SMW Consultants, Ltd.
Writing & Consulting Medical Research
Kylliäisentie 9, FIN-21620 Kuusisto, FINLAND.

The Most Common Statistical Tests in Biomedicine:

Basic Concepts

Foreword

Running the statistical analysis using appropriate tests is the prerequisite of obtaining accurate and meaningful results out of your study. Selecting and running statistical tests correctly are not straight-forward tasks to accomplish, however. Although common sense helps a lot e.g. by recognising what is a qualitative variable and which is quantitative, you still need experience and expertise how to do that. This is best left for professionals who have acquired these skills, and people who are purely medical writers should preferably rely on their expertise.

Under optimal conditions, however, this type of co-operation between medical writers and statisticians could have a significant impact towards improving the quality of research reporting.

Data are in different form

The first thing to check before starting any analysis is to examine what type of data you are at hand. The data variables recorded in any single study (be it clinical, experimental or epidemiological) can be of various types, and each type of data needs to be treated differently and different tests while making the statistical analyses. The main categories of data are the following:

Qualitative data:
- Nominal
- Ordinal

Quantitative data:
- Interval variables
- Ratio variables
- Discrete variables
- Continuous variables

Distinction between these different data types is essential to select the appropriate statistical tests for making the analyses that are suitable for treatment of the qualitative and quantitative form of data. The suitable tests are then selected in the commonly used statistical software.

Descriptive statistics

As the name implies, descriptive statistics are the set of tests and parameter which are used for description of the data. They are not used to make any statistical testing, but simply to give a rapid overview on which type of material and data we are dealing with. Running the tests for descriptive

statistics is the useful first step to start with analysis of any data file that you review for the first time. Typical examples of descriptive statistics are parameters such as the frequencies, means, medians, standard deviations, standard errors, modes, etc.

In the commonly used software, a wide variety of options are available, when you select Descriptive statistics from the Analysis pop-up menu. Although the use of these descriptive statistics sounds logical and clear-cut, even these can be applied and particularly presented in inappropriate context, however.

Chi-square test

This is the single most common basic statistical test used in analysis of practically any data set at some stages. The chi-square test is used to evaluate whether proportions of certain categories are different in different groups. Typically, the data output is a table, where the proportions of different categories (no practical limit) are presented, stratified according to different groups (no limit how many).

The chi-squared test is by no means limited to analysis of simple 2 by 2 tables, but can be used to compare several groups and several categories of outcome variables. When more than two groups are compared, the test gives the overall difference among the groups, but does not tell us which groups differ from which other groups.

For chi-square test, the data must be ordered or unordered numeric categorical variables (ordinal or nominal levels of measurement). Some software provide automatic recode machinery to convert string variables to numeric variables. To perform chi-square test, you do not need to require assumptions about the shape of the underlying distribution. The data are assumed to be a random sample. The expected frequencies for each category should be at least 1, and to give reliable results, no more than 20% of the categories should have expected frequencies of less than 5.

Fisher's exact test

Generally speaking, the chi-square test works well only when the sample size is large. Caution in interpretation should be exercised when you find in the data output that some of the cells in your table contain values less than 5. Fisher's exact test is an inbuilt option in the chi-square test, to be used for the same general purpose, i.e., to test the hypothesis that some proportions of interest differ between the groups. This test is particularly suitable for small sample sizes.

While performing chi-square test with your software, you normally select chi-square to calculate the Pearson chi-square, the likelihood-ratio chi-square, Fisher's exact test, and Yates' corrected chi-square (continuity correction) for tables with two rows and two columns. In many software, like SPSS, Fisher's exact test is computed automatically for any 2 by 2 tables, when a table has a cell with an expected frequency of less than 5. In such cases, you should also use the p value indicated by Fisher's exact test to report the significance level of your result.

Student's t-test

In addition to chi-square test, Student's t-test is one of the most frequently used statistical tests. All tests based on these principles are frequently called simply as t-tests. Two main modifications of t-test exists: 1) paired samples t-test and 2) independent samples t-test. Both types are used to test the hypothesis that some variable differs between two groups. When the groups are paired, e.g. a case and its matched control, or the same subjects are measured twice (before and after treatment), the paired samples t-test is particularly suitable.

Paired samples t-test: To run a paired test, you need to specify two quantitative variables (interval- or ratio-level of measurement). For a matched-pairs or case-control study, the response for each test subject and its matched control subject must be in the same case in the data file. The test assumptions are that the observations for each pair should be made under the same conditions. The mean differences should be normally distributed. Variances of each variable can be equal or unequal.

Independent-samples t-test: Test procedure compares means for two groups of cases. Ideally, for this test, the subjects should be randomly assigned to two groups, so that any difference in response is due to the treatment (or lack of treatment) and not to other factors. This is not the case if you compare e.g. the average income for males and females, because a person is not randomly assigned to be a male or a female. In such situations, you should ensure that differences in other factors are not masking or enhancing a significant difference in means. Differences in average income may be influenced by factors such as education and not by sex alone.

An example of the appropriate use of independent samples t-test is a study, where patients with high blood pressure are randomly assigned to a placebo group and a treatment group. The placebo subjects receive an inactive pill and the treatment subjects receive a new drug that is expected to lower blood pressure. After treating the subjects for two months, the two-sample t test is used to compare the average blood pressures for the placebo group and the treatment group. Each patient is measured once and belongs to one group.

The software calculates the statistics for each variable: sample size, mean, standard deviation, and standard error of the mean, and for the difference in the means: mean, standard error, and confidence intervals.

ANOVA and ANCOVA

ANOVA is an acronym for ANalysis Of VAriance. Together with those listed above, ANOVA belongs among the most frequently used statistical tests. When appropriately used, it gives valuable information about the results of your study. Different modifications of ANOVA are available, depending on the purpose of use and type of data. One-way ANOVA is used to test the hypothesis that some variable of interest differs among groups. Two-way ANOVA can test for differences among groups, and controls for other categorical variables. ANCOVA (analysis of covariance) can control for continuous variables. ANOVA is very much an equivalent test to multiple linear regression (see below).

There are important assumptions that limit the use of ANOVA, however, and these must be recognised before running the test. The outcome variable measured in the test must be continuous, interval variable. ANOVA also requires that the data are normally distributed and that the variance of the data is the same in all groups. There is no restriction on the number of groups that can be analysed by ANOVA. In the case with only 2 groups, ANOVA is equivalent to the t-test (see above).

One-way ANOVA: The one-way procedure produces a one-way analysis of variance for a quantitative dependent variable by a single factor (independent) variable. ANOVA is used to test the hypothesis that several means are equal. This technique is a direct extension of the two-sample t-test (see above).

In addition to determining that differences exist among the means, you may want to know which means differ. There are two types of tests for comparing means: a priori contrasts and post hoc tests. Contrasts are tests set up before running the experiment, and post hoc tests are run after the experiment has been conducted. You suspect that doughnuts absorb fat in various amounts when they are cooked, and want to set up an experiment to test the absorbance of three types of fat: peanut oil,

corn oil, and lard. ANOVA is a suitable test for this. You also know that peanut oil and corn oil are unsaturated fats, whereas lard is a saturated fat. Along with determining whether the amount of fat absorbed depends on the type of fat used, you could set up an a priori contrasts to determine whether the amount of fat absorption differs for saturated and unsaturated fats.

Running one-way ANOVA procedure with your software, you can: a) validate the assumption of variance equality; b) obtain the ANOVA table and results, c) visually inspect the group means, d) perform custom contrasts, tailored to your specific hypotheses; e) compare each mean to every other mean, assuming variance equality or not, and f) perform two types of robust analysis of variance.

Two-way ANOVA: Two-way ANOVA is an extension of one-way ANOVA which allows us to assess the effects on the outcome variable of more than one categorical variable simultaneously. One powerful use of this is to control for confounding variables, in very much the same way as you can do with multiple regression analysis (see below).

In addition to using two-way ANOVA for testing and adjusting for confounding variables, the test has other applications as well. One such application pertinent for clinical trials is to test for the potential sequence effects in crossover study designs. In such settings, we might want tom test not only the difference between drug A and drug B, but also whether the results were affected by the sequence in which the patients received drug A and drug B. In fact, the name two-way is a bit misleading, because the two-way ANOVA is not limited for testing of only two variables, but, in fact it is possible to include multiple variables in the model, and thus adjust for several different confounding variables simultaneously.

ANCOVA: This is an acronym for ANalysis of COVAriance. This test is otherwise very similar to two-way ANOVA, but it allows the adjustments for continuous variables as well. When such a continuous variable (e.g. blood pressure, heart rate, weight, height, etc) is added among the covariates of the two-way ANOVA model, and it becomes an ANCOVA. In practical terms, you just run like a two-way ANOVA to adjust for a confounding variable, but this time the variable is not categorical but continuous instead. ANCOVA is equivalent to multiple regression analysis, where you can include both categorical and continuous variables as covariates.

Where the assumptions of ANOVA (and t-tests) are not satisfactorily met, you need to consider using the non-parametric equivalent tests (Wilcoxon signed rank test for paired data and the Wilcoxon rank sum test for independent data)(see below).

Confidence intervals

Before proceeding to these non-parametric tests, however, let's pay some attention to a couple of other issues that are important in running the statistical tests. In presenting practically any data, it has become popular to give the results and their confidence intervals (CI). Although you can usually define the limits of these CIs yourself, the usual practise is to give the 95% confidence intervals (95%CI).

What is a confidence interval? A confidence interval tells us the range of values (lower and upper bound) in which we are likely to find our observation of interest at the population level, i.e., among the normal people outside our cohort. This information helps us to determine, how important the results, e..g. an effect of a drug, might be in normal population. If we were testing a drug that reduces the serum cholesterol levels to a certain extent in our study group, we would be also interested to estimate, what would be the average fall in cholesterol level, if this same drug would be given to all possible patients with similar characteristics to those in our cohort (i.e., the population

mean). The CI would tell us the lower and upper bound of these values with 95% (or 99% if so wish) probability in the population.

CIs can be calculated for practically any type of statistical results, be it means, proportions, rates, rate ratios, Odds Ratios, etc, using specific formulas available in statistical textbooks. Most statistical software calculate 95%CI as the default for most of the tests, but for some, you still need to consult the textbook to obtain the specific formulas to do these calculations manually with your calculator.

It is evident that confidence intervals are in most cases far more informative than p (probability) values, and the use of 95%CI should be encouraged in presentations of all your statistics where appropriate.

Dummy variables

The use of dummy variables is another subject that is frequently encountered in running statistical tests, particularly those analysing case-control (drug-placebo) studies. These are also known as indicator variables, and their use is of particular importance while entering categorical variables as covariates in various statistical techniques, most notably in multiple linear regression analysis.

To test the efficacy of e.g. a set of two drugs compared with placebo in a multiple regression model, you need to create a set of dummy variables that unequivocally identify these two drugs (A and B) from the placebo (drug C) in your test. This is simply done by giving an identifier 0 (for placebo) and 1 (for drug A and B), but arranged in two groups (A and B), so that you have a combination of variables: 0-0, 0-1, 1-0 that accurately identify the three substances. Thus, you have 3 groups and 2 dummy variables in this case. If you have N groups, you naturally need N-1 dummy variables.

To follow a good practice in your analyses, however, you should always examine the overall significance of your collection of dummy variables first, before looking at the significance of individual dummy variables. This can be done using F statistics (in linear regression or ANOVA), and the likelihood ratio (L-R) test for e.g. in logistic regression.

Linear regression

Linear regression is a technique that gives us valuable information about the relationship between two variables. The variable which is our outcome of interest is called dependent variable, and the other variable predicting it is called independent variable in linear regression analysis.

Examples where linear regression is particularly suitable are easy to find in clinical trials and biomedical research in general. A frequently used example is the age-dependence of our systolic blood pressure. We know that blood pressure increases with age, but the question is, whether this increase is statistically significant in a particular subset of people, who are studied. Linear regression is the appropriate test to answer this. Like ANOVA, also linear regression is based on the assumption that the dependent variable should be approximately normally distributed.

Linear regression estimates the coefficients of the linear equation, involving one or more independent variables, that best predict the value of the dependent variable. For example, you can try to predict a salesperson's total yearly sales (the dependent variable) from independent variables such as age, education, and years of experience.

In another example, not so close to clinical medicine, we might ask, whether the number of games won by a basketball team in a season would be related to the average number of points that the team

scores per game? In a scatter plot, we can see that these two variables are linearly related. The same is true with the number of games won by our team and the average number of points scored by the opponent, which are also linearly related, but naturally these variables have a negative relationship. As the number of games won increases, the average number of points scored by the opponent decreases. With linear regression, you can model the relationship of these variables and those in any settings alike. In this example, a good linear regression model should be able to even predict how many games the teams will win. Or, what should be the blood pressure when you are 20 years older?

Multiple regression

Sometimes the simple linear regression testing the relationship between dependent and independent variables is not enough to model your data, which might contain more than just these two variables. The extension of linear regression is known as multiple linear regression analysis, because it can include more than two variables. Like in the former, also multiple regression analysis can accept only 1 outcome (or dependent) variable, but the model accepts an unlimited number of independent variables as predictors.

In their execution, these two tests are quite similar. Areas of application and interpretation of the results are different, however. Interpretation of the multiple regression analysis is sometimes tricky, and needs a common sense to evaluate, whether the results are meaningful and rational. When used e.g. for adjustment of confounding variables, where the change in y variable expected per a unit increase in the relevant x variable (when all the other x variables are held constant) is viewed, the result can sometimes be completely meaningless.

In any analysis where smoking is included among the independent variables, the model must be absolutely controlled for the confounding effect of smoking. Several examples of erroneous interpretation of the results are available, when the confounding of smoking was not controlled in the analysis. Alcohol is not an important causal factor of lung cancer, whereas smoking is, the vice versa being true with liver cancer, etc. Usually age and sex are factors that should be controlled for confounding in multiple regression analysis as well.

Correspondence between t-tests, ANOVA and regression

As evident from the account above, there are many similarities between the t-tests, ANOVA and regression analyses. They all are tests used for analysing a continuous, normally distributed dependent (outcome) variable. Although provided as separate tests by the statistical packages, regression and ANOVA are equivalent to each other, and the t-tests simply represent a special applications of the two.

Slight differences are found exclusively in our habitual uses of these three tests. Thus, we have used to apply ANOVA to examine the effect of categorical variables on an outcome (continuous) variable, whereas linear regression is used to examine the effect of continuous variables on a continuous dependent variable. As described above, however, also ANOVA can be modified to include continuous variables, to become ANCOVA. Multiple regression can accept categorical variables if entered as dummy variables. The t-test is exactly equivalent to an ANOVA, when only two groups are analysed, or to regression analysis containing a single dummy variable. Because of their equivalence also in the underlying mathematical equations, these three tests should give exactly the same p values as well.

To use parametric or non-parametric tests?

This distinction is important to make, before you select the statistical tests to be used in analysis of your data. Indeed, all statistical hypothesis testing follows the division into these two categories: parametric and non-parametric tests.

In this context, a clear distinction should be made between parameter and variable, which, unfortunately are frequently misused in reporting medical research. It is essential to remember that the appropriate term to describe the individual data records in your data file is ***variable***, not parameter. You have measured body weight, height, body mass index, blood pressure, etc, and once subjected to he statistical analysis, all these are entered in the model as variables, not as parameters. Discussing the regression models (above), we learned that we have both dependent variables and independent variables, and we never call them parameters.

Parameters, on the other hand, describe a mathematical function. In the statistics used to analyse medical research, this function is almost always a frequency distribution. Normal distribution can be described by its mean (M) and standard deviation (SD). The M and SD are the ***parameters*** of the normal distribution. On the other hand, a binomial distribution is described by the sample size and the probability of an individual success; those are its parameters. Parameters describe the population distribution, which can only be estimated, in contrast to sample distribution, which we can directly measure in our study.

Accordingly, a parametric test is any test that relies on the properties of a particular distribution. We already learned that ANOVA and t-tests assume that the basic data have a normal distribution. A frequently forgot principle is that all statistical analyses are based on testing some hypothesis, frequently called a null hypothesis. As the name implies, parametric tests are used to study the hypotheses on the parameters. Thus, t-tests are used to test the hypothesis that two samples have the same population mean. When you get a p value far above the accepted significance level of p=0.05, you have no reason to reject this null hypothesis. On the other hand, when your p values are clearly below p=0.05, you can safely reject the null hypothesis by concluding that the means of the two samples are significantly different, e.g. p=0.003.

On the other hand, non-parametric tests do not make any assumptions about the distribution of your data. They are therefore more powerful when your data do not follow the normal distribution. They are more applicable in testing hypotheses about the samples as a whole, instead of sample means.

The rest of this text is used to give a brief account of these non-parametric tests.

The Wilcoxon rank sum test

The Wilcoxon rank sum test is a non-parametric equivalent of the independent samples t-test (see above). This test is used to test the hypothesis that two independent samples have come from the same population. As mentioned above, being a non-parametric test, Wilcoxon rank sum test need only limited assumptions to be fulfilled concerning the distribution of the data. It assumes that the shape of the distribution is similar in the two groups, which is particularly important in cases, where you want to provide evidence that the median is significantly different between the groups.

If you have recorded your data in ordinal form, as frequently is the case in clinical trials recording the treatment responses in Likert scale (much worse, worse, no change, better, much better), you cannot use t-tests or ANOVA to compare these data. Instead, you should select a non-parametric test like Wilcoxon rank sum test to analyse the data of this format. The test assigns a different rank to each data in the two groups to be compared, staring from rank 1 for the smallest value, and

ending up by adding the ranks together in both groups. This is what the test name implies: rank sum test.

The Mann–Whitney U test

Some of the statistical software packages do not provide the Wilcoxon rank sum test as a the non-parametric equivalent of the independent samples t-test, but use the Mann–Whitney U test instead.

The Mann-Whitney U test is equivalent to the Wilcoxon rank sum test, despite slight differences in the underlying mathematics. Accordingly, any p values obtained with one of these tests shall be identical to those generated from the other test, if the same data set is being analysed.

The Wilcoxon signed rank test

The Wilcoxon signed rank test is a non-parametric equivalent of the paired samples t-test (see above). It is used to test the hypothesis that two paired samples have come from the same population. Because the test is non-parametric, it is not sensitive to violations in the normal distribution of the data.

The test is particularly suitable in analysing the data where continuous variables are recorded in cases and their matched controls (matched pair), or in individual subjects e.g. before and after a treatment of interest. The data must be a continuous variable (e.g., any commonly recorded laboratory test), which usually are some way off being normally distributed. Paired samples t-test would be inappropriate because of this violation of the assumption of normal distribution.

This test belongs among the repertoire of all commonly used statistical software packages as a default non-parametric equivalent of paired samples t-test.

To Conclude...

With the panel of tests presented in brief above, one can cover a vast majority of the statistical analysis needs in regular clinical studies and other types of biomedical research. These tests are also provided as default by all the common statistical software packages. Anything that goes beoynd these "standard" tests is highly irregular and necessitates a careful consideration for feasibility and expert opinion from professional statisticians.