JCD

REVIEW ARTICLE

A Review of Inferential Statistical Methods Commonly Used in Medicine

Kingshuk Bhattacharjee[a]

a Assistant Manager, Medical Services, Biocon Limited, Bangalore, India.
Email: kingshuk.bhattacharjee@biocon.com

Abstract

Inferential statistics is increasingly being used to manage the uncertainties in health and medicine fields by measuring and minimising them. However, descriptive statistics is familiar and advantageous to many of us; a gap is always felt at the clinician end in considering the use of inferential statistics in their basic research work. The aim of this article is to focus on some of the commonly used statistical methods, their interpretation, and the test of their significance. The list of tests may not be exhaustive but may be sufficient to cover the commonly used methods that need attention.

Keywords: ***inferential statistics, test of significance, common statistical tests***

Introduction

In the current era of evidence-based medicine, inferential statistics is the backbone of medical research. *Inferential* statistics is increasingly being used to manage the uncertainties in health and medicine by measuring and minimising them. However, descriptive statistics is familiar and advantageous to many of us; a gap is always felt at the clinician end in considering the use of inferential statistics in their basic research work. The aim of this article is to focus on an explanation of some of the commonly used statistical methods, their interpretation, and the test of their significance.

A brief understanding of the scale of measurement is must before understanding of statistical tests as choosing appropriate statistical tests mostly depends on the data, which are measured in different scales.

Understanding the Scale of Measurement

A scale is an instrument used to measure the characteristics, and they may be quantitative or qualitative. The following types of measurement scales can be identified in everyday practice:[1]

Nominal scale: Not all the characteristics are measured in quantities. Sites of injury, organ affected, religion, blood group, gender,and so on are qualities but are still considered as measurements. There is no notion of higher or smaller as they do not follow any specific order. The only way to represent them is by counts. Categories are said to be dichotomous when the assessment of characteristics is divided into two groups (yes/no and male/female), whereas the categories are said to be polytomous if the numbers of categories are divided into more than two, such as NAFLD Grade 1, Grade 2, and Grade 3. All these measurements belong to the purview of the nominal scale.[1]

Metric scale: These are the characteristics that can be measured exactly in terms of quantities. Duration of disease, count of blood cells, and pulse rate are few examples. Metric scales are seldom categorised into interval and ratio scales. The absolute zero is not present in an interval scale. Hence for characteristics such as body temperature and intelligence quotient on an interval scale, the differences are meaningful, but ratios are irrelevant. One cannot express a body temperature of 105°F as 5% higher than 100°F. In contrast, in a ratio scale, a zero point is meaningfully assigned. Here, it is correct to conclude that the duration of survival of 6 years is twice as much as 3 years.[1]

Ordinal scale: Many characteristics that should be measured in terms of quantity but are ended up being measured in terms of what is called as an ordinal scale (following an order). Age is a classic example of this scale, and it can be accurately measured in years, but its categorisation into child, adult, and old age may be satisfactory in many cases. Similarly smoking can be easily measured in terms of number of cigarettes smoked but often measured as none, light, moderate, and heavy smoking. The reasons can be many from serving the purpose in clinical situations to absence of an accurate measuring device. Disease severity when measured as none, mild, moderate, and severe is also an example of ordinal scale.[1]

Semi-ordered scale: There are certain characteristics that fall between the nominal scale and the ordinal scale. An example of semi-ordered scale is the classification of malignancy as definitely absent, probably absent, uncertain, probably present, and definitely present. These categories are partly nominal and partly ordinal.[1]

Understanding Null Hypothesis and Alternative Hypothesis

Once familiar with the different scale of measurements, an understanding of the null hypothesis is required to make correct inferences of the statistical tests. Let us take an analogy of a court setting. When a convict is first presented in a court, the ideal initial assumption likely to be set by the judge is that the convict is innocent. It is followed by the prosecutor providing evidence against the innocence of the convict, and depending on the evidence, the judgement is drawn. In the case of statistical decisions too, the initial assumption is that there is no difference or no correlation between groups (for quantitative data), no association between groups (for categorical data), and so on (Table 1). This initial assumption is called the null hypothesis and is denoted by H_0. The null hypothesis is then subjected to scrutiny by conducting a study on the sample observations drawn from the population. Depending on the evidence provided by the sample, H_0 can or cannot be rejected.[2]

Alternative hypothesis

If the null hypothesis is rejected ($p<0.05$), then the alternative hypothesis is accepted. This alternative hypothesis, represented by H_1, is the opposite of H_0. *A statistical significance is said to be achieved when the H_0 is rejected, that is, $p<0.05$.*

Type I error

Let us consider a situation when there is no real difference between the groups drawn from the

Table 1 | Null hypothesis of commonly used statistical method

Statistical tests	Null hypothesis (H0)
Unpaired t test / Mann–Whitney U test	No difference in mean or location (median) between the two groups
ANOVA (F test) / Kruskal–Wallis H test	No difference in mean or location (median) between the three or more groups (considered together)
Chi-square test / Fischer's exact test	No association between the qualitative (categorical) variables under consideration
Paired t test / Wilcoxon's signed rank test	No difference between the before and after measurements (measured on two occasions) done on the same set of subjects/related subjects
Repeated-measures ANOVA / Friedman's ANOVA	No difference between the measurements (repeated on three or more occasions) done on the same set of subjects/related subjects
Pearson's correlation / Spear man's correlation	No correlation between the two sets of quantitative variables under study
Regression	Model in question cannot explain any relationship between dependent and predictor variables

population, but the sample data strongly disagree to show a difference. *In this case, the null hypothesis is undesirably rejected and a false-positive conclusion is reached. This is called type I error or alpha error*. Generally, it is set at 5%.[1]

Type II error

The second type of error is failing to reject the null hypothesis when it is false. *This error occurs when a study fails to detect a real difference.* The probability of this type of error is denoted by β. Generally,it is set at 20%.[1]

Power

The complementary of the probability of type II error is known as statistical power and is denoted by $(1-\beta)$. It is the probability of correctly rejecting the null hypothesis when it is false. *In other words, it is the probability of getting a statistically significant result when a difference really exists*.[1]

Assessing the Distribution of Data

The distribution of data is another important aspect to be considered while choosing appropriate statistical tests, and most tests are based on the inherent assumption of shape of the distribution of data. The data may follow a Gaussian distribution (also called as parametric or normal distribution), which has a bell-like pattern when plotted as a histogram. For the normally distributed data, the mean, the median, and the mode coincide at the centre.[1] It can also be checked if mean ±1 SD (standard deviation) covers 67% of the values and mean ±2 SD covers 95% of the values. The range should be roughly 6 SD. *The Gaussianity can also be assessed by statistical tests, namely Anderson–Darling, Shapiro–Wilk, and Kolmogorov–Smirnov tests.* These tests can easily be computed by statistical software packages. The H_0 of these tests is the data that are normally distributed. Hence, if p value is <0.05, then H_0 is rejected, and it is concluded that the data are skewed and vice versa. However, if discrepancies arise between the results of different tests, importance should be given to the results of Kolmogorov–Smirnov and Shapiro–Wilk tests.

There is no need to check distribution in the case of nominal and ordinal data because nominal data follow the chi-square distribution and ordinal data do not follow the normal distribution. Therefore, distribution should only be checked in the case of interval and ratio data.

Checking the data for normal distribution

Once the data distribution is revealed by the above-mentioned methods, it becomes simple to use the appropriate statistical test. If the data follow the normal distribution, then the parametric statistical test should be used and vice versa (Table 2).

When the task is to identify any statistical significant difference between unpaired groups (two unrelated groups); one has to consider different statistical tests for numerical (quantitative data) and categorical or qualitative data. For quantitative data, it has to be determined in advance if the variable under consideration is following Gaussian distribution or not by suitable methods as described earlier. Parametric tests should be used for data following Gaussian distribution (Tables 2 and 3). It is advisable to use non-parametric tests for distributions not following Gaussian or when one is not sure of the data distribution.[3] For analyzing significant differences between more than two groups of numerical data, an ANOVA or a Kruskal–Wallis test is used based on the distribution of data (Tables 2 and 3). If an ANOVA or a Kruskal–Wallis test returns a statistically significant result, it should be followed by a post-hoc test to

Table 2 | Parametric and non-parametric forms of commonly used statistical tests

Parametric tests (for normally distributed data)	Equivalent non-parametric tests (for skewed data)
One sample *t*-test	Wilcoxon test
Unpaired *t*-test	Mann–Whitney *U* test
Paired *t*-test	Wilcoxon Signed rank test/ sign test
One-way ANOVA	Kruskal–Wallis *H* test
Repeated-measures ANOVA	Friedman's ANOVA
Pearson's correlation	Spear man's rank correlation
Simple linear regression or nonlinear regression	Non-parametric regression

determine exactly between which two data sets the difference exists.[3] Different types of post-hoc tests are available, namely Bonferroni's method and Tukey's method. Nevertheless, caution has to be exercised not to repeatedly apply *t*-test or Mann–Whitney *U* test to a multiple group situation, which may inflate the possibility of type 1 error. To analyse any association between the two categorical variables, chi-square test provides a the appropriate result for a cell count of five or more. There is an alternative test, called pooled chi-square test, where adjacent cells may be merged to achieve a cell count of five; however, this should be done without compromising the biological relevance. Fischer's exact test has a tremendous use for analyzing associations in a 2×2 contingency table where the cell count is less than five.[4,5]

Many a times, we come across situations where we have to enquire if any statistically significant difference exists between paired groups. Pairing refers to data sets obtained by repeated measurements of the same variable under investigation. Pairing may also occur if subject groups are different, but values in one group are related to the values in the other group in some way (e.g. twin studies, sibling studies and parent–offspring studies)[3] Once again, a repeated measure study of three or more occasions necessitates post-hoc tests to explore pair wise comparison.[3] There might be some cases where the paired measurements are qualitative in nature. Let us consider a scenario where participants' smoking status are recorded as yes/no (dichotomous) at baseline and after an intervention on smoking cessation, the responses are recaptured. To determine any change in the smoking status change after an intervention, appropriate method would be McNemar's test. The McNemar's test is used to determine whether there are differences on a dichotomous-dependent variable between two related groups. It can be considered to be similar to the paired-samples *t*-test, but for a dichotomous rather than a continuous dependent variable.[6] For a qualitative repeated measurement on more than two occasions, cochrane *Q* test may be used.

Table 3 | Statistical procedures for testing hypothesis on means or locations

Setup	Conditions	Main criterion
One sample	Comparison with a prespecified value – Gaussian distribution Standard deviation known Standard deviation unknown	Gaussian Z Student *t*
Comparison between paired groups (before and after measurement)	Paired and Gaussian	Paired *t*-test
	Paired and non-Gaussian (any *n*)	Sign test
	Paired and non-Gaussian ($5 \leq n \leq 19$)	Wilcoxon signed-ranks (W_S)
	Paired and non-Gaussian ($20 \leq n \leq 29$)	Standardised W_S
Comparison between two different unrelated groups	$n \geq 30$	Student *t* test (unpaired *t*-test)
	Unpaired and Gaussian Equal and unequal variances	Student *t* test (unpaired *t*-test)
	Unpaired and non-Gaussian n_1 and n_2 between (4, 9)	Wilcoxon rank-sum (W_R)
	Unpaired and non-Gaussian n_1 and n_2 between (10, 29)	Standardised W_R
	Unpaired and non-Gaussian n_1 and $n_2 \geq 30$	Student *t*(unpaired *t* test)
Comparison between three or more different groups	One-way layout Gaussian	ANOVA *F*
	Non-Gaussian	Kruskal–Wallis *H* test
	Two-way layout Gaussian	Two-way ANOVA *F*
Comparison between related groups (measured on three or more occasions)	Gaussian	Repeated-measures ANOVA
	Non-Gaussian	Fried man's ANOVA

Adapted from Medical Biostatistics. (3rd edn.). Chapman & Hall/CRC Biostatistics Series; 2013: Summary Tables

For measuring the quantitative relationship between two numerical variables, there are correlation tests that express the strength of association as a correlation coefficient. While regression expresses the nature of relationship, correlation measures degree of relationships. It is known that the relationship of smoking to cotinine level is strong but practically nil with intra ocular pressure. It is to be noted that in this setting, only a linear relationship is being considered, and there is a total of only two variables under consideration at a time. Depending on the distribution of data (Gaussian or skewed), a Pearson's correlation or Spearman's rank correlation method is used (Table 1). The correlation coefficients vary in magnitude from −1 (perfect negative correlation) to 1 (perfect positive correlation). A correlation coefficient between two variables is depicted by a positive (moving in same direction) or a minus sign (moving in opposite direction), whereas a magnitude of 0 means no linear correlation at all. In general, a correlation greater than 0.9 in absolute value can be considered strong, between 0.6 and 0.9 as moderate, between 0.4 and 0.6 as weak and below 0.4 as almost non-existent.[1] However, the absence of a linear correlation does not rule out the possibility of non-linear trend, which should be explored by suitable non-linear methods. A caution should also be exercised not to extrapolate a good correlation into causality.

Relationships are inherent in medicine and health. The primary purpose of studying such relationships is to predict the value of one or more variables (called as dependent or outcome variable) with the help of other variables (called as explanatory, independent or predictor variables). Furthermore, it is helpful to understand the underlying mechanisms in such a relationship, particularly by studying the effect of alteration in value of one variable when other conditions do not change.[7] Here comes the utility of regression analysis. There is much confusion with the nomenclature of different regression setups. A regression is called a simple or multiple regression, depending on the presence of one or more predictor variables in the regression model. For one and more than one outcome variable, it is a called as a univariate or multivariate regression model, respectively. Utility of different regression setups is outlined in Table 4.

Checking agreement between data sets is another frequently encountered scenario in medicine. It can be a comparison of a new screening test with a gold standard test or agreement between ratings or scores given by different observers. For agreement between numerical variables, the agreement may be expressed quantitatively by intra class correlation coefficient or graphically by constructing a Bland–Altman plot. A Bland–Altman plot is constructed by plotting the difference between the two variables in *y*-axis against the average of two variables in *x*-axis. In the case of qualitative data, the Cohen's kappa followed by a Bowker test is regularly used. If Bowker test is non significant ($p>0.05$), it indicates that the two modalities/raters have the same propensity to select categories. If it is significant ($p<0.05$), it means that the modalities are selecting the categories in differing proportions. The kappa varies from 0 (no agreement at all) to 1 (perfect agreement): (kappa < 0.3 –poor agreement, kappa = 0.3–0.5 –fair agreement, kappa = 0.5–0.7 –moderate agreement, kappa = 0.7–0.9– good agreement, and kappa > 0.9 –excellent agreement).

To understand whether there is a significant difference between time to event trends or survival plots, a survival analysis is commonly used. The survival analysis method is used to analyse the data

Table 4 | Methods for studying nature of relationship

Dependent variable(s)	Independent variable(s)	Statistical method(s)
Quantitative	Qualitative	ANOVA
Quantitative	Quantitative	Quantitative regression
Quantitative	A mixture of qualitative and quantitative	ANCOVA
Qualitative (dichotomous)	Qualitative or quantitative or mixture	Binary logistic regression
Qualitative (polytomous)	Qualitative or quantitative or mixture	Multinomial logistic regression
Qualitative	Quantitative	Discriminant analysis

Adapted from *Medical Biostatistics* (3rd edn.). Chapman & Hall/CRC Biostatistics Series; 2013: Summary Tables.

on captured duration.[8] The endpoint for such analysis could be death or any event that can occur after a specific period, which is characterised by censoring data, meaning that a sizeable proportion of the original study subjects may not reach the endpoint in question by the time the study ends.[3] Data sets for survival trends are always considered to be highly skewed. If there are two groups, then the applicable tests are Cox Mantel test, Gehan's (generalised Wilcoxon) test or log rank test. In case of more than two groups, Peto and Peto's test or log-rank test can be applied to search significant difference between the time to event trends.[3]

Conclusion

Although many statistical methods are available, the above-mentioned tests in this review are commonly used. It is fully understood from the above discussion that before selecting any test, one must take into consideration the distribution of the data (Gaussian or skewed). Furthermore, there are certain basic assumptions that must be fulfilled,namely the sample selected from the population is random. Finally, when considering statistical options for specific research designs and data characteristics, it is important to weigh the advantages and disadvantages of available statistical methods before selecting a method or methods.

References

1. Indrayan A. Confidence intervals, principles of tests of significance, and sample size. In: Abhaya Indrayan (ed.). *Medical Biostatistics* (3rd edn.). Chapman & Hall/CRC Biostatistics Series; 2013. pp. 373–443.
2. Parikh MN Hypothesis testing and choice of statistical tests. In: Parikh MN, Hazra A, Mukherjee J, Gogtay N (eds.). *Research Methodology Simplified: Every Clinician a Researcher*. New Delhi: Jaypee Brothers; 2010. pp. 121–128.
3. Nayak BK, Hazra A. How to choose the right statistical test? *Indian JO phthalmol*. 2011; 59(2):85–86.
4. Sabin C The theory of linear regression and performing a linear regression analysis. In: Petrie A, Sabin C (eds). *Medical Statistics at a Glance* (2nd edn.). London: Blackwell Publishing; 2005. pp. 70–73.
5. Wang D, Clayton T, Bakhai A. Analysis of survival data. In: Wang D, Bakhai A (eds). *Clinical Trials: A Practical Guide to Design, Analysis and Reporting*. London: Remedica; 2006. pp. 235–252.
6. McDonald JH. Handbook of Biological Statistics (3rd edn.). Baltimore, Maryland: Sparky House Publishing; 2014. Choosing the right test; pp. 293–296.
7. Relationships: Quantitative Data. In: Abhaya Indrayan. Medical Biostatistics (3rd ed.). Chapman & Hall/CRC Biostatistics Series; 2013:603–661.
8. Survival Analysis. In: *Abhaya Indrayan. Medical Biostatistics* (3rd edn.). Chapman & Hall/CRC Biostatistics Series; 2013:699–732.

Keep your face always toward the sunshine - and shadows will fall behind you.
— Walt Whitman