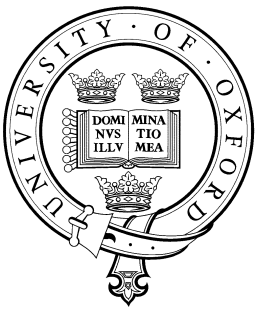
ELEMENTARY STATISTICS: BRIEF NOTES AND DEFINITIONS



Department of Statistics
University of Oxford

October 2015

ANALYSIS OF VARIANCE- The sum of squared deviations from the mean of a whole sample is split into parts representing the contributions from the major sources of variation. The procedure uses these parts to test hypotheses concerning the equality of mean effects of one or more factors and interactions between these factors. A special use of the procedure enables the testing of the significance of regression, although in simple linear regression this test may be calculated and presented in the form of a t-test.

ASSOCIATION- A relationship between two or more variables.

BAYES' RULE or BAYES' THEOREM- Consider two types of event. The first type has three mutually exclusive outcomes, A, B and C. The second type has an outcome X that may happen with either A, B or C with probabilities that are known: $P(X|A)$, $P(X|B)$ and $P(X|C)$, respectively. The probability that X happens is then

$$P(X) = P(A).P(X|A) + P(B).P(X|B) + P(C).P(X|C)$$

and Bayes' Theorem gives the probability that A happens given X has happened as

$$P(A|X) = \frac{P(A).P(X|A)}{P(X)}$$

Bayes' Rule uses this probability to determine the best action in such situations.

BIAS- An estimate of a parameter is biased if its expected value does not equal the population value of the parameter.

BINOMIAL PROBABILITY DISTRIBUTION- In an experiment with n independent trials, where each event results in one of two mutually exclusive outcomes that happen with constant probabilities, p and q respectively, this distribution specifies the probabilities that $r = 0, 1, 2, ..., n$ outcomes of the first type occur. The distribution is discrete and if $p < 0.5$ is positive skew and if $p > 0.5$ is negative skew. For n not small and p near to 0.5, the distribution may be approximated by the normal distribution. For n large and p small, the distribution may be approximated by the Poisson distribution. In many experiments, due to the heterogeneity of subjects, the Binomial fitted to the mean using $\bar{x} = np$ does not fit well as the sample variance exceeds the theoretical variance (npq). The Binomila generalizes to the Multinomial when events have more than two outcomes.

BOX PLOT or BOX-AND-WHISKER PLOT- Graph representing the sample distributiuon by a rectangular box covering the quartile range and whiskers at each end of the box indicating central dispersion beyond the quartiles. Outliers are shown as separate points in the tails.

CATEGORICAL DATA- Observations that indicate categories to which individuals belong rather than measurements or values of variables. Often such data consists of a summary that shows the numbers or frequencies of individuals in each category. This form of data is known as a 'frequency table' or, if there are two or more categorised features, as 'a contingency table'.

CENTRAL LIMIT THEOREM- If independent samples of size n are taken from a population the distribution of the sample means (known as the sampling distribution of the mean) will be approximately normal for large n. The mean of the sampling distribution is the population mean and its variance is the population variance divided by n. Note that there is no need for the distribution in the population sampled to be normal.

The theorem is important because it allows statements to be made about population parameters by comparing the results of a single experiment with those that would be expected according to some null hypothesis.

CHI-SQUARE- χ^2 The name of a squared standard normal variable that has zero mean and unit variance but is more often used for the sum of v independent squared standard normal variables. The mean of this quantity is v, and this is termed the "degrees of freedom". Common uses include:

1. to test a sample variance against a known value;
2. to test the homogeneity of a set of sample means where the population variance is known;
3. to test the homogeneity of a set of proportions;
4. to test the goodness-of-fit of a theoretical distribution to an observed frequency distribution;
5. to test for association between categorised variables in a contingency table;
6. to construct a confidence interval for a sample variance

COEFFICIENT OF VARIATION A standardised measure of variation represented by $100\times$ the standard deviation divided by the mean.

COMBINATION- A selection of items from a group.

CONDITIONAL PROBABILITY- Where two or more outcomes are observed together, the probability that one specified outcome is observed given that one or more other specified outcomes are observed.

CONFIDENCE INTERVAL- The interval between the two values of a parameter known as CONFIDENCE LIMITS.

CONFIDENCE LIMITS- Two values between which the true value of a population parameter is thought to lie with some given probability. This probability represents the proportion of occasions when such limits calculated from repeated samples actually include the true value of the parameter. An essential feature of the interval is that the distance between the limits depends on the size of the sample, being smaller for a larger sample.

CONTINGENCY TABLE- A table in which individuals are categorized according to two or more characteristics or variables. Each cell of the table contains the number of individuals with a particular combination of characteristics or values.

CORRELATION COEFFICIENT- A measure of interdependence between two variables expressed a value between -1 and +1. A value of zero indicates no correlation and the limits represent perfect negative and perfect positive correlation respectively. A sample product moment correlation coefficient is calculated from pairs of values whereas rank correlation measures such as 'Spearman's rho', which has population value ρ_S, are calculated after ranking the variables.

COVARIANCE- A measure of the extent to which two variables vary together. When divided by the product of the standard deviations of the two variables, this measure becomes the Pearson (or product-moment)correlation coefficient.

CRITICAL REGION- The set of values of a test statistic for which the Null Hypothesis is rejected. Often the region is disjoint, comprising two tails of the sampling distribution of the test statistic, but in a one-tail test the region is simply the one tail of that distribution.

CRITICAL VALUES- The values of a TEST-STATISTIC that bound the CRITICAL REGION of a test.

CUMULATIVE DISTRIBUTION FUNCTION (cdf)- For a random variable X, the cdf is the function $F(x) = P\{X \leq x\}$, so it starts at 0 and ends at 1, and gives the probability of X being no bigger than a given value. The EMPIRICAL cdf of a collection of data $x_1, \ldots, x_n$, also called the CUMULATIVE FREQUENCY, is the proportion of the data no bigger than a given number: $F(x) = \frac{1}{n}\#\{i : x_i \leq x\}$.

DESCRIPTIVE STATISTICS- Quantities that are calculated from a sample in order to portray its main features, often with a population distribution in mind.

DISPERSION- The spread of values of a variable in a distribution. Common measures of dispersion include the variance, standard deviation, range, mean deviation and semi-interquartile range.

EFFECT SIZE- That amount which it is desired to determine as significant when a test is used for the magnitude of a factor effect. If two groups with different means and the same standard deviations are to be compared then the effect size is the diffeence of the means divided by the standard deviation.

EXPECTATION- The mean value over repeated samples. The expected value of a function of a variable is the mean value of the function over repeated samples.

EXPLORATORY DATA ANALYSIS (EDA) The use of techniques to examine the main featurs of data, particularly to confirm that assumptions necessary for the application of common techniques may be made.

FACTOR- In experimental design, this term is used to denote a condition or variable that affects the main variable of interest. Factors may be quantitative (e.g. amount of training, dosage level) or qualitative (e.g. type of treatment, sex, occupation). In multivariate analysis, the term is used to describe a composite group of variables that serves as a meaningful entity for further analysis.

F-TEST- The test statistic is the ratio of two mean sums of squares representing estimates of variances that are expected to be equal on the null hypothesis. In general, F may be used to compare the variances of two independent samples. In an important special case, it is calculated from estimates based upon between and within samples variation. This is the single factor analysis of variance. In more complex designs this idea is extended. In these cases, the F-test is actually testing hypotheses concerning equality of means of groups that are subject to the action of one or more factors.

FREQUENCY TABLE. This table shows, for an ungrouped discrete variable, the values that a variable can take together with the counts of each value in the sample. For a grouped variable, whether discrete or continuous, the counts of sample values that fall within contiguous intervals are shown.

FRIEDMAN TEST- Test for the difference between the medians of groups when matched samples have been taken. It may be seen as the nonparametric counterpart of a two or more factor analysis of variance for a balanced design, but a separate analysis is completed for each factor and no tests for interaction effects are available.

GOODNESS-OF-FIT TEST. Occurrences of values of a variable in a sample are compared to the numbers of such values that would be expected under some assumed probability distribution. These occurrences are often in the form of a frequency table and the commonest tests used are χ^2-tests and Kolmogorov-Smirnov tests.

GROUPED DATA. Where data are represented by counts of values of a variable falling into intervals, the data are referred to as 'grouped'.

HOMOGENEITY OF VARIANCE- The property that is required for parametric tests using groups(populations) that may differ in their means but for the tests to be valid need to have their variances the same.

HYPOTHESIS TEST- A procedure for deciding between two hypotheses, the NULL and the ALTERNATIVE HYPOTHESIS, using a test-statistic calculated from a sample.

INDEPENDENCE- Where two events occur together and the happening of one does not affect the happening of the other, the events are said to be independent. Note that 'event' here may mean the taking of a value by a random variable. If two random variables are independent then they have zero covariance but zero covariance does not imply independence.

INDEX A summary measure that allows comparison of observations made at different points in time or in different populations. Commonly the value is standardised to a base value of 100 so that changes e.g. over time can be seen in percentage form. More complex indices combine several variables to represent some composite phenomenon.

INTERACTION- In an experiment, factors may act independently so that the effect of one factor is the same at all levels of another factor, or may interact so that the effect of one factor depends on the levels of another factor. The term is also used for the relation between characteristics of classification in a contingency table, but this phenomenon is usually called 'association'.

KOLMOGOROV-SMIRNOV TEST- This test compares two sample distributions or one sample distribution with a theoretical distribution.

KRUSKAL-WALLIS TEST- This tests for the difference between the medians of several populations. The total sample values are ranked and the rank sums for the separate samples are used to calculate a χ^2 statistic.

LEVEL OF SIGNIFICANCE- The probability that a test-statistic falls into the critical region for a test when the null hypothesis is true is usually expressed as a percentage and termed the significance level.

LOCATION PARAMETER- Measure indicating the central position of a distribution, e.g., mean, mode, median or mid-range.

MANN-WHITNEY TEST- Test for the difference between the medians of two independent samples that uses the rank sums calculated for the two samples from the ranks in the combined sample.

MATCHED PAIRS TEST- When pairs of units from a population are observed in an experiment to investigate the effect of a single factor, a matched pairs test is used. These pairs are identical in other respects but may differ for the factor of interest. Differences between matched pairs have smaller variance than those between unmatched units and this leads to a more efficient design. Sometimes, subjects are used as their own controls in a 'before-after' experiment and the values from the same subject are a better basis for testing the effect of a factor than values from different subjects.

MEAN- The arithmetic mean of a sample of values is the sum of these values divided by the sample size. The population mean is the expected value of a variable considering its probability distribution.

MEDIAN- The value which divides an ordered sample into two equal halves. In a population, the value that is exceeded, and not exceeded with equal probability.

MODE- The most frequent value in a sample. The population value that has greatest probability. There may be several modes in a distribution if separated values have high frequency (or probability) in relation to their neighbours. This is common for heterogeneous populations where the distribution comprise a mixture of simpler component distributions.

MUTUALLY EXCLUSIVE EVENTS- Events that cannot occur together in a sample.

NON-PARAMETRIC TEST- A test that does not depend on the underlying form of the distribution of the observations. Many such tests use ranks instead of the raw data, others use signs.

NORMAL PROBABILITY DISTRIBUTION- This symmetric continuous distribution arises in many natural situations, hence the name 'Normal' (although it is sometimes known as 'Gaussian'). It is unimodal and bell shaped. Within one standard deviation of the mean lies approximately 68% of the distribution, and within two standard deviations, approximately 95%. The Normal is used as an approximating distribution to many other distributions, both discrete and continuous, including the Binomial and the Poisson in circumstances when these distributions are symmetrical and the means are not small. But the major importance of the Normal is due to the CENTRAL LIMIT THEOREM and the fact that many sampling distributions are approximately normal so that statistical inferences can be made over a wide range of circumstances.

NULL HYPOTHESIS- An assertion or statement concerning the values of one or more parameters in one or more populations. Usually referred to as H_0. Often asserts equality of e.g. means of groups, there being *no difference* between them. The term ALTERNATIVE HYPOTHESIS is used for the statement that is true when the Null Hypothesis is untrue. Often known as H_a. Sometimes called the *Research Hypothesis*.

ONE-TAIL TEST- When the CRITICAL REGION for a test is concentrated in one tail of the distribution of a test-statistic. In particular if the Alternative Hypothesis under test states that a parameter value exceeds some specified value then the Null Hypothesis is tested against this one-sided alternative.

OUTLIER- An improbable value that does not resemble other values in a sample.

PARAMETER- A measure used in specifying a particular probability distribution in a population, e.g. mean, variance.

PERMUTATION- An ordering of the members of a group.

POINT ESTIMATE AND INTERVAL ESTIMATE- A single value calculated from a sample to represent a population parameter is called a point estimate. Where two values specify an interval within which the population value for a parameter is thought to lie these two values constitute an interval estimate.

POISSON DISTRIBUTION- The counted number of events in equal intervals of time or of points in equal units in space follows a Poisson distribution if the events or points are random and have constant probability over time or space and are independent. This distribution is defined for all positive integers and has the special property that its mean equals its variance. Because of this, departures from randomness may be detected: clustering leads to a variance that is larger than the mean whereas systematic evenness or regularity leads to a variance that is smaller than the mean.

POPULATION- As used in statistics, the aggregate of all units upon which observations could ever be made in repeated sampling. Sometimes it is appropriate to think of it as all the values of a variable that could be observed.

POWER OF A TEST- The probability that a statistical test correctly accepts the alternate hypothesis when the null hypothesis is not true. The power of a test is $1-\beta$, where β is the Type II error.

PROBABILITY- An expressed degree of belief or a limiting relative frequency of an event over an infinite sample.

PROBABILITY DISTRIBUTION- The set of values taken by a random variable together with their probabilities. Multivariate distributions define the combinations of values taken by several variables and the probabilities with which they occur.

QUARTILES- The values below which occur 25%, 50% and 75% of the values in a distribution. The median is the second quartile.

RANDOM SAMPLE- A sample in which every member of the population has the same chance of appearing and such that the members of the sample are chosen independently.

RANDOM VARIABLE- A variable that follows a probability distribution.

RANGE- The difference between the largest and smallest value in a sample, or for the population range, in the population.

REPLICATION- The taking of several measurements under the same conditions or combination of factors. This device is used to increase the precision of estimates of effects, by increasing the degrees of freedom for the residual or error sum of squares.

REGRESSION- The dependence of a variable upon one or more variables expressed in mathematical form. Simple linear regression refers to a straight line equation: $y = \alpha + \beta x$ in which the parameter α is the intercept and β the slope of the line.

REPEATED MEASURES- Where each subject is measured under each of a specified set of different conditions or at specified times.

RISK- The probability that an individual has a characteristic or that a particular event will take place. Used in risk analysis of systems to assess the probabilities of important events, e.g. disasters, or in medicine to consider the outcomes from treatment or surgery, or the characteristics of individuals that predispose them to certain diseases.

SAMPLE- A subset of the whole population, usually taken to provide insights concerning the whole population

SAMPLING DISTRIBUTION- The distribution of some specified sample quantity, most commonly the sample mean, over repeated samples of fixed size.

SCALE- The type of measurement or observation may be qualitative or quantitative: usually seperated into nominal (labels), ordinal (ordered categories), interval (equal quantities represented by intervals of the same width) or ratio (true zero and meaningful division of values). Importance: may affect the choice of techniques available for analysis.

SEMI-INTERQUARTILE RANGE- Half the difference between the first and third quartiles.

SIGN TEST- A test that uses the signs of observed values (or more often of differences between pairs of observed values) rather than their magnitude to calculate a test-statistic. Typically, + and − signs are equiprobable on the null hypothesis and a test-statistic is based on the number of +'s as this follows a binomial distribution with $p = 0.5$.

SIGNIFICANCE TEST- A null hypothesis describes some feature of one or more populations. We quantify the strength of disagreement between the data and the null hypothesis. After computing a TEST STATISTIC from the data, we reject the null hypothesis if the test statistic is in a set of values, the CRITICAL REGION for the test, that is unlikely to arise if the null hypothesis is true.

SIGNIFICANCE LEVEL- The probability, usually expressed as a percentage, that the null hypothesis is rejected by a significance test when it is in fact true.

SKEWNESS- A unimodal distribution is skew if the probability of exceeding or not exceeding the mode are not equal.

STANDARD DEVIATION- The square root of the variance

STANDARD ERROR- The standard deviation of the sampling distribution of an estimate of a population parameter.

STANDARD or STANDARDISED NORMAL VARIABLE- A normally distributed variable that has mean zero and variance 1.

STATISTIC- A measure calculated from a sample often used to estimate a population parameter. Note that STATISTICS may refer to several such measures or to the whole subject- the scientific collection and interpretation of observations.

STEM-AND-LEAF PLOT- A form of histogram that displays the actual values in intervals as well as their frequency. Useful EDA technique for showing bias in recording or for comparing two distributions.

STRATIFIED RANDOM SAMPLE- a sample that comprises random samples taken from the strata of a population.

TEST STATISTIC- A quantity calculated from a sample in order conduct a SIGNIFICANCE TEST. The test statistic has a sampling distribution from which tail values as critical values for use in the test.

t TEST. A test using a statistic that is calculated using sample standard deviations in the absence of knowledge of the population values. Most often used to test the difference between two means or two regression line gradients.

TRANSFORMATION- The process of replacing each observation by a mathematical function of it (e.g. log or square root) so that the data may conform to a normal distribution or have some desired property such as homogeneity of variance.

TWO-TAIL TEST- Uses two values to determine the critical region. The null hypothesis is rejected if the test statistic does not lie between these values.

TYPE I AND TYPE II ERRORS- When the null hypothesis is true but it is rejected by a statistical test an error of Type I has occurred. When the alternative hypothesis is true but it is rejected by the test an error of Type II has occurred. These errors are expressed as probabilities and are also referred to as α and β, respectively.

UNBIASED ESTIMATOR- One whose sampling distribution has for its mean the true value of the population parameter being estimated .

UNIMODAL DISTRIBUTION- A distribution that has just one mode.

VARIANCE- The mean of the squared deviations from the mean of a sample. For a population, the expected value of the squared deviation of a value from the mean.

WILCOXON MATCHED-PAIRS TEST- Uses the ranks of differences between pairs to test the hypothesis that the median difference is zero (two-tail test) or that it is greater (or less) than zero (one-tail test).

Z TEST- When the quantity (e.g. the difference between two means) calculated to test a particular hypothesis is believed to be normally distributed and its standard error can be determined from a known standard deviation, the statistic obtained by dividing the quantity by its standard error often follows a Z distribution and so a Z test is used.