates a likelihood that conditions on the matching variable. It is the preferred method for the analysis of nested case-control studies when matching is done at the individual level (there may be more than one control per case). In economic analysis, it is called fixed-effects logit for panel data. See **Preisser & Koch, 1997**.

Confounding variable: A variable that is associated with both the outcome and the exposure variable. A classic example is the relationship between heavy drinking and lung cancer. Here, the data should be controlled for smoking as it is related to both drinking and lung cancer. A positive confounder is related to exposure and response variables in the same direction (as in smoking); a negative confounder shows an opposite relationship to these two variables (age in a study of association between oral contraceptive use and myocardial infarction is a negative confounder). The data should be stratified before analyzing it if there is a confounding effect. **Mantel-Haenszel** test is designed to analyze stratified data to control for a confounding variable. Alternatively, **a multivariable regression model** can be used to adjust for the effects of confounders. See **Bias & Confounding Lecture Note**; a review by **Greenland, 2001**.

Conservative test: A test where the chance of type I error (false positive) is reduced and type II error risk is increased. Thus, these tests tend to give larger *P* values compared to non-conservative (liberal) tests for the same comparison.

Contrast: A contrast is combinations of treatment means, which is also called the main effect in **ANOVA**. It measures the change in the mean response when there is a change between the levels of one factor. For example, in an analysis of three different concentrations of a growth factor on cell growth in cell culture with means m_1, m_2, m_3, against a control value (m_o) without any growth factor, a contrast would be:

$$q = m_o - 1/3 \ (m_1 + m_2 + m_3)$$

The important point is that the coefficients sum to zero (1/1 - 1/3 - 1/3 - 1/3). If the value of the contrast (q) is zero or not significantly different from zero, there is no main effect, i.e., the combined growth factor mean is not different (positive or negative) from the no growth factor mean.

Cook statistics: A diagnostic *influence* statistics in regression analysis designed to show the influential observations. **Cook's distance** considers the influence of the *i* th value on all n fitted values and not on the fitted value of the *i* th observation. It yields the shift in the estimated parameter from fitting a regression model when a particular observation is omitted. All distances should be roughly equal; if not, then there is reason to believe that the respective case(s) biased the estimation of the regression coefficients. Relatively large Cook statistics (or Cook's distance) indicates influential observations. This may be due to a high **leverage**, a large **residual** or their combination. An **index plot** of residuals may reveal the reason for it. The leverages depend only on the values of the explanatory variables (and the model parameters). Cook statistics depends on the residuals as well. Cook statistics may not be very satisfactory in binary regression models. Its formula uses the standardized residuals but the modified Cook statistics uses the **deletion residuals**.

Correlation Coefficient: See **Pearson's correlation coefficient (r)**, **Spearman's rank correlation (rho)** and **Multiple regression correlation coefficient (R^2)**. (**Online Correlation & Regression Calculators** at **Vassar College**.)

Correspondence analysis: In population genetics, a complementary analysis to genetic distances and dendrograms. It displays a global view of the relationships among populations (**Greenacre MJ, 1984; Greenacre & Blasius, 1994; Blasius & Greenacre, 1998**). With its visual output, it supplements more formal inferential analyzes. This type of analysis tends to give results similar to those of dendrograms as expected from theory (**Cavalli-Sforza & Piazza, 1975**), but is more informative and accurate than dendrograms especially when there is considerable genetic exchange between close geographic neighbors (**Cavalli-Sforza et al. 1994**). Cavalli-Sforza et al concluded in their enormous effort to work out the genetic relationships among human populations that two-dimensional scatter plots obtained by correspondence analysis frequently resemble geographic maps of the populations with some distortions (**Cavalli-Sforza et al. 1994**). Using the same allele frequencies that are used in phylogenetic tree construction, **correspondence analysis** can be performed on **ViSta** (**v7.0**)**, VST, Statistica, SAS** but most conveniently on MultiVariate Statistical Package (**MVSP**). Link to **a Tutorial**; **Course Notes**; **StatSoft Textbook**: **Correspondence Analysis Chapter**.

Cox proportional hazards model: A regression method described by D.R. Cox (J Royal Stat Soc, Series B 1972;34:187-220; **JSTOR-UK**) for modeling survival times (for significance of the difference between survival times, **log-rank test** is used). It is also called proportional hazards model because it estimates the ratio of the risks (**hazard ratio** or **relative hazard**). As in any regression model, there are multiple predictor variables (such as prognostic markers whose individual contribution to the outcome is being assessed in the presence of the others) and the outcome variable (e.g., whether the patients survived five years, or died during follow-up, etc). The model assumes that the underlying hazard rate (rather than survival time) is a function of the independent variables and consistent over time (**proportionality assumption**, i.e. the survival functions of the groups are approximately parallel). There is no assumption for the shape and nature of the underlying survival function. Cox's regression model has been the most widely used method in survival data analysis regardless of whether the survival time is discrete or continuous and whether there is censoring (**Lee & Go, 1997**). Cox regression uses the **maximum likelihood method** rather than the **least squares method** (**Online Cox Proportional Hazards Survival Regression**; a **Superlectures** on survival analysis; **Cox's proportional hazards model**; an example of **Model Building**).

Covariate: Generally used to mean explanatory variable, less generally an additional explanatory variable is of no interest but included in the model to adjust the statistical model. More specifically, it denotes an explanatory variable which is unaffected by treatments and has a linear relationship to the response. The intention is to produce more precise estimates of the effect of the explanatory variable of main interest. In the analysis, a model is first fitted using the covariate. Then the main explanatory variable is added and its additional effect is assessed statistically. Whether the use of a covariate is wise (i.e., whether it has a statistically significant influence) can be judged by checking its effect on the residual (error) mean square (variance). If the addition of covariate reduces it remarkably, it will improve the analysis. See also **covariance models**.

Covariance (covariation): It is a measure of the association between a pair of variables: the expected value of the product of the deviation of two random variables from their respective means. It is also called a measure of 'linear dependence' between the two random variables. If the two variables are independent (no linear correlation), then their covariance is zero but for non-zero values covariance is unstandardized (unlike **correlation coefficient**); there is no limit to possible values. Because of this, it is difficult to compare covariances. A negative value means that for small values of X, there are large values of Y (inverse association). It is calculated as the mean sum-of-products using each x_i and y_i values, their means m_x and m_y, and (n-1). (The covariance standardized to lie between -1 and +1 is **Pearson's correlation coefficient**.) See **Wikipedia**: **Covariance**; **NetMBA Statistics**: **Covariance**.

Covariance models: Models containing some quantitative and some qualitative explanatory variables, where the chief explanatory variables of interest are qualitative and the quantitative variables are introduced primarily to reduce the variance of the error terms. [Models in which all explanatory variables are qualitative are called analysis of variance -**ANOVA**- models.] Analysis of covariance -ANCOVA- combines features of ANOVA and regression. It augments the ANOVA model containing the factor effects with one or more additional quantitative variables that are related to the response variable. The intention is to make the analysis more precise by reducing the variance of the error terms. Each continuous quantitative variable added to the ANOVA model is called a **concomitant variable** (and sometimes also covariates). If qualitative variables are added to an ANOVA model to reduce error variance, the model remains to be ANOVA. By adding extra variables, the results are said to be controlled or adjusted for the additional variables (like age or sex).

Apart from the above use of the term, analysis of covariance is more generally used for almost any analysis assessing the relationship between a response variable and a number of explanatory variables. In a multiple regression model, additional variables, which are known not to have any effect on the response variable, such as age and sex, are sometimes added to the model to adjust the response for these variables (age and sex in this case). Such variables are called confounders (or covariates). When the response is normally distributed, this is the preferred method over a simple t-test when the two groups compared differ, say, in their age and sex distribution (or any other confounding variable). The result is then controlled (or adjusted) for age and sex. When such adjustments are made, the regression coefficient for the significant effect variable will be (most probably) different from the one obtained from a univariable model involving only that variable (say the effect of a disease on pulse rate as compared to healthy controls).

Cramer's coefficient of association (C): Also known as contingency coefficient. While Chi-squared is used to determine significance of an association (and varies by sample size for the same association), Cramer's C is a measure of association varying from 0 (no association) to 1 (perfect association) that is independent of the sample size. However, it seldom reaches its upper limit. It allows direct comparison of the degree of association between different contingency tables. It is calculated directly from the Chi-squared value and the total sample size as $(C^2/C^2+N)^{½}$.

Cramer's V: A measure of the strength association for any size of contingency tables. It can be seen as a correction of the Chi-squared value for sample size. The transformation of the chi-squared value provides a value between 0 and 1 for relative comparison of the strength of the association. For a 2x2 table, Cramer's V is equal to the **Phi coefficient**. Cramer's V is most useful for large contingency tables. It can also be used as a global linkage disequilibrium value for multiallelic loci (See **GOLD-Disequilibrium Statistics**; Online **Cramer's V calculation**.)

Crossover design: A clinical trial design during which each subject crosses over from receiving one treatment to another one.

Cross-sectional data: Data collected at one point in time (as opposed to longitudinal/cohort data for example). See **Cross-Sectional Studies Chapter** in **Epidemiology for the Uninitiated**.

Degrees of freedom (df): The number of independent units of information in a sample used in the estimation of a parameter or calculation of a statistic. In the simplest example of a 2x2 table, if the marginal totals are fixed, only one of the four cell frequencies is free to vary and the others will be dependent on this value not to alter the marginal totals. Thus, the df is only 1. Similarly, it can easily be worked out that in a contingency table with r rows and c columns, the df = (r-1)(c-1). In parametric tests, the idea is slightly different that the n bits of data have n degrees of freedom before we do any statistical calculations. As soon as we estimate a parameter such as the mean, we use up one of the df, which was initially present. This is why in most formulas, the df is (n-1). In the case of a two-sample t-test with n_1 and n_2 observations, to do the test we calculate both means. Thus, the df = $(n_1 + n_2 - 2)$. In linear regression, when the linear equation $y = a + bx$ is calculated, two parameters are estimated (the intercept and the slope). The df used up is then 2: df = $(n_1 + n_2 - 2)$. Non-parametric tests do not estimate parameters from the sample data, therefore, df do not occur in them.

In simple linear regression, the df is partitioned similar to the total sum of squares (**TSS**). The df for TSS is N-k. Although there are n deviations, one df is lost due to the constraint they are subject to: they must sum to zero. TSS equals to **RSS** + **ESS**. In one-way ANOVA, the df for RSS is N-2 because two parameters are estimated in obtaining the fitted line. ESS has only one df associated with it. This is because the n deviations between the fitted values and the overall mean are calculated using the same estimated regression line, which is associated with two df (see above). One of them is lost because the of the constraint that the deviations must sum to zero. Thus, there is only one df associated with ESS. Just like TSS = RSS + ESS, their df have the same relationship: N-1 = (N-2) + 1.

Deletion (or deleted) residual: A modified version of the standardized residual, which uses an estimate of s^2 from a regression in which point *i* has been deleted. It is used in calculation of the modified **Cook's distance** instead of standardized residuals. Deletion residuals are also known as **likelihood residuals**.

Descriptive statistics: Summary of available data. Examples are male-to-female ratio in a group; numbers of patients in each subgroup; the mean weight of male and female students in a class, etc. Only when the distribution is symmetric, mean and standard deviation can be used. Otherwise (as in survival data), mean and percentiles/range should be used to describe the data. See GraphPad guide to **Interpreting Descriptive Statistics**.

Deviance: A measure for judging the degree of matching of the model to the data when the parameter estimation is carried out by maximizing the likelihood (as in GLMs). The deviance has asymptotically a Chi-squared distribution with df equal to the difference in the number of parameters in the two models being compared. The total deviance compares the fit of the saturated model to the null model, thus, expresses the total variability around a fitted line which can be decomposed to explained and unexplained (error) variability. The *residual deviance* in a GLM analysis of deviance table corresponds to RSS in an ANOVA table, *and regression deviance* corresponds to ESS. The residual deviance measures how much fit to the data is lost (in likelihood terms) by modeling compared to the saturated model. This will be a small value if the model is good (and it will be zero for the saturated model containing all main effects and all interactions). The regression deviance measures how much better is the model taking into account the explanatory variables compared to the simplest model ignoring all of them and only containing a constant (the mean of the responses). This measurement is again made in terms of log-likelihood and the bigger the regression deviance the better fits the model (i.e., the regression effect is strong). In likelihood terms, the *residual* deviance can be expressed as follows:

$D = -2 [\mathit{ln}\ L_c - \mathit{ln}\ L_s]$ or $D = -2\ \mathit{ln}\ [L_c / L_s]$

where L_c is the likelihood of the current model, and L_s is the likelihood of the saturated model.

Similarly, the *regression* deviance can be expressed as:

$D = -2 [\mathit{ln}\ L_c - \mathit{ln}\ L_n]$ or $D = -2\ \mathit{ln}\ [L_c / L_n]$

where L_n is the likelihood of the null model. The bigger the regression deviance the better the model including this particular variable.

For purposes of assessing the significance of an independent variable, the value of *D* with an without the independent variable is compared (note the nested character of the two sets). This is called the **deviance difference**. If a variable is dropped, the residual deviance difference, if a variable is added, the regression deviance difference is compared with the C^2 distribution.

Deviance difference: In generalized linear modeling, models that have many explanatory variables may be simplified, provided information is not lost in this process. This is tested by the difference in deviance between any two nested models being compared:

G = *D* (for the model without the variable) - *D* (for the model with the variable)

The smaller the difference in (residual) deviance, the smaller the impact of the variables removed. This can be tested by the C^2 test.

In the simplest example (in simple linear regression), if the log-likelihood of a model containing only a constant term is L_0 and the model containing a single independent variable along with the constant term is L_1, multiplying the difference in these log-likelihoods by -2 gives the deviance difference, i.e., $G = -2\ (L_1 - L_0)$. *G* statistics (**likelihood ratio test**) can be compared with the C^2 distribution on df = 1, and it can be decided whether the model with the independent variable has a regression effect (if $P < 0.05$, it has). The same method can be used to detect the effect of the interaction by adding any interaction to the model and obtaining the regression deviance. If this deviance is significantly higher than the one without the interaction in the model, there is interaction [the coefficient for the interaction, however, does not give the **odds ratio** in **logistic regression**].

Discrete variable: A variable of countable number of integer outcomes. Examples include -**ordinal multinomial**- several prognostic outcomes (such as poor, median and good) as a function of treatment modalities, stage of the disease, age etc., or -**multinomial**- people’s choices of hospitals (hospital A, B or C) as a function of their income level, age, education etc. A discrete variable may be **binomial**: diseased or non-diseased in a cohort or case-control study. The nature of the outcome variable as discrete or continuous is crucial in the choice of a regression model (see the **generalized linear model**).

Dummy variables: A binary variable that is used to represent a given level of a **categorical variable**. In genetic data analysis, for example, it is created for each allele at a multiallelic locus. The most common choices as dummy variables are 0 and 1. For each level of the variable, 1 denotes having the feature and 0 denotes all other levels. Also called indicator variables. If an indicator variable is used, the regression coefficient gives the change per unit compared to the reference value. Creating dummy variables on Stata: **Stata** & **UCLA Statistics**: **STATA Dummy Variables**.

Dunn's Test: This test is used when a difference between the groups is found in a non-parametric ANOVA test. Dunn's test is a **post hoc test** that makes pairwise (multiple) comparisons to identify the different group. See **GraphPad Prism Guide**: **K-W and Dunn's Test** (for an example, see **Online K-W Test**; if the link does not work: go to **Stats ToolBoxs**

Homepage and choose Kruskal-Wallis OWAV from Frequency Tables. Dunn's Q scores are presented at the end of the analysis for pairwise difference tests. Another example is presented in an **ANOVA PPT Presentation**).

Dunnett's test: When ANOVA shows a significant difference between groups, if one of the groups is a control (reference) group, Dunnett's Test is used as a ***post hoc*** **test**. This multiple comparison test can be used to determine the significant differences between a single control group mean and the remaining treatment group means in an analysis of variance setting. It is one of the least conservative ***post hoc*** **tests**.

Ecological study: Analyses based on data grouped to the municipal, provincial or national level. See **Ecological Studies Chapter** in **Epidemiology for the Uninitiated**.

Ecological fallacy: The aggregation bias, which is the unfortunate consequence of making inferences for individuals from aggregate data. It results from thinking that relationships observed for groups necessarily hold for individuals. The problem is that it is not valid to apply group statistics to an individual member of the same group. See an essay on **Ecological Fallacy**.

Edwards' test: A statistical test for seasonality that looks for a one-cycle sinusoidal deviation from the null distribution (see **Westerbreek et al, 1998**).

Effect modification: The situation in which a measure of effect changes over values of another variable (the association estimates are different in different subpopulations of the sample). The relative risk or odds ratio associated with exposure will be different depending on the value of the effect modifier. For example if in a disease association study, the odds ratios are different in different age groups or in different sexes, age or sex are effect modifiers. Effect modification is highly related to statistical interaction in regression models. If where an exposure decreases risk for one value of the effect modifier and increases risk for another value of effect modifier, this is called crossover. See **Thompson, 1991** and *Effect Modification* in **Encyclopedia of Biostatistics.**

Eigenvalues (latent values): In multivariate statistics, eigenvalues give the variance of a linear function of the variables. Eigenvalues measure the amount of the variation explained by each principal component (PC) and will be largest for the first PC and smaller for the subsequent PCs. An eigenvalue greater than 1 indicates that PCs account for more variance than accounted by one of the original variables in standardized data. This is commonly used as a cut-off point for which PCs are retained.

EM Algorithm: A method for calculating maximum likelihood estimates with incomplete data. E (expectation)-step computes the expected values for missing data and M (maximization)-step computes the maximum likelihood estimates assuming complete data. It was first used in genetics (**Ceppellini R et al, 1955**) to estimate allele frequency for phenotype data when genotypes are not fully observable (this requires the assumption of HWE and calculation of expected genotypes from phenotype frequencies). For a brief overview, see **ARC CIGMR**: **EM Algorithm**. **Arlequin** implements EM algorithm in haplotype construction and frequency analysis.

Empirical *P* value: A *P* value obtained by a simulation program such as Monte Carlo statistics. This is less likely to be affected by multiple comparisons.

Empirical rule: In variables normally distributed, 68% of the data values are within 1SD of the mean; 95% are within 2SD of the mean; and 99.7% (nearly all) are within 3SD of the mean.

Epidemiologic flaws and fallacies: Beware of confounders, selection bias, response bias, variable observer, Hawthorne effect (changes caused by the observer in the observed values), diagnostic accuracy bias, **regression to the mean**, significance Turkey, nerd of nonsignificance, cohort effect, **ecological fallacy**, Berkson bias (selection bias in hospital-based studies) and others (discussed in M Michael III et al. **Biomedical Bestiary**. Little, Brown and Company, 1984; and in **Bias & Confounding in Molecular Epidemiology**).

Epi-Info: An epidemiologic data management and analysis package freely available from **the CDC website**. Originally a DOS program, the latest version (**3.3.2**) is designed for Microsoft Windows 95-XP (release date: Febr 9, 2005). There are also tutorials available online: **CDC**, **Nebraska University**, **Dalhousie University** and **Henry Ford Health Systems**.

Error terms: Residuals in regression models. Shown as W_i or e_i. Their expected value is zero, thus, they vary around zero with a variance equal to s^2: $N(0, s^2)$. They are assumed to be normally distributed, have equal variance for all fitted values, and independent. Normality is a reasonable assumption in many cases. The assumption of equal variance implies that every observation on the dependent variable contains the same amount of information. The impact of heterogeneous variances is a loss of precision of estimates compared to the precision that would have been realized if the heterogeneous variances had been taken into account. Transformation of the dependent variable may help to homogenize the unequal variances. Correlated errors are most frequent in time sequence data and they also cause the loss in precision in the estimates. See also **residuals**.

Explained (regression) sum of squares (ESS): The measure of between treatments sum of squares (variability) in ANOVA. If the means of treatment groups are different, the ESS would be greater than RSS to yield a high variance ratio. The bigger the ESS, the better explained the data by the model.

Exploratory data analysis: An initial look at the data with minimal use of formal mathematics or statistical methods, but more with an informal graphical approach. Scatter plots, correlation matrices and contingency tables (for binary data) can be used to get an initial idea for relationships between explanatory variables (for collinearity) or between an explanatory variable(s) and a response variable(s) (correlation). In ANOVA, normality can be checked by box-plots. It gives some indication of which variables should be in the model and which one of them should be put into the model first, and whether linear relationship is adequate.

Exponential distribution: The (continuous) distribution of time intervals between independent consecutive random events like those in a Poisson process.

Exponential family: A family of probability distributions in the form

$f(x) = \exp \{a(q)b(x) + c(q) + d(x)\}$

where q is a parameter and a, b, c, d are known functions. This family includes the **normal distribution**, binomial distribution, Poisson distribution and gamma distribution as special cases.

Exposure: In an epidemiologic study, exposure may represent an environmental exposure, an intervention or the presence of a marker (biomarker/genetic marker).

Factor: A categorical explanatory variable with small number of **levels** such that each item classified belongs to exactly one level for that category. If the factor is 'sex', the levels are 'male' and 'female'; if the factor is 'drug received', the levels are 'drug A', 'drug B', 'drug C', etc. A set of factor levels, uniquely defining a single treatment, is called a **cell**. A cell may have just one observation (no replication) or multiple observations (replications).

Factorial experiments: In some data, the explanatory (predictor) variables are all categorical (i.e., **factors**) and the response variable is quantitative. When there are two or more categorical predictor variables, the data are called **factorial**. The different possible values of the factors are often assigned numerical values known as **levels**.

Factorial analysis of variance: An analysis in which the treatments differ in terms of two or more factors (with several levels) as opposed to treatments being different levels of a single factor as in one-way ANOVA.

F distribution: A continuous probability distribution of the ratio of two independent random variables, each having a Chi-squared distribution, divided by their respective degrees of freedom. The commonest use is to assign *P* values to mean square ratios (variance ratios) in ANOVA. In regression analysis, the F-test can be used to test the joint significance of all variables of a model. (**Tables of critical values of t, F and Chi-square**).

Fisher's exact test: An exact significance test to analyze 2x2 tables for any sample size. It is a misconception that it is suitable only for small sample sizes. This arises from the demanding computational procedure for large samples, which is no longer an issue. It is the only test for a 2x2 table when an expected number in any cell is smaller than 5 (**Online Fisher's Test (1)**; **(2)**; **Calculator 3** in **Clinical Research Calculators** at **Vassar**). For an exact test for larger contingency tables, see **Vassar Online** or download **RxC** by Mark Miller.

F test: The F test for linear regression tests whether the slope is significantly different from 0, which is equivalent to testing whether the fit using non-zero slope is significantly better than the null model with 0 slope. See also **mean squares**.

Gambler's ruin: A classical topic in probability theory. It is a game of chance related to a series of Bernoulli trials. There are variations of the game theory associated with problems of the random walk and sequential sampling.

Game theory: The theory of contests between two or more players under specified sets of rules. The statistical aspect is that the game proceeds under a chance scheme such as throwing a die.

Gaussian distribution: Another name for the **normal distribution** (**GraphPad Gaussian Distribution Calculator**).

G Statistics: An application of the **log-likelihood ratio statistics** for the hypothesis of independence in an *r* x *c* contingency table. It can also be used to test goodness-of-fit. The G-test should be preferred over Chi-squared test when for any cell in the table, ½ O-E½ > E. The Chi-squared distribution is usually poor for the test statistics G^2 when N/rc is smaller than five (preferable to the Chi-squared test in Hardy-Weinberg Equilibrium (HWE) test as long as this condition is met). **HyperStat** and **StatXact** perform G statistics (**Online G Statistics**).

General linear model: A group of linear regression models in which the response variable is continuous and normally distributed, the response variable values are predicted from a linear combination of predictor variables, and the linear combination of values for the predictor variables is not transformed (i.e., there is no **link function** as in **generalized linear models**). Linear multiple regression is a typical example of general linear models whereas simple linear regression is a special case of **generalized linear models** with the identity link function.

Generalized linear model (GLM): A model for linear and non-linear effects of continuous and categorical predictor variables on a discrete or continuous but not necessarily normally distributed dependent (outcome) variable. (Note that in the **general linear model**, the dependent (outcome) variable should be normally distributed). Normal, binary (or linear logistic; when the outcome variable is a proportion), binomial or Poisson (when the outcome variable is a count), exponential and gamma (when the outcome variable is continuous and non-negative) models are different versions of generalized linear models. Particular types of models arise by specifying an appropriate ***link function***, variance and distribution. For example, normal linear regression corresponds to an identity link function, constant variance and a normal distribution. Logistic regression arises from a logit link function and a binomial distribution (the variance of the response (npq) is related to its mean (np): variance = mean (1 - (mean/n)). Loglinear models are used for binomial or Poisson counts. Standard techniques for analyzing censored survival data, such as the **Cox regression**, can also be handled within the GLM framework (**Online GLM**, **GLM Website**, **Lecture Notes of a GLM Course**).

Genetic distance: A measurement of genetic relatedness of populations. The estimate is based on the number of allelic substitutions per locus that have occurred during the separate evolution of two populations. Link to a lecture on **Estimating Genetic Distance** and **GeneDist: Online Calculator of Genetic Distance**. The software **Arlequin**, **PHYLIP**, **GDA**, **PopGene** and **SGS** are suitable to calculate population-to-population genetic distance from allele frequencies. See **Basic Population Genetics**.

Genetic Distance Estimation by PHYLIP: The most popular (and free) phylogenetics program **PHYLIP** can be used to estimate genetic distance between populations. Most components of PHYLIP can be run **online**. One component of the package **GENDIST** estimates genetic distance from allele frequencies using one of the three methods: Nei's, Cavalli-Sforza's or Reynold's (see papers by **Nei et al, 1983**, **Nei M, 1996** and a **lecture note** for more information on these methods). GENDIST can be run **online** using the default options (**Nei's genetic distance**) to obtain genetic distance matrix data. The PHYLIP program **CONTML** estimates phylogenies from gene frequency data by maximum likelihood under a model in which all divergence is due to genetic drift in the absence of new mutations (Cavalli-Sforza's method) and draws a tree. The program comes as a freeware as part of PHYLIP or this program can be run **online** with default options. If new mutations are contributing to allele frequency changes, Nei's method should be selected on GENDIST to estimate genetic distances first. Then a tree can be obtained using one of the following components of PHYLIP: **NEIGHBOR** also draws a phylogenetic tree using the genetic distance matrix data (from GENDIST). It uses either Nei's "**Neighbor Joining Method**," or the **UPGMA** (**u**nweighted **p**air **g**roup **m**ethod with **a**rithmetic mean; average linkage clustering) method. Neighbor Joining is a distance matrix method producing an unrooted tree without the assumption of a clock (UPGMA does assume a clock). NEIGHBOR can be run **online**. Other components of PHYLIP that draw phylogenetic trees from genetic distance matrix data are **FITCH** / **online** (does not assume evolutionary clock) and **KITSCH** / **online** (assumes evolutionary clock).

Geometric mean: G = $(x_1.x_2...x_n)^{1/n}$ where n is the sample size. This can also be expressed as antilog ((1/n) S log *x*), which means the antilog of the mean of the logs of each value.

Half-normal plot: A diagnostic test for model inadequacy or revealing the presence of outliers. It compares the ordered residuals from the data to the expected values of ordered observations from a normal distribution. While the full-normal plots use the signed residuals, half-normal plots use the absolute values of the residuals. Outliers appear at the top right of the plot as distinct points, and departures from a straight line mean that the model is not satisfactory. It is appropriate to use a half-normal plot only when the distribution is symmetrical about zero because any information on symmetry will be lost.

Haplotype Relative Risk method: This method uses non-inherited parental haplotypes of affected persons as the control group and thus eliminates the risks and bias associated with using unrelated individuals as controls in case-control association studies, as well as the higher cost (see **Falk & Rubinstein, 1987**; **Knapp, 1993**; **Terwilliger & Ott, 1992**).

Hardy-Weinberg equilibrium (HWE): In an infinitely large population, gene and genotype frequencies remain stable as long as there is no selection, mutation, or migration. For a bi-allelic locus where the gene frequencies are p and q: $p^2+2pq+q^2 = 1$ (see **Hardy-Weinberg parabola**). HWE should be assessed in controls in a case-control study and any deviation from HWE should alert for genotyping errors (**Lewis, 2002**) unless there are biological reasons for any deviation (see **Ineichen & Batschelet, 1975** for the effect of natural selection on Hardy-Weinberg equilibrium). (**Online HWE Analysis**; **HWE and Association Testing for SNPs in Case-Control Studies**; **OEGE: HWE Calculator**; **Population Genetics Notes**).

Harmonic mean: Of a set of numbers (y_1 to y_n), the harmonic mean is the reciprocal of the arithmetic mean of the reciprocal of the numbers [H = N / (1/(y_1 + y_2 + ... y_n))]. The harmonic mean is either smaller than or equal to the arithmetic mean. It is a measure of position.

Hazard function (instantaneous failure rate, conditional failure, intensity, or force of mortality function): The function that describes the probability of failure during a very small time increment (assuming that no failures have occurred prior to that time). Hazard is the slope of the survival curve – a measure of how rapidly subjects are having the event (dying, developing an outcome etc).

Hazard Rate: It is a time-to-failure function used in survival analysis. It is defined as the probability per time unit that a case that has survived to the beginning of the respective interval will fail in that interval. Specifically, it is computed as the number of failures per time units in the respective interval, divided by the average number of surviving cases at the mid-point of the interval.

Hazard Ratio (Relative Hazard): Hazard ratio compares two groups differing in treatments or prognostic variables etc. If the hazard ratio is 2.0, then the rate of failure in one group is twice the rate in the other group. The computation of the hazard ratio assumes that the ratio is consistent over time, and that any differences are due to random sampling. Before performing any tests of hypotheses to compare survival curves, the **proportionality of hazards assumption** should be checked (and should hold for the validity of **Cox's proportional hazard models**). (See also **Log-rank test**).

Hetereoscedastic data: Data that have non-constant (heterogeneous) variance across the predicted values of y. In this case, residual graph will show varying variability across the fitted values. This is a regression diagnostic problem and should be fixed by transforming the data. See also **Homoscedasticity**.

Heuristics: A term in computer science that refers to guesses made by a program to obtain approximately accurate results. Frequently used in phylogenetics and computational biology.

Hierarchical model: In linear modeling, models which always include all the lower-order interactions and main effects corresponding to any interaction they include.

Historical fallacy: The mistake of assuming that an association observed in **cross-sectional data** will be similar to that observed in longitudinal data or vice versa.

Homoscedasticity (homogeneity of variance): Normal-theory-based tests for the equality of population means such as the t-test and analysis of variance, assume that the data come from populations that have the same variance, even if the test rejects the null hypothesis of equality of population means. If this assumption of **homogeneity of variance** is not met, the statistical test results may not be valid. **Heteroscedasticity** refers to lack of homogeneity of variances.

Hotelling's T^2 test: This is a generalization of Student's t-test for multivariate data. Designed to provide a global significance test for the difference between two groups with simultaneously measured multiple dependent/outcome variables and multiple explanatory/independent variables. It can also be used for one group with simultaneously measured multiple dependent outcome variables (another test similar to Hotelling's T^2 test is Mahalanobis's D^2 test). See **multivariate analysis**.

Hypergeometric distribution: A probability distribution of a discrete variable generally associated with sampling from a finite population without replacement. An example may be that given a lot with 25 good units and five faulty. The probability that a sample of five will yield not more than one faulty item follows a hypergeometric distribution.

Index plot: An index plot plots each **residual**, **leverage**, or **Cook's distance** against the corresponding observation (row) number (*i* or index) in the dataset. In many cases, the row number corresponds to the order in which the data were collected. If this is the case, this would be similar to plotting the residuals (or another diagnostic quantity) against time. The index plot is a helpful diagnostic test for normal linear and particularly generalized linear models. Both outliers and influential points can be detected by the index plot. It is particularly useful when the data is in time order so that pattern in the residuals, etc. over time can be detected. If a residual index plot is showing a trend in time, then they are not independent (violation of a major assumption of linear regression).

Inferential statistics: Making inferences about the population from which a sample has been drawn and analyzed.

Influential points: Observations that actually dominate a **regression analysis** (due to high **leverage**, high **residuals** or their combination). The method of ordinary **least squares** gives equal weight to every observation. However, every observation does not have equal impact on the **least squares** results. The slope, for example, is influenced most by the observations having values of the independent variable farthest from the mean. An observation is influential if deleting it from the dataset would lead to a substantial change in the fit of the **generalized linear model**. High-leverage points have the potential to dominate a regression analysis but not necessarily exert an influence (i.e., a point may have high leverage but low influence as measured by **Cook statistics**). Cook statistics is used to determine the influence of a data point on the model.

Interaction: If the effect of one factor depends on the level of another factor, the two factors involved are said to interact, and a contrast involving all these levels is called their interaction. Factors A and B interact if the effect of factor A is not independent of the level of factor B. For example, when there are two main effects on a response variable, if their combined effect is higher than the sum of their main effects due to a bonus (say, the effects of a kind of exercise and a kind of diet on blood lipid levels), they have an interaction (meaning a simple additive model is not sufficient to account for the observed data and a multiplicative term must be added). Briefly, interaction is a deviation from additivity. Also, there would be an interaction between the factors sex and treatment if the effect of treatment was not the same for males and females in a drug trial. Interaction is closely linked with **effect modification** in epidemiology (see **Genetic Epidemiology Glossary**; **Wikipedia**: **Statistical Interaction**).

Intercept: In linear regression, the intercept is the mean value of the response variable when the explanatory variable takes the value of zero (the value of y when x=0).

Interpolation: Making deductions from a model for values that lie between data points. Making deductions for values beyond the data points is called extrapolation and the results are not valid.

Interquartile range (dQ): dQ is a measure of spread and is the counterpart of the standard deviation for skewed distributions. dQ is the distance between the upper and lower quartiles (Q_U-Q_L).

Interval variable (equivalent to **continuous variable**): A quantitative variable measured on a scale with constant intervals (like days, milliliters, kilograms, miles so that equal-sized differences on different parts of the scale are equivalent) where the zero point and unit of measurement are arbitrary. When temperature is measured on two scales, Fahrenheit and Centigrade, the zero points in these two scales do not correspond, and a 10% increase in Fahrenheit (from 50^0 to 55^0) is not a 10% increase in the corresponding Centigrade scale (10^o to 12.8^o = 2.8%); these two measurements cannot be mixed or compared. For estimation of correlation coefficients, data should be interval type (See also **ratio variable** and **variable**).

Kolmogorov-Smirnov two-sample test: A non-parametric test applicable to continuous frequency distributions. It is considered to be the equivalent of the C^2-test for quantitative data and has greater power than the G-statistics or C^2-test for goodness of fit especially when the sample size is small. It can be used to compare two independent groups. The test is based on differences between two cumulative relative frequency distributions (it compares the distributions not the parameters). Thus, the **Kolmogorov-Smirnov** test is also sensitive to differences in the general shapes of the distributions in the two samples such as differences in dispersion, skewness. Its interpretation is similar to that of the Wald-Wolfowitz runs test. **Online Kolmogorov-Smirnov "One-Sample" Test** at **Vassar**.

Kruskal-Wallis test (One-way ANOVA by ranks): It is one of the non-parametric tests equivalent to one-way **ANOVA** that are used to compare *multiple* (k > 2) *independent* samples. This test assesses the hypothesis that the different samples in the comparison were drawn from the same distribution or from distributions with the same median. It can be used to analyze ordinal variables. It is an extension of the **Mann-Whitney (U) test** (for two independent samples). The interpretation of the Kruskal-Wallis test is identical to that of one-way ANOVA, but is based on ranks rather than means (**Online K-W test**; if the link does not work: go to **Stats ToolBoxs Homepage** and choose Kruskal-Wallis OWAV from Frequency Tables).

Kurtosis: Kurtosis is a measure of whether the data are peaked or flat in its distribution relative to a normal distribution (whose kurtosis is zero). Positive kurtosis indicates a 'peaked' distribution and negative kurtosis indicates a 'flat' distribution (data sets with high kurtosis have a distinct peak near the mean and decline rapidly; data sets with low kurtosis tend to have a flat top near the mean rather than a sharp peak) (**Definition of Kurtosis and Skewness**; **Online Skewness-Kurtosis Calculator**). See also **skewness**.

Large sample effect: In large samples, even small or trivial differences can become statistically significant. This should be distinguished from biological/clinical importance.

Least squares method: A method of fitting a straight line or curve based one minimization of the sum of squared differences (residuals) between the predicted and the observed points. Given the data points (x_i, y_i), it is possible to fit a straight line using a formula, which gives the y=a+bx. The gradient of the straight line b is given by $[S(x_i - m_x)(y_i - m_y)]$ / $[(S(x - m_x))^2]$, where m_x and m_y are the means for x_i and y_i. The intercept a is obtained by m_y - bm_x. See **Wikipedia**: **Least Squares**.

Leverage points: In regression analysis, these are the observations that have an extreme value on one or more explanatory variable. The leverage values indicate whether or not X values for a given observation are outlying (far from the main body of the data). A high leverage value indicates that the particular observation is distant from the centre of the X observations. High-leverage points have the potential to dominate a regression analysis but not necessarily influential. If the residual of the same data point and **Cook's distance** are also high, then it is an influential point. See also **influential points**.

Likelihood: The probability of a set of observations given the value of some parameter or set of parameters. For example, the likelihood of a random sample of n observations (x_1 to x_n) with probability distribution f(x; q) is given by: L = P f(x_i; q_0). This function, which applies equally to continuous density and discrete mass functions, is the basis of maximum likelihood estimation.

Likelihood distance test (likelihood residual, deletion residual): This test is based on the change in deviance when one observation is excluded from the dataset. It uses the difference between the log-likelihood of the complete dataset and the log-likelihood when a particular observation is removed. A relatively large difference indicates that the observation involved is an outlier (poorly fitted by the model).

Likelihood ratio test: A general purpose test of hypothesis H_o against an alternative H_1 based on the ratio of two likelihood functions one derived from each of H_o and H_1. The statistics l is given by l = -2 *ln* (L_{H0} / L_{H1}) has approximately a C^2 distribution with df equal to the difference in the number of parameters in the two hypotheses. One application of this test is the **G-test**, which is used in categorical data analysis as a goodness-of-fit or independence test (the tests statistics has a Chi-squared distribution).

Linear expression: A **polynomial** expression with the degree of **polynomial** being 1. It will be something like, f(x)=2x^1+3, but not x^2+2x+4.

Linear logistic model: A linear logistic model assumes that for each possible set of values for the independent (X) variables, there is a probability p that an event (success) occurs. Then the model is that Y is a linear combination of the values of the X variables: Y = b_o + b_1*X_1 + b_2*X_2 + b_3*X_3 + … b_k*X_k, where Y is the logit transformation of the probability p. Logistic in statistical usage stems from logit and has nothing to do with the military use of the word which means the provision of material.

Linear regression models: In the context of linear statistical modeling, 'linear' means linear in the parameters (coefficients), not the explanatory variables. The explanatory variables can be transformed (say, x^2), but the model will still be linear if the coefficients remain linear. When the overall function (Y) remains a sum of terms that are each an X variable multiplied by a coefficient, the function Y is said to be linear in the coefficients. A non-linear model is different in that it has a non-constant slope (a tutorial on **Simple Linear Regression**; see also **Vassar College**; Excel macro for **Linear Correlation & Regression**).

Linkage disequilibrium: Also called gametic association, which is more appropriate. It means increased probability for two or more alleles to be on the same chromosome at the population level. In a population at equilibrium, haplotype frequency is obtained by multiplying the allele frequencies x2. When there is linkage disequilibrium, the observed frequency (say by family analysis or sperm typing) is different from the expected frequency. The difference gives the D value (D for difference), which can be tested for significant difference from 0 by a 2x2 table analysis (for two alleles). The D value can be negative or positive. Linkage disequilibrium can derive from population admixture, tight linkage or elapse of insufficient time for the population to reach equilibrium. A classic example in immunogenetics is the HLA-A1-B8-DR3 haplotype which shows significant linkage disequilibrium extending over 6.5 Mb. Software for LD estimation: **Genetic Data Analysis**, **EH**, **2LD**, **MLD**, **LDA**, **PopGene**, **Online Linkage Disequilibrium Analysis**, **Genotype2LDBlock (online)**. For more, see **Basic Population Genetics**.

Link function: A particular function of the expected value of the response variable used in modeling as a linear combination of the explanatory variables. For example, logistic regression uses a *logit* link function rather than the raw expected values of the response variable; and Poisson regression uses the *log* link function. The response variable values are predicted from a linear combination of explanatory variables, which are connected to the response variable via one of these link functions. In the case of the general linear model for a single response variable (a special case of the generalized linear model), the response variable has a normal distribution and the link function is a simple identity function (i.e., the linear combination of values for the predictor variable(s) is not transformed).

LOD Score: Stands for the logarithm of odds. It is a statistical measure of the likelihood that two genetic markers occur together on the same chromosome and are inherited as a single unit of DNA (co-segregation). The LOD score serves as a test of the null hypothesis of free recombination versus the alternative hypothesis of linkage. Determination of LOD scores requires pedigree analysis and a score of >3 is traditionally taken as evidence for linkage. Linkage is between two genetic loci but not alleles. An example is the linkage between the hemochromatosis gene (HFE) and HLA-A. This means that within the same family all affected subjects will have the same HLA-A allele but not necessarily a particular one, i.e., there will be no recombination between HFE and HLA-A. LOD score has nothing to do with **linkage disequilibrium**.

Log 0 is undefined. If we need to use a log transformation but some data values are 0, the usual way to get round this problem is to add a small positive quantity (such as 1/2) to all the values before taking logs.

Log-rank test (of equality across strata): A non-parametric test of significance for the difference between two survival curves (for categorical data). It is a special application of the **Mantel-Haenszel test**. It can be adjusted for confounders (not preferable to **Cox proportional hazard regression** which is a semi-parametric model), or performed for trend (for details and parametric alternatives, see **Lee & Go, 1997**). It was developed by Mantel and Haenszel as an adaptation of the **Mantel-Haenszel C^2 test**. The other commonly used nonparametric tests for comparison of two survival distributions are the Gehan's Generalized Wilcoxon test (**Gehan 1965a** & **1965b**). The log-rank test is appropriate for survival distributions whose hazard functions are proportional over time, i.e. the two survival curves do not cross (**proportionality assumption**). Otherwise, the Gehan's Wilcoxon test is recommended. The Wilcoxon statistic puts more weight on early deaths compared to the log-rank (**Lee & Go, 1997**). See **UCLA Stata**: **Survival Analysis & Log-rank Test**; **BMJ Statistics Notes**: **Logrank Test**.

Log transformation: This transformation pulls smaller data values apart and brings the larger data values together and closer to the smaller data values (shrinkage effect). Thus, it is mostly used to shrink highly positively skewed data.

Logistic (binary) regression: A statistical analysis most frequently models the relationship between a dichotomous (binary) outcome variable (such as diseased or healthy; dead or alive; relapsed or not relapsed), and a set of explanatory variables of any kind (such as age, HLA type, blood pressure, kind of treatment, disease stage etc). It can also be used when the outcome variable is polytomous (several categories of the prognosis; including ordinal response 'ordinal logistic regression' or 'proportional odds ratio model'), and when there are several outcome variables (multinomial logistic regression - a special class of loglinear models). Analysis of data from case-control studies via logistic regression can proceed in the same way as cohort studies. See **Logistic Regression Lecture Note**, **Online Logistic Regression**, **Logistic Regression with SPSS, STATA** and **SAS**; **Power Calculation for Logistic Regression (including Interaction)**.

Logit transformation: The logit (or logistic) transformation Y of a probability p of an event is the logarithm of the ratio between the probability that the event occurs (p) and that the event does not occur (1-p): Y = ln (p/(1-p)). Thus it is a transformation of a binary (dichotomous) response variable. The logit transformation of p is also known as the *log odds* of p, since it is the logarithm of the **odds**. There are other link functions for binary response variables.

Loglinear model: Multinomial data (from contingency tables) can be fitted by using a generalized linear model with a Poisson response distribution and a log link function. The resulting models for counts in the cells of a contingency table are known as loglinear models in which the logarithm of the expected value of a count variable is modeled as a linear function of parameters. Loglinear models try to model all important associations among all variables. In this respect, they are related to **ANOVA** models for quantitative data. Loglinear modeling allows more than two discrete variables to be analyzed and interactions between any combination of them to be identified. Associations between variables in log-linear models are analogous to interactions in **ANOVA** models. The aim is to find a small model, which achieves a reasonable fit (small residual). Data sets with a binary response (outcome) variable and a set of explanatory variables that are all categorical can be modeled either by **logistic regression** or by loglinear modeling. More on **Loglinear Models** and **Online Loglinear Test** at **Vassar College**.

In loglinear modeling, the counts in the cells of the contingency table are treated as values of the response variable rather than either of the categorical variables defining the rows and columns. A loglinear model relates the distribution of the counts to the values of the explanatory variables (rows and columns), and tests for the presence of an interaction between them. Whether the interaction term is necessary is tested by fitting two models, one without and one with the interaction term, and using the difference between the regression deviance values for the two models (as well as the difference in df). If the SP obtained by the C^2-test is small, the interaction term should be in the model. This corresponds to a significant difference in Fisher's exact test or the C^2-test. For a 2x2 table, the difference between the regression deviances in the two models is the same as the residual deviance in the model with no interaction term. In loglinear modeling, it makes no sense if any main effect is omitted.

In loglinear modeling (as in other GLMs), the saturated model, which includes all interactions, fits the data exactly, and the fitted values are exactly equal to the observed values (the residual deviance is zero). The residual deviance of any other model can be used to test how worse it is compared to the saturated model. A small SP arising from a high residual deviance would mean that it does not fit better than the saturated model (the alternative model is rejected). If the associated SP is not small, then the model is not significantly worse than the saturated model (i.e., the exact fit).

An important constraint on the choice of models to fit, i.e., which terms and which interactions to include, is that whether any marginal totals (row or column) for combinations of terms are fixed in advance. If this applies to any combination of terms, their interaction must be included (the main effects and this particular interaction make up the null model for such data). When only the overall sample size is fixed in advance, there is no constraint on the models that can be fit.

When there are more than two variables, whether any interaction can be omitted from the model can be tested by fitting a model containing all the interactions of a particular order (say, second-order interactions). Then, new models omitting each interaction are fitted. Each new model's regression deviance is compared with the regression deviance of the model containing all interactions. If the resulting difference in regression deviances (and df) results in a large SP, that interaction can be omitted (it is not significantly different but has fewer terms; or it does not significantly contribute to the model).

The loglinear model can be used to find a conditional probability involving the factors in the contingency table for one of the factors chosen as a response variable. When the factor chosen as a response variable is binomial, logistic regression can be used to analyze the same data (a binomial response variable and categorical explanatory variables). Logistic regression substitutes the loglinear model with equal results as long as the fitted loglinear model includes the main effects of the response variable and a saturated model for all the explanatory variables.

Longitudinal data: Data collected over a period of time as in cohort studies. These data are usually analyzed by using **survival analysis** techniques. See **Longitudinal Studies Chapter** in **Epidemiology for the Uninitiated**.

Mann-Whitney (U) test: A non-parametric test for comparing the distribution of a continuous variable between two *independent* groups. It is analogous to the independent two-sample t-test, so that it can be used when the data are ordinal or not normally distributed. The Wilcoxon signed ranks T-test for independent samples is another non-parametric alternative to the t-test in this context (for *paired* samples, **Wilcoxon matched pairs signed rank test** should be used) (online **Mann-Whitney Test**).

Mantel-Haenszel C^2 test (also called Cochran-Mantel-Haenszel (CMH) test): Test for a null hypothesis of no overall relationship in a series of 2x2 tables for stratified data derived either from a cohort or a case-control study. It allows analysis of **confounding** and gives an **adjusted odds ratio** or relative risk. It can be used on categorical or categorized continuous data. The test is only valid when the variance of observed data is ^{3} 5. It is inappropriate when the association changes dramatically across strata (heterogeneity is usually tested by Breslow-Day test). It is, however, applicable for sparse data sets for which asymptotic theory does not hold for G^2. The test statistics, M^2, has approximately a Chi-squared distribution with df = 1 (see a review by **Stefano & Ezio, 2007**; online **Mantel-Haenszel Test** for a single table). Mantel & Haenszel's 1959 JNCI paper is now a **citation classic**.

Markov Chain Monte Carlo (**MCMC, random walk Monte Carlo) methods**: See **Wikipedia**: **MCMC**; **MCMC Applet**; **Markov Chain Simulation Applets**; **Buffon's Needle Applet**; **Monte Carlo Methods Links**; **MCMC Tests of Genetic Equilibrium (Lazzeroni & Lange, 1997)**; **Markov Chain Monte Carlo in Practice (Book)** and **Virtual Labs: Markov Chain** (requires **MathPlayer**). See also **Metropolis-Hastings algorithms**.

Maximum likelihood: This method is a general method of finding estimated values of parameters. It yields values for the unknown parameters, which maximize the probability of obtaining the observed values. The estimation process involves considering the observed data values as constants and the parameter to be estimated as a variable, and then using differentiation to find the value of the parameter that maximizes the likelihood function. First a likelihood function is set up which expresses the probability of the observed data as a function of the unknown parameters. The maximum likelihood estimators of these parameters are chosen to be those values, which maximize this function. The resulting estimators are those, which agree most closely with the observed data. This method works best for large samples, where it tends to produce estimators with the smallest possible variance. The maximum likelihood estimators are often biased in small samples (see **maximum likelihood**). Another method for point estimation is the **method of moments**.

McNemar's test: A special form of the Chi-squared test used in the analysis of paired (not independent) proportions. This non-parametric test compares two correlated dichotomous responses and finds its most frequent use in situations where the same sample is used to find out the agreement (concordance) of two diagnostic tests or difference (discordance) between two treatments. If the pairs of data points are the measurements on two matched people (such as affected and unaffected siblings) in a case-control study, or two measurements on the same person, the appropriate test for equality of proportions is the McNemar's test. It can be used to assess the outcome of two treatments applied to the same individuals or the significance of the agreement between two detection methods of a physical sign. If there are more than two periods of data collection (such as pretest, posttest and follow-up), **Cochran's Q test** should be used

(**Online McNemar's Test (1)** **(2)** **(3)** **(4)**.

Mean (or average): A measure of location for a batch of data values; the sum of all data values divided by the number of elements in the distribution. Its accompanying measure of spread is usually the **standard deviation**. Unlike the **median** and the **mode**, it is not appropriate to use the mean to characterize a skewed distribution (see also **standard deviation**) (**Online Calculator for Mean**).

Mean squares: A sum of squares divided by its associated df is a mean square. In **ANOVA**, the regression (explained) mean square is **ESS**/k-1, and the residual (error) mean square is **RSS**/N-k. Note that the mean squares are not additive, i.e., they do not add up to **TSS**/N-1. Importantly, the residual (error) mean square is an unbiased estimator of the variance (s^2) in ANOVA. The regression (explained) mean square equals to the variance only when the slope (b) is zero. Their ratio (mean square ratio = regression MS / residual MS) is therefore provides a test for the null hypothesis that b = 0. Large F values support the alternative hypothesis that the slope is not zero. This is the basis of the **F test** in **ANOVA**.

Measurement type: The data may be measured in the following scales: nominal, ordinal, interval or ratio scales (known as Stevens' typology). The scale of the measurements may be (other than nominal scale measurements) either continuous or discrete, and either bounded or unbounded.

Measures of association: These measures include the **Phi coefficient of association**, **Cramer's contingency coefficient (C) and V**, Kendall's tau-b and (Stuart's) tau-c, Somers' D (a modification of Kendall's tau-b), Yule's Q, gamma, Spearman's rank correlation coefficient (rho), Pearson's correlation coefficient, lambda (symmetric and asymmetric), uncertainty coefficients (symmetric and asymmetric), Guttman's coefficient of predictability (lambda)/Goodman-Kruskal's lambda, Goodman-Kruskal's gamma and Goodman-Kruskal's tau (concentration coefficient). See also **Measures of Effect-Size & Association**; **Measures of Association for Cross-tabulations**; Measures on **SAS**, **STATA**, **SYSTAT**; **Ennis, 2001**; **Morton, 2001**.

Measures of central tendency: These are parameters that characterize an entire distribution. These include **mode**, **median** and **mean**.

Median: Another measure of location just like the **mean**. The value that divides the frequency distribution in half when all data values are listed in order. It is insensitive to small numbers of extreme scores in a distribution. Therefore, it is the preferred measure of central tendency for a skewed distribution (in which the mean would be biased) and is usually paired with the **interquartile range (dQ)** as the accompanying measure of spread. See Martin Bland's page for calculation of **confidence intervals for a median**.

Median test: This is a crude version of the **Kruskal-Wallis ANOVA** in that it assesses the difference in samples in terms of a contingency table. The number of cases in each sample that fall above or below the common median is counted and the Chi-square value for the resulting 2xk samples contingency table is calculated. Under the null hypothesis (all samples come from populations with identical medians), approximately 50% of all cases in each sample are expected to fall above (or below) the common median. The median test is particularly useful when the scale contains artificial limits, and many cases fall at either extreme of the scale. In this case, the median test is the most appropriate method for comparing samples (**Online Median Test**).

Meta-analysis: A systematic approach yielding an overall answer by analyzing a set of studies that address a related question. This approach is best suited to questions, which remain unanswered after a series of studies. Meta-analysis provides a weighted average of the measure of effect (such as odds ratio). The rationale is to increase the power by analyzing the sets of data. The selection of studies to include in a meta-analysis study is the main problem with this approach. **Funnel Plot** is an informal method to assess the effect of publication bias in this context. See also **Introduction to Meta-Analysis** by the Cochrane Collaboration; **Meta-Analysis in Epidemiology** by Stroup et al (2000); **Methods for Meta-Analysis in Medical Research** by AJ Sutton; **Introduction to Meta-Analysis** by Borenstein et al (2009), and **Online Meta-Analysis Tests**.

Metropolis-Hastings algorithms: These algorithms are a class of Markov chains which are commonly used to perform large scale calculations and simulations in Physics and Statistics. See **Metropolis-Hastings Applet**. See **Markov Chain Monte Carlo methods**.

Mode: The observed value that occurs with the greatest frequency. The mode is *not* influenced by small numbers of extreme values.

Model building: The traditional approach to statistical model building is to find the most parsimonious model that still explains the data. The more variables included in a model (overfitting), the more likely it becomes mathematically unstable, the greater the estimated standard errors become, and the more dependent the model becomes on the observed data. Choosing the most adequate and minimal number of explanatory variables helps to find out the main sources of influence on the response variable, and increases the predictive ability of the model. Ideally, there should be more than 10 observations for each variable in the model. The usual procedures used in variable selection in regression analysis are: univariate analysis of each variable (using C^2 test), stepwise method (backward or forward elimination of variables; using the deviance difference), and best subsets selection. Once the essential main effects are chosen, interactions should be considered next. As in all model building situations in biostatistics, biological considerations should play a role in variable selection.

Monte Carlo trial: Studying a complex relationship difficult to solve by mathematical analysis by means of computer simulations. An online book on **Resampling Statistics**, and software (**CLUMP**, downloadable from **clump22.zip**) to do Monte Carlo statistics for case-control association studies.

Multicolinearity: In multiple regression, two or more X variables are colinear if they show strong linear relationships. This makes estimation of regression coefficients impossible. It can also produce unexpectedly large estimated standard errors for the coefficients of the X variables involved. This is why an exploratory analysis of the data should be first done to see if any collinearity among explanatory variables exists. Multicolinearity is suggested by non-significant results in individual tests on the regression coefficients for important explanatory (predictor) variables. Multicolinearity may make the determination of the main predictor variable having an effect on the outcome difficult.

Multiple regression: To quantify the relationship between several independent (explanatory) variables and a dependent (outcome) variable. The coefficients (a, b_1 to b_i) are estimated by the **least squares** method, which is equivalent to **maximum likelihood** estimation. A multiple regression model is built upon three major assumptions:
1. The response variable is normally distributed,
2. The residual variance does not vary for small and large fitted values (constant variance),
3. The observations (explanatory variables) are independent.
Multiple regression is the prototype for **general linear models** because the response variable should be normally distributed and there is no **link function**, whereas, simple linear regression is a special case for **generalized linear models**. The extension of multiple regression to multivariate data analysis is called canonical correlation (**Online Multiple Regression**; **Reference Guide on Multiple Regression**).

Multiple regression correlation coefficient (R^2 - R-squared): R^2 is a measure of the proportion of variability explained by, or due to the regression (linear relationship) in a sample of data (equal to the percentage of variance accounted for). It is a measure of the effect of X in reducing the uncertainty in predicting Y (the distance of data points to the best fir line). The R^2 statistic is given by R^2 = s.s. (regression) / s.s. (total) [i.e., ESS/TSS] where s.s. is the sum of squares. It is also called the **coefficient of determination**. When all observations fall on the fitted regression line, then s.s. (residual) = 0; and ESS=TSS; thus, $R^2 = 1$. When the fitted regression line is horizontal ($b = 0$), then s.s. (regression) = 0 and $R^2 = 0$. The adjusted R^2_a is different in that it takes into account the sample size and the number of explanatory variables. The fit of a model should never be judged from the R^2 value even if the statistical significance is achieved (for example, a high value may arise when the relationship between the two variables is non-linear; or a small R^2 may be significant due to large sample effect). The square root of the R^2 is the **correlation coefficient (r)** and has the same sign (minus or plus) as the slope. The r value does not have a clear-cut interpretation as the R^2 in linear regression. The R^2 and r are simply descriptive measures of the degree of linear association between X and Y in the sample observations (see a **Cautionary Note on R^2**; **Logistic Regression and R^2**; **Interpretation of r and R^2**; **Statistical Significance of Correlation Coefficient**; **GraphPad Guide to Correlation Parameters**; **Statistical Comparison of Two Correlation Coefficients**).

Multiplication rule: The probability of two or more statistically independent events all occurring is equal to the product of their individual probabilities. If the probability of having trisomy 21 is *a*, and trisomy 6 is *b*, assuming independence of these two events, for a baby the probability of having both trisomies is (*a* x *b*). One of the most critical errors of judgment in the use of independence assumption relates to a court case in the UK (**Watkins, 2000**). (See also **addition rule**.)

Multivariable analysis: As opposed to univariable analysis, statistical analysis performed in the presence of more than one explanatory variable to determine the relative contributions of each to a single event is (or should be) called multivariable analysis (in practice, however, it is called univariate and multivariate analysis more frequently). It is a method to simultaneously assess contributions of multiple variables or adjust for the effects of confounders. Multiple linear regression, multiple logistic regression, proportional hazards analysis are examples of multivariable analysis, which has no similarity whatsoever to **multivariate analysis** (see also **Peter TJ, 2009**). See a review on **Multivariable Methods by MH Katz** (book on **Multivariable Analysis by MH Katz**).

Multivariate analysis: Methods to deal with more than one related 'outcome/dependent variable' (like two outcome measures from the same individual) simultaneously with adjustment for multiple confounding variables (covariates). When there is more than one dependent variable, it is inappropriate to do a series of univariate tests. **Hotelling's T^2** test is used when there are two groups (like cases and controls) with multiple dependent measures (may be more than two), and multivariate analysis of variance (**MANOVA**) is used for more than two groups. Wilks' lambda is the most common parameter used to quantitate the significance of the result of a multivariate procedure (others are Roy's largest criterion, Hotelling-Lawley trace, and Pillai-Bartlett trace). Unfortunately, the word 'multivariate' is most frequently used instead of '**multivariable' analysis** (which means multiple independent/explanatory variables but one outcome/dependent variable; see also **Peter TJ, 2009**)). Canonical correlation, cluster analysis, factor analysis (FA) and principal component analysis (PCA) are types of multivariate analysis. PCA aims at reducing a large set of variables to a small set that still contains most of the information in the large set. While PCA summarizes or approximates the data using fewer dimensions (to visualize it, for example), FA provides an explanatory model for the correlations among the data. See an **Online Multivariate Statistics Book**, **Multivariate Statistical Methods: A Primer by BF Manly** and **MultiVariate Statistical Package-MVSP**.

Multivariate analysis of variance (MANOVA): An extension of **Hotelling's T^2** test to more than two groups with related 'multiple' outcome measures. Groups are compared on all variables simultaneously as a global test (rather than one-by-one as ANOVA does). See also **multivariate analysis**; **MANOVA on Stata**.

Natural (raw) residuals: The difference between the observed (Y_i) and fitted values (Y_i-hat) of the response variable in regression. To obtain the **standardized residual**, this value is divided by its estimated standard deviation. Other residual types are Pearson residual, deviance residual and **deletion residual**. In all cases, a relatively large residual indicates an observation poorly fitted by the model. The one that is most closely related to maximum likelihood estimation (i.e., **generalized linear models**) is the standardized deviance residual (which is equal to the standardized Pearson residual in normal linear models).

Negative predictive value: Probability of a true negative as in a person identified healthy by a test is really free from the disease (see also **positive predictive value)**.

Nested model: Models that are related where one model is an extension of the other.

Nominal variable: A qualitative variable defined by mutually exclusive unordered categories such as blood groups, races, sex etc. (see also **ordinal variable**).

Nonlinear Regression: Regression analysis in which the fitted (predicted) value of the response variable is a nonlinear function of one or more X variables. **GraphPad Guide to Nonlinear Regression**, **Introduction to Nonlinear Regression**, **A GraphPad Practical Guide to Curve Fitting**.

Nonparametric methods (distribution free methods): Statistical methods to analyze data from populations, which do not assume a particular population distribution. **Mann-Whitney U test**, **Kruskal-Wallis test** and **Wilcoxon's (T) test** are examples. Such tests do not assume a distribution of the data specified by certain parameters (such as mean or variance). For example, one of the assumptions of the Student's t-test and ANOVA is normal distribution of the data. If this is not valid, a non-parametric equivalent must be used. If a wrong choice of test has been made, it does not matter very much if the sample size is large (a non-parametric test can be used where a parametric test might have been used but a parametric test can only be used when the assumptions are met). For a small sample size, non-parametric tests tend to give a larger *P* value. In general, parametric tests are more robust, more complicated to compute and have greater power efficiency. Parametric tests compare parameters such as the mean in t-test and variance in F-test as opposed to non-parametric tests that compare distributions. Nonparametric methods are most appropriate when the sample sizes are small. In large (e.g., $n > 100$) data sets, it makes little sense to use nonparametric statistics (a tutorial on **Parametric vs Nonparametric Methods**; **Review of Nonparametric Tests** in **Intuitive Biostatistics; Nonparametric Tests for Ordinal Data** at **Vassar**).

Normal distribution (Gaussian distribution) is a model for values on a continuous scale. A normal distribution can be completely described by two parameters: mean (m) and variance (s^2). It is shown as C ~ N(m, s^2). The distribution is symmetrical with mean, mode, and median all equal at m. In the special case of m = 1 and $s^2 = 1$, it is called the standard normal distribution. See **Normal Distribution (1)**, **(2)**, **(3)** & **(4)**.

Normal probability plot of the residuals: A diagnostic test for the assumption of normal distribution of residuals in linear regression models. Each residual is plotted against its expected value under normality. A plot that is nearly linear suggests normal distribution of the residuals. A plot that obviously departs from linearity suggests that the error distribution is not normal.

Null model: A model in which all parameters except the intercept are 0. It is also called the intercept-only model. The null model in linear regression is that the slope is 0, so that the predicted value of Y is the mean of Y for all values of X. The F test for the linear regression tests whether the slope is significantly different from 0, which is equivalent to testing whether the fit using non-zero slope is significantly better than the null model with 0 slope.

Number needed to treat: The reciprocal of the reduction in absolute risk between treated and control groups in a clinical trial. It is interpreted as the number of patients who need to be treated to prevent one adverse event. See also **Number Needed to Treat (NNT) Guide by Bandolier (1) (2)**; **Interpreting Diagnostic Tests; NNT Calculator**; **GraphPad NNT Calculator**; **EpiMax Table Calculator**; **Evidence-based Medicine Toolbox** and articles by **Cook & Sackett, 1995**; **Wu, 2002**; **Barratt, 2004**.

Odds: The odds of a success is defined as the ratio of the probability of a success to the probability of a failure (p/(1-p)). If a team has a probability of 0.6 of winning the championship, the odds for winning is 0.6/(1-0.6) = 3:2. Similarly, the odd in a case-control study is the frequency of the presence of the marker divided by the frequency of absence of the marker (in cases or controls separately). The link function logit (or logistic) is the $\log_e$ of the odds.

Odds multiplier: In logistic regression, b = $\log_e$ (odds ratio), thus exp b = odds ratio. For a continuous (explanatory) variable, exp b is called the odds multiplier and corresponds to the odds ratio for a unit increase in the explanatory variable. The odds multiplier of the coefficient is the odds ratio for its level relative to the reference level. If x increases from a to b by c, the odds multiplier becomes exp (cb). The resulting value shows the proportional change in the odds associated with x = b relative to x = a. It follows that for binary variables where x can only get values of 0 and 1, exp b = odds ratio.

Odds ratio (OR): Also known as relative odds and approximate relative risk. It is the ratio of the odds of the risk factor in a diseased group and in a non-diseased (control) group (the ratio of the frequency of presence / absence of the marker in cases to the frequency of presence / absence of the marker in controls). The interpretation of the OR is that the risk factor increases the odds of the disease 'OR' times. OR is used in retrospective case-control studies (**relative risk** (RR) is the ratio of proportions in two groups which can be estimated in a prospective -cohort- study). These two and **relative hazard** (or **hazard ratio**) are measures of the strength/magnitude of an association. As opposed to the *P* value, these do not change with the sample size. OR and RR are considered interchangeable when certain assumptions are met, especially for large samples and rare diseases. Odds ratio is calculated as ad/bc where a,b,c,d are the entries in a 2x2 contingency table (hence the alternative definition as the cross-product ratio). In logistic regression, the coefficient b corresponds to the $\log_e$ of the odds ratio. There are statistical methods to test the homogeneity of odds ratios (**Online Odds-Ratio & 95% CI Calculation**; **Odds Ratio-Relative Risk Calculation** (**Calculator 3**) in **Clinical Research Calculators** at **Vassar**).

Offset: A fixed, already known regression coefficient included in a **generalized linear model** (which does not have to be estimated).

Omnibus test: If the chi-square test has more than one degree of freedom (larger than 2x2 table), it is called an 'omnibus' test, which evaluates the significance of an overall hypothesis containing multiple sub-hypotheses (these multiple sub-hypotheses then need to be tested using follow up tests).

One-way ANOVA: A comparison of several groups of observations, all of which are independent and subjected to different levels of a single treatment (such as cells exposed to different dosage of a growth factor). It may be that different groups were exposed to the same treatment (different cell types exposed to a new agent). The main interest focuses on the differences among the means of groups.

Ordinal variable: An ordered (ranked) **qualitative/categorical** variable. The degree of HLA matching (one, two, three or four antigen matching in two loci) in transplant pairs, or HLA sharing in parents (one-to-four shared antigens) are ordinal variables although the increments do not have to be equal in magnitude (see **interval** and **ratio variables**). An ordinal variable may be a categorized quantitative variable. When two groups are compared for an ordinal variable, it is inappropriate to use ordinary Chi-squared test but the **trend test** or its equivalents must be used.

Outcome (response, dependent) variable: The observed variable, which is shown on y axis. A statistical model shows this as a function of predictor variable(s).

Outlier: An extreme observations that is well separated from the remainder of the data. In regression analysis, not all outlying values will have an influence on the fitted function. Those outlying with regard to their X values (high **leverage**), and those with Y values that are not consistent with the regression relation for the other values (high **residual**) are expected to be influential. The test the influence of such values, the **Cook statistics** is used.

Overdispersion: Dispersion is a measure of the extent to which data are spread about an average. Overdispersion is the situation that occurs most frequently in Poisson and binomial regression when variance is much higher than the mean (normally, it should be similar). It is evident with a high (>2) residual mean deviance (which should normally be around one) and the presence of too many outliers. The reasons for overdispersion may be outliers, misspecification of the model, variation between the response probabilities and correlation between the binary responses. It distorts standard error and confidence interval estimations. In the analysis, overdispersion may be taken into account by estimating a dispersion parameter.

Overfitting: In a **multivariable model**, having more variables than can be justified from sample size. The statistical rule of thumb is to have at least ten subjects for each variable investigated.

Overmatching: When cases and controls are matched by an unnecessary non-confounding variable, this is called overmatching and causes underestimation of an association. For example, matching for a variable strongly correlated with the exposure but not with the outcome will cause loss of efficiency. Another kind of overmatching is matching for a variable which is in the causal chain or closely linked to a factor in the causal chain. This will also obscure evidence for a true association. Finally, if a too strict matching scheme is used, it will be very hard to find controls for the study. See **BMJ Statistics Notes**: **Matching**.

Parameter: A numerical characteristic of a population specifying a distribution model (such as normal or Poisson distribution). This may be the mean, variance, degrees of freedom, the probability of a success in a binomial distribution, etc.

Parsimonious: The simplest plausible model with the fewest possible number of variables.

Pearson's correlation coefficient (r): A measure of the strength of the 'linear' relationship between two quantitative variables. A major assumption is the normal distribution of variables. If this assumption is invalid (for example, due to outliers), the non-parametric equivalent **Spearman's rank correlation** should be used. The **r** represents C^2 obtained from the 2x2 table, corrected for the total sample size. It can then be calculated as $\pm(C^2/N)^{1/2}$. This formula is equivalent to covariance divided by the product of the standard deviations of the two variables. The **correlation coefficient**, r, can take any value between -1 and +1; 0 meaning no "linear" relationship (there may still be a strong non-linear relationship). It is the absolute value of r showing the strength of relationship. An associated *P* value can be computed for the statistical significance (a small *P* value does not necessarily mean a strong relationship). The square of the r is r^2 (r-squared or **coefficient of determination**) which corresponds to the variance explained by the correlated variable (see **GraphPad Guide to Correlation Parameters and Interpretation of r**). R^2 is also used in regression analysis (see **multiple regression correlation coefficient**; **Online Correlation & Regression Calculators** at **Vassar College**; **Excel Macro for Linear Correlation & Regression)**.

Pharmacoepidemiology: Application of epidemiological reasoning, methods and knowledge to the study of the uses and effects (beneficial and adverse) of drugs in human populations. A relatively new field in epidemiology becoming more closely related to pharmacogenetics. See a review on statistical analysis of pharmacoepidemiological case-control studies (**Ashby, 1998**).

Phi coefficient: A measure of association of two variables calculated from a contingency table as $(X^2 / N)^{1/2}$. Its value varies between 0 (no association) and 1 (strongest association) for 2x2 tables where it is an accurate statistics (for larger tables, **Cramer's V** is more accurate). In a way, it is a corrected Chi-squared value for the number of observations. See **Calculator 3** in **Clinical Research Calculators** at **Vassar**.

Poisson distribution: The probability distribution of the number of (rare) occurrences of some random event in an interval of time or space. Poisson distribution is used to represent distribution of counts like number of defects in a piece of material, customer arrivals, insurance claims, incoming telephone calls, or alpha particles emitted. A transformation that often changes Poisson data approximately normal is the square root. See **Poisson Distribution (QuickTime)**; **GraphPad Poisson Probability Calculator**.

Poisson regression: Analysis of the relationship between an observed count with a Poisson distribution (i.e., outcome variable) and a set of explanatory variables. In general it is appropriate to fit a Poisson model to the data if the sample size is > 100 and the mean for the occurrence of the event is <0.10xN.

Polynomial: A sum of multiples of integer powers of a variable. The highest power in the expression (n) is the degree of the polynomial. When n=1, for example, $f(x)=2x^1+3$, this is a linear expression. If n=2, it is quadratic (for example, $x^2 + 2x + 4$); if n=3, it is cubic, if n=4, it is quartic and if n=5, it is quintic.

Polytomous variable: A variable with more than two levels. If there are two levels it is called dichotomous (as in the most common form of **logistic regression**).

Population: The population is the universe of all the objects from which a sample could be drawn for an experiment.

Population attributable risk: The proportion of a disease in a specified population attributable to a specific factor (such as a genetic risk factor).

Population stratification (substructure): An example of 'confounding by ethnicity' in which the co-existence of different disease rates and allele frequencies within population sub-sections lead to an association between the two at a whole population level. Differing allele frequencies in ethnically different strata in a single population may lead to a spurious association or mask an association by artificially modifying allele frequencies in cases and controls when there is no real association (for this to happen, the subpopulations should differ not only in allele frequencies but also in baseline risk to the disease being studied). Case-control association studies can still be conducted by using genomic controls (**Devlin, 1999**; **Pritchard, 1999**) even when population stratification is present. The software **STRUCTURE & STRAT** or **ADMIXMAP** can be used to analyze case-control data with genomic control. See **Cardon & Palmer, 2003** for an example of spurious association due to population stratification. See also **Genetic Epidemiology**.

Positive predictive value: Probability of a true positive as in a person identified as diseased by a test is really diseased (see also **negative predictive value**).

Post hoc test: A test following another one. The most common example is performing multiple comparisons between groups when the overall comparison between groups shows a significant difference. For example, when an **ANOVA** analysis yields a small *P* value, *post hoc* tests (such as Newman-Keuls, Duncan's or **Dunnett's** tests) are done to narrow down exactly which pairs differ most significantly (similarly, **Dunn's test** is done in a non-parametric **ANOVA** setting) (**GraphPad Post ANOVA Test Calculator**). In genetic association studies, multiple comparisons are justified only when performed as a *post hoc* test following a significant deviation in overall gene/marker frequencies (see **HLA and Disease Association Studies**).

Power of a statistical test: See **Statistical Power**.

Predictor (explanatory, independent) variable: The variable already in hand in the beginning of an experiment or observation and whose effect on an outcome variable is being modeled.

Predictive value: The probability that a person with a positive test if affected by the disease (positive predictive value) or the probability that a person with a negative test does not have the disease (negative predictive value). Estimation requires sensitivity, specificity and disease prevalence.

Prevented fraction: The amount of a health problem that actually has been prevented by a prevention strategy in real world.

Probability: The ratio of the number of likely outcomes to the number of possible outcomes.

Probability density function: When a curve is used to model the variation in a population and the total area between the curve and the x-axis is 1, then the function that defines the curve is a probability density function.

Probability distribution function: A function which gives for each number x, the probability that the value of a continuous random variable X is less than or equal to x. For discrete random variables, the probability distribution function is given as the probability associated with each possible discrete value of the variable.

Probability vector: Any vector with non-negative entries whose sum is equal to 1.0. See **Wikipedia**.

Proportional odds ratio: When the response variable in an ordered/ordinal logistic regression model has more than two ordered response categories, odds ratio obtained for each category is called a proportional odds ratio. See **UCLA Stata**: **Ordinal Logistic Regression**; **Lecture note on Logistic Regression**.

***P* value (SP = significance probability)**: The *P* value gives the probability that the null hypothesis is correct; therefore, if it is a small value (like <0.05), null hypothesis is rejected. More technically, it is the probability of the observed data or more extreme outcome would have occurred by chance, i.e., departure from the null hypothesis when the null hypothesis is true. In a genetic association study, the *P* value represents the probability of error in accepting the alternative hypothesis (or rejecting the null hypothesis) for the presence of an association. For example, the *P* level of 0.05 (i.e., 1/20) indicates that assuming there was no relation between those variables whatsoever (the null hypothesis is correct), and we were repeating experiments like ours one after another, we could expect that approximately in every 20 replications of the experiment, there would be one in which the relation between the variables in question would be equal to or more extreme than what has been found. In the interpretation of a *P* value, it is important to know the accompanying measure of association and the biological/clinical significance of the significant difference. However small, a *P* value does not indicate the size of an effect (odds ratio/relative risk/hazard ratio do). A *P* value >0.05 does not necessarily mean lack of association. It does so only if there is enough power to detect an association. Most statistical nonsignificance is due to lack of power to detect an association (poor experimental design). Both 'p' and 'P' are used to indicate significance probability but the international standard is *P* (capital italic). See **Interpreting Statistical *P* Values**; **Interpreting Nonsignificant *P* values** in **Intuitive Biostatistics**.

Qualitative: Qualitative (**categorical**) variables define different categories or classes of an attribute. Examples are gender, blood groups or disease states. A qualitative (categorical) variable may be **nominal** or **ordinal**. When there are only two categories, it is termed binary (like sex, dead or alive).

Quantitative: Quantitative variables are variables for which a numeric value representing an amount is measured. They may be discrete (for example, taking values of integers) or continuous (such as weight, height, blood pressure). If a quantitative variable is categorized, it becomes an **ordinal variable**.

R^2 (R-squared): See **Multiple regression correlation coefficient** and **Pearson's correlation coefficient (for r^2)**.

Random sampling: A method of selecting a sample from a target population or study base using simple or systematic random methods. In random sampling, each subject in the target population has equal chance of being selected to the sample. Sampling is a crucially important point in selection of controls for a case-control study. By randomization, systematic effects are turned into error (term), and there is an expected balancing out effect: known and unknown factors that might influence the outcome are assigned equally to the comparison groups. One disadvantage of randomization is generation of a potentially large error term. This can be avoided by using a **block design**. See **Wikipedia**: **Random Sampling**.

Randomized (complete) block design: An experimental design in which the treatments in each block are assigned to the experimental units in random order. Blocks are all of the same size and each treatment appears in the same number of times within each block (usually once). A different level of the factor is assigned to each member of the block randomly. The data can be analyzed using the paired t-test (when there are two units per block) or by randomized block ANOVA (in blocks of any size). The results are substantially more precise than a completely randomized design of comparable size. In studies with a block design, more assumptions are required for the model: no interactions between treatments and blocks, and constant variance from block to block.

Ratio variable: A quantitative variable that has a zero point as its origin (like 0 cm = 0 inch) so that ratios between values are meaningfully defined. Unlike the **interval variables**, which do not have a true zero point, the ratio of any two values in the scales is independent of the unit of measurement. For example, 2/12 cm has the same ratio as the corresponding values in inch (but the same cannot be said for 2/12 Celsius and 2/12 Fahrenheit which are interval variables).

Receiver operating characteristics (ROC) curve analysis: Also called discrimination statistics. See **ROC** in **Clinical Research Calculators** and **Difference Between the Areas Under Two ROC Curves** at **Vassar**, **ROC101** by Tom Fawcett; **ROC Analysis by Obuchowski NA, 2005**; **Cook NR, 2007**. See also the **Supplementary Data File** for **Mamtani, 2006** for the use of **Stata in ROC analysis**.

Regression diagnostics: Tests to identify the main problem areas in regression analysis: normality, common variance and independence of the error terms; outliers, influential data points, collinearity, independent variables being subject to error, and inadequate specification of the functional form of the model. The purpose of the diagnostic techniques is to identify weaknesses in the regression model or the data. Remedial measures, correction of errors in the data, elimination of true outliers, collection of better data, or improvement of the model, will allow greater confidence in the final product. See also **error terms**, **residuals** (including **likelihood distance test**), **leverages** and **Cook statistics**.

Regression modeling: Formulating a mathematical model of the relationship between a response (outcome, dependent) variable, Y, and a set of explanatory (predictor, independent, regressor) variables, x. Depending on the characteristics of the variables, the choice of model can be simple linear regression, multiple regression, logistic (binary) regression, Poisson regression, etc. In any regression problem, the key quantity is the mean value of the outcome variable, given the value of the independent variable(s). This quantity is called the conditional mean and expressed as "E (Y½x)" where Y is the response (outcome), x is the explanatory (predictor) variable. The question is whether the variable(s) in question tells us more about the outcome variable than a model that does not include that variable. In other words, whether the coefficient of the variable(s) is zero and the outcome is equal to a constant (which is the mean for Y) or not. The aim of model building is to arrive at a meaningful (say, biologically relevant) and parsimonious model that explains the data. The model may be linear if the parameters are linear, or nonparametric if the parameters are not linear. No matter how strong is the statistical relationship between x and Y, no cause-and-effect pattern is necessarily implied by the regression models. See **Regression Applet**.

Regression towards the mean: See the explanation and a **simulation** at **Rice Virtual Lab in Statistics**; and a **Lecture Note**.

Relative risk (RR): Also known as **risk ratio**. The RR shows how many times more or less the individuals with the risk factor are likely to get the disease relative to those who do not have the risk factor. RR gives the strength of association in prospective cohort studies. It cannot be estimated in retrospective case-control studies, and its use to describe the cross-product ratio (as frequently done in HLA association studies) is inappropriate. See **Calculator 3** in **Clinical Research Calculators** at **Vassar**. See also **odds ratio**.

Repeated measures design: In this design, the same experimental unit is subjected to the different treatments under consideration at different points in time. Each unit, therefore, serves as a block. If for example, two different treatments and placebo treatment are applied to the same patient sequentially, this is a repeated measures design. See also **cross-over design**.

Resampling statistics: Data-based simulation procedures that sample (with replacement) repeatedly from observed data to generate empirical estimates of results that would be expected by chance. Examples include **bootstrapping** and permutation tests. See also **Online Resampling Book**.

Residuals: Residuals reflect the overall badness-of-fit of the model. They are the differences between the observed values of the outcome variable and the corresponding fitted values predicted by the regression line (the vertical distance between the observed values and the fitted line). In a regression analysis, a large residual for a data point indicates that the data point concerned is not close to its fitted value. If there are too many large (standardized) residuals either the model fitted is not adequate or there is **overdispersion** of the data. Ideally, the residuals should have constant variance along the line. This can be checked by a normal probability plot of the residuals. In the plot of residuals against the explanatory variable (or the fitted values), there should not be any pattern if the assumption of constant variation is met, i.e., residuals do not tend to get larger as the variable values get larger or smaller (see also **likelihood distance test**).

Residual plot: A graph that plots residuals against fitted values. It is used to check equal variance assumption of the error terms in linear regression. Residual analysis for logistic regression is more difficult than for general linear regression models because the responses Y_i can take only the values 0 and 1. Consequently, the residuals will not be normally distributed. Plots of residuals against fitted values or explanatory variables will be uninformative. Residual plots are generally unhelpful for **generalized linear models** (where **index plots** and **half-normal plots** are preferred).

Residual (error) sum of squares (RSS): The measure of within treatment groups sum of squares (variability) in ANOVA. It is the deviation around the fitted regression line. The sum of squared differences between each observed Y value (Y_i) and the fitted Y value (Y_i -hat) equals to the residual (error) sum of squares. See also **degrees of freedom** and **mean squares**.

Regression (explained) sum of squares (ESS): The measure of between treatment groups sum of squares (variability) in ANOVA. It is the deviation of fitted regression value around mean. The sum of squared differences between each fitted Y value (Y_i -hat) and the overall mean of the Y values equals to the explained (regression) sum of squares. The sum of ESS and RSS gives the total sum of squares **(TSS)** which equals to the sum of squared differences between each observed Y value (Y_i) and the overall mean (total deviation). See also **degrees of freedom** and **mean squares**.

Risk ratio (relative risk): The risk ratio is the percentage difference in classification between two groups obtained as the ratio of two risks or proportions. For example, the proportion of people recovering after stroke with one treatment equals 0.10, while the proportion after a different treatment equals 0.16. The risk ratio equals 0.625 (0.10/0.16); 37.5% ((1-0.625)*100 or (0.16-0.10)/0.16) fewer patients treated by the first method recover. The risk ratio takes on values between zero ('0') and infinity. One ('1') is the neutral value and means that there is no difference between the groups compared. See also **relative risk**.

Robustness: A statistical test or procedure is robust when violation of assumptions has little effect on the results. Student's t-test, for example, is robust against departures from normality.

R project for statistical computing: R is a language and environment for statistical computing and graphics which can be seen as a different implementation of the S language. R and a comprehensive set of programs written for a variety of statistical analysis are all available as Free Software. See the **R Project Website** & **List of Contributed R Packages**.

Sample size determination: Mathematical process of deciding how many subjects should be studied (at the planning phase of a study). Among the factors to consider are the incidence or prevalence of the condition, the magnitude of difference expected between cases and controls, the power that is desired and the allowable magnitude of type I error (pre-determined significance probability). **Sample size calculator, sample size calculation**.

SAS (Statistical Analysis System): A comprehensive computer software system for data processing and analysis. It can be used for almost any type of statistical analysis. Produced by **SAS Institute**. See **SAS Learning Resources (UCLA)**; **SAS Tutorials**; **Getting Started with SAS Enterprise Guide (Free Online Course)**; **SAS e-Learning**; **SAS Genetic**

Software and **Genetic Data Analysis**.

Saturated model: A model that contains as many parameters as there are data points. This model contains all main effects and all possible interactions between factors. For categorical data, this model contains the same number of parameters as cells and results in a perfect fit for a data set. The (residual) deviance is a measure of the extent to which a particular model differs from the saturated model.

Scales of measurement: The type of data is always one of the following four scales of measurement: nominal, ordinal, interval, or ratio. Each of these can be discrete or continuous.

Schoenfeld residual test: One of the diagnostic tests to check the proportionality assumption (covariates are time independent) in proportional hazard modeling. A variation is the use of scaled Schoenfeld residuals (see **Tests of Proportionality in SAS, Stata, R and SPLUS**).

Sensitivity: Sensitivity is the proportion of true positives that are correctly identified by a diagnostic test. Those that produce few false negatives have higher sensitivity. See also **specificity** (**Sensitivity and Specificity by Altman & Bland. BMJ 1994**), **Interpreting Diagnostic Tests** and **DAG-STAT**.

Sign test: A test based on the probabilities of different outcomes for any number of pluses and minuses, i.e., observations below or above a prespecified value. The sign test can be used to investigate the significance of the difference between a population median and a specified value for it, or between the observed sex/transmission ratio and the 50:50 expected value. It can also be used for paired data. This time, the differences between the pairs will be either negative or positive, and the smaller of the two total negatives or positives plus the total number of pairs will form the test statistics. For example, when the total number is 20, if the number for the less frequent sign is 5 or smaller, $P <$ 0.05 (two-tailed). A sign test in disguise is **McNemar's test**, which is used for paired data for dichotomous response.

Simple linear regression model: The linear regression model for a normally distributed outcome (response) variable and a single predictor (explanatory) variable. The straight line models the mean value of the response variable for each value of the explanatory variable. The major assumption is constant variation of residuals along the fitted line which points out that the model is equally good across all x values. The null hypothesis stating that the explanatory variable has no effect on the response (in other words, the slope of the fitted line is zero) can be tested statistically. The two main aims of regression analysis are to predict the response and to understand the relationships between variables. As in all linear models, the error term (shown as W_i or e_i) is additive (as opposed to multiplicative, i.e., $y_i = a + bx_i + e_i$) and independent, and they are assumed to have a normal distribution. As an exception, the simple linear regression is a special case for **generalized linear models**.

Skewness: The degree of (lack of) asymmetry about a central value of a distribution. A distribution with many small values and few large values is positively (right) skewed (long tail in the distribution curve or stemplot is to the right); the opposite (left tail) is negatively (left) skewed. The measures of location median, midinterquartile range (midQ) and midrange decrease in this order for a left-skewed distribution. (**Definition of Kurtosis and Skewness**; **Online Skewness-Kurtosis Calculator**; see also **kurtosis**).

Sparseness: A contingency table is sparse when many cells have small values. When N is the total sample size, and r and c are the number of rows and columns, N / rc is an index of sparseness. Smaller values refer to more sparse tables. Sparse tables often contain zero values (empty cell).

Spearman's rank correlation: A non-parametric **correlation coefficient (rho)** that is calculated by computing the **Pearson's correlation coefficient (r)** for the association between the ranks given to the values of the variables involved. It is used for ordinal data and interval/ratio data. It is *not* appropriate to take the square of Spearmen’s correlation coefficient rho to obtain **coefficient of determination** (r^2). It is also possible to **compare two correlation coefficients**. See also **Pearson's correlation coefficient**.

Specificity: Specificity is the proportion of true negatives that are correctly identified by the test. Those that produce few false positives have higher specificity. See also **sensitivity** (**Sensitivity and Specificity by Altman & Bland. BMJ 1994**), **Interpreting Diagnostic Tests** and **DAG-STAT**.

Square root transformation: Usually used for highly positively skewed data, but especially in transforming Poisson counts to normality.

Standard deviation: Like **variance**, the standard deviation (SD) is a measure of spread (scatter) of a set of data. Unlike variance, which is expressed in squared units of measurement, the SD is expressed in the same units as the measurements of the original data. It is calculated from the deviations between each data value and the sample mean. It is the square root of the variance. For different purposes, n (the total number of values) or n-1 may be used in computing the variance/SD. If you have a SD calculated by dividing by n and want to convert it to a SD corresponding to a denominator of n-1, multiply the result by the square root of n/(n-1). If a distribution's SD is greater than its **mean**, the mean is inadequate as a representative measure of central tendency. For normally distributed data values, approximately 68% of the distribution falls within ± 1 SD of the mean, 95% of the distribution falls within ± 2 SDs of the mean, and 99.7% of the distribution falls within ± 3 SDs of the mean (empirical rule). SD should not be confused with the **standard error of the mean (SEM)**, which is the SD of the sampling distribution of a statistics and quantifies how accurately the mean is known (See **Normal Distribution**; **Online Calculator for Standard Deviation**).

Standard error: The standard error (SE) or as commonly called the standard error of the mean (SEM) is a measure of the extent to which the sample mean deviates from the true but unknown population mean. It is the **standard deviation** (SD) of the random sampling distribution of means (i.e., means of multiple samples from the same population). As such, it measures the precision of the statistic as an estimate of a population. The (estimated) SE/SEM is dependent on the sample size. It is inversely related to the square root of the sample size:

(estimated) SE = SD / $(N)^{1/2}$

The true value of the SE can only be calculated if the SD of the population is known. When the sample SD is used (as almost always), it is an estimate and should be called estimated standard error (ESE). When the sample size is relatively large (N ³ 100), the sample SD provides a reliable estimate of the SE.

Standard residual: The standardized **residual** value (observed minus predicted divided by the square root of the residual **mean square**).

Stata: A powerful statistical package particularly useful for epidemiologic and longitudinal data management and analysis. It is mainly a command driven program produced by **Stata Corporation**. See the list of **Stata Capabilities**, **Stata Starter Kit** with **Learning Modules** by **UCLA**; **Tutorial by University of Essex**; **Tutorial by Princeton University**; **Stata Highlights by Notre Dame University**; **Tutorial by Carolina Population Center**; **Stata Refresher by Syracuse University**; **Genetic Data Analysis on Stata** and **Stata Programs for Genetic Epidemiologists**.

Statistical Power: The probability that a test will produce a significant difference at a given significance level is called the power of the test. This is equal to the probability of rejecting the null hypothesis when it is untrue, i.e., making the correct decision. It is 1 minus the probability of a type II error. The true differences between the populations compared, the sample size and the significance level chosen affect the power of a statistical test. Ideally, power should be at least 0.80 to detect a reasonable departure from the null hypothesis. See power calculators: **PS: Power and Sample Size Calculation**; **G*Power 3**; **General Statistical Calculators Including a Power Calculator** (**UCLA**); **Statistical Power Calculator for Frequencies**; **Retrospective Power Calculation**; **General Power Calculator**; **Power Calculation for Logistic Regression (including Interaction)**; **Genetic Power Calculator**; **Power for Genetic Association Analyses (PGA)**; **Quanto** (sample size and power calculation for association studies of genes, gene-environment or gene-gene interactions).

Stepwise regression model: A method in multiple regression studies aimed to find the best model. This method seeks a model that balances a relatively small number of variables with a good fit to the data by seeking a model with high R^2_a (the most parsimonious model with the highest percentage accounted for). The stepwise regression can be started from a null or a full model and can go forward or backward, respectively. At any step in the procedure, the statistically most important variable will be the one that produces the greatest change in the log-likelihood relative to a model lacking the variable. This would be the variable, which would result in the largest likelihood ratio statistics, *G* (a high percentage accounted for gives an indication that the model fits well). See also **multiple regression correlation coefficient - R^2**.

Stochastic model: A probability model that includes chance events in the form of random measurement error or uncertainty. In a deterministic model, however, random error is inconsequential or nonexistent. See **Wikipedia**: **Stochastic Modeling**.

Stratum (plural strata): When data are stratified according to its characteristics, each subgroup is a stratum.

Student's t-test: A parametric test for the significance between means (**two-samples t-test**) or between a mean and a hypothesized value (**one-sample t-test**). One assumption is that the observations must be normally distributed, and the ratio of variances in two samples should not be more than three. If the assumptions are not met, there are non-parametric equivalents of the t-test to use (see for example, **Wilcoxon's Test**). It is inappropriate to use the t-test for multiple comparisons as a ***post hoc*** **test**. The t-test for independent samples tests whether or not two means are significantly different from each other but only if they were the only two samples taken (**Tutorials on t-test (1)** & **(2)**; **Online t-test (GraphPad)**; **Online t-test**; **Online One-Sample t-test**; **Two-Sample t-test Power Calculator; Online t-test for Samples with Unequal Variance** at **Vassar**). (**Tables of critical values of t, F and Chi-square**).

Subgroup analysis: Analysis of subgroups of a sample either because of a prior hypothesis (gender or age-specific effect/association) or as a fishing expedition / data dredging. This practice increases type I error rates. See a commentary by **Sleight, 2000**.

Survival Analysis: See **Superlectures** on 'Survival Analysis'; **A Primer on Survival Analysis (J Nephrol 2004)**; **Tutorials on Survival Analysis in Br J Cancer 2003: Part I - II - III - IV**; '**Understanding Survival Curves**' at NMDP website; **Survival Curves** and **Comparing Survival Curves** in **Intuitive Biostatistics**; **BMJ Statistics Notes**: **Time to Event (Survival) Data - Survival Probabilities (the Kaplan-Meier method) - Logrank Test**; Survival Analysis by **STATA** and **SAS**; **Power Calculator for Survival Outcomes**, **PS: Power and Sample Size Calculation**. **Online survival analysis** at **Vassar College**. For a comprehensive review, see **Lee & Go, 1997**.

Survival function: A time to failure function that gives the probability that an individual survives past a time point (does not experience an event like death, metastasis, conception etc)). Where the event is death, the value of the survival function at time T is the probability that a subject will die at some time greater than T. The survival function always has a value between 0 and 1 and is nonincreasing.

Synergism: A joint effect of two treatments being greater than the sum of their effects when administered separately (positive synergism) or the opposite (negative synergism).

Theta (q): Used to denote recombination fraction (in statistical genetics).

Transformations (ladder of powers): Transformation deals with non-normality of the data points and non-homogeneous variance. The power transformations form the following ladder: ..., x^{-2}, x^{-1}, $x^{-1/2}$, log x, $x^{1/2}$ **;** x^{1}, x^{2}, x^{3}, ... Provided $x > 1$, powers below 1 (such as $x^{1/2}$ or log x) reduce the high values relative to the low values as in positively skewed data, whereas, powers above 1 (such as x^{2}) have the opposite effect of stretching out high values relative to low ones, as in negatively skewed data. All power transformations are monotonic when applied to positive data (they are either increasing or decreasing, but not first increasing and then decreasing, or vice versa). The **square root transformation** often renders Poisson data approximately normal.

Transmission Disequilibrium Test (TDT): A family-based study to compare the proportion of alleles transmitted (or inherited) from a heterozygous parent to a disease-affected child. Any significant deviation from 0.50 in transmission ratio implies an association (**Spielman, 1993** & **1994**).

Treatment: In experiments, a treatment is what is administered to experimental units (explanatory variables). It does not have to be a medical treatment. Fertilizers in agricultural experiments; different books and multimedia methods in teaching; and chemotherapy of bone marrow transplantation in the treatment of leukemia are examples of treatments in regression analysis.

Trend test for counts and proportions: A special application of the Chi-squared test (with a different formula) for ordinal data tabulated as a 2xk table. It should be used when the intention is not just to compare the differences between the two groups but to see whether there is a consistent trend towards decrease or increase in the difference between the groups. An example is the association of parental HLA sharing (one-to-four antigens in two loci) with fetal loss in a case-control study (those with recurrent miscarriages and normal fertile couples). A frequent application is the analysis of dose-response relationships. The Chi-squared test for trend has one degree of freedom. The associated *P* value obtained by the trend test is always smaller than the corresponding *P* value of an ordinary Chi-squared test. The trend test for counts and proportions is called Cochrane-Armitage trend test. Alternative tests for the analysis of trend are **Wilcoxon-Mann-Whitney test** or the t-test with use of ordered scores (See **Cochrane-Armitage Trend Test**; **Analysis of Association, Confounding and Interaction**; **Trend for Binomial Outcome** in the manual of Epi Info; **Epi Info Freeware for Trend Test** (Trend Test in StatCalc); **InStat** fully functional demo version for Trend Test; **Trend Tests in Stata**).

t-statistics: Defined as difference of sample means divided by standard error of difference of sample means (see **Student's t-test**).

Two-way ANOVA: This method studies the effects of two factors (with several levels) separately (main effect) and, if desired, their effect in combination (interaction).

Type I error: If the null hypothesis is true but we reject it this is an error of first kind or type I error (also called a error). This results in a false positive finding.

Type II error: If the null hypothesis is accepted when it is in fact wrong, this is an error of the second kind or type II error (also called b error). This results in a false negative result.

Unreplicated factorial: A single replicate of a 2^{k} design (where each of k factors of interest has only two levels).

Variable: Some characteristic that varies among experimental units (subjects) or from time to time. A variable may be **quantitative** or **categorical**. A quantitative variable is either **discrete** (assigning meaningful numerical values to observations: number of children, dosage in mg) or **continuous** (such as height, weight, temperature, blood pressure; also called **interval variable**). A categorical variable is either **nominal** (assigning observations to categories: gender, treatment, disease subtype, groups) or **ordinal** (ranked variables: low, median, high dosage). Conventionally, a random variable is shown by a capital letter, and the data values it takes by lower case letters.

Variance: The major measure of variability for a data set. To calculate the variance, all data values, their mean, and the number of data values are required. It is expressed in the squared unit of measurement. Its square root is the **standard deviation**. It is symbolized by s^2 for a population and S^2 for a sample (**Online Calculator for Variance and Other Descriptive Statistics**).

Variance ratio: Mean square ratio obtained by dividing the mean square (regression) by mean square (residual). The variance ratio is assessed by the F-test using the two degrees of freedom (k-1, N-k).

Wald test: A test for the statistical significance of a regression coefficient. It is obtained by comparing the maximum likelihood estimate of the slope parameter (expected b_1) to an estimate of its standard error. The resulting ratio (*W*), under the hypothesis that b_1 = 0, will follow a standard normal distribution. The two-tailed *P* value will be found from the Z table corresponding to *P* (ç Z ç > *W*). It is not more reliable than the **likelihood ratio test** (**deviance difference**).

Welch-Satterthwaite t-test: The Welch-Satterthwaite t-test is an alternative to the pooled-variance t-test, and is used when the assumption that the two populations have equal variances seems unreasonable. It provides a t statistic that asymptotically approaches a t-distribution, allowing for an approximate t-test to be calculated when the population variances are not equal (**Online Welch's Unpaired t-test**; **t-Test Assuming Unequal Sample Variances** at **Vassar**).

Wilcoxon *matched pairs* signed rank T-test: A non-parametric significance test analogous to paired t-test. Most suitable for *paired* **ordinal** or **interval/ratio variables** (**Online Wilcoxon's Matched-Pairs Signed-Ranks (W) Test** at **Vassar College**; **Online Wilcoxon's Matched-Pairs Signed-Ranks Test**). It is normally used to test for the significance of the difference between the distributions of two independent samples (repeated measures or matched pairs) instead of t-test when normality is violated or in doubt (**Wilcoxon's Two Independent Samples Test**). See also **Wilcoxon Ranked Sum Test**.

William's correction (for **G statistics**)**:** This is equivalent to Yates' continuity correction for Chi-squared test but used in likelihood ratio (G) statistics for 2x2 tables (**Online G Statistics** with William's correction).

Woolf-Haldane analysis: A method first described by Woolf and later modified by Haldane for the analysis of 2x2 table and relative incidence (**relative risk**) calculation. It is the preferred method for relative risk calculation when one of the cells has a zero using the formula: RR = (2a+1)(2d+1) / (2b+1)(2c+1). Since it is a modification of the cross-product ratio, it should be called **odds ratio**. For details and references, see **Statistical Analysis in HLA and Disease Association Studies**.

Yates's correction: The approximation of the Chi-square statistic in small 2x2 tables can be improved by reducing the absolute value of differences between expected and observed frequencies by 0.5 before squaring. This correction, which makes the estimation more conservative, is usually applied when the table contains only small observed frequencies (<20). The effect of this correction is to bring the distribution based on discontinuous frequencies nearer to the continuous Chi-squared distribution. This correction is best suited to the contingency tables with fixed marginal totals. Its use in other types of contingency tables (for independence and homogeneity) results in very conservative significance probabilities. This correction is no longer needed since exact tests are available.

Z score: The Z score or value expresses the number of standard errors by which a sample mean lies above or below the true population mean. Z scores are standardized for a distribution with mean = 0 and standard deviation = 1 (**Corresponding *P* values for Z**; **Vassar**: **Z to P Calculator**). The Z-statistics is defined as difference of sample proportions divided by standard error of difference of sample proportions (**Online Two-Sample Z-Test**, **Online One-Sample Z-Test**).

Major Resources in Biostatistics

Armitage P & Colton T. Encyclopedia of Biostatistics. Volumes 1-8. John Wiley & Sons, 2005

Bland M. An Introduction to Medical Statistics. 3rd Edition. Oxford Medical Publications, 2000

Campbell MJ & Machin D. Medical Statistics: A Common Sense Approach. Wiley, 2002

Daly LE & Bourke GJ. Interpretation and Uses of Medical Statistics. 5th Edition. Blackwell Scientific Publications, 2000

Motulsky H. Intuitive Statistics. OUP, 1995

Norman GR & Streiner DL. PDQ Statistics. BC Decker, 1997

Rosner B. Fundamentals of Biostatistics. 5th Edition. Duxbury Press, 1999

Sokal RR, Rohlf FJ. Biometry. 3rd Edition. WH Freeman & Company, 1994

Zar JH. Biostatistical Analysis. 4th Edition. Prentice Hall, 1998

Elston, Olson & Palmer. Biostatistical Genetics and Genetic Epidemiology. Wiley, 2002

Internet Links

Extensive Epidemiology and Biostatistics Links   Understanding the Fundamentals of Epidemiology

Epidemiology & BioStatistics Super Lectures   Epidemiology-ResearchEasy

Centre for Evidence-based Medicine

Clinical Epidemiology & Evidence-Based Medicine Glossary (1) (2)   Evidence-based Practice

Glossary of Statistical Terms (Berkeley)

A Glossary for Multilevel Analysis

Epidemiology - Biostatistics Board Review

Online Handbook of Biological Statistics (PDF)   Essential Statistics in Biology

Commonly Used Statistical Tests   Interpreting Diagnostic Tests (DAG-STAT)

Rice Virtual Lab in Statistics (including simulations)   Statistic Simulations

StatPrimer: Statistics for Public Health Practice   StatNotes: An Online Statistics Textbook

Improving Medical Statistics   Medical Statistics Misadventures   Introductory Biostatistics (e-Medicine)

Downloadable Statistical Books / Papers

ONLINE STATISTICAL ANALYSIS

A Compilation of Online Analyses

Concepts and Applications of Inferential Statistics & Online Statistics Site (Vassar): TOC (IE users)

Clinical Research Calculators at Vassar

Statistical Online Computational Resource (SOCR)

EpiMax Table Calculator   Evidence-based Medicine Toolbox   OpenEpi-Epidemiologic Calculators

Globally Accessible Statistical Procedures - GASP

The Chinese University of Hong Kong Statistical Tools Pages   GraphPad QuickCalcs   PHYLIP Online

StatCrunch Online Statistics   Online Statistical Analysis   HyperStat Statistics Online   Wessa

Online LD Analysis   Genotype2LDBlock (online)

Simple SNP Data Analysis (incl HWE): SNPStats & HWA

Partition for Online Bayesian Analysis   Free Statistics on the Web

TEXTBOOKS

Epidemiology Textbook   Principles of Epidemiology

Statistics at Square One   Statistics Notes: BMJ CMAJ Radiology

SCOR: Probability & Statistics eBOOK & Educational Materials

Reference Guide on Statistics (& Glossary)   Data Analysis BriefBook (Contents)

SticiGui (Statistics Tools for Internet & Classroom Instruction with a Graphical User Interface)

Introductory Statistics (DW Stockburger)

StatsDirect Help STATISTICA Glossary Statistical Tables

Learning by Simulations WISE (Web Interface for Statistical Education): Tutorials

SMART (Explorapedia of Statistical & Mathematical Techniques)

InStat Guide to Choosing the Right Test The Prism Guide to Interpreting Statistical Results Resampling

Applying the Right Statistics: Analyses of Measurement Studies by M Bland Seven Common Errors

Statistics to Use AS/A2 Mathematics & Statistics Modules

Multimedia Statistics Statistical Animal Models STATA (Tutorial) SAS (e-Learning)

JMP JMP GENOMICS WINKS GENSTAT R S-PLUS NCSS PASS DAG-STAT

SPSS (Training; Multimedia Tutorial) SYSTAT SigmaStat STATISTICA

InStat Epi Info ViSta 6.4 MVSP LISREL StatsDirect

PAST (Paleontological Statistics Software Package for Education and Data Analysis) (Hammer, 2001)

Statistics Software Discussion List Subscription Services

Animated Glossary Virtual Laboratories in Probability and Statistics

Statistical Terms Statistics Glossary (1) (2) MV Stat Glossary Statistics.Com Glossary

Discussion Groups: MedStats AllStat

Links to Mathematics & Statistics Sites

***Address for bookmark*: http://www.dorak.info/mtd/glosstat.html**

M.Tevfik DORAK, MD, PhD

Last updated on 2 December 2010

Genetics Population Genetics Genetic Epidemiology Bias & Confounding Evolution HLA MHC Homepage